PENN INSTITUTE for ECONOMIC RESEARCH
UNIVERSITY of PENNSYLVANIA

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
23-007

Behavioral Foundations of Model Misspecification

J. AISLINN BOHREN
University of Pennsylvania

DANIEL N. HAUSER
Aalto University and Helsinki GSE

May 18, 2023

Behavioral Foundations of Model Misspecification[*]

J. Aislinn Bohren[†] Daniel N. Hauser[‡]

May 18, 2023

We link two approaches to biased belief formation: non-Bayesian updating rules and model misspecification. Each approach has advantages: updating rules transparently capture the underlying bias and are identifiable from belief data; misspecified models are 'complete' and amenable to general analysis. We show that misspecified models can be decomposed into an updating rule and forecast of anticipated future beliefs. We derive necessary and sufficient conditions for an updating rule and forecast to have a misspecified model representation, show the representation is unique, and construct it. This highlights the belief restrictions implicit in the misspecified model approach. Finally, we explore two ways to select forecasts— introspection-proof and naive consistent—and derive when a representation of each exists.

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

[*] We thank Nageeb Ali, Cuimin Ba, Renee Bowen, Sylvain Chassang, Hanming Fang, Mira Frick, Kevin He, Ryota Iijima, Alex Imas, Jawaad Noor and seminar and conference participants at various institutions for helpful comments and suggestions. Cuimin Ba and Marcus Tomaino provided excellent research assistance. Bohren gratefully acknowledges financial support from NSF grant SES-1851629.

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

[‡] Aalto University and Helsinki GSE; Email: daniel.hauser@aalto.fi

1 Introduction

Extensive work in economics and psychology has documented individuals' systematic biases and errors when interpreting information and forming beliefs. A rich literature has explored how to model such inaccurate updating. Two modeling approaches are commonly used: the 'non-Bayesian' approach and the 'misspecified model' approach. In the non-Bayesian approach, a particular bias is parameterized with an updating rule that maps signal realizations to posterior beliefs (e.g. under- and overreaction in Epstein, Noor, and Sandroni (2010).) In the misspecified model approach, a subjective model of the signal generating process describes how individuals interpret signals; the individual forms beliefs using Bayes rule with respect to this model 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 adapt existing methods to this approach. Moreover, a misspecified model is 'complete': in addition to specifying how an agent forms beliefs, it also pins down how an agent forms expectations before observing information, which can be relevant for ex-ante decisions and strategic interaction. Finally, the approach is amenable to analysis in a general context. A large literature establishes general learning properties for misspecified models (e.g. Bohren and Hauser (2021); Fudenberg, Lanzani, and Strack (2022); Frick, Iijima, and Ishii (2020b)) 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. Additionally, this is the approach often used in empirical work, as an updating rule can be identified from belief data (Danz, Vesterlund, and Wilson 2022). Importantly, however, the approach is incomplete: 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 expectations. For example, Rabin and Schrag (1999); Epstein et al. (2010) study how confirmation bias and over/undereaction, 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 less 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 *forecast* of her future beliefs is the other component of belief formation needed to pin down a unique representation. Importantly for empirical work, a 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 forecast to be jointly represented and constructs this unique representation. Finally, we explore how to select a forecast to pair with an updating rule. From the perspective of the misspecified model approach, these results clarify the belief formation restrictions implicit in using this approach, decompose the model into empirically identifiable components, and highlight how these components isolate the forms of bias that a given model induces. From the perspective of the non-Bayesian approach, these results provide guidance on how to incorporate a given form of bias into more complex decision problems (e.g. strategic settings, settings with an ex-ante decision before information arrives) and yield a set of off-the-shelf tools that can be used to immediately establish important results such as the convergence of beliefs.

We now describe our setting in more detail. We focus on an informational environment in which an agent learns about a hidden state from a signal. The non-Bayesian approach consists of an updating rule mapping each signal realization to a posterior belief and a 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 forecast—and the *retrospective* bias—how the agent reasons about realized information via the updating rule.[2] 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. Similarly, a misspecified model represents a forecast when the predicted

[1] See Chambers and Lambert (2021); Karni (2020) for methods to elicit an agent's prediction of her own future belief and Manski and Neri (2013) for a method to elicit an agent's prediction of others' beliefs.

[2] Benjamin, Rabin, and Raymond (2016) drew a similar distinction between how an agent retrospectively versus prospectively processed information in models of non-Bayesian updating; we discuss the conceptual differences with our distinction in Section 1.1.

distribution of the posterior belief derived from the misspecified model is equal to the forecast.

We first derive individual necessary and sufficient conditions for an updating rule or a forecast to be represented. The condition for the updating rule is quite mild: it must be *responsive*, 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. It is satisfied by many updating rules commonly used in the literature (e.g. overreaction (Epstein et al. 2010), partisan bias (Bohren and Hauser 2021), confirmation bias (Rabin and Schrag 1999)). The condition for a 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 show that these conditions are a necessary implication of Bayesian updating, as required in a misspecified model; the more innovative aspect is to show that these conditions are also sufficient.

We then bring these results together 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* condition is needed. This condition requires the set of posteriors prescribed by the updating rule to be equal to the support of the forecast. In other words, any set of posterior beliefs that the agent anticipates with positive probability must arise with positive probability given her updating rule. This condition is mild for sufficiently rich signal spaces. A given updating rule can therefore be paired with many different forecasts, and similarly for a given forecast. This shows that a misspecified model can feature very different forms of prospective and retrospective bias—neither component places much restriction on the form of the other. Together, these conditions clarify the belief formation restrictions implicit in using the misspecified model approach: (i) responsive updating rules, (ii) plausible 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 forecast satisfying (ii) and (iii) have a misspecified model representation.

Importantly, such a representation is unique. From the perspective of the misspecified model approach, this establishes that a misspecified model can be uniquely decomposed into the prospective and retrospective biases that it induces. The prospective bias reflected in the forecast captures errors in anticipating future belief formation; the retrospective bias reflected in the updating rule captures how an agent misinterprets information after it arrives. Every misspecified model is uniquely identified by these two components, and they describe all biases that the model induces. 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 forecast together pin down all behavioral implications of a misspecified model, in that 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 forecast uniquely pins down a misspecified model that can be used for analysis. Moreover, the chosen form of retrospective bias places very little structure on the choice of prospective bias, and vice versa.

Finally, our main result provides a method to construct the misspecified model that represents a desired updating rule and forecast. This construction provides a simple formula that can be easily used in applications.

Since the majority of the non-Bayesian updating literature focuses on updating rules, we next provide guidance on how to select a forecast to pair with a given updating rule. We propose two reasonable choices, both of which use the correctly specified model to impose structure on anticipated beliefs. We first consider the accurate forecast, where the agent's anticipated distribution of her posterior belief is equal to the true ex-ante distribution. In other words, the agent exhibits no prospective bias. The accurate forecast automatically satisfies *no unexpected beliefs.* Therefore, if it is plausible, then it has a (unique) representation. Moreover, the corresponding misspecified model satisfies a property we call *introspection-proof.* This property ensures that even with an infinite amount of data, a misspecified agent would not observe inconsistencies with her model. While the introspection-proof property provides a natural constraint in many settings, the accurate forecast is not plausible for many common updating rules. This means that many common updating rules cannot be represented by an introspection-proof misspecified model.

We next define a forecast that captures a natural analogue to the naiveté assumption used in many behavioral settings. The *naive consistent* forecast corresponds to the accurate forecast of an agent who uses Bayes rule to update beliefs. An agent with a naive consistent forecast believes that she will correctly interpret information in the future—she believes that she has no retrospective bias. But when she actually updates her beliefs, she uses a biased updating rule.[3] The naive consistent forecast is plausible by definition. Therefore, if it satisfies the *no unexpected beliefs* property, then it has a (unique) representation. In contrast to introspection-proofness, a naive consistent representation exists for many common updating rules.

[3] Benjamin et al. (2019) study an updating rule that features base rate neglect and close their model with an assumption that results in a naive consistent forecast.

In two applications, we demonstrate how our results yield novel insights in specific settings. The first shows how discrimination can emerge endogenously due to self-image concerns. Consider a dual-selves model in which a manager learns about her own and others' ability (high or low) from a signal, which includes a test score and a group identity. The first self selects an updating rule to interpret this signal and the second self uses this updating rule to form beliefs. The second self has an intrinsic value for believing that he is a high type and an instrumental value for learning the types of others. He must use the same updating rule to interpret his own and others' signals. A natural constraint to place on the first self's choice of updating rule is that any bias will be undetectable by the second self, i.e. the updating rule can be represented by an introspection-proof misspecified model. Self-image concerns lead the first self to select an updating rule that exhibits motivated reasoning, in that the manager inflates his interpretation of the test score for workers in his own demographic group. The introspection-proof constraint places an endogenous upper bound on the magnitude of this bias. It also leads the chosen updating rule to shade down the interpretation of the test score for members of the other demographic group. We show that such self-image concerns can generate biased stereotypes and inaccurate statistical discrimination (Bordalo, Coffman, Gennaioli, and Shleifer 2016; Bohren, Haggag, Imas, and Pope forthcoming).

In the second application, we show that conceptually similar prospective and retrospective biases can lead to very different predictions about behavior. Consider a firm searching for a new technology. After observing a signal of a technology's productivity, the firm chooses whether to adopt this technology or to continue searching. We compare how over- and underprecision impact this search decision, depending on whether the bias emerges prospectively or retrospectively. A firm that interprets signals correctly but has an overprecise forecast searches inefficiently many alternatives, as it overestimates the value of future information. In contrast, a firm with an overprecise updating rule and a low search cost searches too few alternatives, as it overestimates the accuracy of current positive information. Analogous insights hold for underprecision. Together, these applications demonstrate how our results can be used to harness the advantages of both the misspecified model approach and non-Bayesian updating approach when studying how biased beliefs impact economic decisions.

We close with several extensions of our setting. First, we characterize the set of updating rules that have a prior-free representation, in the sense that the same misspecified model represents the updating rule at all prior beliefs. Second, we allow for the possibility that an agent also has a misspecified prior and derive an analogue of our main result. Finally, we discuss how time inconsistency can emerge in a dynamic version of our framework.

1.1 Literature Review

There has been renewed interest in using model misspecification as a tool for capturing behavioral biases.[4] 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, (Molavi 2019; Bohren and Hauser 2021; Fudenberg, Lanzani, and Strack 2021; Frick et al. 2020b; Esponda, Pouzo, and Yamamoto 2021), and explored questions of robustness to perturbations of the model (Frick, Iijima, and Ishii 2020a; Bohren and Hauser 2021). 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 the non-Bayesian approach can be represented as a misspecified model, allowing for analysis using these general results.

Another strand of literature seeks to provide a foundation for model misspecification (Ba 2022; Fudenberg and Lanzani 2022; Gagnon-Bartsch, Rabin, and Schwartzstein 2018; He and Libgober 2021; Frick, Iijima, and Ishii 2021; Fudenberg et al. 2022). One of the main classes of models we consider—introspection-proof models—are naturally robust to many of these criteria. This condition—which requires that the misspecified agent correctly anticipates the unconditional distribution of signals—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 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) show that if an agent updates using Bayes rule, then the prior belief is in the interior of the convex hull of the set of posterior beliefs. In Lemma 1, we provide a minor extension of this result that applies to the class of updating rules and misspecified models we consider. Molavi (2021) 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 both the condition in Shmaya and Yariv (2016) and our conditions, as he allows the misspecified model to put positive probability on signals outside of the support of the correctly specified model. A similar result follows from our characterization under slightly more restrictive conditions to account for our stronger requirement on the support of the misspecified model. Augenblick and Rabin (2021) provide conditions on the movement of posterior beliefs over time to

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

test for (correctly specified) Bayesian updating.

There is also a related literature on non-Bayesian updating. A number of recent papers provide foundations for general classes of non-Bayesian updating rules and draw parallels between the structure of non-Bayesian and Bayesian updating (Epstein, Noor, and Sandroni 2008; Lehrer and Teper 2017; Cripps 2018; Chauvin 2020; Zhao 2022; Jakobsen 2022). In contrast, we characterize the properties of updating rules that emerge from Bayesian updating with respect to a misspecified model of the signal process. Other work characterizes properties of specific non-Bayesian updating rules. For example, He and Xiao (2017) describe a class of updating rules that distorts the prior likelihood and signal likelihood terms in Bayes rule in a specific way. They show that this class of updating rules satisfies processing consistency in that sequential and simultaneous signal processing lead to the same posterior. Benjamin et al. (2019) study an updating rule that captures base rate neglect by distorting the prior likelihood ratio. As in our paper, they highlight prospective beliefs as a necessary model component in many economic settings and pin them down by assuming that an agent believes she will use Bayes rule to update in the future. This is similar in spirit to our naive consistent forecast.

Benjamin et al. (2016) first highlighted the need to distinguish between how an agent retrospectively processes information she has already observed versus prospectively predicts she will process information in models of non-Bayesian updating. In their context, this distinction is drawn in relation to how an agent groups multiple signals for processing. They highlight how different retrospective versus prospective groupings can lead to time-inconsistency, which arises because realized signals change how an agent interprets future signals. 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): an agent's bias in anticipating what her belief will be after observing this signal versus the bias in actual updating after observing the signal. Such an agent can be time-consistent. In Section 6.3, we discuss how time-inconsistency can emerge in a dynamic version of our setting with a sequence of signals. Similar to Benjamin et al. (2016), time-inconsistency stems from the difference between which model(s) an agent anticipates versus actually uses at future information sets. This notion of bias in anticipated versus actual model is conceptually distinct from our notion of bias in anticipated versus actual processing of a signal within a given model.[5]

Much of the literature on biased belief formation focuses on either a prospective or a retrospective bias. The work on misspecified causal graphs (Spiegler 2016) and Berk-Nash equilibrium (Esponda and Pouzo 2016) take a prospective perspective, focusing on

[5] The former relates to the property of processing consistency studied in He and Xiao (2017); updating rules that satisfy their condition for processing consistency result in the same prospective and retrospective beliefs.

how an agent (incorrectly) predicts what will happen after she has made her decision. In contrast, papers such as Heidhues et al. (2018); Levy et al. (2022) as well as much of the behavioral literature that models and empirically documents specific updating biases (see Benjamin (2019) for a survey) focus on retrospective biases. When modeling even simple economic decisions, such as the search application in Section 5.2, or strategic interactions, such as those studied in Bohren and Hauser (2021); He (2022); Frick et al. (2021), both prospective and retrospective biases play a role in determining beliefs and behavior. In Bohren and Hauser (2023), we show how our decomposition can be used to determine the way in which retrospective and prospective biases differentially impact an optimal lending contract.

Our work also relates to the literature on Bayesian persuasion. Our main result requires a version of Bayes plausibility (Kamenica and Gentzkow 2011) for misspecified models. de Clippel and Zhang (2022) develop an analogue of Bayes plausibility in a persuasion setting where the receiver uses a non-Bayesian updating rule.[6] Their condition is both technically and conceptually distinct from ours: it characterizes the set of possible distributions over posteriors that a correctly-specified sender could induce, while our plausibility condition describes the set of distributions over posteriors that a misspecified agent could hold about her future belief. The forecast in our decomposition is analogous to the unconditional distribution over posterior beliefs often characterized in the Bayesian persuasion literature (Kamenica 2019). This unconditional distribution also plays an important role in the literature that describes general measures of the value/cost of information (Frankel and Kamenica 2019; Caplin, Dean, and Leahy 2022; Pomatto, Strack, and Tamuz Forthcoming; Mensch 2018).

Recent work on the wisdom of the crowd focuses on how higher order beliefs impact prediction and identification. Prelec, Seung, and McCoy (2017) show that knowing both an agent's posterior belief and her belief about others' beliefs can yield more accurate predictions. Prelec and McCoy (2022); Libgober (2023) show that if many agents draw signals from the same information structure, then knowing both an agent's posterior belief about the state and her posterior belief about the distribution of others' beliefs identifies the information structure. This relates to our insight that eliciting an updating rule on its own is insufficient to identify a unique misspecified model (i.e. subjective information structure)—it must be paired with a component describing the distribution over beliefs.

[6] Alonso and Câmara (2016); Lee, Lim, and Zhao (2023) also stusdy communication games with biased receivers.

2 Model

2.1 The Informational Environment.

We study belief updating in the following informational environment. Suppose nature selects one of N states of the world $\omega \in \Omega \equiv \{\omega_1, \omega_2, \ldots, \omega_N\}$ according to prior distribution $p \equiv (p_1, ..., p_N) \in \Delta(\Omega)$, which we assume to be strictly interior. An agent observes a signal of the state drawn from a measurable space $(\mathcal{Z}, \mathcal{F})$, where $\mathcal{Z}$ is an arbitrary set with element z and $\mathcal{F}$ is a σ-algebra defined on $\mathcal{Z}$. To ensure that densities exist, we define a σ-finite reference measure ν on $(\mathcal{Z}, \mathcal{F})$; we will assume all subsequent measures are absolutely continuous with respect to ν.[7] Let $\mu_i \in \Delta(\mathcal{Z})$ be the true probability measure on $\mathcal{Z}$ in state ω_i. Assume that μ_i and μ_j are mutually absolutely continuous for each $i, j = 1, ..., N$ and μ_i is absolutely continuous with respect to ν for all $i = 1, ..., N$.[8] This ensures that no signal perfectly rules out a state.[9] Let $\Delta^*(\mathcal{Z})$ denote the set of all probability measures that are mutually absolutely continuous with respect to μ_1 (note this also implies the measures are mutually absolutely continuous with respect to μ_i for $i \neq 1$). Finally, let $\mu \equiv \sum_{i=1}^N p_i \mu_i$ denote the unconditional measure on $\mathcal{Z}$.

Our main analysis focuses on belief updating following a single signal realization or "batch" of realizations. The framework naturally extends to studying a dynamic signal process, albeit with more cumbersome notation (see Section 6.2). The signal space is rich enough to capture many different common signal structures used in the literature, including real-valued continuous signals ($\mathcal{Z} \subseteq \mathbb{R}$ and ν is the Lebesgue measure), finite signals ($\mathcal{Z} \subseteq \mathbb{R}$ is finite and ν is the counting measure), multidimensional signals, causal graphs, Markov signals, and signal distributions that are neither continuous nor discrete (e.g. mixture distributions).[10]

2.2 Modeling Errors in Belief Updating

We are interested in exploring the relationship between two approaches used to model behavioral biases and errors in belief-formation: (i) a "non-Bayesian" approach that

[7] When $\mathcal{Z}$ is not finite, this introduces a number of measure-theoretic and topological complications. A standard tool to resolve these complications is to define a reference measure that dominates the other measures in the model. This allows us to consider multiple types of signal spaces within the same framework, such as settings where the signal measures have densities and settings where the signal is not a real-valued continuous random variable. Note that our set-up is the finite state version of the misspecified parametric environment from Kleijn and van der Vaart (2006).

[8] Given our assumptions, one could set $\nu = \mu_i$ for any i or $\nu = \mu$. We chose to separate these objects to maintain a reference measure that is independent of the state and prior.

[9] Note this implies that $\frac{d\mu_i}{d\nu}(z) = 0$ if and only if $\frac{d\mu_j}{d\nu}(z) = 0$ except on a set of ν-measure 0, so that signals that lead to a Bayesian posterior that places probability zero on a state or signals for which the Bayes posterior is not defined are a probability 0 events.

[10] This set-up can also capture signals that are multiple draws from an urn (Rabin 2002), signals that are up to K realizations of some process (He 2022), and signals that are a realization of a Brownian motion (Fudenberg, Romanyuk, and Strack 2017).

consists of defining an arbitrary updating rule and/or a prediction about future beliefs; and (ii) a "misspecified Bayesian" approach that derives beliefs from Bayesian updating with respect to a misspecified model. We introduce each approach in turn, then discuss the relative advantages and disadvantages of each approach.

The Non-Bayesian Approach. This approach, often used in the behavioral learning literature (e.g. see Benjamin (2019) for review), describes how an agent forms a posterior belief after observing each possible signal realization—that is, an *updating rule*. A second component of belief formation—an agent's *forecast*, or prediction of what future beliefs will be—is needed in many economic settings. The updating rule determines how an optimal action depends on the signal realization for decisions that occur after the signal is observed, whereas the forecast guides pre-signal action choices by pinning down the likelihood of different post-signal actions. The general definitions of updating rules and forecasts we outline below nest specific updating rules and forecasts used in non-Bayesian approaches to belief-formation.

An updating rule specifies how an agent forms beliefs after observing each signal realization. An agent uses *updating rule* $h(z)$ if, for each $i = 1, ..., N$, the agent assigns probability $h(z)_i$ to state ω_i after observing signal realization $z \in \mathcal{Z}$.[11]

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

We restrict attention to updating rules that do not interpret any signals as perfectly ruling out a state and map a certain prior belief to a certain posterior belief: $h(z)_i = 0$ iff $p_i = 0$ and $h(z)_i = 1$ iff $p_i = 1$. A special case of an updating rule is Bayesian updating with respect to the true family of measures $(\mu_i)_{\omega_i \in \Omega}$. Given a signal realization $z \in \mathcal{Z}$, this corresponds to

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

with $0/0 = 0$ by convention.[12]

We refer to bias that arises from the updating rule as *retrospective bias*, since it arises following the signal realization. An updating rule can capture many common biases studied in the literature. For example, suppose $\Omega = \{\omega_1, \omega_2\}$ and define the

[11] Given our focus on belief updating following a single signal realization, this definition of an updating rule is for a fixed prior. In Section 6.2, we define an updating rule as a mapping from the signal and prior to a posterior in order to study a dynamic signal process.

[12] This describes an equivalence class of updating rules that differ on a set of measure 0 with respect to ν (and thus with respect to all distributions considered). Fix h_B as some arbitrary member of this class.

biases with respect to the belief that the state is ω_2, i.e. $h(z)_2$. Partisan bias in favor of ω_2 is captured by $h(z)_2 = h_B(z)_2^\alpha$ for some $\alpha \in (0,1)$, a counting updating rule is captured by $\mathcal{Z} = \{\omega_1, \omega_2\}^K$ for some $K \in \mathbb{N}$ and $h(z)_2 = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}_{z_k=\omega_2}$, confirmation bias is captured by $h(z)_2 \geq h_B(z)_2$ if $p_2 \geq 1/2$ and $h(z)_2 \leq h_B(z)_2$ if $p_2 \leq 1/2$, $h(z)_2 = \alpha p_2 + (1-\alpha)h_B(z)_2$ captures linear underreaction for $\alpha \in (0,1)$ and overreaction for $\alpha > 1$, $\frac{h(z)_2}{h(z)_1} = \frac{p_2}{p_1}\left(\frac{d\mu_2}{d\mu_1}(z)\right)^\beta$ captures geometric overreaction for $\beta > 1$ and underreaction for $\beta \in (0,1)$, and base rate neglect is captured by $\frac{h(z)_2}{h(z)_1} = \left(\frac{p_2}{p_1}\right)^\alpha \frac{d\mu_2}{d\mu_1}(z)$ for some $\alpha \in (0,1)$.

While an updating rule captures how an agent forms beliefs retrospectively after observing the signal, it does not specify what an agent thinks prospectively about what her posterior belief will be. Such prospective beliefs are a crucial component of many economic settings where there is an ex-ante decision before the signal is observed (e.g. what information to acquire or pay attention to, whether to pursue a new project before learning about its profitability). an agent must also predict how she will form future beliefs. This is captured by the agent's *forecast*, which specifies a distribution over posterior beliefs. The forecast is also a necessary component in settings with strategic interaction and social learning.

Definition 2 (Forecast)**.** *A* forecast $\hat{\rho}$ *is a Borel probability measure over $\Delta(\Omega)$ for which there exists a measurable $g : \mathcal{Z} \to \Delta(\Omega)$ such that $\mu \circ g^{-1}$ and $\hat{\rho}$ are mutually absolutely continuous.*

The second part of the definition describes a condition to ensure that the forecast is compatible with the signal. The space of posteriors cannot be "larger" than the space of signal realizations, since each signal realization maps to a unique posterior. In the case of a finite support $\mathcal{Z}$, this condition is straightforward—it requires that the cardinality of the support of the forecast is less than or equal to the cardinality of $\mathcal{Z}$. In the case of an infinite $\mathcal{Z}$, the condition is a bit more nuanced—it uses mutual absolute continuity to relate the measure-zero sets of the forecast to the measure zero sets of the information structure.

For a given updating rule h, we define the *accurate* forecast with respect to h as

$$\rho_h(X) \equiv \mu(\{z : h(z) \in X\}) \tag{2}$$

for any Borel set $X \in \Delta(\Omega)$. This is well-defined since h is measurable. We denote the special case of the accurate forecast with respect to Bayes rule as $\rho_B(X) \equiv \mu(\{z : h_B(z) \in X\})$.

Bias can also enter through the forecast. We refer to such bias as *prospective bias*, since it stems from a prediction of what the signal will be. For example, suppose

$\Omega = \{\omega_1, \omega_2\}$ and denote the posterior belief by the belief that the state is ω_2. When the accurate forecast with respect to Bayes rule is uniform on $[0,1]$, then overprecision is captured by a distribution that overweights extreme beliefs and underweights intermediate beliefs, while underprecision overweights intermediate beliefs and overweights extreme beliefs.

Given that updating rules are more frequently the object of focus in the non-Bayesian learning literature, one goal of this paper is to construct reasonable forecasts and analyze how they interact with different updating rules. In this vein, we construct two classes of forecasts with compelling properties in Section 4.

The Misspecified Model Approach. This approach defines an agent's subjective model of the signal process. Posterior beliefs and predictions of posterior beliefs are both pinned down by this model and Bayes rule.

A *misspecified model* is a family of subjective measures over the signal space that is not equal to the family of true measures. We focus on misspecified models where μ_i and $\hat{\mu}_i$ are mutually absolutely continuous for all $i = 1, ..., N$.[13]

Definition 3 (Misspecified Model)**.** *A* misspecified model *corresponds to* $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ *such that there exists an* $\omega_i \in \Omega$ *where* $\hat{\mu}_i \neq \mu_i$*.*

An agent with a misspecified model uses Bayes rule as defined in Eq. (1) to form her posterior belief with respect to her subjective measures. Mutual absolute continuity with respect to the correct model implies that no set of signal realizations that the misspecified model assigns zero probability occur with positive probability under the correctly specified model, and that the misspecified model does not assign positive probability to sets of signal realizations that occur with probability zero under the correctly specified model. It also implies that $\hat{\mu}_i$ and $\hat{\mu}_j$ are mutually absolutely continuous for each $i, j = 1, ..., N$, since μ_i and μ_j are mutually absolutely continuous. Let $\hat{\mu} \equiv \sum_{i=1}^N p_i \hat{\mu}_i$ denote the subjective unconditional signal measure (note this depends on the prior).

It follows directly from Bayes rule and mutual absolute continuity that a misspecified model induces an updating rule. Specifically, $(\hat{\mu}_i)_{\omega_i \in \Omega}$ induces posterior belief

$$\frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z)}{\sum_{j=1}^N p_j \frac{d\hat{\mu}_j}{d\nu}(z)} \tag{3}$$

that the state is ω_i following signal realization z. A model also induces a forecast, which is the unconditional distribution of posteriors according to the model. Specifically, $(\hat{\mu}_i)_{\omega_i \in \Omega}$

[13] This implies that $\frac{d\mu_i}{d\nu}(z) = 0$ iff $\frac{d\hat{\mu}_i}{d\nu}(z) = 0$ except on a set of ν-measure 0. It also implies that $\hat{\mu}_i$ is absolutely continuous with respect to ν for all $i = 1, ..., N$.

induces forecast

$$\hat{\mu}\left(\left\{z : \left\{\frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z)}{\sum_{j=1}^N p_j \frac{d\hat{\mu}_j}{d\nu}(z)}\right\}_{\omega_i \in \Omega} \in X\right\}\right) \tag{4}$$

that the posterior belief is in Borel set $X \in \Delta(\Omega)$.

In order to focus on errors in interpreting signals, this set-up implicitly assumes that the agent has a correctly specified prior belief. The misspecified learning literature has also studied settings with a misspecified prior (Fudenberg et al. 2017). In Section 6.1, we augment the model to also include a subjective prior and derive an analogue of our main result.

2.3 Defining a Representation

The goal of this paper is to connect these two approaches. Specifically, we seek to characterize when different updating rules and forecasts can be represented as a misspecified model. To this end, we define what it means for a misspecified model to represent an updating rule or a forecast. In particular, a model represents an updating rule if it prescribes the same posterior belief as the updating rule following each signal realization, and a model represents a forecast if it prescribes the same ex-ante distribution over posterior beliefs as the forecast.

Definition 4 (Representing Updating Rules and Forecasts)**.**

1. *An updating rule h is represented by misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega}$ if, for every signal $z \in \mathcal{Z}$, an agent who uses Bayes rule to update her posterior with respect to this misspecified model forms the beliefs prescribed by the updating rule μ-almost everywhere:*

$$\frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z)}{\sum_{j=1}^N p_j \frac{d\hat{\mu}_j}{d\nu}(z)} = h(z)_i. \tag{5}$$

2. *A forecast $\hat{\rho}$ is represented by misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega}$ if, for every Borel set $X \subset \Delta(\Omega)$:*

$$\hat{\mu}\left(\left\{z : \left(\frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z)}{\sum_{j=1}^N p_j \frac{d\hat{\mu}_j}{d\nu}(z)}\right)_{\omega_i \in \Omega} \in X\right\}\right) = \hat{\rho}(X). \tag{6}$$

Given that we also focus on characterizing when a representation is unique, we next formalize our notion of uniqueness. If an updating rule maps multiple signal realizations to the same posterior belief and can be represented by a given misspecified model, then any other model that shifts mass between these signal realizations will also represent this updating rule. However, the difference between these models is trivial in an economic sense, since they all prescribe the same distribution over realized beliefs and they all induce the same forecast. Therefore, we define the following notion of *essential uniqueness* to capture the idea that the representation is unique in terms of the model features that are relevant for beliefs about the state and decisions.

Definition 5 (Essentially Unique Representation)**.** *An updating rule h has an* essentially unique *representation if all misspecified models representing h are equivalent when restricted to sets of signal realizations in the σ-algebra generated by h, i.e. $\mathcal{F}_h \equiv \{Z \in \mathcal{F} : Z = h^{-1}(X)$ for some Borel set $X \subset \Delta(\Omega)\}$.*

Informally, an updating rule has an essentially unique representation when any misspecified model representing the updating rule is equivalent on the sets of signal realizations that map to the same posterior belief.

2.4 Discussion of Two Approaches

A fundamental aspect of behavioral learning models, which separates them from most fully rational models, is the distinction between "prospective" and "retrospective" belief formation (see, e.g., Benjamin et al. (2016, 2019)). The way a behavioral agent forecasts her future behavior may be in some sense different from how she formed beliefs in the past. This is common in the literatures on time consistency, projection bias, reference dependence, and self-control, and relates to the two components of our behavioral learning set-up. We formalize retrospective bias in the form of an updating rule and prospective bias in the form of a forecast. While misspecified models are generally time-consistent, misspecification allows for inconsistency with respect to predicted versus actual beliefs that is similar in spirit to time inconsistency. In particular, in misspecified settings, the distribution an agent expects her future beliefs and behavior to be drawn from is fundamentally different from the distribution her past behavior was actually drawn from. See Section 6.3 for further discussion of time consistency in our framework.

The updating rule approach is often used to model a specific form of bias or belief-updating error. In general, this literature chooses a reasonable parameterization for a bias, and studies how this parameterization impacts beliefs and behavior. In contrast, the misspecified model approach is often applied to general learning environments that can capture a range of biases within the same framework. For example, recent work in the misspecified learning literature establishes general convergence results for a large class of misspecified models (Bohren and Hauser 2021; Frick et al. 2020b; Fudenberg et al. 2021). Connecting these approaches makes it straightforward to apply the tools developed in the misspecified learning literature to extend the results from the updating rules literature to a larger set of parameterizations of a given bias. For instance, in Bohren and Hauser (2019), we use these tools to generalize the learning results from Rabin and Schrag (1999) to a larger set of updating rules that capture the conceptual features of confirmation bias. This establishes that the qualitative insights of Rabin and Schrag (1999) do not rely on their specific parameterization of confirmation bias or choice of information structure (i.e. binary signals).

To a large extent, the theoretical and empirical literature on behavioral biases has focused on updating rules, which are a simple way to define and express biases. But updating rules are 'incomplete' in that on their own, they do not pin down all aspects of belief formation required for economic analysis. Since a misspecified model of belief formation is complete, in the sense that it describes all aspects of the environment necessary for analysis, mapping updating rules into misspecified models makes it possible to study the implications of a given bias in a richer set of economic environments.

3 Representing Updating Rules and Forecasts

This section derives our main representation result. We first establish a necessary and sufficient condition for an updating rule to be represented by a misspecified model, and analogously for a forecast. We then establish a necessary and sufficient condition on an updating rule and forecast pair for it to be jointly represented by a misspecified model and show that this model is essentially unique.

3.1 Representing Updating Rules

We begin by fixing an updating rule and characterizing when it can be represented by a misspecified model. An important feature of Bayesian updating is that the posterior belief is equal to the prior in expectation. Therefore, given the set of posterior beliefs induced by the updating rule, it must be possible to find a misspecified model that satisfies this property. We use this martingale property of beliefs to characterize necessary and sufficient conditions for there to exist a misspecified model that represents the updating rule.

Let $\mathcal{N}(h) \equiv \operatorname{supp} \rho_h$ denote the support of the accurate forecast ρ_h for updating rule h (that is, the set of posteriors that arise from h), and let

$$S(h) \equiv \operatorname{rel\,int}(\operatorname{Conv} \mathcal{N}(h)) \tag{7}$$

denote the relative interior of the convex hull of this support.[14] We say that an updating rule is *responsive* if the prior belief lies inside this set of posterior beliefs.

Definition 6 (Responsive Updating Rule). *An updating rule is* responsive *if* $p \in S(h)$.

[14] Recall that the relative interior of a set S is the set of points that are on the interior of S within its affine hull.

Many non-Bayesian updating rules considered in the literature are responsive, including all of the examples discussed in Section 2.2. It is not satisfied for pathological updating rules such as one in which beliefs move towards a given state following all signal realizations. It can also be violated at certain parameters in updating rules that capture base rate neglect (Benjamin et al. 2019) and cognitive noise (Woodford 2020).[15]

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

Lemma 1 (Updating Rule Representation)**.** *There exists a model* $(\hat{\mu}^i)_{\omega_i \in \Omega}$ *with* $\hat{\mu}_i \in \Delta^*(\mathcal{Z})$ *that represents updating rule* h *if and only if* h *is responsive.*

This result extends Lemma 1 from Shmaya and Yariv (2016) to a more general signal space.[16] Some care must be taken here, 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 reference measure ν and has non-zero Radon-Nikodym derivatives. The space of posterior beliefs has more structure than the signal space, which we leverage via $S(h)$ for this characterization.

The condition in Lemma 1 is very weak. As discussed above, it holds for many of the non-Bayesian updating rules that have been considered in the literature. Therefore, this result establishes that most non-Bayesian updating rules of interest can be represented by a misspecified 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. As we illustrate in the following example, there are often many distinct misspecified models that represent a given updating rule and each representation induces a different forecast. Therefore, the choice of representation determines the prospective bias. Different representations can lead to different predictions precisely when a forecast is needed to close the model.

Example 1. *Consider binary state space* $\Omega = \{L, R\}$ *with a flat prior* $p_1 = 1/2$ *and signal space* $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$*. In a slight abuse of notation, when the state space is binary we can 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 forecast as a distribution* $\hat{\rho}$ *over a set of probabilities that the state is* R*. Note* $|\operatorname{supp} \hat{\rho}| \leq 4$ *since a*

[15] The key feature of these updating rules that leads to a violation is that the bias manipulates the prior belief. In Section 6, we extend our framework to allow for a misspecified prior and show that such updating rules are responsive with respect to this misspecified prior.

[16] In Shmaya and Yariv (2016), $S(h)$ is the relative interior of the convex hull spanned by posteriors. Our set $S(h)$ is the analogue of this set with the additional measurability restrictions necessary for this to be well-defined on infinite signal spaces.

signal cannot map to multiple beliefs. 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 $\sum_{k=1}^4 m_{\omega,k} = 1$.

A responsive updating rule maps at least one signal to a posterior above the prior and one signal to a posterior below the prior, i.e. $\min_k h(z_k) < 1/2 < \max_k h(z_k)$*. Given a responsive updating rule h, any solution $(m_1, m_2, m_3, m_4) \in \Delta$ to $\sum_{k=1}^4 h(z_k)m_k = 1/2$ pins down a model with $m_{R,k} = 2h(z_k)m_k$ and $m_{L,k} = 2(1-h(z_k))m_k$ for $k = 1, \dots, 4$ that represents h.*[17] *Aside from knife-edge cases, $\sum_{k=1}^4 h(z_k)m_k = 1/2$ has multiple solutions, and therefore, h has multiple representations. For example, if $h(z_1) = .1$, $h(z_2) = .2$, $h(z_3) = .8$ and $h(z_4) = .9$, then $(.2, .3, .3, .2)$ and $(.1, .4, .4, .1)$ are both solutions (in fact, there are a continuum of solutions). Note that each model induces a unique forecast, which assigns probability $m_k = m_{R,k}/2 + m_{L,k}/2$ to posterior belief $h(z_k)$.*

Appendix D.2 provides an additional example of the construction and non-uniqueness of a misspecified model representation of the non-Bayesian updating rule of over- and underreaction in Epstein et al. (2010).

3.2 Representing Forecasts

We next develop an analogous result to Lemma 1 for 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 7 (Plausible Forecast)**.** *A forecast $\hat{\rho}$ is* plausible *if $\int_{\Delta(\Omega)} x_i d\hat{\rho}(x) = p_i$ for each $\omega_i \in \Omega$.*

In other words, a forecast is *plausible* if the expected posterior, taken with respect to 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 misspecified model, it must be plausible—a misspecified agent does not believe that she is systematically biased. This condition is also sufficient for a representation to exist—for any plausible forecast, it is possible to find a misspecified model that induces it. This is the analogue in our setting of the result in Kamenica and Gentzkow (2011), who show that a distribution over posteriors can be induced by some information structure if and only if it satisfies the martingale property.

[17] 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.

Lemma 2 (Existence of a Forecast Representation)**.** *There exists a model* $(\hat{\mu}^i)_{\omega_i \in \Omega}$ *with* $\hat{\mu}_i \in \Delta^*(\mathcal{Z})$ *that represents forecast* $\hat{\rho}$ *if and only if* $\hat{\rho}$ *is plausible.*

The condition in Lemma 2 is relatively strong compared to Lemma 1. Unlike the updating rule, which needs very little structure to be consistent with a misspecified model, a forecast must satisfy a strong requirement of Bayesian learning. However, while plausibility rules out many forecasts (e.g. forecasts that systematically slant posteriors towards one state), it still allows for a broad class of forecasts, as illustrated below in Example 2. Moreover, by also allowing for a misspecified prior, a broader class of prospective biases than those that satisfy plausibility with respect to the correct prior can be represented by a misspecified model. We explore this extension in Section 6.1.

As in the case of updating rules, a forecast on its own generally does not identify a unique misspecified model. In fact, a continuum of misspecified models can be consistent with a given forecast. Each model is associated with a different induced updating rule. Therefore, the choice of model to represent a given forecast determines the retrospective bias. Different models that represent the same forecast can lead to very different predictions depending on the updating rule they induce.

The following two examples demonstrate the notion of plausible forecasts and provide an illustration of the multiplicity of representations.

Example 1 (continued)**.** *Return to the set-up introduced in Section 3.1 with* $\Omega = \{L, R\}$, $p_1 = 1/2$, *and* $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$. *A forecast* $\hat{\rho}$ *is plausible if* $\sum_{x \in \operatorname{supp} \hat{\rho}} x\hat{\rho}(x) = 1/2$. *For example, the forecast* $\hat{\rho} = \{.5, .5\}$ *with support* $\{x, 1-x\}$ *for some* $x \in (0, .5)$ *is plausible since* $.5x + .5(1-x) = .5$. *One such model that represents this forecast is* $m_{R,1} = x/2$, $m_{R,2} = x/2$, $m_{R,3} = (1-x)/2$ *and* $m_{R,4} = (1-x)/2$ *in state* R, *and similarly for state* L *substituting* $1-x$ *for* x.[18] *This model induces updating rule* $h(z_1) = h(z_2) = x$ *and* $h(z_3) = h(z_4) = 1-x$.[19] *Alternatively, the model* $m_{R,1} = x/3$, $m_{R,2} = x/3$, $m_{R,3} = x/3$ *and* $m_{R,4} = 1-x$ *in state* R, *and similarly for state* L *substituting* $1-x$ *for* x, *also represents* $\hat{\rho}$. *This model induces a different updating rule: it maps* $\{z_1, z_2, z_3\}$ *to posterior* x *and* z_4 *to posterior* $1-x$. *In fact, for any updating rule that assigns at*

[18] To see that this model represents $\hat{\rho}$, note that from Bayes rule, it induces posterior belief $m_{R,k}/(m_{R,k} + m_{L,k})$ following signal z_k. This simplifies to posterior belief x following z_1 and z_2 and posterior belief $1-x$ following z_3 and z_4. Therefore, it induces forecast $\hat{\rho}(x) = \hat{\mu}(\{z_1, z_2\}) = (m_{R,1} + m_{L,1})/2 + (m_{R,2} + m_{L,2})/2 = .5$ and $\hat{\rho}(1-x) = \hat{\mu}(\{z_3, z_4\}) = .5$ by an analogous calculation, as desired.

[19] In fact, any $\alpha \in (0, 1)$ pins down a model that represents $\hat{\rho}$ with signal distribution $m_{R,1} = \alpha x$, $m_{R,2} = (1-\alpha)x$, $m_{R,3} = \alpha(1-x)$ and $m_{R,4} = (1-\alpha)(1-x)$ in state R, and similarly for state L substituting $1-x$ for x. For each α, the corresponding model induces updating rule $h(z_1) = h(z_2) = x$ and $h(z_3) = h(z_4) = 1-x$. Therefore, all models in this class induce the same forecast and updating rule, and hence, their difference is economically irrelevant. This motivates our notion of essential uniqueness in Definition 5.

least one signal to each posterior x and $1-x$, it is possible to find a model that induces this updating rule and represents $\hat{\rho}$. As discussed above, different updating rules induce different retrospective biases. For example, if the updating rule generated by the correct model maps $\{z_1, z_2\}$ to posterior x, then mapping $\{z_1, z_2, z_3\}$ to x corresponds to slanting information towards state L, whereas mapping $\{z_1, z_3\}$ to x corresponds to inverting the interpretation of z_2 and z_3.

Example 2. *Suppose there are two equally likely states of the world $\Omega = \{L, R\}$. Let $\mathcal{Z} = [0,1]$ and $\mathcal{F}$ be the Borel σ-algebra, and let 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 probability density function of the forecast:*

$$d\hat{\rho}_\theta(x) = \frac{x_L^{\theta-1}(1-x_L)^{\theta-1}}{\Gamma(\theta)^2/\Gamma(2\theta)} \tag{8}$$

for $\theta > 0$, where $x = (x_L, x_R)$ is a posterior belief.[20] *This corresponds to the family of beta distributions with mean $1/2$. Any forecast from this family is plausible since $\int_{\Delta(\Omega)} x_i \, d\hat{\rho}_\theta(x) = 1/2$ for $\omega_i \in \Omega$.*

To illustrate the multiplicity of representations, consider the case of $\theta = 1$. This corresponds to the uniform forecast, i.e. $d\hat{\rho}_1(x) = 1$. 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 $\hat{\rho}_1$.[21] *From Bayes rule, this model induces updating rule $h(z)_R = d\hat{\mu}_R(z)/(d\hat{\mu}_R(z) + d\hat{\mu}_L(z)) = z^\gamma$. Each value of γ captures a different level of retrospective bias: as γ increases, the updating rule slants information more towards state R.*

3.3 Decomposition

As shown above, an updating rule or 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 this place on the set of forecasts that are compatible with it for a representation? In other words, does fixing a retrospective bias restrict the set of feasible prospective biases, and vice versa? Second, given an updating rule and forecast that are jointly compatible with a representation, are these two parts sufficient to pin down a unique representation, or does a model contain additional restrictions on belief formation? Our next result answers these questions.

A necessary condition for a forecast to be compatible with a given updating rule, in

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

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

that the pair can be jointly represented by a misspecified model, is that the support of the updating rule and the forecast are the same.

Definition 8 (No 'unexpected' beliefs)**.** *An updating rule h and forecast $\hat{\rho}$ satisfy* no unexpected beliefs *if $\hat{\rho}$ is mutually absolutely continuous with the accurate forecast with respect to h, ρ_h.*

In other words, it is not possible for an agent to arrive at an entirely unexpected posterior. Additionally, she cannot assign positive probability to a set of posteriors that will never eventuate. Any set of posteriors that the agent anticipates holding with positive probability also arise with positive probability given her updating rule. Importantly, this does not rule out the possibility that the agent has an incorrect expectation about her posterior. The predicted and actual probabilities of holding a given set of posteriors can differ—and indeed do whenever the agent exhibits prospective bias.

Our main result shows that 'no unexpected beliefs', together with the conditions for the updating rule and forecast to be individually represented (i.e. responsive and plausible), are necessary and sufficient for determining whether an updating rule and a forecast can be jointly represented by a misspecified model. Moreover, plausibility and 'no unexpected beliefs' imply that the updating rule is responsive—and hence, this condition is redundant. We also show that the representation is unique and provide a construction.

Theorem 1 (Decomposition)**.** *Consider an updating rule h and a forecast $\hat{\rho}$. There exists a model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ that represents h and $\hat{\rho}$ if and only if (i) $\hat{\rho}$ is plausible and (ii) h and $\hat{\rho}$ satisfy no unexpected beliefs. When such a representation exists, it is essentially unique and satisfies*

$$\hat{\mu}_i(Z) = \frac{1}{p_i} \int_Z h(z)_i d\hat{\rho}(h(z)) \tag{9}$$

for any measurable set of signal realizations $Z \in \mathcal{F}_h$ and $i = 1, ..., N$. This model is misspecified unless $h = h_B$ μ-almost everywhere and $\hat{\rho} = \rho_B$.

This result shows that if we take a forecast and an updating rule that satisfy plausibility and no unexpected beliefs, then we can find a misspecified model to represent it. It also answers the reverse question: If a forecast and an updating rule are induced by a misspecified model, then what properties must they satisfy? The answer is that the forecast must be plausible and the pair must satisfy no unexpected beliefs. Together, this tells us that not only are plausibility and no unexpected beliefs necessary consequences of the misspecified model approach, but they encompass *all* of the belief formation restrictions implicit in using this approach.

This result has several important theoretical and empirical implications. From a theoretical perspective, it shows that the updating rule and the forecast are the "essential" components of a misspecified model. Together they uniquely pin down a complete model for analysis and capture all features of a misspecified model that impact behavior (e.g. how behavior will depart from that of a correctly specified agent). Thus, a misspecified model can be decomposed into the two forms of bias it induces: the prospective bias through the forecast and the retrospective bias through the updating rule.

Moreover, these components are largely independent from each other: aside from 'no unexpected beliefs', the forecast places no further restrictions on which updating rules it can be paired with and vice versa. Given that 'no unexpected beliefs' is a relatively mild condition, especially for sufficiently rich signal spaces, this shows that a given retrospective bias does not place very strong restrictions on the prospective bias that a misspecified model induces. For instance, optimistic updating does not imply optimistic forecasting. This is an appealing property of the misspecified model approach, as it shows that it can be used to capture the interaction between different natural biases.

Second, the result provides a powerful tool for the construction of models of biased belief formation. Rather than needing to specify a family of conditional probability distributions—which is potentially quite complicated and removed from the conceptual biases of interest—we can simply write down a reasonable parameterization of the desired retrospective and prospective biases and use these components to construct a model. Section 5 illustrates this in two applications.

On the empirical side, the updating rule and the forecast can both be identified from belief data (see e.g. Danz et al. (2022) 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. Finally, relatively simple parameterizations of updating rules or forecasts are often used in empirical analysis. When one would like to connect the estimates from such analysis with a misspecified model—for instance, to capitalize on the rich set of theoretical results that have been developed for misspecified models—then one needs to simply make sure that the desired parameterization satisfies the given conditions.

Before outlining the proof of Theorem 1, we present a technical corollary. The forecast doesn't place structure on how mass is allocated within a set of signal realizations $h^{-1}(x)$ that induce the same posterior x. This is why the construction in Eq. (9) is for the sigma-algebra generated by h, i.e. $\mathcal{F}_h$. The following result constructs a representation on the underlying sigma-algebra on the signal space $\mathcal{Z}$, i.e. $\mathcal{F}$. It uses the correctly specified unconditional signal distribution to allocate mass within sets of signal realizations that map to the same posterior. Specifically, the likelihood allocated to a subset of signals that map to a given posterior is equal to the true likelihood of this subset of signals relative to the true likelihood of the set of signals that map to this posterior. This is a simple way to ensure that the constructed misspecified model is mutually absolutely continuous with the correctly specified model, as required by our set-up.

Corollary 1 (Construction of Representation)**.** *Consider an updating rule h and a forecast $\hat{\rho}$, and let ρ_h be the accurate forecast for h. The following model represents h and $\hat{\rho}$:*

$$\hat{\mu}_i(Z) = \frac{1}{p_i} \int_Z h(z)_i \frac{d\hat{\rho}}{d\rho_h}(h(z))\, d\mu(z) \tag{10}$$

for any measurable set of signal realizations $Z \in \mathcal{F}$ and $i = 1, ..., N$. This model is equal to Eq. (9) on $\mathcal{F}_h$.

Note that while other constructions are possible, all such models are equivalent on $\mathcal{F}_h$, as required for essential uniqueness.

Intuition for proof of Theorem 1. To establish this result, we first prove an intermediate result that significantly simplifies the process of finding misspecified model(s) to represent a given updating rule. Given either an unconditional distribution $\hat{\mu}$ or a state-contingent distribution $\hat{\mu}_i$ in state ω_i, we establish a necessary and sufficient condition for this distribution to be part of a misspecified model representing the updating rule. Moreover, if a model that includes this distribution exists, we show that this single distribution *uniquely* pins down the remainder of the model—in other words, all of the other state-contingent distributions. When the condition is not satisfied, then the updating rule is incompatible with the given distribution and it cannot be part of a representation.

This intermediate result implies a restriction on how a forecast and updating rule must jointly behave over measure 0 sets in order to be represented by the same misspecified model. Specifically, to be compatible with an updating rule, a forecast cannot place positive probability on a set of posteriors that are associated with a measure zero set of signals under the updating rule. This corresponds to the no unexpected beliefs condition. It is straightforward to see why this condition is necessary to find a misspecified model to jointly represent the forecast and updating rule. We show that this condition is also sufficient, and therefore, is the only joint requirement on the updating rule and forecast for such a representation to exist.

The following example illustrates the no unexpected beliefs condition and Theorem 1.

Example 1 (continued)**.** *Recall $\Omega = \{L, R\}$, $p_1 = 1/2$, and $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$. Con-*

sider the plausible forecast $\hat{\rho} = \{.5, .5\}$ with support $\{x, 1-x\}$ for some $x \in (0, .5)$. Then any updating rule with $h(z) \in \{x, 1-x\}$ for all $z \in \mathcal{Z}$ satisfies no unexpected beliefs. Consider $h(z_1) = h(z_2) = x$ and $h(z_3) = h(z_4) = 1-x$. Given that h maps $\{z_1, z_2\}$ to the same posterior and similarly for $\{z_3, z_4\}$, the σ-algebra generated by h is $\mathcal{F}_h = \{\emptyset, \{z_1, z_2\}, \{z_3, z_4\}, \mathcal{Z}\}$. From Eq. (9), h and $\hat{\rho}$ have an essentially unique representation that satisfies $\hat{\mu}_R(\{z_1, z_2\}) = x$ and $\hat{\mu}_R(\{z_3, z_4\}) = 1-x$ in state R and $\hat{\mu}_L(\{z_1, z_2\}) = 1-x$ and $\hat{\mu}_L(\{z_3, z_4\}) = x$ in state L. Applying Corollary 1, one such representation is given by $\hat{\mu}_i(z_k) = \left(\frac{\mu(z_k)}{\mu(z_1)+\mu(z_2)}\right)\hat{\mu}_i(\{z_1, z_2\})$ 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 $k = 3, 4$ and $\omega_i \in \{L, R\}$. Note that this construction uses the true relative likelihood of signals that map to the same posterior under h to pin down the subjective probability of these signals in the representation.

4 Selecting Forecasts

Given that updating rules are much more frequently studied in the literature, a natural next question is how to choose a forecast. Sections 4.1 and 4.2 explore two natural conditions to place on a forecast—introspection-proof and naive consistency—in order to select a representation with certain desirable properties when retrospective bias is the primary focus. The introspection-proof condition selects the forecast that is accurate with respect to the given updating rule—it shuts down any prospective bias. Naive consistency selects the forecast that is accurate with respect to Bayesian updating—it describes an agent who is naive about anticipating her retrospective bias. Combined with an updating rule, each condition uniquely selects a representation, provided one exists. Both conditions are similar in spirit to other conditions that have been used in the literature, but the introspection-proof condition is much more restrictive than naive consistency—it is not satisfied for many common updating rules, and therefore no such representation exists, while naive consistency selects a representation that broadly exists. A forecast can also be chosen to explicitly capture a particular form of prospective bias. In Section 4.3, we highlight a representation that shuts down any retrospective bias in order to focus on prospective bias.

4.1 Introspection-Proof Models

A common concern with the misspecified model approach is that, if an agent observes a lot of data (i.e. many independent draws of the signal), she may observe a pattern that is highly unlikely under her misspecified view of the world. For example, she may observe an extreme violation of the law of large numbers. In such a scenario, introspection may lead the agent to eventually realize that she is misspecified. Motivated by this concern, we define the following notion of an introspection-proof model.

Definition 9 (Introspection-Proof Model)**.** *A model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ with induced unconditional measure $\hat{\mu}$ is* introspection-proof *if $\hat{\mu}(Z) = \mu(Z)$ for all measurable sets of signal realizations $Z \in \mathcal{F}$.*

In an introspection-proof model, the agent's predicted distribution of the signal is equal to the true distribution. Any distortion in belief formation stems from the agent's misperception of the relative likelihood of signal realizations in different states. In such a model, the empirical frequency of signal realizations that the agent observes will be in line with her expectation given her model. Therefore, she is no more likely than a correctly specified agent to observe "unexpected" patterns in the data and come to question her model.

A natural implication of an introspection-proof model is that an agent accurately forecasts her future beliefs. To see this, note that under the misspecified model approach, the agent correctly predicts how she will form her posterior belief following each signal realization (prospective bias stems from the agent misperceiving the likelihood of each signal realization, not from misperceiving her mapping from signal realizations to posterior beliefs). When she also correctly predicts the likelihood of each signal, as required in an introspection-proof model, then she must correctly predict the likelihood of each posterior belief. In other words, an introspection-proof model has no prospective bias. This makes an introspection-proof model a natural choice when a researcher is interested in shutting down prospective bias in order to isolate the impact of a given form of retrospective bias.

Given this property, in any *introspection-proof representation* of an updating rule h—that is, a representation by an introspection-proof model—the updating rule must be paired with the accurate forecast for h, i.e. $\hat{\rho} = \rho_h$. From Theorem 1, we know that an updating rule and forecast can be jointly represented if and only if the forecast is plausible and the forecast and updating rule satisfy no unexpected beliefs. When $\hat{\rho} = \rho_h$, no unexpected beliefs is trivially satisfied. Therefore, the necessary and sufficient condition for an updating rule to have an introspection-proof representation is that the accurate forecast is plausible. It immediately follows from Theorem 1 that when such a representation exists, it is unique and defined by Eq. (11) as outlined in the following proposition.

Proposition 1 (Introspection-Proof Representation)**.** *Consider an updating rule h. There exists an introspection-proof model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ that represents h if and only if the accurate forecast ρ_h is plausible. When such a representation exists, it is unique and* *defined by*

$$\hat{\mu}_i(Z) = \frac{1}{p_i} \int_Z h(z)_i \, d\mu(z) \tag{11}$$

for any measurable set of signal realizations $Z \in \mathcal{F}$ and $i = 1, ..., N$. This model is misspecified unless $h = h_B$ μ-almost everywhere.

Substituting an accurate forecast into the construction of a misspecified model in Theorem 1 yields the introspection-proof model constructed in Eq. (11). Note that this model is unique on the original sigma-algebra $\mathcal{F}$, and not just on the sigma-algebra $\mathcal{F}_h$ generated by h as in Theorem 1. This is because the introspection-proof condition, $\hat{\mu} = \mu$, together with the other conditions uniquely pins down the signal distribution at all signal realizations, including those that induce the same posterior.

The requirement that the accurate forecast is plausible is quite restrictive. Recall that a plausible forecast $\hat{\rho}$ must satisfy $\int_{\Delta\Omega} x_i \, d\hat{\rho}(x) = p_i$ for all i. By change of variables and imposing the introspection-proof condition $\hat{\mu} = \mu$, this becomes $\int_{\mathcal{Z}} h(z)_i \, d\mu(z) = p_i$. So the accurate forecast is plausible only if the updating rule averages to the prior under the *true* unconditional signal distribution.[22] 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 h_B, i.e. $\int_{\mathcal{Z}} h_B(z)_i \, d\mu(z) = p_i$.

Despite this restrictive condition, an introspection-proof representation exists for some forms of retrospective bias. An introspection-proof model must preserve the "center of mass" of beliefs but can otherwise arbitrarily distort the spread of these beliefs. This makes it possible to represent conceptual biases in updating that distort the variance of posterior beliefs, such as conservatism or overreaction, as we illustrate below in Example 3. On the other hand, biases that distort the mean of posterior beliefs, such as partisan bias that systematically slants beliefs in one direction, can never have such a representation. We illustrate this below in Example 4. This condition also requires a certain amount of complexity in how the updating rule distorts beliefs, which prevents many simple updating rules from having an introspection-proof representation. For example, the updating rule implied by the canonical Grether regressions (Grether 1980) and commonly used in other empirical work (i.e. $\frac{h(z)_2}{h(z)_1} = \frac{p_2}{p_1}(\frac{d\mu_2}{d\mu_1}(z))^\beta)$ does not have such a representation.

Introspection-proof representations have important implications for empirical work.

[22] A natural class of biases that may appear to satisfy this condition are those that either over- or underestimate the precision of information, in the sense that the corresponding misspecified model is Blackwell ranked with respect to the true model. But this is not the case. In Appendix C, we provide examples of misspecified models that are Blackwell less informative than the true model but not introspection-proof and misspecified models that are Blackwell more informative than the true model but not introspection-proof.

From the perspective of a researcher who only observes signal realizations, an agent who forms beliefs using an updating rule that admits an introspection-proof representation is indistinguishable from a correctly-specified Bayesian agent. Therefore, this agent will pass any test designed to detect Bayesian updating. In contrast, if an updating rule does not have an introspection-proof representation, then with sufficient data, the analyst will be able to reject the hypothesis that the agent is a correctly-specified Bayesian.

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 an introspection-proof condition pin down the misspecified model.

In Section 5.1, we show how the introspection-proof condition is a natural constraint to impose in a dual-selves model where the first self selects an updating rule for the second self. The first self selects an updating rule that exhibits motivated reasoning, but the introspection-proof constraint limits the magnitude of this bias.

Examples. The following example illustrates how to determine whether an introspection-proof representation exists and construct one when it does.

Example 1 (continued)**.** *Recall $\Omega = \{L, R\}$, $p_1 = 1/2$, and $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$. Consider updating rule $h(z_1) = h(z_2) = x$ and $h(z_3) = h(z_4) = 1-x$ for some $x \in (0, 1)$. The accurate forecast corresponds to $\rho_h(x) = \mu(z_1)+\mu(z_2)$ and $\rho_h(1-x) = \mu(z_3)+\mu(z_4)$, where μ is the correct unconditional model. An introspection-proof representation of h exists if this forecast is plausible, i.e. $x(\mu(z_1)+\mu(z_2))+(1-x)(\mu(z_3)+\mu(z_4)) = 1/2$. Note that this is satisfied if either an equal mass of signals map to each posterior, $\mu(z_1) + \mu(z_2) = 1/2$ and $\mu(z_3) + \mu(z_4) = 1/2$, or the signal is perceived to be uninformative, $x = 1/2$. To construct an introspection-proof representation of h, suppose the correct unconditional distribution is $\mu = (.2, .3, .3, .2)$. From Eq. (11), the unique introspection-proof represen-*

tation is $\hat{\mu}_R = (.4x, .6x, .6(1-x), .4(1-x))$ and $\hat{\mu}_L = (.4(1-x), .6(1-x), .6x, .4x)$.[23]

The next example shows that a common parameterization of conservatism has an introspection-proof representation.

Example 3 (Conservatism)**.** *Consider a common updating rule for conservatism where the posterior belief is a weighted average of the Bayesian posterior and the prior, $h(z) = \lambda h_B(z) + (1-\lambda)p$ for some $\lambda \in (0,1)$ (Epstein et al. 2010; Hagmann and Loewenstein 2019; Gabaix 2019). From Eq. (11), this updating rule is represented by the introspection-proof misspecified model $\hat{\mu}_i \equiv (1-\lambda)\mu_i + \lambda\mu$. Note that the second term in this sum depends on the prior, and hence, the constructed model will vary with the prior. See Section 6.2 for further discussion of prior-dependent versus independent representations.*

The final example shows that a common parameterization of partisan bias does not have an introspection-proof representation.

Example 4 (Partisan Bias)**.** *Consider binary state space $\Omega = \{L, R\}$. As in Example 1, let h denote the mapping from the set of signal realizations to the posterior belief that the state is R and let $\hat{\rho}$ denote the distribution over the posterior belief that the state is R. Consider updating rule $h(z) = h_B(z)^\alpha$ for $\alpha \in (0,1)$. This updating rule exhibits R-partisan bias: after any signal realization, the agent places higher probability on state R than a correctly specified agent. Under the accurate forecast, $\int_0^1 x\, d\rho_h(x) = \int_{\mathcal{Z}} h(z)\, d\rho_h(h(z)) = \int_{\mathcal{Z}} h(z)\, d\mu(z)$, where the first equality follows from change of variables and the second follows from a property of the accurate forecast. But $\int_{\mathcal{Z}} h(z)\, d\mu(z) > \int_{\mathcal{Z}} h_B(z)\, d\mu(z) = p_2$, where the equality follows from Bayes-plausibility (i.e. h_B is plausible). Therefore, the accurate forecast cannot be plausible. This argument clearly applies more generally to any bias that systematically skews posterior beliefs in one direction.*

Alternative Notions of Introspection-Proof. The notion of introspection-proof model in Definition 9 is relatively strong, in that it requires the subjective unconditional signal measure to exactly match the correct unconditional measure. It is possible to define conceptually similar notions with weaker requirements. For example, one could restrict attention to models in which the means of the subjective and correct unconditional signal measures match, but allow these measures to differ on other dimensions (e.g. variance) that may be less salient or harder to observe. Alternatively, one could select the representation for an updating rule that is "closest" to introspection-proof un-

[23] To see that an introspection-proof representation can be misspecified, suppose the true model is $\mu_R = (.4x', .6x', .6(1-x'), .4(1-x'))$ and $\mu_L = (.4(1-x'), .6(1-x'), .6x', .4x')$ for some $x' \in (0,1)$. For any x', this model generates unconditional distribution $\mu = (.2, .3, .3, .2)$. This representation is misspecified when $x \neq x'$.

der some natural measure of distance. Additionally, the notion we define is with respect to the unconditional signal distribution. Analogous results hold when the subjective conditional distribution in a fixed state is required to match the true distribution in this state. See Appendix B for further exploration of these notions.

4.2 Naive Consistent Forecasts

When an agent exhibits bias at a future decision point, a common question is how she anticipates this future bias. The two cases typically explored in the literature are that an agent is sophisticated, in that she accurately anticipates her future bias, and that an agent is naive, in that she believes she will have no future bias. The introspection-proof condition in the previous section is in line with the first case. In this section, we develop the notion of naive consistency to capture the latter case.

The naive consistent forecast captures an agent who naively predicts that she will update beliefs correctly in the future, but when the information arrives, she interprets it with bias.

Definition 10 (Naive Consistent Forecast)**.** *An agent has a* naive consistent forecast *if she has an accurate forecast with respect to the Bayesian updating rule, $\hat{\rho} = \rho_B$, but she updates with bias, $h \neq h_B$ with μ-positive probability.*

An agent who has a naive consistent forecast assigns the same probability as a correctly specified agent to the possibility that she will form a posterior belief in set $X \subset \Delta(\Omega)$. But when information arrives, she actually arrives at a posterior belief in set X with a different probability than the correctly specified agent. In the absence of guidance on the nature of the prospective bias, the naive consistent forecast in some sense shuts down this channel. Before information is realized, an agent using a naive consistent forecast will make exactly the same decisions as a Bayesian agent. Therefore, the agent's bias only alters behavior that occurs after the signal is observed.

The *naive consistent representation* of an updating rule corresponds to the misspecified model that induces this updating rule and the naive consistent forecast. Since ρ_B is the distribution of posteriors that a correctly specified Bayesian generates, it is always plausible. 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 consistent representation. It immediately follows from Theorem 1 that when such a representation exists, it is essentially unique and defined by Eq. (12) as outlined in the following proposition.

Proposition 2 (Naive Consistent Representation)**.** *Consider an updating rule h and assume $h \neq h_B$ with μ-positive probability. There exists a misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ that represents h and induces the naive consistent forecast ρ_B if and only if h*

and ρ_B satisfy no unexpected beliefs. When such a representation exists, it is essentially unique and defined by

$$\hat{\mu}_i(Z) = \mu_i(\{z : h_B(z) \in h(Z)\}) \tag{12}$$

for any measurable set of signal realizations $Z \in \mathcal{F}_h$ and $i = 1, ..., N$.[24]

In the naive consistent representation, an analogue of the naive consistent forecasting property holds in each state: for each ω_i, the naive consistent representation induces a forecast over the posterior that is equal to the forecast of a correctly specified Bayesian. This is a consequence of Lemma 3 and is straightforward to see from Bayes rule.

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 is possible to satisfy for many commonly used updating rules, including all of the examples in Section 2.2. Therefore, in contrast to the introspection-proof representation, a naive consistent representation broadly exists for many retrospective biases of interest.

Naive consistency is analogous to common naiveté assumptions made in many behavioral models (e.g. models of time inconsistency (O'Donoghue and Rabin 1999)). The assumption has previously been used to pin down prospective beliefs in models of biased learning such as for the case of base rate neglect (Benjamin et al. 2019) and social learning with partisan bias and overreaction (Bohren and Hauser 2021). We impose it in Section 5.2 to study how retrospective overprecision impacts search behavior. It has also informally been made in many less detailed behavioral models. Therefore, formalizing how to capture naive consistency in the misspecified model approach shows that we can consistently and rigorously impose such an assumption in this more complete framework.

Examples. The first example illustrates how to construct the naive consistent forecast and corresponding representation for a discrete signal space.

Example 1 (continued). *Recall $\Omega = \{L, R\}$, $p_1 = 1/2$, and $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$. As in Section 4.1, consider the updating rule $h(z_1) = h(z_2) = x$ and $h(z_3) = h(z_4) = 1 - x$ for some $x \in (0, 1)$. Suppose the correct model induces updating rule $h_B(z_1) = h_B(z_2) = h_B(z_3) = x$ and $h_B(z_4) = 1 - x$. Then $\rho_B(x) = \mu(\{z_1, z_2, z_3\})$ and $\rho_B(1 - x) = \mu(z_4)$, where μ is the correct unconditional model. Given that the updating rule h induces set of posteriors $\{x, 1 - x\}$, which is equal to the support of ρ_B, h and ρ_B satisfy no unexpected beliefs. Therefore, a naive consistent representation of h exists. From Eq. (12), this representation is unique on $\mathcal{F}_h = \{\{z_1, z_2\}, \{z_3, z_4\}, \mathcal{Z}\}$ and satisfies*

[24] Alternatively, one could write this representation in analogous form to Eq. (9) as $\hat{\mu}_i(Z) = \frac{1}{p_i}\int_Z h(z)_i d\rho_B(h(z))$ and pin down a full representation by substituting ρ_B for $\hat{\rho}$ in Eq. (10).

$\hat{\mu}_i(\{z_1, z_2\}) = \mu_i(\{z_1, z_2, z_3\})$ and $\hat{\mu}_i(\{z_3, z_4\}) = \mu_i(z_4)$ for $\omega_i \in \{L, R\}$. We can complete the construction of the representation by applying Corollary 1.

In Example 4, we showed that an updating rule that captured partisan bias did not have an introspection-proof representation. We next show that this updating rule does have a naive consistent representation.

Example 4 (Partisan Bias, cont.)**.** *Return to the set-up introduced in Section 4.1 with $\Omega = \{L, R\}$ and $h(z) = h_B(z)^\alpha$ for $\alpha \in (0,1)$. Recall this updating rule slants beliefs towards state R. Consider prior $p_1 = 1/2$ and signal space $\mathcal{Z} = [0,1]$ with true measures $\mu_R(z) = z^2$ and $\mu_L(z) = 2z - z^2$ (in a slight abuse of notation written in cdf form). This induces unconditional measure $\mu(z) = (\mu_R(z) + \mu_L(z))/2 = z$, updating rule $h_B(z) = \frac{1}{1+d\mu_L/\mu_R(z)} = z$ and forecast $\rho_B(x) = Pr(z : h_B(z) \leq x) = \mu(x) = x$ (again in cdf form). The accurate forecast with respect to h is $\rho_h(x) = Pr(z : h(z) \leq x) = Pr(z : z \leq x^{1/\alpha}) = \mu(x^{1/\alpha}) = x^{1/\alpha}$. Given that ρ_B and ρ_h are mutually absolutely continuous, h and the naive consistent forecast ρ_B satisfy no unexpected beliefs. Therefore, a naive consistent representation of h exists. This representation is unique since each signal maps to a unique posterior (i.e. $\mathcal{F}_h = \mathcal{F}$). From Eq. (12), the naive consistent representation is $\hat{\mu}_i(z) = \mu_i(z^\alpha)$ for $\omega_i \in \{L, R\}$.*

4.3 Retrospectively Correct Models

We next consider a representation in which prospective bias is the main focus. Introspection-proof and naive consistent representations both shut down prospective bias in order to isolate the implications of retrospective bias. Analogously, we can consider situations in which an agent correctly interprets signals (i.e. uses the Bayesian updating rule $h_B(z)$) but has a biased forecast. A *retrospectively correct model* shuts down any retrospective bias and only allows for prospective bias.

Definition 11 (Retrospectively Correct Model)**.** *A misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ is* retrospectively correct *if it induces $h_B(z)$, i.e. for all $\omega_i \in \Omega$,*

$$\frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z)}{\sum_{j=1}^N p_j \frac{d\hat{\mu}_j}{d\nu}(z)} = h_B(z)_i \tag{13}$$

μ-almost everywhere.

A misspecified agent with a retrospectively correct model makes the same decisions as a correctly specified agent after information arrives, but can behave differently ex-ante.

The *retrospectively correct* representation of a forecast corresponds to the misspecified model that induces this forecast and the Bayesian updating rule. While introspection-proof and naive consistent representations pin down a forecast with respect to the correctly specified model (the forecast is either accurate with respect to the agent's updating rule or the Bayesian updating rule), a retrospectively correct representation pins down the updating rule with respect to the correctly specified model. The following corollary immediately follows from Theorem 1.

Corollary 2. *Consider a forecast $\hat{\rho}$. There exists a retrospectively correct model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ that represents $\hat{\rho}$ if and only if $\hat{\rho}$ is plausible and h_B and $\hat{\rho}$ satisfy no unexpected beliefs. When such a representation exists, it is essentially unique and satisfies Eq. (9) setting $h = h_B$.*

This establishes that many forecasts are consistent with the Bayesian updating rule. An agent can form 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.

In Section 5.2, we use a retrospectively correct representation to study how prospective overprecision impacts search behavior. The following example illustrates a retrospectively correct representation of a forecast that exhibits overprecision.

Example 1 (continued)**.** *Recall $\Omega = \{L, R\}$, $p_1 = 1/2$, and $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$. Suppose the correct model induces updating rule $h_B(z_1) = .1$, $h_B(z_2) = .2$, $h_B(z_3) = .8$ and $h_B(z_4) = .9$ and forecast $\rho_B = (.1, .4, .4, .1)$ with $\operatorname{supp} \rho_B = \{.1, .2, .8, .9\}$. Consider a forecast $\hat{\rho} = (.4, .1, .1, .4)$ with $\operatorname{supp} \hat{\rho} = \{.1, .2, .8, .9\}$. Note this forecast is plausible and satisfies no unexpected beliefs with respect to h_B. Therefore, there exists a retrospectively correct representation of $\hat{\rho}$. In this representation, the agent exhibits overprecision prospectively, since $\hat{\rho}$ puts more weight on the extreme posteriors and less weight on the interior posteriors relative to the accurate forecast ρ_B. From Eq. (9), the retrospectively correct representation of $\hat{\rho}$ is $\hat{\mu}_R = (.08, .04, .16, .72)$ and $\hat{\mu}_L = (.72, .16, .04, .08)$. Note this model is misspecified, since from h_B and ρ_B, the correctly specified model is $\mu_R = (.02, .16, .64, .18)$ and $\mu_L = (.0.18, .64, .16, .02)$.*

5 Applications

The following two applications demonstrate the results from Sections 3 and 4. In the first, we show how the introspection-proof condition is a natural constraint to impose in a dual-selves model with self-image concerns. This requirement moderates the magnitude of motivated reasoning bias exhibited by the updating rule that the first self selects. In the second, we show how our decomposition can clarify the impact that a misspecified model has on an agent's search decisions. Whether the bias induced by the misspecified model emerges ex-ante versus ex-post to information arrival plays a key role in determining how it impacts search behavior.

5.1 Optimal Bias with Self-Image Concerns

Overview. Consider a dual-selves model where a manager first chooses an updating rule to interpret information about ability, then uses this updating rule to evaluate himself and other workers. The manager observes the group identity—male or female—and a productivity signal for workers as well as himself. He seeks to accurately evaluate workers and also derives utility from his perception that he is of high ability. This self-image concern leads to the selection of an updating rule that exhibits motivated reasoning—it overestimates the ability of male workers. Requiring the updating rule to be introspection-proof bounds the magnitude of this motivated reasoning, relative to no such constraint. Moreover, it leads the manager to compensate for overestimating the ability of male workers by underestimating the ability of female workers, despite group identity being orthogonal to productivity. In contrast, without the introspection-proof constraint, the manager does not distort beliefs about female workers. Therefore, self-image concerns interact with introspection-proof belief formation to generate inaccurate beliefs about both in-group and out-group workers, whereas self-image concerns in isolation only lead to inaccurate beliefs about in-group workers. This illustrates how motivated reasoning can be a driver of discrimination stemming from inaccurate beliefs (Bohren et al. forthcoming; Eyting 2023).[25]

Set-up. Suppose a worker has either low or high ability, $\omega_w \in \{L, H\}$, drawn with equal probability. A manager observes a two-dimensional signal $z_w = (y_w, t_w)$ for the worker but not the worker's ability. The first dimension $y_w \in \{b, g\}$ provides information about the worker's ability, with distribution $Pr(g|H) = Pr(b|L) = \alpha > 1/2$. We refer to this as the worker's test performance. The second dimension is the worker's group identity $t_w \in \{M, F\}$, male or female, which we assume is independent of (y_w, ω_w) and distributed according to $q \equiv Pr(M)$. Analogous to the worker, the manager has unobserved ability $\omega_m \in \{L, H\}$ drawn with equal probability and observes his own test performance $y_m \in \{b, g\}$, which has the same distribution as the worker's test performance. Without loss of generality, assume that the manager's group identity is $t_m = M$, and therefore the manager's own two-dimensional signal is $z_m = (y_m, M)$. The manager's ability and signal are independent of the worker's ability and signal.

The manager's first self chooses an updating rule $h(z)$ for interpreting all signals. The second self observes the realized signals, uses this rule to update his beliefs about his own ability to $h(z_m)$ and the worker's ability to $h(z_w)$, and selects evaluation $a \in [0, 1]$ for the worker. Given that the state is binary, in a slight abuse of notation we let $h(z)$ denote the manager's subjective probability that ability is high following signal z.

[25] Eyting (2023) shows that motivated reasoning leads to distorted belief formation, and hence, inaccurate statistical discrimination in an experimental setting.

The manager cares about accurately predicting the worker's ability and his self-image, captured by the second self's belief $h(z_m)$ that he is of high ability,

$$u(a, \omega_w, z_m) = h(z_m) - c(\mathbb{1}_{\{\omega_w = H\}} - a)^2, \tag{14}$$

where $c > 1/2q(1-\alpha)$ to ensure that the manager puts sufficient weight on accurately evaluating the worker.[26] Each self maximizes the expectation of Eq. (14), where the first self takes this expectation with respect to the correctly specified model before signals are realized and the second self takes this expectation with respect to the chosen updating rule h after signals are realized.

Given updating rule $h(z)$, it is straightforward to see that the second self will choose evaluation $a^*(z_w) = h(z_w)$. Therefore, the first self chooses an updating rule to maximize

$$E[h(z_m) - c(\mathbb{1}_{\{\omega_w = H\}} - h(z_w))^2]. \tag{15}$$

Given that the manager must choose the same updating rule to interpret his own and the worker's signals, the choice of updating rule influences both the payoff from self-image and the payoff from evaluation accuracy. Self-image concerns lead the manager to exhibit motivated reasoning, i.e. to choose an updating rule that inflates the interpretation of test performance for male workers. The desire for accuracy prevents this motivated reasoning from becoming too extreme. This is the key trade-off in selecting an updating rule.

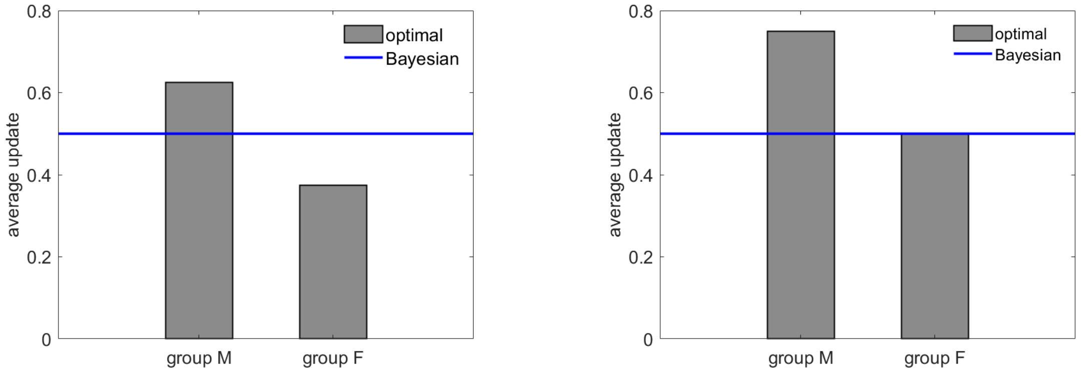
Optimal IP Updating Rule. Suppose the first self wishes to select an updating rule such that, after evaluating a sufficiently large number of workers, the second self does not observe a pattern of signals that is at odds with what he expects—in other words, an updating rule that has an introspection-proof representation. From Proposition 1, an updating rule has an introspection-proof representation if

$$\sum_{y \in \{b,g\}} \frac{1}{2}(qh(y, M) + (1-q)h(y, F)) = \frac{1}{2}. \tag{16}$$

In order to inflate self-image and simultaneously satisfy the introspection-proof condition, the manager must compensate for inflating the test performance of males by deflating the test performance for females. This leads the manager to overestimate the ability of male workers and underestimate the ability of female workers, relative to the Bayesian updating rule $h_B(z)$.

Proposition 3. *The optimal introspection-proof updating rule $h^*(z)$ inflates the inter-*

[26] This condition ensures that the manager does not choose an updating rule that maps a noisy signal into a certain belief about ability.



(a) Introspection-Proof

(b) Unconstrained

FIGURE 1. Optimal average update by group ($\alpha = .7$, $c = 4$, $q = .5$).

pretation of both test outcomes for group M, $h^*(y, M) = h_B(y, M) + \frac{1-q}{2cq}$ for $y \in \{b, g\}$, and deflates the interpretation of both test outcomes for group F, $h^*(y, F) = h_B(y, F) - \frac{1}{2c}$ for $y \in \{b, g\}$.

The optimal introspection-proof updating rule features inaccurate beliefs about both groups. These inaccurate beliefs endogenously emerge from the interaction between self-image concerns and the introspection-proof constraint. Therefore, in settings where managers evaluate a sufficiently large pool of workers such that consistency with the underlying signal distributions is a reasonable requirement, self-image concerns can lead to inaccurate beliefs about other groups even though the manager derives no intrinsic payoff benefit from this distortion.[27] Fig. 1(a) illustrates this result.

The magnitude of the motivated reasoning is decreasing in the share q of male workers. This is because when the hiring pool is more similar to the manager, it becomes more costly for the manager to distort information in a way that improves his self-image, as this distortion leads to a bigger loss from inaccurately evaluating workers. In contrast, the optimal distortion for female workers is independent of their frequency in the population. As the share of male workers increases, the distortion against female workers is less costly since they comprise a smaller share of workers, but it is also less beneficial as a means to balance the distortion against males, since less distortion is desired. It turns out that these two forces exactly balance for the linear-quadratic payoff form in Eq. (14).

Optimal Unconstrained Updating Rule. Without the introspection-proof constraint, self-image concerns still lead the manager to inflate male test performance.

[27] Heidhues, Kőszegi, and Strack (2023) show that overconfidence about own ability can lead to a similar pattern of inaccurate beliefs towards an in-group versus out-group.

However, there is no incentive to distort test performance for females, as it is not necessary to balance the distortion against males.

Proposition 4. *The optimal unconstrained updating rule h_U inflates the interpretation of both test outcomes for group M, $h_U(y, M) = h_B(y, M) + \frac{1}{2cq}$ for $y \in \{b, g\}$, and accurately interprets both test outcomes for group F, $h_U(y, F) = h_B(y, F)$ for $y \in \{b, g\}$.*

In contrast to the optimal introspection-proof updating rule, the optimal unconstrained updating rule only features inaccurate beliefs about the manager's group. However, it features a larger signal distortion for this group relative to the introspection-proof updating rule. This is because without the introspection-proof constraint, distorting test performance is only costly for the manager when he is hiring male workers; the manager stands to lose less from distorting his belief about his own ability, as he does not have to compensate by also distorting the perception of female workers. Thus, the introspection-proof constraint serves as a natural moderator to the magnitude of motivated reasoning bias that can emerge. It also leads to more accurate evaluations overall.[28] Fig. 1(b) illustrates this result.

5.2 Search with Over- and Underprecision

Overview. We explore how overprecision and underprecision—overestimating or underestimating the precision of signals—impact search decisions. Prospective overprecision leads to excess search, while retrospective overprecision leads to too little. The insight reverses for underprecision. This demonstrates that whether a bias manifests ex-ante or ex-post relative to the arrival of information is a key determinant of how it impacts behavior.

Set-Up. A firm is considering whether to adopt one of two new technologies, $j \in \{1, 2\}$. Technology j has either low or high value, $\omega^j \in \{L, H\}$ with $Pr(\omega^j = H) = p \in (0, 1)$. 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 model (μ_L, μ_H) but believes that the signal is drawn from misspecified model $(\hat{\mu}_L, \hat{\mu}_H)$. The signals are independent and perceived to be independent across technologies. Let h denote the updating rule and $\hat{\rho}$ denote the forecast induced by the misspecified model, where in a slight abuse of notation $h(z)$ is the subjective probability that the technology is high

[28] Although the unconstrained case results in less belief distortion for females, the higher distortion for males dominates: the expected loss $E((\mathbb{1}_{\omega=H} - h(z_w))^2)$ from using h_u is $1 - \alpha^2 - (1-\alpha)^2 + 1/8qc^2$, which is larger than the expected loss from using h^*, $1 - \alpha^2 - (1-\alpha)^2 + (1-q)/(8qc^2)$.

value after observing realization z and $\hat{\rho}$ is the subjective distribution over the posterior belief that the technology is high value.

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 $c \in (0, p)$ to search each technology. The firm always searches the first technology since $p > c$. After observing signal realization z^1, it searches the second technology if

$$c < \int_{h(z^1)}^{1} (x - h(z^1))\, d\hat{\rho}(x), \tag{17}$$

where we assume the firm does not search when indifferent.

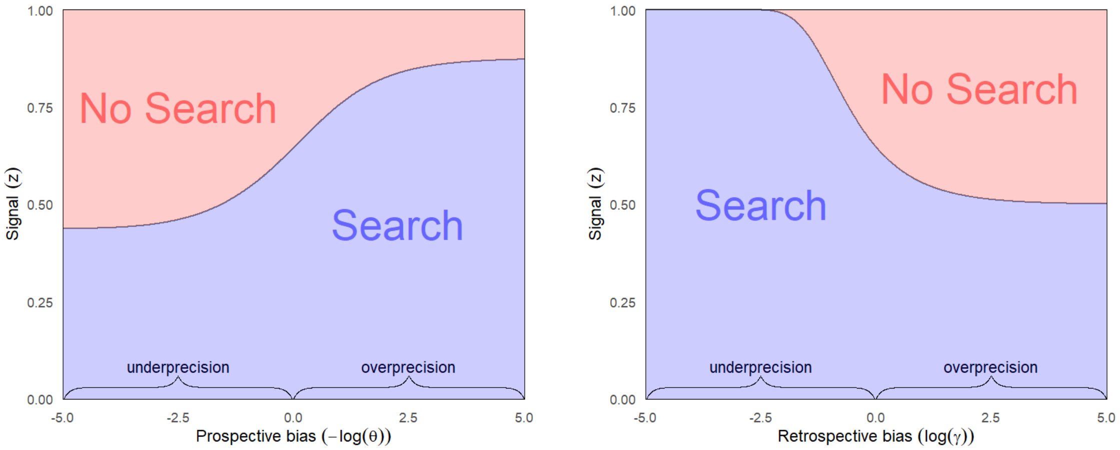
The decomposition of the misspecified model into an updating rule and a forecast isolates the forms of bias it induces. Moreover, as can be seen in Eq. (17), it highlights the way each component of belief formation impacts search behavior. We next examine how search decisions differ based on whether bias enters via the updating rule versus the forecast.

Search with Prospective Bias. First suppose that the firm has prospective bias but no retrospective bias, i.e. $h = h_B$. The firm has an *overprecise* forecast if $\hat{\rho}$ is a mean-preserving spread of the accurate forecast ρ_B, and has an *underprecise* forecast if ρ_B is a mean-preserving spread of $\hat{\rho}$. When the firm has an overprecise forecast, it overweights the likelihood of receiving a very precise signal when it searches the second technology. From Eq. (17), it follows that such a firm searches the second technology too often. In contrast, an underprecise forecast overweights the likelihood of receiving a relatively uninformative signal, which leads to too little search. We formalize this in the following proposition.

Proposition 5. *Let $\mathcal{Z}_S$ denote the set of signal realizations following which a correctly specified firm searches the second technology. Then an overprecise forecast leads to search following signal realizations in a superset of $\mathcal{Z}_S$, while an underprecise forecast leads to search following signal realizations in a subset of $\mathcal{Z}_S$.*

Even though ex-post the firm interprets information correctly, its prospective bias leads to inefficiency.

Search with Retrospective Bias. Now suppose that the firm has retrospective bias and uses the naive consistent forecast, i.e. $\hat{\rho} = \rho_B$. The firm exhibits *overprecision* if it interprets signals as more precise then they actually are, $h(z) < h_B(z)$ when $h_B(z) < 1/2$, $h(z) > h_B(z)$ when $h_B(z) > 1/2$, and $h(z) = 1/2$ when $h_B(z) = 1/2$. It exhibits *underprecision* if it interprets signals as less precise, $h(z) \in (h_B(z), 1/2)$ when $h_B(z) < 1/2$, $h(z) \in (1/2, h_B(z))$ when $h_B(z) > 1/2$, and $h(z) = 1/2$ when $h_B(z) = 1/2$.



(a) Prospective Bias: $h = h_B, \hat{\rho} \sim \text{Beta}(\theta, \theta)$

(b) Retrospective Bias: $h(z)/(1-h(z)) = (h_B(z)/(1-h_B(z))^{\gamma}, \hat{\rho} = \rho_B$

FIGURE 2. Search Decisions with Over- and Underprecision ($p = 1/2$, $c = 1/16$, $\rho_B \sim U[0,1]$).

When the firm has an overprecise updating rule, it overreacts to the signal about the first technology. Relative to an unbiased updating rule, the firm is (weakly) less likely to search the second technology following a good signal realization and (weakly) more likely following a bad realization. Whether more or less search emerges overall depends on the cost of search: for a sufficiently low cost, there is less search and for a sufficiently high cost, there is more. We formalize this in the following proposition.

Proposition 6. *Let $\mathcal{Z}_S$ denote the set of signal realizations following which a correctly specified firm searches the second technology. There exists a $\bar{c} > 0$ such that:*

1. *For $c < \bar{c}$, an overprecise updating rule leads to search following signal realizations in a subset of $\mathcal{Z}_S$ and an underprecise updating rule leads to search following signal realizations in a superset of $\mathcal{Z}_S$.*
2. *For $c > \bar{c}$, an overprecise updating rule leads to search following signal realizations in a superset of $\mathcal{Z}_S$ and an underprecise updating rule leads to search following signal realizations in a subset of $\mathcal{Z}_S$.*

In contrast, as shown above, overprecise forecasting leads to (weakly) more search after any signal realization relative to an unbiased forecast. Fig. 2 illustrates how the decision to search the second technology depends on the level of prospective or retrospective bias.

Taken together, these results show that whether overprecision emerges prospectively versus retrospectively leads to qualitatively different predictions of how the bias impacts search behavior: prospective overprecision leads to more search while the retrospective overprecision leads to less. Therefore, the timing of when the bias emerges has important implications for economic behavior.

6 Extensions

6.1 Misspecified Prior

Recent work on biased learning has also allowed for a misspecified prior (e.g. Fudenberg et al. (2017); Bohren et al. (forthcoming)). Our framework easily extends to such settings. Let $\hat{p}$ denote the subjective prior. The following result is an analogue of Theorem 1. It pins down a unique prior and model of the signal process to represent an updating rule and forecast pair.

Theorem 2 (Decomposition with Misspecified Priors)**.** *Consider an updating rule h and a forecast $\hat{\rho}$. There exists a prior $\hat{p} \in \Delta(\Omega)$ and a model of the signal $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ that represents h and $\hat{\rho}$ if and only if h and $\hat{\rho}$ satisfy no unexpected beliefs. When such a representation exists, it is essentially unique and satisfies*

$$\hat{p}_i = \int_{\Delta(\Omega)} x_i \, d\hat{\rho}(x) \tag{18}$$

and

$$\hat{\mu}_i(Z) = \frac{1}{\hat{p}_i} \int_Z h(z)_i d\hat{\rho}(h(z)) \tag{19}$$

for any measurable set of signal realizations $Z \in \mathcal{F}_h$ and $i = 1, ..., N$. This model is misspecified unless $\hat{p} = p$, $h = h_B$ μ-almost everywhere, and $\hat{\rho} = \rho_B$.

The key difference from Theorem 1 is that, rather than requiring the forecast to be plausible, Theorem 2 uses plausibility to pin down the unique subjective prior. Given this subjective prior, the misspecified model of the signal process is as in Theorem 1.[29]

A wider range of forecasts can be represented by a model with a misspecified prior, as this relaxes the restrictive plausibility condition. Additionally, a wider range of updating rules can be represented by an introspection-proof model with a misspecified prior. In fact, *all* updating rules now have such a representation.[30]

[29] Analogous to Corollary 1, one can use the correctly specified model to pin down a representation on $\mathcal{F}$.

[30] Note that in order for the predicted empirical frequencies of signals to match their true empirical frequencies, the predicted empirical frequencies of states do not match their true empirical frequencies. So although the representation is introspection-proof with respect to the signal distribution, it is not introspection-proof with respect to the state distribution.

6.2 Prior-Independent Representations.

In order to consider dynamic updating for a sequence of signals or comparative statics with respect to the prior, we next extend the definitions of an updating rule and a forecast to allow them to depend on the prior. Specifically, updating rule $h(z,p)$ specifies a posterior belief for each signal realization $z \in \mathcal{Z}$ and prior $p \in \Delta(\Omega)$, and the forecast $\hat{\rho}(x,p)$ specifies the likelihood of each posterior belief x at each prior $p \in \Delta(\Omega)$. Theorem 1 pins down the misspecified model that represents the updating rule and forecast at each prior.

In this expanded framework, an important question is whether an updating rule has a representation that is independent of the prior.

Definition 12 (Prior-Independent Representation)**.** *An updating rule $h(z,p)$ has a* prior-independent representation *if there exists a model $(\hat{\mu}_i)_{\omega_i \in \Omega}$ that represents $h(z,p)$ at all $p \in \Delta(\Omega)$.*

When this property holds, the model representing the updating rule does not vary with the prior belief. This is a conceptually appealing property for biases in which an agent is inherently Bayesian but has a mistaken understanding of the information generating process that does not depend on her current belief. For example, biases such as overreaction and optimism are not intrinsically linked to the agent's current belief. In contrast, the property is conceptually at odds with biases in which the agent's current belief influences her perception of information. For example, an agent's current belief is a key component of confirmation bias, and therefore, an updating rule exhibiting confirmation bias is naturally represented by a model that varies with the prior. As we will discuss below, the property is also at odds with some biases in which an agent is non-Bayesian.

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. When this condition holds, then any model that represents an updating rule at some prior p also represents the updating rule at any other prior p'—and can therefore form a prior-independent representation.

Proposition 7 (Prior-Independent Representation)**.** *Fix an updating rule $h(z,p)$ that is responsive at all $p \in \Delta(\Omega)$. Then $h(z,p)$ has a prior-independent representation if and only if*

$$\frac{p_i}{p_j}\frac{h(z,p)_j}{h(z,p)_i} \tag{20}$$

is independent of p *for all* $p \in \Delta(\Omega)$, $z \in \mathcal{Z}$, *and* $i, j = 1, ..., N$. *When this holds, then any model that represents* $h(z,p)$ *at prior* p *also represents* $h(z,p)$ *at all other priors* $p' \in \Delta(\Omega)$.[31]

This result has an important implication for empirical work. When an updating rule has a prior-independent representation, then identifying the updating rule at one prior pins down the updating rule at all priors.

Many well-known parameterizations of common biases have prior-independent representations. For example, the updating rule capturing geometric overreaction in Section 2.2 and the partisan bias updating rule in Example 4 both have prior-independent representations (see Appendix D.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. There are many biases that are naturally parameterized in a way that inherently varies with the prior. For example, the direction of confirmation bias and the magnitude of base rate neglect depend on the prior. Therefore, updating rules that do not have a prior-independent representation are essential for capturing the essence of these biases (see Page 11 for examples of such updating rules). While less obvious, the linear parameterization of over/underreaction in Epstein et al. (2010) (see Example 3) and the posterior parameterization of partisan bias in Example 4 only admit prior-dependent representations (see Appendix D.1). In the former, even though the over/underreaction parameter is independent of the prior, the additivity of the non-Bayesian updating rule differs structurally from the multiplicative form of Bayes rule. Therefore, it can only be represented in a framework that imposes Bayesian updating by allowing the model to vary with the prior. In the latter, distorting the Bayesian posterior, rather than the signal likelihood, links the magnitude of the bias to the prior even though the parameter is independent of the prior. Similarly, the misspecified causal models from Spiegler (2020) only admit prior-dependent representations.[32]

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

[31] Whenever an updating rule h has at least two representations at some prior p, then trivially, a prior-dependent representation exists. To see this, suppose Eq. (20) holds and consider two models that represent h at prior p. Then both models represent h at all priors. To form a prior-dependent representation, select one model to represent h at a subset of priors and select the other model to represent h at the remaining priors.

[32] While prior-independent representations lend themselves to more straightforward dynamic analysis, prior-dependent representations are still tractable. For example, recent work in the literature on misspecified learning establishes general convergence results in settings where the model varies with the prior (Bohren and Hauser 2021; Frick et al. 2020b).

due to the dependence of the forecast on the prior. This brings us to the following result, which establishes a desirable property of the naive consistent forecast.

Proposition 8 (NCF and Prior-Independence)**.** *Fix an updating rule $h(z,p)$ that has a prior-independent representation. Then the unique representation of $h(z,p)$ and the naive consistent forecast is prior-independent.*

We already know that, by definition, the naive consistent forecast is consistent with the forecast induced by the correctly specified model in a one-period setting. In a dynamic setting with a sequence of signals, the naive consistent forecast 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-τ posterior belief for any $\tau > t$. The representation of the naive consistent forecast and an updating rule with a prior-independent representation induces a period-t forecast over period-τ posterior beliefs that is equal to the period-t forecast of period-τ posterior beliefs in the correctly specified model.

6.3 Time Inconsistency and Prior-Dependent Representations

Time inconsistency is a key property of many dynamic behavioral models. In terms of belief distortions, time inconsistency 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. This means that the model an agent believes she will use in future periods differs from the model she actually uses. A prior-dependent representation is a natural way to capture this property.

Consider the following setting. 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 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 prior p, she believes that she will use the forecast and updating rule induced by model $(\hat{\mu}_i(\cdot;p))_{\omega_i \in \Omega}$ in all future periods. This can lead to dynamically inconsistent behavior. 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.

Prior-dependent models do not always lead to time inconsistency. When the agent accurately anticipates how her model varies with the prior, she will be time consistent. For example, when the correctly specified model varies with the prior—as in active and social learning environments—it is prior-dependent but clearly also time consistent. Alternatively, a biased agent who is sophisticated about her bias and accurately predicts how her future updating rule and forecast vary with her future belief is time consistent. If an agent has a prior-independent representation and believes she will continue to use this same updating rule and forecast at future beliefs, then behavior is time consistent even if the agent is not aware of her bias.

7 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 components of belief formation that are relevant for decision-making—the updating rule and the forecast—and highlights the belief formation restrictions implicit in using the misspecified model approach. Moreover, it demonstrates how one can 'complete' an updating rule through the construction of a forecast. We also identify two natural paths for constructing such a forecast—the introspection-proof model and the naive consistent forecast—and provide necessary and sufficient conditions for these models to exist. Taken together, these results provide a method to embed belief formation biases into economic decision problems when the updating rule on its own is incomplete. They also highlight the importance of eliciting a forecast as well as the commonly measured updating rule in empirical work, as both components of belief formation play a key role in many economic settings.

A Proofs

A.1 Proofs from Section 3

Proof of Lemma 1. (If:) Let $F \equiv \{x : x_i = \int_{\mathcal{Z}} h(z)_i \, d\hat{\mu}(z), \, \hat{\mu} \in \Delta^*(\mathcal{Z})\}$. We first show that $\overline{F} = \overline{S}(h)$, which implies that $S(h) = \operatorname{rel\,int} F$ 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 $\overline{S}(h)$ is a compact convex set, there is a set of $K \leq N$ $a_i \in \bar{S}(h)$ s.t. $\sum_{j=1}^{K} \lambda_j a_j = x$, $\lambda_j > 0$, $\sum_{j=1}^{K} \lambda_j = 1$. 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

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

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 $z \in h^{-1}(B_\delta(a_i))$. Then 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. So, we can work directly with points in F.

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 $\gamma \in \Delta^*(\mathcal{Z})$ absolutely continuous with respect to ν s.t. $q = \int_{\mathcal{Z}} h(z)_i \, d\gamma(z)$. 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 $\hat{\mu}$. This is a contradiction. Therefore, we can represent this with a misspecified model.

(Only If:) Take a measure $\hat{\mu} \in \Delta^*(\mathcal{Z})$. 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 and the associated function $g : \mathcal{Z} \to \Delta(\Omega)$. Let $\rho_g = \mu \circ g^{-1}$. Define the measure

$$\hat{\mu}(Z) = \int_Z \frac{d\hat{\rho}}{d\rho_g}(g(z)) \, d\mu(z).$$

Now note that

$$\int_{\mathcal{Z}} g(z)_i d\hat{\mu}(z) = \int_{\Delta(\Omega)} x_i d\hat{\rho}(x) = p_i$$

so the misspecified model

$$\hat{\mu}_i(Z) = \frac{1}{p_i} \int_Z g(z)_i \, d\hat{\mu}(z)$$

is a misspecified model with unconditional signal distribution $\hat{\mu}$. This misspecified model has forecast $\hat{\rho}$ by construction of $\hat{\mu}$ and the change of variables formula.

(Only If:) Fix a misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega}$. Let $h(z)$ be the updating rule defined by Bayes rule with respect to this misspecified model. Then if $\hat{\rho}(X) = \hat{\mu}(h^{-1}(X))$ is a forecast, it is, by definition, the forecast represented by the misspecified 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, for any i

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

so it is a plausible forecast. □

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

Lemma 3.

1. *Updating rule h can be represented by a misspecified model with unconditional signal distribution $\hat{\mu} \in \Delta^*(\mathcal{Z})$ iff*

$$\int_{\mathcal{Z}} h(z)_i \, d\hat{\mu}(z) = p_i \tag{21}$$

for all i. If a representation exists, then for any state ω_i with $p_i > 0$, $\hat{\mu}_i(Z) = \frac{1}{p_i} \int_Z h(z)_i \, d\hat{\mu}(z)$ for any measurable set of signal realizations $Z \subset \mathcal{F}$.

2. *Updating rule h can be represented by a misspecified model with conditional signal distribution $\hat{\mu}_j \in \Delta^*(\mathcal{Z})$ in state ω_j iff*

$$\int_{\mathcal{Z}} \frac{h(z)_i}{h(z)_j} d\hat{\mu}_j(z) = \frac{p_i}{p_j} \tag{22}$$

for all $\omega_i \in \Omega$. If a representation exists, then for any state 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 $Z \subset \mathcal{F}$.

The first part of this result is reminiscent of the well-known Bayes plausibility condition from the literature on communication games (Kamenica and Gentzkow 2011)—that is, the posterior belief must be a martingale with respect to the prior. The second part follows from the well-known condition that the likelihood ratio of the probability of state ω_i to state ω_j is a martingale with respect to the distribution in state ω_j—here, with respect to the subjective distribution $\hat{\mu}_j$. In either case, once one distribution is fixed, this distribution in conjunction with the updating rule either pins down the entire set of conditional signal distributions or violates Bayes-plausibility, and therefore, cannot be part of a misspecified model that represents the updating rule.

Lemma 3 simplifies the process of selecting a model to represent an updating rule. In particular, since specifying either the unconditional signal measure or one of the state-contingent signal measures uniquely pins down the remainder of the misspecified model, a condition that selects an essentially unique such measure will also select an essentially unique misspecified model.

Proof of Lemma 3. Fix an updating rule h.

Part 1: ($\Rightarrow$) 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 $\int_{\mathcal{Z}} h(z)_i \, d\hat{\mu}(z) = p_i$.

($\Leftarrow$) 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. ($\Rightarrow$) 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

$$\int_{\mathcal{Z}} \frac{h(z)_i}{h(z)_j} d\hat{\mu}_j(z) = \frac{p_i}{p_j}.$$

($\Leftarrow$) 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 $h(z)$. WLOG assume $j = 1$. Then, any family of misspecified models with updating rule h and conditional signal distribution $\hat{\mu}_1$ must solve

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

$\hat{\mu}_1$ a.s. so this model is unique. □

Proof of Theorem 1. We first prove the sufficiency of the conditions. To do so requires a proof of Corollary 1. We then prove the necessity of the conditions.

Sufficiency: To establish sufficiency of the conditions, it is convenient to establish the model defined in Corollary 1 represents h and $\hat{\rho}$. 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

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

It remains to show that this induces the desired forecast and updating rule. That is, we must establish Corollary 1.

Proof of Corollary 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. It remains to show that $\hat{\mu} \circ h^{-1}(X) = \hat{\rho}(X)$

for any Borel set X. For any Borel set X, note that

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

Therefore, for any Borel set X, $\hat{\mu}(h^{-1}(X)) = \sum p_i \hat{\rho}_i(X) = \hat{\rho}(X)$. Therefore $(\hat{\mu}_i)_{\omega_i \in \Omega}$ is a model that induces the desired forecast.

Necessity: We use Lemmas 2 and 3 to establish the necessity of the absolute continuity and plausibility conditions and the uniqueness of the representation. The forecast must be plausible by Lemma 2. Suppose there exists a Borel set X such that $\rho_h(X) > 0$ but $\hat{\rho}(X) = 0$ and a misspecified 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 $\rho_h(X) = 0$ but $\hat{\rho}(X) > 0$ is a contradiction. Therefore, ρ_h and $\hat{\rho}$ must be mutually absolutely continuous.

Uniqueness of the representation for sets in $\mathcal{F}_h$ 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 $\hat{\mu}(Z) = \hat{\rho} \circ h^{-1}(Z)$. 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

$$\hat{\mu}_i(Z) = \int_Z h(z)_i d\hat{\mu}(z) = \int_Z h(z)_i d\hat{\rho}(h^{-1}(z)).$$

so these conditional measures are unique. □

A.2 Proofs from Section 4

Proof of Proposition 1. Suppose h is an updating rule such that the accurate forecast is plausible. This implies that

$$\int_{\mathcal{Z}} h(z)_i \, d\mu = p_i \text{ for all } \omega \in \Omega.$$

Then by Lemma 3, there exists a misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega}$ that induces unconditional distribution μ over $\mathcal{Z}$ and is represented by updating rule $h(z)$. By the proof of Lemma 3,

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

describes a misspecified model that induces the desired distribution and updating rule. Moreover, as argued before any misspecified model must solve

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

There is at most one Radon-Nikodym derivative that solves this equation, and thus, the misspecified model is unique.

Now suppose that $(\hat{\mu}_i)_{\omega_i \in \Omega}$ describes an introspection proof misspecified model that induces updating rule h and has unconditional distribution μ. By the above logic, the Radon-Nikodym derivative $\frac{d\hat{\mu}^i}{d\mu}(z) = \frac{1}{p_i} h(z)_i$. This implies that

$$\hat{\mu}_i(\mathcal{Z}) = \int_{\mathcal{Z}} \frac{1}{p_i} h(z)_i \, d\mu(z) = 1,$$

which in turn implies that $\int_{\mathcal{Z}} h(z)_i \, d\hat{\mu}(z) = \int_{\mathcal{Z}} h(z_i) \, d\mu(z) = p_i$, so the accurate forecast is plausible. Therefore, the desired condition holds. □

Proof of Proposition 2. (If:) The existence of a misspecified model with naive consistent forecast follows immediately 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 $\hat{\mu}_i$, Eq. (9), and the naive consistency of the forecast

(Only If:) Let $\rho_B = \mu(h^{-1}(X))$ be the accurate Bayesian forecast. Suppose there exists a naive consistent 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 that $\mu(h^{-1}(X)) = 0$. But, this then implies that $\mu(h_B^{-1}(X)) = 0$ which is a contradiction. A similar argument applies to the case where $\rho_B(X) = 0$ but $\hat{\rho}(X) > 0$. □

A.3 Proofs from Section 5

Proof of Proposition 3. Fix the manager's expected self-image, $\gamma \equiv E(h(z_m)|M) = (h(g, M) + h(b, M))/2$. The larger γ, the more the test scores for group identity M need to be inflated on average. In order to maintain the introspection-proof constraint, this requires on average a lower interpretation of test scores for group identity F, $(h(g, F) + h(b, F))/2 = \frac{1-2q\gamma}{2(1-q)}$. For a given γ, the first self chooses an updating rule to maximize

$$-E[(\mathbb{1}_{\omega_w = H} - h(z_w))^2],$$

where the expectation is taken with respect to the true distribution over z_w, subject to the constraint that the self-image is indeed equal to γ, $\frac{1}{2}(h(g,M)+h(b,M))=\gamma$ and that the updating rule is introspection-proof, $\frac{1}{2}(h(g,F)+h(b,F))=\frac{1-2q\gamma}{2(1-q)}$. This is solved by:

$$\begin{aligned}
h^*(g,M;\gamma) &= \alpha+\gamma-1/2\\
h^*(b,M;\gamma) &= 1-\alpha+\gamma-1/2\\
h^*(g,F;\gamma) &= \alpha+\frac{q}{1-q}\left(\frac{1}{2}-\gamma\right)\\
h^*(b,F;\gamma) &= 1-\alpha+\frac{q}{1-q}\left(\frac{1}{2}-\gamma\right).
\end{aligned}$$

To choose the optimal γ, the first self maximizes

$$\max_{\gamma\in[0,1]} \gamma - cE[(1_{\omega=H}-h^*(z_w;\gamma))^2].$$

This is solved by $\gamma^*=\frac{1}{2}+\frac{1-q}{2qc}$. This leads to the IP-updating rule in Proposition 3. □

Proof of Proposition 4. Fix the manager's expected self-image, $\gamma\equiv E(h(z_m)|M)=(h(g,M)+h(b,M))/2$. Similar to the derivation for Proposition 3, the optimal updating rule in terms of γ is

$$\begin{aligned}
h^*(g,M;\gamma) &= \alpha+\gamma-1/2\\
h^*(b,M;\gamma) &= 1-\alpha+\gamma-1/2\\
h^*(g,F;\gamma) &= \alpha\\
h^*(b,F;\gamma) &= 1-\alpha.
\end{aligned}$$

This leads to the optimal $\gamma^*=\frac{1}{2cq}+\frac{1}{2}$, which is higher than in the introspection-proof case. □

Proof of Proposition 5. Consider how $\int_{h_B(z)}^1(x-h_B(z))\,d\hat{\rho}(x)$ varies with $\hat{\rho}$. We can recast this as the utility a consumer with utility function $u(x)=\max\{0,x-h_B(z)\}$ receives from the lottery $\hat{\rho}$. Since $u(x)$ is a convex function, if $\hat{\rho}$ is a mean-preserving spread of $\hat{\rho}'$ then $E_{\hat{\rho}}(u(x))\geq E_{\hat{\rho}'}(u(x))$. Therefore, the right hand side of Eq. (17) is higher under $\hat{\rho}$ than $\hat{\rho}'$, so as the degree of overprecision increases, the firm searches following a larger set of signal realizations. □

Proof of Proposition 6. Let $f(q)\equiv\int_q^1(x-q)\,d\rho_B(x)$. This is clearly a decreasing function. The firm searches the second technology following any z^1 such that $f(h(z^1))>c$. Define $\bar{c}\equiv f(1/2)$. Suppose $c<\bar{c}$. Then for any h, a firm that uses updating rule h

searches for all z^1 such that $h(z^1) < 1/2$. If the firm is retrospectively underprecise, then for every z^1 such that the underprecise firm does not search following z^1, i.e. $f(h(z^1)) \leq c$, we have $h(z^1) \geq 1/2$, and therefore, $h(z^1) < h_B(z^1)$. This implies $f(h_B(z^1)) < f(h(z^1)) \leq c$, and therefore, the unbiased firm also does not search. Therefore, the underprecise firm searches following signal realizations in a superset of $\mathcal{Z}_S$. If the firm is retrospectively overprecise, then for every z^1 such that the unbiased firm does not search, $f(h_B(z^1)) \leq c$, by analogous reasoning we have $f(h(z^1)) < f(h_B(z^1)) \leq c$ and the overprecise firm also does not search. Therefore, the overprecise firm searches for signal realizations in a subset of $\mathcal{Z}_S$. The logic for $c > \bar{c}$ is analogous. □

A.4 Proofs from Section 6

Proof of Proposition 7. (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

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

So the condition from observation 1 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'

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

□

Proof of Proposition 8. Fix a prior p and let $(\hat{\mu}_i)_{\omega_i \in \Omega}$ be the essentially unique representation of $h(z,p)$ and the naive consistent forecast ρ_B at prior p. It follows from Proposition 7 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 the naive consistent forecast ρ_B at p', as for any Borel set X,

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

□

Proof of Theorem 2 This follows immediately from Theorem 1. If $\hat{\rho}$ and ρ are mutually absolutely continuous then the forecast $\hat{\rho}$ is plausible with respect to the prior $\hat{p}$. So, under prior $\hat{p}$ the conditions of Theorem 1 are satisfied and a misspecified model exists and is essentially unique. Moreover, under any other prior the forecast is not plausible. Therefore the misspecified prior and model that jointly represents $\hat{\rho}$ and h is unique.

If $\hat{\rho}$ and ρ are not mutually absolutely continuous, then by Theorem 1 the forecast and updating rule cannot be jointly represented. Therefore, the absolute continuity condition is necessary and sufficient for the existence of a unique misspecified prior and an essentially unique misspecified model that jointly represent h and $\hat{\rho}$. □

B Extensions

B.1 Almost Introspection-Proof.

Given a misspecified model, it is natural to ask (i) how far away is the forecast it induces from the true distribution over misspecified posteriors, and (ii) how far away is the forecast it induces from the "optimal" forecast for the given updating rule. A natural way to formalize these questions is in terms of divergences.

Definition 13. *Fix a misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega}$. Let $\hat{\rho}$ and h be the updating rule and forecast induced by this misspecified model. $\hat{\rho}$ is the KL-optimal forecast for updating rule h if it minimizes $\min_{\hat{\rho}^*} D(\hat{\rho}^* || \mu \circ h^{-1})$ across all forecasts that can represented by a misspecified model that induces $h(z)$.*

The KL-optimal forecast provides a natural benchmark for in some sense quantifying the additional prospective distortions induced by a misspecified model.

Before characterizing the KL optimal forecast $\hat{\rho}^*$, it is convenient to think about the following natural exercise. Even if no introspection-proof representation exists, perhaps a natural model to represent an updating rule would be the one that in some sense did the best against any sort of test for misspecification the agent could construct. To formalize this, let $T_n : \mathcal{Z}^n \to \Delta\{0, 1\}$ be a test, a mapping from a realized sequence of signals to a 0 or 1. We say a sequence passes the test if the realization of this random variable is 1, and it fails otherwise. Let $\mathcal{T}_n$ the set of all tests for samples of size n.

Given a test T_n, we can ask how effective it is at detecting misspecification. That is, T_n is a hypothesis test for the binary hypothesis

$$\begin{aligned} &H_0 : \hat{\mu} \\ &H_1 : \mu. \end{aligned}$$

Let

$$\begin{aligned} \beta_\alpha^n = \sup_{T_n \in \mathcal{T}_n} & -\frac{\ln Pr_\mu(T_n = 1)}{n} \\ \text{s.t. } & Pr_{\hat{\mu}}(T_n = 1) \geq 1 - \alpha \end{aligned}$$

so $e^{-n\beta}$ is the probability of failing the detect misspecification when the true data generating process is $(\mu_i)_{\omega_i \in \Omega}$.

Using this, we can define another class of misspecified models:

Definition 14. *Given an updating rule h, a misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ that represents h is* α-introspection proof *if across all possible representations of h it minimizes* $\liminf_{n \to \infty} \beta_\alpha^n$.

That is, given any hypothesis test that rejects the misspecified model with probability less than α if it was true, the α-introspection proof model minimizes the worst-case probability of rejection under the true distribution as n grows large. The α-introspection proof misspecified model is in some sense the representation of h that makes it hardest to detect misspecification. By the Chernoff-Stein lemma, for any $\alpha \in (0,1)$ and $\hat{\mu}$, $\lim_{n \to \infty} \beta_\alpha^n = D(\hat{\mu}||\mu)$, where D is KL-divergence, so we can reformulate this problem as

$$\begin{aligned} \min_{\hat{\mu} \in \Delta^*(\mathcal{Z})} & \; D(\hat{\mu}||\mu) \\ \text{s.t. } & \int h_i(z) \frac{d\hat{\mu}}{d\nu}(z) d\nu(z) = p_i \text{ for all i.} \end{aligned}$$

So the α-introspection proof misspecified model is the model induced by the KL-optimal forecast.

Using tools from information theory, we can then characterize the α-introspection proof misspecified model.

Proposition 9. *Let $\psi_h : \mathbb{R}_+^N \to \mathbb{R}$ be the joint moment generating function of posteriors $\psi_h(\lambda) = E_\mu(e^{\lambda \cdot h(z)})$. Given an updating rule h, the α-introspection proof misspecified model is given by:*

$$\frac{d\hat{\mu}_i}{d\nu}(z) = \frac{1}{p_i} h_i(z) \exp(\lambda \cdot h(z) - \log \psi_h(\lambda)) \frac{d\mu}{d\nu}(z)$$

where $\lambda \in \mathbb{R}^n$ *solves* $\int (p_i - h(z)_i)e^{\lambda \cdot h(z)} d\mu(z) = 0$ *for each* i.[33] *This model has KL-divergence* $p \cdot \lambda - \log \psi_h(\lambda)$ *from the truth.*

Proof. Our goal is to solve the program

$$\min_{\hat{\mu} \in \Delta^*(\mathcal{Z})} D(\hat{\mu}||\mu)$$

$$\text{s.t.} \int h_i(z) d\hat{\mu}(z) = p_i \text{ for all } i$$

If this min exists, we can apply our tools to construct a misspecified model that induces unconditional distribution $\hat{\mu}$, and by the Chernoff-Stein Lemma this is the α introspection-proof misspecified model. Let $\frac{d\hat{\mu}^*}{d\nu}(z) = \exp(\lambda \cdot h(z) - \log \psi_h(\lambda)) \frac{d\mu}{d\nu}(z)$. Under this measure $E_{\hat{\mu}^*}(h(z)_i)$ satisfies

$$E_{\hat{\mu}^*}(h(z)_i) = \frac{1}{\psi_h(\lambda)} \int_{\mathcal{Z}} h(z)_i \exp(\lambda \cdot h(z)) d\mu(z) = \frac{1}{\psi_h(\lambda)} \int_{\mathcal{Z}} p_i \exp(\lambda \cdot h(z)) \, d\mu(z) = p_i,$$

where the first equality follows from the definition of λ. In addition, $\hat{\mu}^*$ is non-negative and integrates to 1 by the definition of ψ_h, so $\hat{\mu}^*$ satisfies the constraints. The misspecified model described in the statement of the theorem is then simply the model induced jointly by $\hat{\mu}^*$ and h (see Lemma 3). To see that this model is a minimizer, note that for any feasible $\hat{\mu}$,

$$\begin{aligned} D(\hat{\mu}||\mu) &= E_{\hat{\mu}}(\log \frac{d\hat{\mu}}{d\hat{\mu}^*}(z) \frac{d\hat{\mu}^*}{d\mu}(z)) \\ &= E_{\hat{\mu}}(\log \frac{d\hat{\mu}}{d\hat{\mu}^*}(z)) + E_{\hat{\mu}}(\log \frac{d\hat{\mu}^*}{d\mu}(z)) \\ &= D(\hat{\mu}||\hat{\mu}^*) + E_{\hat{\mu}}(\lambda \cdot h(z) - \log \psi_h(\lambda)) \\ &= D(\hat{\mu}||\hat{\mu}^*) + \lambda \cdot p - \log \psi_h(\lambda) \\ &\geq \lambda \cdot p - \log \psi_h(\lambda) = D(\hat{\mu}^*||\mu) \end{aligned}$$

so $\hat{\mu}^*$ is a minimizer. □

The updating rule h pins down the exponential family that the α-introspection proof misspecified model belongs to while the true distribution determines the exact representative of this family. Applying the change of variables formula, this also characterizes the KL-optimal forecast, which satisfies for any $x \in \Delta(\Omega)$

$$\frac{d\hat{\rho}^*}{d\rho_h}(x) = \exp(\lambda \cdot x - \log E_{\rho_h}(\exp(\lambda \cdot x))),$$

[33] Since h is a bounded random variable ψ_h exists. λ solves $\max_\lambda p \cdot \lambda - \log \psi_h(\lambda)$, which has a solution iff the $h(z)$ is responsive.

where λ is the λ from the proposition.

B.2 State-Dependent Introspection-Proof

We motivated our notion of introspection-proofness as robustness of the misspecified model to infinite independent draws of the state and the signal. A natural, related notion, would be to instead fix the true state of the world ω_i and then require the misspecified model to be robust to observing infinite conditionally independent draws of z.

Definition 15. *A family of misspecified models $(\hat{\mu}_i)_{\omega_i \in \Omega}$ representing updating rule $h(z)$ is an ω_i* introspection-proof model relative to ω_j *if for all $Z \in \mathcal{F}$, $\hat{\mu}_j(Z) = \mu_i(Z)$.*

This restriction requires there to exist some state ω_j where the observed frequencies of different signals matches the truth. As with introspection-proofness, this condition is enough to pin down a unique misspecified model that represents a given updating rule.

Proposition 10. *Fix an updating rule h. This can be represented by an ω_i-introspection-proof misspecified model relative to ω_j, $(\hat{\mu}_k)_{\omega_i \in \Omega}$ if and only if for all $k \in \{1, 2, \dots N\}$*

$$\int_{\mathcal{Z}} \frac{h(z)_k}{h(z)_j} \, d\mu_i(z) = \frac{p_k}{p_j}.$$

Moreover, if this representation exists, for any k and any $Z \in \mathcal{F}$,

$$\hat{\mu}_k(Z) = \int_Z \frac{p_j}{p_k} \frac{h(z)_k}{h(z)_j} \, d\mu_i(z).$$

Proof. Note that this satisfies the introspection proof condition as

$$\hat{\mu}_j(Z) = \int_Z \frac{p_j}{p_j} \frac{h(z)_j}{h(z)_j} \, d\mu_i(z) = \mu_i(Z).$$

It follows immediately from Lemma 3 that $(\hat{\mu})_{\omega_i \in \Omega}$ represents h and induces distribution $\hat{\mu}_j$. □

This condition is once again a variation of the martingale property of beliefs—in this case, the requirement that the likelihood ratio is a martingale with respect to the true data generating process. While on the surface it appears very similar to the original introspection-proof condition, this condition is in fact much less restrictive.

C Comparison to Blackwell's Order

A common way of modeling biases, especially in the motivated reasoning literature, is to take a correctly specified model and have the agent imperfectly recall signals by adding noise to the signal distribution. This makes the agent perceive information as being drawn from a Blackwell less informative distribution. At first glance, it may seem like this may be connected with our notion of introspection-proof models. In this section, we demonstrate that these are distinct concepts.

Roughly, an information structure is Blackwell more informative than another information structure if and only if it is a mean-preserving spread of the distribution of posteriors, which is equivalent to the existence of a garbling matrix. A garbled distribution in general induces different probabilities of each signal realization, as it combines signals to make them less precise. In contrast, it is difficult to combine signals in a way that is introspection-proof, as the agent still observes a draw from the original signal space. In this section, we formally show that these concepts are distinct by providing examples in which a misspecified model is Blackwell ranked with respect to the true model but not introspection-proof, and introspection-proof but not Blackwell ranked with respect to the true model.

Consider a finite signal space $\mathcal{Z} = \{z_1, z_2, \dots z_K\}$ and let Q be a $N \times K$ matrix with $(Q)_{ij} = \mu_i(\{z_j\})$. Define $\hat{Q}$ analogously. In this framework, Q and $\hat{Q}$ capture models. Model $\hat{Q}$ is Blackwell less informative than Q iff there exists an $K \times K$ stochastic matrix M s.t. $QM = \hat{Q}$. The definition of introspection-proof corresponds to $pQ = p\hat{Q}$, where p is the (row) vector of priors as defined in Section 2. Proposition 1 establishes that introspection-proof is equivalent to the the requirement that $HQ'p' = \hat{H}Q'p'$, where H is the matrix with $H_{ij} = h_B(z_j)_i$ and $\hat{H}$ is the matrix with $\hat{H}_{ij} = h(z_j)_i$.

To see that a misspecified model can be Blackwell ranked with respect to the true model but not be introspection-proof, consider the models

$$Q = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}, \hat{Q} = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}.$$

Then $\hat{Q}$ is a garbling of Q (use $M = (7/8, 1/8; 1/24, 23/24)$) and therefore, Blackwell less informative. But $\hat{Q}$ is not introspection-proof with respect to Q for any interior prior, as unlike garbling information, the unconditional probabilities of each signal must be the same under Q and $\hat{Q}$, e.g. for z_1,

$$p_1 \frac{2}{3} + (1 - p_1)\frac{1}{4} = p_1 \frac{3}{4} + (1 - p_1)\frac{1}{4},$$

which only holds at $p_1 = 0$.

However, the introspection-proof condition does not preclude a model from being Blackwell ranked with respect to the true model. To see that a model can be Blackwell ranked and introspection-proof, consider prior $p_1 = 1/2$ and model

$$\hat{Q} = \begin{pmatrix} \frac{3}{4} - \tau & \frac{1}{4} + \tau \\ \frac{1}{4} + \tau & \frac{3}{4} - \tau \end{pmatrix}.$$

for $\tau \in [0, 1/4]$. Then model $\hat{Q}$ is introspection-proof with respect to Q and is also Blackwell less informative than Q.

To see that models that are not Blackwell ranked with respect to the true model can also be introspection-proof, consider

$$Q = \begin{pmatrix} \frac{2}{8} & \frac{3}{8} & \frac{2}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{2}{8} & \frac{3}{8} & \frac{2}{8} \end{pmatrix}, \hat{Q} = \begin{pmatrix} \frac{5}{16} & \frac{5}{16} & \frac{5}{16} & \frac{1}{16} \\ \frac{1}{16} & \frac{5}{16} & \frac{5}{16} & \frac{5}{16} \end{pmatrix}.$$

Then Q and $\hat{Q}$ are not Blackwell ranked but $\hat{Q}$ is introspection-proof with respect to Q.

D Additional Examples

D.1 Examples of Prior-Independent and Prior-Dependent Representations

We show that the updating rule modeling geometric overreaction in Section 2.2 and the updating rule modeling partisan bias in Bohren and Hauser (2021) satisfy the condition in Proposition 7, and therefore, have a prior-independent representation. We also show that the updating rule modeling over/underreaction in Epstein et al. (2010) (see Example 3) and the updating rule modeling partisan bias in Example 4 do not satisfy the condition in Proposition 7, and therefore, do not have a prior-independent representation.

The geometric overreaction updating rule from Section 2.2 corresponds to

$$\frac{h(z,p)_i}{h(z,p)_j} = \frac{p_i}{p_j}\left(\frac{d\mu_i}{d\mu_j}(z)\right)^{\gamma}.$$

It is straightforward to see that it is possible to factor out the prior from this expression, and therefore, it has a prior-independent representation. Consider the parameterization of partisan bias from Bohren and Hauser (2021). There are two states, $|\Omega| = 2$. Normalize the signal to be the posterior probability of ω_1 following a flat prior, $z = \frac{d\mu_1}{d\nu}(z)/(\frac{d\mu_2}{d\nu}(z) + \frac{d\mu_1}{d\nu}(z))$, with support $\mathcal{Z} \subset [0, 1]$. Consider updating rule

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

where $\alpha \in (0, \infty)$ is the partisan bias parameter. 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.

In contrast, the model of over/underreaction in Example 3 does not satisfy the condition in Proposition 7, as

$$\frac{p_j}{p_i}\frac{h(z,p)_i}{h(z,p)_j} = \frac{\frac{d\mu_i}{d\nu}(z) + \sum_{k=1}^N p_k \frac{d\mu_k}{d\nu}(z)}{\frac{d\mu_j}{d\nu}(z) + \sum_{k=1}^N p_k \frac{d\mu_k}{d\nu}(z)}$$

clearly depends on the prior. Similarly, in the model of partisan bias in Example 4,

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

where $h_B(z,p)_1 \equiv \frac{p_1 \frac{d\mu_1}{d\nu}(z)}{p_1\frac{d\mu_1}{d\nu}(z)+p_2\frac{d\mu_2}{d\nu}(z)}$. This expression also clearly depends on the prior.

D.2 Linear Under- and Overreaction

The following example illustrates the multiplicity of representations for the updating rule for under- and overreaction defined in Epstein et al. (2010),

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

for some $\alpha \in (-\infty, 1]$. We can use Lemma 3 to find misspecified models that represent this updating rule. First consider a misspecified model that induces a subjective unconditional distribution that is equal to the true unconditional distribution, $\hat{\mu} = \mu$. Then $\hat{\mu}$ satisfies Eq. (21) as $\int_{\mathcal{Z}} h_B(z)_i \, d\hat{\mu}(z) = \int_{\mathcal{Z}} h_B(z)_i \, d\mu(z) = p_i$ by standard argument, and therefore, $\int_{\mathcal{Z}} (\alpha h_B(z)_i + (1-\alpha)p_i) \, d\hat{\mu}(z) = p_i$. In this case, the subjective distribution in state ω_i must be equal to:

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

In other words, it is completely pinned down by the true unconditional measure μ, the Bayesian updating rule h_B, and the under- or overreaction parameter α.

This representation is not unique. To illustrate this, consider a setting with $|\Omega| = 2$, $\mathcal{Z} = [0,1]$, $p = 1/2$, a uniform true unconditional signal distribution, and $|h_B(z)_1 - \frac{1}{2}|$ symmetric about $z = 1/2$. Then the model that induces subjective unconditional pdf $d\hat{\mu}(z) = 3/2 - 6(z-1/2)^2$ (in a slight abuse of notation, using $d\hat{\mu}$ to denote the pdf) also satisfies $\int_{\mathcal{Z}} h_B(z)_i d\hat{\mu}(z) = 1/2$, and therefore, $\int_{\mathcal{Z}} (\alpha h_B(z)_i + (1-\alpha)/2) \, d\hat{\mu}(z) = 1/2$. In the first representation, the agent correctly anticipates the frequencies of different signals but underreacts to them, while in the second representation, the agent underestimates the frequency of "extreme" signal realizations. Given that h_B is monotone, this means that in addition to underreacting to the signal, the agent also anticipates that she will observe signal realizations that are, on average, less informative than the signal realizations she actually observes.

References

Alonso, R. and O. Câmara (2016): "Bayesian persuasion with heterogeneous priors," *Journal of Economic Theory*, 165, 672–706.

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.*

Augenblick, N. and M. Rabin (2021): "Belief movement, uncertainty reduction, and rational updating," *The Quarterly Journal of Economics*, 136, 933–985.

Ba, C. (2022): "Robust Model Misspecification and Paradigm Shifts," *PIER 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., K. Haggag, A. Imas, and D. G. Pope (forthcoming): "Inaccurate Statistical Discrimination: An Identification Problem," *Review of Economics and Statistics.*

Bohren, J. A. and D. N. Hauser (2019): "Social Learning with Model Misspecification: A Framework and a Characterization," *PIER Working Paper 18-017*, available at SSRN: https://ssrn.com/abstract=3236842 or http://dx.doi.org/10.2139/ssrn.3236842.

——— (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.*

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

Caplin, A., M. Dean, and J. Leahy (2022): "Rationally inattentive behavior: Characterizing and generalizing Shannon entropy," *Journal of Political Economy*, 130, 1676–1715.

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," .

Danz, D., L. Vesterlund, and A. J. Wilson (2022): "Belief Elicitation and Behavioral Incentive Compatibility," *American Economic Review*, 112, 2851–83.

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

Epstein, L., J. Noor, and A. Sandroni (2008): "Non-Bayesian updating: a theoretical framework," *Theoretical Economics*, 3, 193–229.

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.

Eyting, M. (2023): "Why do we Discriminate? The Role of Motivated Reasoning," Mimeo.

Frankel, A. and E. Kamenica (2019): "Quantifying information and uncertainty," *American Economic Review*, 109, 3650–80.

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

——— (2020b): "Stability and Robustness in Misspecified Learning Models," .

——— (2021): "Welfare comparisons for biased learning," .

Fudenberg, D. and G. Lanzani (2022): "Which misperceptions persist?" *Available at SSRN 3709932*.

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

——— (2022): "Selective Memory Equilibrium," *Available at SSRN 4015313*.

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

Gabaix, X. (2019): "Behavioral inattention," in *Handbook of Behavioral Economics: Applications and Foundations 1*, Elsevier, vol. 2, 261–343.

Gagnon-Bartsch, T., M. Rabin, and J. Schwartzstein (2018): "Channeled attention and stable errors," .

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

Hagmann, D. and G. Loewenstein (2019): "Persuasion With Motivated Beliefs," .

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 (2021): "Evolutionarily stable (mis) specifications: Theory and applications," .

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.

Heidhues, P., B. Kőszegi, and P. Strack (2023): "Overconfidence and Prejudice," Mimeo.

Jakobsen, A. M. (2022): "Coarse Bayesian updating," Mimeo.

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

Kamenica, E. (2019): "Bayesian persuasion and information design," *Annual Review of Economics*, 11, 249–272.

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.

Kleijn, B. J. and A. W. van der Vaart (2006): "Misspecification in infinite-dimensional Bayesian statistics," *The Annals of Statistics*, 837–877.

Lee, Y.-J., W. Lim, and C. Zhao (2023): “Cheap talk with prior-biased inferences,” *Games and Economic Behavior.*

Lehrer, E. and R. Teper (2017): “The dynamics of preferences, predictive probabilities, and learning,” .

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

Libgober, J. (2023): “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.

Manski, C. F. and C. Neri (2013): “First-and second-order subjective expectations in strategic decision-making: Experimental evidence,” *Games and Economic Behavior*, 81, 232–254.

Mensch, J. (2018): “Cardinal representations of information,” *Available at SSRN 3148954.*

Molavi, P. (2019): “Macroeconomics with learning and misspecification: A general theory and applications,” .

——— (2021): “Tests of Bayesian Rationality,” *arXiv preprint arXiv:2109.07007.*

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.

Pomatto, L., P. Strack, and O. Tamuz (Forthcoming): “The cost of information,” *American Economic Review.*

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

Prelec, D., H. S. Seung, and J. McCoy (2017): “A solution to the single-question crowd wisdom problem,” *Nature*, 541, 532–535.

Rabin, M. (2002): “Inference by believers in the law of small numbers,” *The Quarterly Journal of Economics*, 117, 775–816.

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.

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

Zhao, C. (2022): “Pseudo-Bayesian updating,” *Theoretical Economics*, 17, 253–289.