The Behavioral Foundations of Model Misspecification: A Decomposition[*]

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

November 16, 2022

A growing literature in economics seeks to model how agents process information and update beliefs. In this paper, we link two common approaches: (i) defining an updating rule that specifies a mapping from prior beliefs and the signal to the agent's subjective posterior, and (ii) modeling an agent as a Bayesian learner with a misspecified model. The updating rule approach has a more transparent conceptual link to the underlying bias being modeled, while the misspecified model approach is 'complete,' in that no further assumptions on belief-updating are necessary to analyze the model, and has well-developed solution concepts and convergence results. We show that any misspecified model can be decomposed into two objects that summarize the biases it introduces: the updating rule captures how the agent interprets realized information, while the forecast captures how the agent anticipates future information. We derive necessary and sufficient conditions for a forecast and updating rule pair to be represented by a misspecified model. This provides conceptual guidance for which model to select to represent a given bias. Finally, we consider two natural ways to select forecasts: introspection-proofness and naive consistency. We demonstrate how introspection-proofness places a natural bound on the magnitude of bias in an application with motivated reasoning, and how naive consistency impacts a firm's ability to screen consumers in a credit market application.

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

[*] We thank Cuimin Ba, 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 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

A rich empirical literature in economics and psychology has documented the ways in which individuals exhibit systematic biases and make errors in how they interpret information and form beliefs. A complementary theoretical literature examines how to model such biased updating. In general, the literature employs two distinct approaches. The first, which we refer to as the 'non-Baysian' approach, consists of parameterizing a particular bias with an *updating rule* that describes how each signal realization maps to a posterior belief.[1] The second is the "misspecified model" approach, which describes a model of the signal generating process that the individual uses to interpret the signal. The individual forms beliefs from Bayesian updating with respect to this model, but the model may be wrong.[2]

In this paper, we seek to connect these two approaches. In particular, we first ask 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. We then explore what other components besides the updating rule are needed to pin down a unique representation, and explore the forms of bias contained in these other components.

We explore these questions in a general informational environment in which an agent learns about an unknown state from a signal process. An updating rule corresponds to a mapping from each signal realization to a posterior belief, while a misspecified model corresponds to a family of distributions over the signal space, one for each state. The final piece needed for our analysis is a forecast, which is the agent's belief about the distribution of her posterior belief.

Our main result shows that a misspecified model can be decomposed into the two classes of bias that it induces: (i) the prospective bias, which corresponds to how an agent anticipates she will form beliefs before observing the signal; and (ii) the retrospective bias, which corresponds to how the agent misinterprets information after she observes the signal. The latter is captured by the updating rule, while the former is captured by the forecast. Every misspecified model can be decomposed into these two parts. Further, any forecast that satisfies a condition we call plausibility—the requirement that, from the agent's perspective, the expected future belief is equal to the prior—and any updating rule together identify an essentially unique misspecified model. This provides a convenient formulation for the misspecified model in terms of the biases it induces. Further, it establishes that prospective biases do not place much structure

[1] For example, Epstein, Noor, and Sandroni (2008) writes down a parametric updating rule to capture under- and overreaction.

[2] For example, Bohren (2016) models naive learning as a misspecified model of other agents.

on retrospective biases and vice versa: a given updating rule can be paired with many different forecasts, and similarly for a given forecast. Therefore, specifying one piece of the decomposition does not restrict the other piece. Finally, it establishes that together, the updating rule and forecast pin down all behavioral implications of the misspecified model, in that the chosen model imposes no further restrictions on behavior beyond those implied by the induced forecast and updating rule.

This decomposition provides a natural tool for constructing misspecified models to represent a desired updating rule. In general, as we show, an updating rule can be represented by a multiplicity of misspecified models. Therefore, our decomposition sheds light on the necessary second piece to find a unique representation i.e. the forecast. Motivated by this insight, we examine reasonable choices of forecast in different economic decision-problems.

We first consider the accurate forecast, where the agent's subjective distribution of her posterior belief is equal to the true ex-ante distribution. If this accurate forecast is plausible, then it identifies a unique misspecified model. Moreover, the corresponding misspecified model satisfies a property called 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—and therefore, not representable—for many common updating rules. This means that many simple updating rules cannot be represented by an introspection-proof misspecified model.

We then define a forecast that captures a natural analogue to the naivete assumption commonly used in many behavioral settings. The naive consistent forecast is equal to the accurate forecast of an agent who uses Bayes rule to update beliefs. A agent with a naive consistent forecast behaves as-if all updates will be formed correctly but when she actually updates her beliefs, she does so incorrectly using a biased updating rule. In other words, the agent believes that she will update without bias in the future, but interprets her past information with bias. We also identify necessary and sufficient conditions for a naive consistent forecast to be represented, and show that again this representation is unique. The condition in this case is quite mild—informally, it requires the naive consistent forecast to have the same support as the accurate forecast. In contrast to introspection-proofness, a naive consistent representation exists for many common updating rules.

The benefits to connecting the updating rules and misspecified model approaches are fourfold. First, there is a large literature in economics and statistics that seeks to establish general properties of Bayesian updating with a misspecified model. Connecting the misspecified model approach to non-Bayesian updating rules provides the analyst with a set of off-the-shelf tools that can be used to immediately establish, for instance, convergence of beliefs when agents are non-Bayesian. Second, this linkage helps clarify the conceptual connection between the form of the misspecification and the behavioral bias that the misspecification induces. Third, this approach provides guidance on how to incorporate a behavioral bias into more complex decision problems, including strategic settings where agents must draw inference about both the underlying state and the behavior of others, and settings where agents must make ex-ante decisions before information arrives. Model misspecification allows us to apply the choice frameworks and game-theoretic tools that were largely developed with respect to an expected utility framework to settings with biased updating in a straightforward way.[3] Finally, an updating rule and forecast are relatively straightforward to measure empirically by eliciting beliefs either before or after information is observed, whereas a misspecified model is more complicated to measure. Therefore, distilling a misspecified model into the two components that are empirically identifiable provides an indirect way to measure such misspecified models.

We close with two applications to demonstrate how our results can be used to derive novel economic insights. The first shows how discrimination can emerge endogenously due to self-image concerns, which lead to motivated reasoning when interpreting information from others with a shared group identity. In this dual-selves model, the first self selects an updating rule that the second self uses to evaluate herself and others. A natural constraint to place on the first self's choice of updating rule is that the bias will be undetectable by the second self, i.e. the updating rule is consistent with an introspection-proof misspecified model. We show that this places an endogenous upper bound on the magnitude of the motivated reasoning bias that emerges. It also leads to discrimination in the sense that the chosen updating rule inflates signals for others who share the same group identity, and compensates for this inflation by shading down signals for members of the other group identity.

In the second application, an overconfident entrepreneur decides whether to purchase access to a line of credit to fund future investment. After observing a signal of future returns, the entrepreneur then decides how much to borrow. A credit contract consists of an origination fee to open the line of credit and an interest rate at which the amount borrowed will be repaid. We derive how the optimal credit contract varies in terms of the magnitude of optimism bias and the chosen forecast. In a mixed population where some

[3] A bit of caution here. There is a true data generating process that misspecified agents may place zero weight on. So justifications for equilibrium concepts that rely on limiting behavior of repeated play under a common prior assumption do not translate directly to misspecified models. Esponda and Pouzo (2016) provide an alternative to Bayes-Nash equilibrium that addresses these concerns in equilibrium settings.

agents have optimism bias and others are unbiased, the naive consistent forecast prevents the lender from engaging in second degree price discrimination on the origination fee, since bother types of agents make the same decisions ex-ante. We show that the optimal interest rate is increasing in either the share of biased agents or the magnitude of their bias. In turn, the optimal origination fee is decreasing in both of these parameters. In other words, the lender takes advantage of the biased agent's propensity to over-borrow by charging a higher interest rate—and charges a lower origination fee to draw these borrowers into the contract. In contrast, other forms of forecasts—for example, either over- or underestimating the precision of future beliefs—lead to fundamentally different optimal contracts.

1.1 Literature Review

Model misspecification has received renewed interest in recent years. Esponda and Pouzo (2016) developed a solution concept, Berk-Nash equilibrium, for studying model misspecification in strategic settings.[4] The literature has for the most part focused on characterizing the asymptotic of misspecified Bayesian learning in a variety of general settings (Bohren and Hauser 2021; Fudenberg, Lanzani, and Strack 2020; Frick, Iijima, and Ishii 2020; Heidhues, Koszegi, and Strack 2018; Esponda, Pouzo, and Yamamoto 2019). This paper shows how the updating rules approach can be converted to misspecified models that these results can be applied to.

A number of recent papers on non-Bayesian updating draw parallels between the structure of non-Bayesian rules and Bayes rule. Chauvin (2020), Epstein et al. (2008), Cripps (2018), Lehrer and Teper (2017), and Zhao (2022) provide foundations for general classes of non-Bayesian updating rules and link them to Bayesian updating. In contrast, we study what updating rule are the immediate result of Bayesian updating over a misspecified model. He and Xiao (2017) describe a class of updating rules induced by replacing the likelihood in Bayes rule with an object called the pseudo-likelihood and distorting the prior probability. They provide necessary and sufficient conditions for an updating rule in this class to process signals the same whether they arrive sequentially or at the same time. Fudenberg, Lanzani, and Strack (2022) shows that the long-run outcomes that can arise with selective memory coincide with outcomes in a learning problem with perfect memory but a misspecified model and vice-versa. This provides an alternative foundation for Berk-Nash equilibrium as a consequence of selective memory.

Our main result includes a version of Bayes plausibility (Kamenica and Gentzkow 2011). Recent work by de Clippel and Zhang (2019) studies Bayesian persuasion in setting where the receiver updates according to a function that maps from the true

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

posterior to an incorrect posterior. They develop a version of Bayes plausibility and concavification arguments for these update rules. Their Bayes plausibility condition characterizes the set of possible distributions over posteriors a correctly specified sender can induce, while our plausibility condition describes the possible distributions over posteriors a biased updater could believe their own posterior will be drawn from.[5]

In addition to Bayes plausibility, a number of papers provide characterizations of Bayesian updating in terms of the behavior of posteriors. Augenblick and Rabin (2021) provide tests on the movement of beliefs over time to detect (correctly specified) Bayesian updating. Shmaya and Yariv (2016) show that if an agent updates using Bayes rule then the prior belief belong to the interior of the convex hull of posteriors. We provide a minor extension of this result that can be applied to our class of updating rules and misspecified models. Molavi (2021) shows that any distribution over posteriors satisfying very mild assumptions can be induced via Bayes rule with respect to a misspecified model. This 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 more stringent requirements on the support of the model.

There is a literature that seeks to provide a foundation for the emergence of a misspecified model as a robust phenomenon (Ba 2021; Fudenberg and Lanzani 2022; Gagnon-Bartsch, Rabin, and Schwartzstein 2018; He and Libgober 2021; Frick, Iijima, and Ishii 2021). Our approach is complementary to this literature in that we provide tools to analyze the updating rules that result from these theories. One of the main classes of models we consider, introspection-proof models, identifies a class of models that 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 (2019), and solution concepts like cursed equilibrium (Eyster and Rabin 2005), behavioral equilibrium (Esponda 2008), and analogy expectation equilibrium (Jehiel 2005).

Our characterization draws a distinction between the prospective bias of the agent – how the agent reasons about information yet to be realized– and the retrospective model – how the agent reasons about realized information. Similar distinctions have previously been highlighted in specific non-Bayesian settings in Benjamin, Rabin, and Raymond (2016); Benjamin, Bodoh-Creed, and Rabin (2019); He and Xiao (2017). The distinction

[5] Other papers on communication games with biased receivers include Alonso and Câmara (2016); Lee, Lim, and Zhao (2020).

they draw describes the time inconsistency properties of the updating rules. In contract, we distinguish between how a time consistent misspecified agent effectively has the potential to make two types of mistakes, mistakes anticipating how they'll update and mistakes actually updating. We argue that all biases induced by misspecified models can in some sense be uniquely described by the differences between this kind of prospective and retrospective reasoning. In Section 7.2 we briefly discuss a way to incorporate time inconsistencies into our setting.

Much of the literature on misspecification uses the misspecified model to capture either a prospective or retrospective bias. The work on misspecified causal graphs ((Spiegler 2016)), and Berk-Nash equilibrium (Esponda and Pouzo 2016) take a largely prospective perspective, focusing on identifying how an agent (incorrectly) predicts the world will act once they've made their decision. In contrast, papers like Heidhues et al. (2018); Levy, Razin, and Young (2022) as well as much of the behavioral work that documents and models specific biases in updating (see Benjamin (2019) for a survey) focus on retrospective biases. When modeling even simple economic decisions, like the environment in Section 6.2, or interactions between economic agents, such as those studied in Bohren and Hauser (2021); He (2020); Frick et al. (2021), both prospective and retrospective biases play a role.

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 $\mathcal{F} \subseteq 2^{\mathcal{Z}}$ 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.[6] 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

[6] 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).

to ν for all $i = 1, ..., N$.[7] This ensures that no signal perfectly rules out a state.[8] 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 $\mathcal{Z}$.

This set-up 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).[9]

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 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.* When there is an ex-ante decision before the signal is observed, an agent must also predict what future beliefs will be. We refer to this as a *forecast,* which describes a predicted distribution over the posterior belief. The posterior belief determines how the 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. In addition to ex-ante decision-making, the forecast is a necessary component for strategic interaction and social learning.Our general definitions of an updating rule and a forecast nest specific updating rules and forecasts used in these non-Bayesian approaches to belief-formation.

An updating rule specifies how an agent forms beliefs after observing each signal

[7] 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.

[8] 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.

[9] 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 2020), and signals that are a realization of a Brownian motion (Fudenberg, Romanyuk, and Strack 2017).

realization.

Definition 1 (Updating Rule)**.** *An* updating rule $h : \mathcal{Z} \to \Delta(\Omega)$ *is a function that maps a signal realization to a posterior belief over the state space such that* $z \mapsto h(z)$ *is measurable and not constant* ν*-almost everywhere.*

An agent uses *updating rule* $h(z)$ if, fixing prior p, for each $i = 1, ..., N$, the agent assigns probability $h(z)_i$ to state ω_i after observing signal realization $z \in \mathcal{Z}$.[10] 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.[11]

An updating rule can capture many common biases studied in the literature. For example, suppose $\Omega = \{\omega_1, \omega_2\}$ and define the 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 $\alpha \in (0, 1)$. We refer to bias that arise from the updating rule as *retrospective bias,* since it arises following the signal realization.

A *forecast* is an agent's prediction of how she will form beliefs after observing the signal—that is, it is a distribution over posterior beliefs. In order for the forecast to be compatible with the signal, the space of posteriors cannot be "larger" than the space of signal realizations. 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

[10] We start with a fixed prior, but one could instead define an updating rule as a mapping from the signal and the prior to a posterior in order to study dynamics or comparative statics with respect to the prior. Our framework and analysis naturally extends to this set-up, albeit with more cumbersome notation. See Section 7.1 for the formal treatment of this more general set-up.

[11] This defines 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).

measure zero sets of the information structure.

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.*

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. 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. 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 over-weights extreme beliefs and underweights intermediate beliefs, while underprecision overweights intermediate beliefs and overweights extreme beliefs. We provide more specific examples of forecasts in Section 6.2. We refer to bias that arises from the forecast as *prospective bias*, since it stems from a prediction of what the signal will be.

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 Section 5.

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 $i = 1, ..., N$.[12]

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 arise with probability zero under the misspecified model occur with positive probability un-

[12] 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 $i = 1, ..., N$.

der the correctly specified model, and that the misspecified model does not place positive probability on 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. A model also induces a forecast, which is the unconditional distribution of posteriors according to the model. Specifically, $(\hat{\mu}_i)_{\omega_i \in \Omega}$ 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.

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, and when this representation is unique. To this end, we formalize what it means for a misspecified model to represent an updating rule, in the sense that the model prescribes the same posterior beliefs as the updating rule following each signal realization, and for the model to represent a forecast, in the sense that the model prescribes the same forecast over posterior beliefs.

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 ν-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*

$X \subset \Delta(\Omega)$:

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

If an updating rule maps a positive measure of signal realizations to the same posterior belief and can be represented by a given misspecified model, then any other model that shifts mass between the signal realizations that map to the same posterior will also represent this updating rule. However, the difference between these models is trivial in an economic sense 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 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. $\mathcal{P}_Z \equiv \{Z \in \mathcal{F} : h(Z) = X$ for all Borel sets $X \subset \Delta(\Omega)\}$.*

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 Comparison of 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. The way a behavioral decision-maker 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. This motivates the two components of our behavioral learning set-up: we formalize this retrospective bias in the form of an updating rule, or updating rule, and this prospective bias in the form of a forecast. While misspecified models are generally time-consistent, misspecification allows for a stochastic version of this phenomena. 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.

The updating rules approach is often used to model a specific form of bias or belief-updating error in a specific context. In general, this literature studies beliefs and behavior for specific parameterizations of a bias. In contrast, the model misspecification approach is often applied to general learning environments and is used to simultaneously model a range of biases within the same framework. For example, recent work in the literature on learning with model misspecification establish general convergence results and show that in many situations, agents' behavior converges to a Berk-Nash equilibrium (Bohren and Hauser 2021; Frick et al. 2020; Fudenberg et al. 2020). Connecting these approaches makes it straightforward to apply the tools developed in the misspecified learning literature to generalize the stylized results from the updating rules literature to a larger set of parameterizations of a given bias. For instance, 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 their specific choice of information structure (i.e. binary signals).

To a large extent, the literature on behavioral biases has focused on updating rules, which allow for 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. On the other hand, a misspecified model of belief formation is complete, in the sense that it describes all aspects of the environment necessary for analysis. Therefore, mapping updating rules into the misspecified model approach makes it possible to study the implications of a given bias in a much richer set of economic environments.

3 Simple Example

Consider a binary state space $\varOmega = \{L, R\}$ with a flat prior $Pr(R) = 1/2$, and a 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 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_k^ω for each signal z_k in each state ω, i.e. $\hat{\mu}^\omega = (m_1^\omega, m_2^\omega, m_3^\omega, m_4^\omega)$ with $\sum_{k=1}^4 m_k^\omega = 1$.

We first show that when an updating rule is considered in isolation, if it can be represented by a misspecified model then this model is generally not unique. As we will show in Lemma 1, a very mild condition determines whether an updating rule h can be represented by a misspecified model. In this example, the condition requires the updating rule to map at least one signal to a posterior above the prior and similarly one signal to a posterior below the prior, i.e. $\min_k h(z_k) < 1/2 < \max_k h(z_k)$. In fact, a continuum of misspecified models represent such an h: any solution $(m_1, m_2, m_3, m_4) \in \varDelta$ to $\sum_{k=1}^4 h(z_k) m_k = 1/2$ pins down a model with signal distribution $m_k^R = 2h(z_k) m_k$ in state R and signal distribution $m_k^L = 2(1-h(z_k))m_k$ in state L that represents h.[13] Aside from knife-edge cases, $\sum_{k=1}^4 h(z_k)m_k = 1/2$ has multiple solutions. Each corresponding model induces a unique forecast, which assigns probability $m_k = m_k^R/2 + m_k^L/2$ to posterior belief $h(z_k)$.

The forecast determines the prospective bias. Therefore, for a given updating rule, the chosen model to represent it will pin down the prospective bias through the induced forecast. Different models that represent the same updating rule can lead to very different predictions depending on the forecast they induce.

We develop a similar result for forecasts. As we will show in Lemma 2, a forecast $\hat{\rho}$ in this example can be represented by a misspecified model if and only if it yields an expected posterior equal to the prior, i.e. $\sum_{x \in \text{supp}\,\hat{\rho}} x\hat{\rho}(x) = 1/2$. This condition, which we refer to as plausibility, ensures that the decision maker believes that their prior captures all their current uncertainty about the state. For example, the forecast $\hat{\rho} = \{.5, .5\}$ with support $\{x, 1-x\}$ for some $x \in (0, .5)$ can be represented by a misspecified model since $.5x + .5(1-x) = .5$ satisfies the plausibility condition. One such model is $m_1^R = x/2$, $m_2^R = x/2$, $m_3^R = (1-x)/2$ and $m_4^R = (1-x)/2$ in state R, and similarly for state L substituting $1-x$ for x.[14] This model induces updating rule $h(z_1) = h(z_2) = x$ and $h(z_3) = h(z_4) = 1-x$.[15] Again, multiple models can represent a given forecast. The model $m_1^R = x/3$, $m_2^R = x/3$, $m_3^R = x/3$ and $m_4^R = 1-x$ 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 $1-x$.[16] In fact, for any updating rule that assigns at least one signal to each posterior x and $1-x$, it is possible to find a misspecified model that induces this updating rule and represents $\hat{\rho}$.

The updating rule determines the retrospective bias. Therefore, which model is chosen to represent a given forecast determines the retrospective bias through the induced updating rule. For example, if the updating rule generated by the correct model maps

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

[14] To see that this model represents $\hat{\rho}$, note that from Bayes rule, it induces posterior belief $m_k^R/(m_k^R + m_k^L)$ 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 $\hat{\rho}(x) = \hat{\mu}(\{z_1, z_2\}) = (m_1^R + m_1^L)/2 + (m_2^R + m_2^L)/2 = .5$ and $\hat{\rho}(1-x) = \hat{\mu}(\{z_3, z_4\}) = .5$ by an analogous calculation, as desired.

[15] In fact, any $\alpha \in (0,1)$ pins down a model that represents $\hat{\rho}$ with signal distribution $m_1^R = \alpha x$, $m_2^R = (1-\alpha)x$, $m_3^R = \alpha(1-x)$ and $m_4^R = (1-\alpha)(1-x)$ 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 defined in Definition 5.

[16] To see that this model represents $\hat{\rho}$, note that from $h(z_1) = h(z_2) = h(z_3) = x$ and $h(z_4) = 1-x$, it induces forecast $\hat{\rho}(x) = (m_1^R + m_2^R + m_3^R)/2 + (m_1^L + m_2^L + m_3^L)/2 = .5$ and similarly $\hat{\rho}(1-x) = .5$.

$\{z_1, z_2\}$ 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. Therefore, different models that represent the same forecast can lead to very different predictions depending on the updating rule they induce.

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 relevant information about the decision environment?

Our first main result answers these questions. We establish a necessary and sufficient condition for a forecast to be compatible with a given updating rule, in that the pair can be jointly represented by a misspecified model, and vice versa. This condition is quite mild: in this example, any plausible forecast and updating rule can be jointly represented provided that the support of the forecast is equal to the image of the updating rule, i.e. given h, $\operatorname{supp} \hat{\rho} = \{h(z_1), h(z_2), h(z_3), h(z_4)\}$. Therefore, representing a given retrospective bias does not place very strong restrictions on the set of prospective biases that can arise alongside it, and vice versa. Further, we establish that this representation is unique. Hence, the updating rule and forecast jointly pin down a complete model for analysis.

Given that updating rules are much more frequently studied in the literature, we next turn to the question of how to select a forecast to pair with a given updating rule. We focus on two classes of forecasts that have desirable properties in relation to the correct model: introspection-proof forecasts and naive consistent forecasts. For a given updating rule h, introspection-proofness imposes the requirement that the forecast is correct with respect to h. Suppose the updating rule generated by the correct model maps $\{z_1, z_2\}$ to posterior x and $\{z_3, z_4\}$ to $1 - x$ and consider the updating rule that maps $\{z_1, z_2, z_3\}$ to x and z_4 to $1-x$. Then the introspection-proof forecast corresponds to $\hat{\rho}(x) = \mu(z_1) + \mu(z_2) + \mu(z_3)$ and $\hat{\rho}(1-x) = \mu(z_4)$, where μ is the correct unconditional model, as this is the accurate probability of each posterior given the biased updating rule. Naive-consistency imposes the requirement that the forecast is equal to the accurate forecast for an agent with the correct model. In this example, this corresponds to $\hat{\rho}(x) = \mu(z_1) + \mu(z_2)$ and $\hat{\rho}(1-x) = \mu(z_3) + \mu(z_4)$, since this is the correct forecast given an unbiased updating rule.

4 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 derive an analogous result for a forecast. We then establish a necessary and sufficient condition on an updating rule and forecast pair for them to be jointly represented by a misspecified model, and shows that this model is essentially unique.

4.1 Representing Updating Rules

We begin by fixing an updating rule and characterizing when it can be represented by a misspecified model. Let $\mathcal{N}(h) \equiv \operatorname{supp} \rho_h$ denote the support of the accurate forecast ρ for updating rule h, which is well defined, 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.[17] An important feature of Bayesian updating is that the posterior belief is equal to the prior in expectation. We use this martingale property of beliefs to characterize necessary and sufficient conditions for there to exist a misspecified model that represents an updating rule.

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

This result extends Lemma 1 from Shmaya and Yariv (2016) to a more general signal space.[18] 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.[19] The space of posterior beliefs has more structure than the signal space, which we leverage for this characterization. In particular, given an updating rule h, if the prior belief does not lie in the set $S(h)$ as defined in Eq. (7), then it is impossible for the martingale property to hold for any "full support" measure. Therefore, the prior belief must lie in $S(h)$ for it to be possible to represent h. It also turns out that this condition is sufficient for the prior to be the center of mass for *some* distribution.

The condition in Lemma 1 is very weak. For all practical purposes, it only rules

[17] 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.

[18] 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.

[19] These conditions prevent probability 0 events from occurring with positive probability and rule out misspecified models that, for instance, create atoms by placing positive probability on signals that lie in the support but occur with probability 0 under the correctly specified model.

out pathological updating rules such as an updating rule that increases the posterior probability of some state ω following all possible signal realizations. Therefore, this result establishes that most updating rules of interest can be represented by a misspecified model. However, this representation is generally not essentially unique. As we saw in the example in Section 3, there are often many misspecified models that represent a given updating rule. A natural next question is which representation one should select, which we address in Section 5.

4.2 Representing Forecasts

We next develop an analogous result to Lemma 1 for forecasts. A forecast is *plausible* if the expected posterior, taken with respect to the agent's forecast, is equal to the prior.

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

In order for the forecast to be represented by a misspecified model, it must be plausible. In fact, this is both a necessary and sufficient condition for a forecast to be represented by a misspecified model.

Lemma 2 (Existence of a Forecast Representation)**.** *There exists a misspecified 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.*

Plausibility is a necessary property of Bayesian updating: a Bayesian agent always believes that on average, her posterior will be equal to her prior. In other words, even a misspecified Bayesian agent does not believe that she is systematically biased. Unlike the updating rule, which needs very little structure to be consistent with a misspecified model, a forecast must satisfy this strong requirement of Bayesian learning. However, while the plausibility requirement rules out many forecasts, it still allows for a broad class of forecasts as we illustrate in the following example.

Example 1. *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* $x = (x_L, x_R) \in \Delta$ *is a posterior belief. This corresponds to the family of beta distributions with mean* $1/2$.[20] *Any forecast from this family is plausible since*

[20] Note that these are indeed forecasts, as $g(z) = (z, 1-z)$ satisfies the mutually absolutely continuous

$\int_{\Delta(\Omega)} x_i \, d\hat{\rho}_\theta(x) = 1/2$ *for* $i = L, R$. *Therefore, any such forecast can be represented by a misspecified model.*

As in the case of updating rules, a forecast on its own does not necessarily identify a unique misspecified model. In fact, a continuum of misspecified models can be consistent with a give forecast. For example, consider the case of $\theta = 1$ in Example 1. This corresponds to the uniform forecast, i.e. $d\hat{\rho}(x) = 1$. For any $\gamma > 0$, the misspecified 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 $\hat{\rho}$.[21] From Bayes rule, this model induces posterior belief $d\hat{\mu}^R(z)/(d\hat{\mu}^R(z) + d\hat{\mu}^L(z)) = z^\gamma$. Each value of γ captures a different level of retrospective bias: as γ increases, the updating rule slants information more towards state R.

4.3 Decomposition

As shown above, an updating rule or a forecast on its own does not identify a unique misspecified model. In the next result, we show that an updating rule and a forecast jointly identify a misspecified model that is essentially unique. In other words, a misspecified model can be decomposed into a "prospective bias", the forecast, and a "retrospective bias", the updating rule. We also show that neither component imposes much structure on the other.

Given that we focus on misspecified models that are mutually absolutely continuous with respect to the correctly specified model, we must place some restriction on how the forecast and updating rule jointly behave over measure 0 sets. Specifically, 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 h. This corresponds to the subjective forecast $\hat{\rho}$ being mutually absolutely continuous with the accurate forecast ρ_h. It is straightforward to see why this condition is necessary to find a misspecified model to represent the forecast and updating rule. It turns out that it is also sufficient, and therefore, is the only joint requirement on the updating rule and forecast for such a representation to exist.

Theorem 1 (Decomposition). *Consider a forecast $\hat{\rho}$ and an updating rule h. Let ρ_h be the accurate forecast for h. There exists a misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ that represents h and $\hat{\rho}$ if and only if $\hat{\rho}$ is plausible and mutually absolutely continuous with ρ_h. If such a model exists, it is essentially unique and defined by*

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

condition.

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

for any measurable $Z \subset \mathcal{Z}$, where μ is the unconditional measure over signals in the correct model.

This result shows that the updating rule and the forecast are the "essential" components of a misspecified model: together they completely pin down a misspecified model. It also shows that these components are largely independent of each other: aside from the mild restriction that the forecast and updating rule have the same measure zero sets, the forecast does not place restrictions on the updating rule and vice versa. Thus, a misspecified model is fully pinned down by the retrospective and prospective biases that it induces, and these two forms of bias are largely separate properties of the model—they do not contain overlapping restrictions. For instance, optimistic updating does not imply optimistic forecasting. This insight has appealing consequences for economic modeling, as it allows for the interaction between different natural biases within the same misspecified model.

This representation provides a powerful tool for the construction of models of biased learning, as it reduces a misspecified model into two components that transparently relate to the conceptual properties the model seeks to capture. Rather than specifying a family of conditional probability distributions—which is potentially quite complicated and removed from the conceptual bias of interest—one can simply write down a reasonable parameterization of the desired retrospective and prospective biases. Together these biases completely capture how a misspecified agent's behaviour will depart from that of a correctly specified agent.

It may, at first glance, appear odd that the correctly specified model appears in the representation in Eq. (9). This is innocuous: since the forecast doesn't place structure on the unconditional distribution over signals that induce the same posterior, using the correctly specified distribution to determine randomizations over each of the elements of $h^{-1}(x)$ is a simple way to ensure that the correctly specified and misspecified models are mutually absolutely continuous.

Intuition for proof. We first prove an intermediate result that significantly simplifies the process of finding misspecified model(s) to represent a given updating rule. Given either a state-contingent distribution $\hat{\mu}_i$ in state ω_i or the unconditional distribution $\hat{\mu}$, we establish a necessary and sufficient condition for this distribution to be part of a misspecified model representing a given updating rule. Moreover, if a model that includes this distribution exists, then 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 measure and it cannot be part of a model that represents the updating rule.

Theorem 1 shows that, together with a forecast, the updating rule identifies a unique misspecified model. However, it also indicates that a plethora of forecasts induce the same updating rule, and each pair is represented by a different misspecified model. If one's goal is to select a natural misspecified model to represent an updating rule, then Theorem 1 provides no guidance on how to do so. This motivates the remainder of the paper, in which we explore which forecasts to pair with an updating rule in order to construct misspecified model representations with certain desirable properties.

5 Selecting Forecasts

In this and the following sections, we use the decomposition into forecasts and updating rules to identify natural restrictions on the forecast that provide conceptual guidance for which model to select. The first condition—introspection-proofness—imposes structure on how the misspecified model relates to the correctly specified model. The second condition—naive consistent forecasting—is a condition on the agent's belief about how he will form beliefs in the future. Each condition uniquely selects a misspecified model when such a model exists.

5.1 Introspection-Proof Models

A common concern with the use of misspecified models as a modeling tool is that, given a large number of observations, an agent may observe a pattern that is incredibly unlikely under her misspecified view of the world. For example, she may observe an extreme violation of the law of large numbers. Therefore, if an agent forms her view of the world by observing a lot of data—in the context of this framework, an infinite sequence of independent draws of the signal and state—one might worry that the agent could, through introspection, come to realize that she is misspecified. Motivated by this concern, we define the following notion of an introspection-proof model.

Definition 7 (Introspection-Proof Model)**.** *A misspecified 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 $Z \in \mathcal{F}$.*

Using the tools developed in Theorem 1, we establish a necessary and sufficient condition for an updating rule and forecast to have an introspection-proof representation—namely, the forecast must be plausible and accurate with respect to the updating rule. When such a representation exists, the following result also constructs the corresponding model.

Proposition 1. *Fix an updating rule h. There exists an introspection-proof misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$ that represents h and the accurate forecast ρ_h if and only if*

ρ_h is plausible. If this representation exists, then it is unique and defined by

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

for all measurable $Z \in \mathcal{F}$. There is no introspection-proof misspecified model that represents h and a forecast $\hat{\rho} \neq \rho_h$.

This result follows from Theorem 1 and the observation that $\hat{\rho}^{IP} = \rho_h$, and therefore, trivially $\hat{\rho}^{IP}$ and ρ_h are mutually absolutely continuous. The requirement that ρ_h is plausible is quite restrictive. Recall that a plausible forecast ρ satisfies $\int_{\Delta\Omega} x_i \, d\rho(x) = p_i$ for all i, which by the change of variables formula becomes $\int_{\mathcal{Z}} h_i(z) \, d\mu = p_i$. So the accurate forecast is plausible only if the updating rule is on average equal to the prior under the correctly specified signal distribution.[22]

When an updating rule is represented by an introspection-proof misspecified model, then the agent observes *exactly* the distribution of signal realizations that she expects, given her model of the world. Regardless of the information that she selects to validate her model, she does not observe violations that would cause her to second guess this worldview. Alternatively, introspection-proofness can be viewed as a robustness criteria for the updating rule: from the perspective of an analyst who only observes signal realizations, any agent who updates using an updating rule that admits an introspection-proof misspecified model is indistinguishable from a correctly-specified Bayesian agent, and therefore, this agent will naturally pass any tests that the analyst designs to detect Bayesianism. If an updating rule can't be represented by an introspection-proof model, then with infinite data the analyst will be able to reject that the agent is a correctly-specified Bayesian.

The condition required the forecast to be plausible is reminiscent of the Bayes-plausibility condition in Kamenica and Gentzkow (2011). It requires that the forecast of the expected posterior belief from the updating rule is equal to the prior, where the expectation is taken with respect to the *true* unconditional signal distribution. If this condition does not hold, then it is impossible for the true distribution over posterior beliefs to be equal to the forecast, and thus it can't be a forecast from an agent using a misspecified model. An agent's posterior beliefs must satisfy the martingale property with respect to the subjective unconditional signal distribution. Under the introspection-proof condition, this simplifies to Eq. (10). Moreover, Eq. (10) is sufficient to construct

[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 E, we provide examples of misspecified models that are Blackwell less informative than the true model but not introspection-proof and Blackwell more informative than the true model but not introspection-proof.

an introspection-proof misspecified model and it uniquely pins down such a model.

Introspection-proofness provides the researcher with a natural choice of misspecified model to represent a given updating rule. An introspection-proof misspecified model must preserve the "center of mass" of beliefs, but otherwise has the freedom to arbitrarily distort the spread of these beliefs. This makes it possible to represent conceptual biases such as conservatism in belief updating or overreaction to new information with an introspection-proof misspecified model, as we illustrate in the following example.

Example 2 (Conservatism)**.** *Consider a common updating rule for conservatism in belief updating, $h(z) = \lambda h_B(z) + (1-\lambda)p$ for some $\lambda \in (0,1)$ (Epstein et al. 2008; Hagmann and Loewenstein 2019; Gabaix 2019). In other words, the posterior belief is a weighted average of the Bayesian posterior and the prior. 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.*

On the other hand, updating rules that systematically shift beliefs in one direction, such as partisan bias, can never be paired with forecasts that satisfy this condition, as any reasonable parameterization of such a bias must shift the center of mass of beliefs.

Example 3 (Partisan Bias)**.** *Consider the following model of partisan bias. There are two states of the world $\omega \in \{\omega_1, \omega_2\}$ and an agent who updates according to update rule $h(z)_1 = (h_B(z)_1)^2$ and $h(z)_2 = 1-(h_B(z)_1)^2$. In this model, after any signal the decision maker has beliefs that are more favorable to state ω_2 than the correctly specified agent. Under the accurate forecast ρ_h*

$$\int_0^1 x_i \, d\rho_h(x) = \int_{\mathcal{Z}} h(z)_i \, d\rho_h(h(z)) = \int_{\mathcal{Z}} h(z)_i \, d\mu(z).$$

But, $\int_{\mathcal{Z}} h(z)_1 \, d\mu < \int_{\mathcal{Z}} h_B(z)_1 \, d\mu = p_1$, so the accurate forecast is not plausible. This argument clearly holds not only here, but more generally for any bias that systematically skews the updates in one direction.

Moreover, this condition requires a certain amount of complexity in how the updating rule distorts updates which prevents many simple updating rules from satisfying it. For example, the canonical model of overreaction cannot satisfy it.

Similar approaches to introspection-proof models have been used in existing work to construct plausible restrictions on the space of misspecified models being considered. For example, Spiegler (2016) uses a similar condition to connect misspecified causal graphs—as opposed to updating rules—to a misspecified model. He requires a condition resembling introspection-proofness on each link of the graph, which pins down a misspecified probability distribution over the outcome of interest. Mailath and Samuelson (2019) study a model of omitted variable bias, where the set of omitted variables combined with an introspection-proof condition pin down the misspecified model agents use.

Alternative Notions of Introspection. This notion of introspection-proofness is relatively strong, in that it requires the subjective unconditional signal measure to exactly match the correct unconditional measure. With a bit more structure on the signal space, one could do a conceptually similar exercise with weaker requirements. For example, one could require that mean of the subjective unconditional signal measure matches the mean of the correct unconditional signal measure, but allow the subjective unconditional signal measure to differ from the correct unconditional signal measure on other dimensions, such as the variance, that may be harder to detect than differences in means. We explore alternative definitions of introspection-proof in Appendix D.

5.2 Naive Consistent Forecasts

Another natural forecast is one in which the agent naively predicts that she will form accurate beliefs in the future. This property is analogous to common naivete assumptions made in many behavioral models (e.g. models of time inconsistency), and has previously been made in models of biased learning such as Benjamin et al. (2019); Bohren and Hauser (2021). We say a forecast exhibits naive consistent forecasting when the agent's forecast of future beliefs is identical to the true forecast of future beliefs when an agent updates using Bayes rule.

Definition 8 (Naive Consistent Forecast)**.** *Given prior p, the* naive-consistent forecast *is defined by the accurate forecast with respect to the Bayesian updating rule h_B, $\hat{\rho}^{NCF} \equiv \rho_B$.*

Informally, naive consistent forecasting requires that for any Borel set of posteriors X, the agent using forecast $\hat{\rho}^{NCF}(X)$ places the same probability on the posteriors lying in X as the accurate forecast with respect to the Bayesian updating rule $h_B(z)$. So, an agent who exhibits naive consistent forecasting places the same probability as a correctly specified agent that she will form a posterior belief in set X. That is, the forecast $\hat{\rho}^{NCF}$ is the distribution of posteriors that a correctly specified agent would generate. As this forecast is the forecast a correctly specified Bayesian has, it is always plausible.

Using Theorem 1, we establish that mutual absolute continuity of $\hat{\rho}^{NCF}$ and ρ_h is the only property that an updating rule must satisfy to be represented by a misspecified model that exhibits naive consistent forecasting.

Proposition 2 (Naive Consistent Representation)**.** *Fix an updating rule $h(z)$, and the naive consistent forecast $\hat{\rho}^{NCF}$. There exists a misspecified model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})^N$*

that represents $h(z)$ and $\hat{\rho}^{NCF}(X)$ if and only ρ_h and $\hat{\rho}^{NCF}$ are mutually absolutely continuous. If this representation exists, it is essentially unique and defined by

$$\hat{\mu}_i(Z) = \mu_i(\{z : h_B(z) \in h(Z)\}) \tag{11}$$

for all $Z \in \mathcal{P}$.

The absolute continuity condition is relatively mild: it simply requires that the set of posteriors that are possible under the updating rule are the same as the set of posteriors that the agent believes are possible under the forecast. Formally, $\mu(\{z : h(z) \in X\}) = 0$ if and only if $\hat{\rho}^{NCF}(X) = 0$. Naive consistent forecasting also implies that in each state ω_i, an analogue of the naive consistent forecasting property holds. In each state ω_i, a naive consistent forecast predicts that the posterior will be in set X with the same probability as a correctly specified agent predicts that the posterior will be in this set. This is a consequence of Lemma 3 and is straightforward to see from Bayes rule.

In Bohren and Hauser (2021), we impose naive consistent forecasting to study social learning in the presence of overreaction or partisan bias.

Example 4 (Overreaction)**.** *Consider a binary state space $\omega \in \{\omega_1, \omega_2\}$ and an agent who updates using the updating rule defined by*

$$\frac{h(z)_2}{1 - h(z)_2} = \frac{p}{1-p}\left(\frac{\frac{d\mu_R}{d\nu}(z)}{\frac{d\mu_L}{d\nu}(z)}\right)^{\gamma}.$$

in state ω_2, with complementary probability $h(z)_1 = 1 - h(z)_2$ in state ω_1 and overreaction parameter $\gamma > 1$. If the random variable $h(z, 1/2)_2$ is continuous with support $[0, 1]$, then this admits a misspecified model that exhibits naive consistent forecasting at all priors.

This naive consistent representation is particularly convenient in the social learning game studied in Bohren and Hauser (2021). In that game, a sequence of short-lived agents see a private signal z_t with they interpret using the updating rule and choose an action $a_t \in \{L_1, L_2, R_1, R_2\}$. An agent receives utility $u(a, \omega)$ from their action choice where $u(L_1, \omega_1) > u(L_2, \omega_1) > u(R_2, \omega_1) > u(R_1, \omega_1)$ and $u(R_1, \omega_2) > u(R_2, \omega_2) > u(L_2, \omega_1) > u(L_1, \omega_1)$. If all agents use a naive consistent forecast, then at any belief the update following each action is exactly the update that an agent would make in the correctly specified game. That is, at any prior p_t, the perceived probability of each action is equal to the probability that a correctly specified player would take that action, the update a player would make if, for instance, they didn't take into account that themselves and others were interpreting information using the updating rule $h(z)$, but instead believed that all agents were updating correctly. This does not imply that the learning dynamics in this problem are the same, or even similar to, that of the correctly specified *social learning game. Since agents are in fact using a biased updating rule to interpret information, the realized action frequency is different from the frequency that would be realized in the game with correctly specified agents. As we show in the paper, this can cause beliefs to fail to converge.*

5.3 Biased Forecasts with Accurate Updating

Introspection-proof models and naive consistent forecasting both pin down a forecast with respect to the correctly specified signal distribution. That is, the forecast is either accurate with respect to the agent's updating rule or with respect to the Bayesian updating rule. This isolates the retrospective bias from any prospective bias by having the agent naively form forecasts as-if they were correctly specified. One can also consider situations in which an agent correctly interprets signals (i.e. uses the Bayesian updating rule $h_B(z)$) but has a biased forecast. These situations such down any retrospective biases and only allow for prospective biases. The following definition formalizes this notion of a retrospectively correct model.

Definition 9 (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 $\omega_i \in \Omega$,*

$$\frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z)}{\sum_{j=1}^N p_j \frac{d\hat{\mu}_j}{d\nu}(z)} = h_B(z)_i \tag{12}$$

ν-almost everywhere.

The following corollary follows immediately from Theorem 1.

Corollary 1. *Fix a forecast $\hat{\rho}$. There exists a retrospectively correct model $(\hat{\mu}_i)_{\omega_i \in \Omega} \in \Delta^*(\mathcal{Z})$ that represents $\hat{\rho}$ if and only if $\hat{\rho}$ and ρ_B are mutually absolutely continuous and $\hat{\rho}$ is plausible.*

This establishes that many forecasts are consistent with Bayesian updating. An agent can form very wrong predictions about their future beliefs, but still update correctly after observing a signal. Therefore, the misspecified model approach can be used to capture prospective biases without needing to also allow for retrospective bias.

6 Applications

We next provide three applications to demonstrate the results from Sections 4 and 5. In the first application, we start with the updating rule approach and show how the introspection-proof condition is a natural requirement to impose when selecting an updating rule in a dual-selves model with self-image concerns. In the second application, we start with the misspecified model approach and show how the updating rule and forecast decomposition yields insight into the way the different components of bias induced by the misspecified model impact the design of lending contracts. Continuing with this environment, in the final application we allow for borrower heterogeneity in updating and show how second degree price discrimination is not possible when agents use the naive consistent forecast.

6.1 Optimal Bias with Self-Image Concerns

The first application is a dual-selves model with self-image concerns, where a manager first chooses an updating rule to interpret information about ability, then uses this updating rule to evaluate himself and workers. We show that the introspection-proof constraint places a natural upper bound on the level of motivated reasoning that the manager exhibits. Moreover, the manager compensates for overestimating the ability of workers sharing a group identity with the manager by underestimating the ability of workers from the other group identity, despite group identity being orthogonal to productivity. In contrast, without the introspection-proof constraint, the manager does not distort beliefs for the other group identity. Therefore, self-image concerns in combination with introspection-proof updating leads to inaccurate beliefs about workers from all group identities, whereas self-image concerns on their own only lead to inaccurate beliefs about workers that share a group identity with the manager.

Set-up. Suppose a manager evaluates a worker. The worker has either low or high ability, $\omega_w \in \{L, H\}$, drawn with equal probability. The manager selects evaluation $a \in [0, 1]$ for the worker. Before evaluating the worker, the manager observes a two-dimensional signal $z_w = (y_w, t_w)$. 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 $t_w \in \{M, F\}$, which we assume is independent of (y_w, ω_w) and distributed according to $q \equiv Pr(M)$. This group identity can be interpreted as a demographic variable that is readily observed from interacting with the worker.

Analogous to the worker, the manager has ability $\omega_m \in \{H, L\}$ drawn with equal probability. The manager also observes his own test performance $y_m \in \{b, g\}$, which has the same distribution as the worker's test performance y_w. Without loss of generality, assume that the manager's group identity is $t_m = M$, and therefore the manager's two-dimensional signal is $z_m = (y_m, M)$. The manager's ability and signal are independent of the worker's ability and signal.

We consider a dual-selves model where the manager's first self chooses an updating rule for interpreting the signal, and the second self uses this rule to update his beliefs about his own ability and the worker's ability then evaluates the worker. Before the signals are realized, the first self chooses an updating rule h for the second period self to use. 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. After the signals are realized, the second self updates his belief about his own ability to $h(z_m)$ and his belief about the worker's ability to $h(z_w)$, then chooses evaluation a.

The manager cares about his self-image, captured by the second self's belief that he is of high ability, and accurately forecasting the worker's ability,

$$u(a, \omega_w, z_m) = h(z_m) - c(\mathbb{1}_{\{\omega_w = H\}} - a)^2, \tag{13}$$

where $c > 1/2q(1-\alpha)$ to ensure that the manager puts sufficient weight on accurately evaluating the worker.[23] Each self maximizes his expected utility, 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, 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 h to maximize

$$E[h(z_m) - c(\mathbb{1}_{\{\omega_w = H\}} - h(z_w))^2]. \tag{14}$$

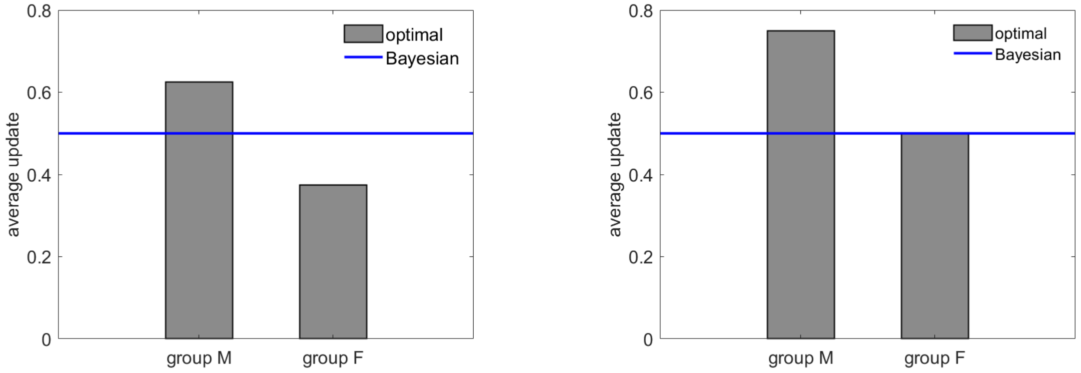
Given that the manager must choose the same updating rule to interpret his own and the worker's signals, the choice of updating rule influences both the payoff from self-image and the payoff from the accuracy of the evaluation. Self-image concerns lead the manager to exhibit motivated reasoning, i.e.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

The Optimal IP Updating Rule. The first self may wish to select an updating rule such that the second self does not observe a pattern of signals that, after evaluating a sufficiently large number of workers, is at odds with his forecast about his beliefs—in other words, an introspection-proof updating rule. We next characterize the optimal introspection-proof updating rule and compare it to the optimal updating rule without this constraint.

From Proposition 1, an updating rule has an introspection-proof representation if

$$\sum_{y \in \{b,g\}} \frac{1}{2}(qh(y, M) + (1-q)h(y, F)) = \frac{1}{2}. \tag{15}$$

[23] 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 ($\alpha = .7$, $c = 4$, $q = .5$).

In order to inflate self-image and simultaneously satisfy the introspection-proof condition, which requires consistency with the observed signal distribution, the manager must compensate for overestimating the ability of group M workers by deflating the interpretation of test performance for group F workers, thereby underestimating their ability. Given Bayesian updating rule $h_B(g,t) = \alpha$ and $h_B(b,t) = 1-\alpha$, this leads to the following result.

Proposition 3. *The optimal introspection-proof updating rule inflates the interpretation 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 updating rule features inaccurate beliefs about both groups that endogenously emerge from the interaction between self-image concerns and the introspection-proof constraint. Fig. 1(a) illustrates this updating rule.

The optimal distortion for group M is decreasing in q: when the hiring pool is more similar to the manager in terms of group identity, the manager uses a less biased updating rule for group M, resulting in more accurate evaluations. This is because 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 inflating the evaluation of the larger share of group M workers. In contrast, the optimal distortion for group F is independent of q: as q increases, distortion is less costly for this group since it comprises a smaller share of workers, but also less beneficial as a means to balance the distortion against group M since less distortion against group M is desired. It turns out that these two forces exactly balance for the linear-quadratic payoff form Eq. (13).

The Optimal Unconstrained Updating Rule. When the updating rule is not constrained to be introspection-proof, self-image concerns still lead the manager to inflate the interpretation of test performance for members of group M. However, there is no reason to distort the information for group F. This leads to the following result.

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

The optimal updating rule only features inaccurate beliefs about the manager's group. This contrasts with the optimal introspection-proof updating rule, in which the introspection-proof constraint forces the overestimation of own group ability to be counterbalanced by underestimating the ability of the other group. Therefore, in settings where agents 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.

Without the discipline of the introspection-proof constraint, distorting self-image is only costly for the manager when he is hiring type M workers. This leads to a higher level of signal distortion for group M relative to the optimal introspection-proof updating rule. Thus, the introspection-proof constraint serves as a natural moderator to the magnitude of the motivated reasoning bias that can emerge. Without this constraint, the manager stands to lose less from distorting his belief about his ability, as he does not have to compensate for this distortion by also distorting the perception of group F. Fig. 1(b) illustrates the optimal unconstrained updating rule. Although there is less belief distortion for group F, the higher distortion for group M dominates and leads to less accurate evaluations overall.[24]

6.2 Lending Contracts with Bias

In this application, we show how the decomposition in Theorem 1 can be used to determine how the retrospective and prospective biases induced by a misspecified model impact the optimal lending contract. Each form of bias has a distinct and intuitive impact on the structure of the optimal contract. Using a parameterized family of under- and overconfident forecasts, we then show that a lender leverages overconfident forecasts by charging a high upfront price for a favorable interest rate, and leverages underconfident forecasts by offering an upfront discount then subsequently charging a high interest

[24] The expected loss $E((\mathbb{1}_{\omega=H} - h(z_w))^2)$ from the evaluation using the optimal unconstrained updating rule is $1 - \alpha^2 - (1-\alpha)^2 + 1/8qc^2$, which is larger than the expected loss from using the optimal introspection-proof updating rule, $1 - \alpha^2 - (1-\alpha)^2 + (1-q)/(8qc^2)$.

rate.

The Entrepreneur's Borrowing Problem. Consider a setting in which a lender offers an entrepreneur access to capital. The entrepreneur has a project that is either low or high quality, $\omega \in \{L, H\}$, drawn with equal probability. In period $t = 0$, the entrepreneur chooses whether to pay origination fee $c > 0$ to open a line of credit with the lender. In period $t = 1$, the entrepreneur receives signal $z \sim \mu^{\omega}$ with support $\mathcal{Z} \subseteq [0, 1]$ about the quality of her project. After observing the signal, if the entrepreneur opened a line of credit then she chooses an amount $I \geq 0$ to borrow at rate $r > 0$. If the entrepreneur did not open a line of credit, then she cannot borrow, $I = 0$. The entrepreneur invests all of the money she borrows in the project. In state L, the project leads to return $g(I, L) = 0$ for any level of investment I. In state H, the return is increasing in the level of investment by the entrepreneur, $g(I, H) = 2\sqrt{I}$. After realizing returns, the entrepreneur pays back her loan. The entrepreneur's payoff is

$$g(I, \omega) - (1 + r)I - c * \mathbb{1}_{opencredit} \tag{16}$$

The entrepreneur has a misspecified model of the signal process, captured by $\hat{\mu}^L(z)$ and $\hat{\mu}^H(z)$. We let h denote the induced updating rule, where in a slight abuse of notation $h(z)$ is the entrepreneur's subjective probability that quality is high following signal z, and let $\hat{\rho}$ denote the induced forecast, where in a slight abuse of notation $\hat{\rho}(x)$ is the probability of posterior x that the state is H. From Theorem 1, $\hat{\rho}$ is plausible and mutually absolutely continuous with ρ_h.

Suppose the entrepreneur has posterior belief $x \in [0, 1]$ that the return is high after observing the signal. Then she chooses an investment level to maximize her ex-post expected return minus the loan repayment,

$$\max_{I \geq 0} 2x\sqrt{I} - (1 + r)I. \tag{17}$$

This yields optimal investment strategy $I^*(x; r) = x^2/(1 + r)^2$. Therefore, when the entrepreneur uses updating rule h to form her posterior belief, she chooses investment level $h(z)^2/(1+r)^2$ following signal realization z. The entrepreneur chooses to open a line of credit if her ex-ante expected return minus the loan repayment exceeds the origination fee, or, substituting $I^*(x; r)$ into Eq. (17), $E_{\hat{\rho}}[x^2]/(1 + r) \geq c$, where $E_{\hat{\rho}}$ denotes the expectation with respect to forecast $\hat{\rho}$. Therefore, the entrepreneur's updating rule influences her chosen investment level following the signal, whereas her forecast influences her credit decision before observing the signal.

The Optimal Contract. The lender is risk-neutral and it costs the lender I to lend I units of capital. The lender has a correctly specified model of the signal process and a correct model of the entrepreneur's model. This induces forecast ρ_h over the entrepreneur's posterior belief. The lender offers a contract that specifies an origination fee $c \in \mathbb{R}$ and a borrowing rate $r \in \mathbb{R}$ to maximize his expected revenue subject to the constraint that the entrepreneur chooses to open a line of credit,

$$\max_{c,r\in\mathbb{R}} c + rE_{\rho_h}[I^*(x;r)] \qquad \text{s.t. } E_{\hat{\rho}}[x^2]/(1+r) \geq c.$$

From Theorem 1, we know that a misspecified model is fully pinned down by its induced updating rule and forecast. We next show that the optimal contract can be described as a function of the expectation and variance of these two objects. Let $V_{\hat{\rho}} \equiv \int_0^1 x^2\, d\hat{\rho} - 1/4$ denote the variance of entrepreneur's forecast of his posterior belief, $V_h \equiv \int_0^1 h(z)^2 d\mu - (\int_0^1 h(z)d\mu)^2$ denote the true variance of the entrepreneur's posterior belief, and $m_h \equiv \int_0^1 h(z)d\mu$ denote the true expectation of the entrepreneur's posterior belief. The expectation of the entrepreneur's forecast is $m_{\hat{\rho}} = 1/2$ since the forecast is plausible.

Proposition 5 (The Optimal Contract)**.** *The optimal interest rate is*

$$r^*(h,\hat{\rho}) = \frac{V_h - V_{\hat{\rho}} + m_h^2 - 1/4}{V_h + V_{\hat{\rho}} + m_h^2 + 1/4} \tag{18}$$

and the optimal origination fee is $c^*(h,\hat{\rho}) = (V_{\hat{\rho}} + 1/4)/(1 + r^*(h,\hat{\rho}))$.

Fixing an updating rule, the more informative the entrepreneur expects her signals to be as measured by the variance of her forecast, the lower the optimal interest rate and the higher the optimal origination fee. As the variance of the forecast increases, the entrepreneur has a higher value for the lending product since she expects to have more precise information before making an investment decision, and therefore, the lender can charge a higher origination fee. Further, the lender finds it optimal to charge a lower interest rate since the entrepreneur's benefit from the low interest rate is proportional to her chosen investment, and this chosen investment is convex in the posterior belief. Therefore, when the entrepreneur expects more extreme posterior beliefs, she overestimates the value of a low interest rate and is willing to pay a higher fee to enter such contracts. In contrast, the higher the expectation or the variance of the entrepreneur's actual posterior belief, the higher the interest rate and the lower the origination fee. This is because the entrepreneur's investment strategy is increasing and convex in her posterior belief—and therefore, higher average beliefs or, fixing the average, higher variance leads to higher expected investment and hence, the lender earns higher revenue from interest.

We next show that when the forecast is introspection proof, then the lender charges the entrepreneur the same interest rate as that which she would charge a correctly specified agent. This is because the entrepreneur correctly anticipates the mean and variance of her posterior belief, $V_{\hat{\rho}} = V_h$ and $m_{\hat{\rho}} = m_h$, as is the case for a correctly specified agent, and when this property holds then the optimal interest rate is zero.

Corollary 2 (Introspection-Proof Optimal Contract)**.** *When the forecast $\hat{\rho}$ and updating rule h can be represented by an introspection-proof misspecified model, then the optimal interest rate is the same as that charged to a correctly specified agent, $r^*(h, \hat{\rho}) = 0$.*

From these results, we see that the retrospective and prospective biases induced by a misspecified model have a fundamentally different impact on decision-making and contract design. Therefore, our decomposition provides a crucial tool for understanding how the forms of bias induced by a misspecified model impact economic behavior.

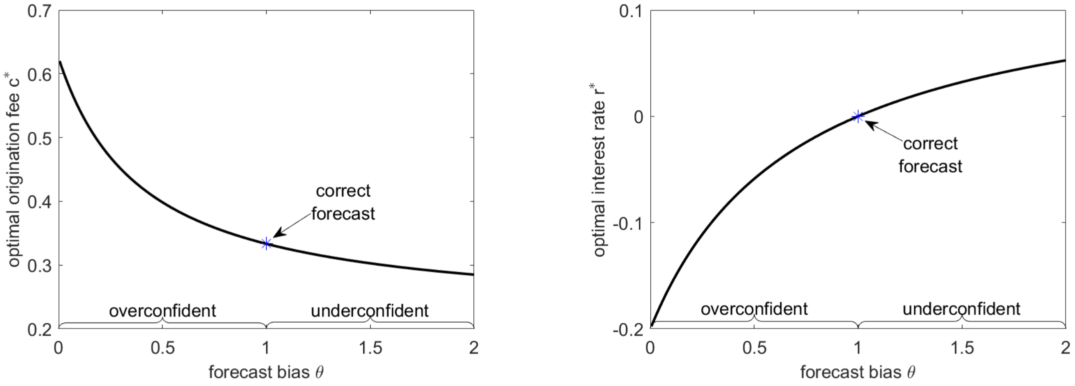
Overconfident and Underconfident Forecasts. We next fix the updating rule as Bayes rule, $h = h_B$, so that there is no bias in updating, and explore how the optimal contract differs with the bias in the forecast. An entrepreneur with an overconfident forecast is offered a contract with an origination fee that is significantly higher than what an entrepreneur with an unbiased forecast would be willing to accept and a negative interest rate. In contrast, an entrepreneur with a sufficiently underconfident forecast is offered a contract with an origination fee that is approximately zero and a positive interest rate. Therefore, an updating rule on its own does not significantly restrict the range of optimal contract terms—depending on the forecast, the optimal contract can feature very different origination and borrowing costs.

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:

$$d\hat{\rho}_\theta(x) = \frac{x^{\theta-1}(1-x)^{\theta-1}}{\Gamma(\theta)^2/\Gamma(2\theta)} \tag{19}$$

for $\theta > 0$ and $x \in [0,1]$. This corresponds to the family of beta distributions with mean 1/2. Suppose that the accurate forecast with respect to Bayes rule is uniform, i.e. $d\rho_B = 1$. Then $\theta = 1$ corresponds to the accurate forecast (and also the naive consistent forecast, since these are equal when $h = h_B$). For $\theta > 1$, as θ increases the entrepreneur is increasingly underconfident about the precision of her information in that she places more mass on intermediate posteriors and less mass on extreme posteriors relative to the accurate forecast. For $\theta < 1$, as θ decreases the entrepreneur is increasingly overconfident about the precision of her information in that she places more mass on low and high posteriors and less mass on intermediate posteriors relative to the accurate forecast.

From Proposition 5, the optimal interest rate is $r^*(h_B, \hat{\rho}_\theta) = \frac{\theta-1}{7\theta+5}$ and the optimal



(a) Origination fee c^*

(b) Interest rate r^*

Figure 2. Optimal contract.

origination fee is $c^*(h_B, \hat{\rho}_\theta) = \frac{(\theta+1)(7\theta+5)}{8(2\theta+1)^2}$ (Fig. 2 illustrates this optimal contract).[25] When the lender has an accurate forecast, $\theta = 1$, the optimal interest rate is zero, $r^*(h_B, \rho_B) = 0$, and the optimal origination fee is $c^*(h_B, \rho_B) = 1/3$. We compare this benchmark to the optimal contracts for under- and overconfident forecasts.

When the entrepreneur is overconfident, i.e. $\theta < 1$, the lender offers a negative interest rate, $r^*(h_B, \hat{\rho}_\theta) < 0$ and a higher origination fee than the accurate benchmark, $c^*(h_B, \hat{\rho}_\theta) > 1/3$. The overconfident borrower believes she will have very precise information to utilize when choosing how much to borrow in the future. In contrast, the lender knows that the entrepreneur overestimates the frequency of the signal realizations for which the she will borrow a large amount (i.e. the realizations for which the negative interest rate is very costly to the lender). Therefore, the entrepreneur overestimates the benefit of a negative interest rate. The lender leverages this forecasting bias by charging a high upfront price for a very favorable interest rate.

In contrast, when the entrepreneur is underconfident, i.e. $\theta > 1$, the lender offers the entrepreneur an up-front discount via the origination fee, $c^*(h_B, \hat{\rho}_\theta) < 1/3$, and a positive interest rate, $r^*(h_B, \hat{\rho}_\theta) > 0$. The lender knows that the entrepreneur underestimates the frequency of the signal realizations that induce the entrepreneur to borrow a large amount, and therefore, the entrepreneur underestimates the future cost of the positive interest rate. Therefore, the lender offers an upfront discount in order to induce the entrepreneur to enter the contract, then subsequently profits from the positive interest rate. For high enough θ, the optimal origination fee approaches zero.

[25] This follows from $V_{h_B} = 1/12$ when $d\rho_B = 1$ and $V_{\hat{\rho}} = 1/(8\theta + 4)$.

6.3 Lending Contracts with Heterogeneous Entrepreneurs

In this next application, we modify the lending framework from Section 6.2 to consider a setting where entrepreneurs have heterogeneous updating biases: some exhibit optimism bias, in which they overestimate future returns, and others are unbiased. We show that a key property of naive consistent forecasting is that when all entrepreneurs use this forecast, then the lender cannot engage in second degree price discrimination and screen by updating rule. We also illustrate how the updating rule approach makes it straightforward to derive comparative statics on how the optimal contract varies with the extent of the bias: more biased types or more extreme bias leads to cheaper access to credit but higher borrowing costs.

Suppose that there are two types of entrepreneurs. A share $\alpha \in (0, 1)$ are unbiased and update using Bayes rule, i.e. h_B. The remaining $1 - \alpha$ exhibit optimism bias in that they overestimate the probability of high quality after observing the signal, $h(z) = h_B(z)^\gamma$ for some $\gamma \in (0, 1)$, where lower γ corresponds to more severe bias. The lender cannot observe whether a given entrepreneur is biased, but knows the frequency of biased entrepreneurs α in the population and the extent of their bias γ. All entrepreneurs use the naive consistent forecast, $\hat{\rho} = \rho_B$. As discussed in Section 5, this is a natural form of unbiased forecast.

Naive consistent forecasting means that, before observing the signal of returns, both the unbiased and biased entrepreneurs have the same prediction about future beliefs, and therefore, the same expectation about their future borrowing behavior. This leads them to select the same option from any menu of contracts, which means that the lender cannot screen between the two types e.g. engage in second degree price discriminate. Therefore, the lender offers a single contract to both types. In contrast, other forecasts lead to different predictions about future beliefs for each type. Therefore, it would be possible to induce different types to choose different contracts.

The share of biased entrepreneurs and the extent of their bias impacts the terms of the optimal contract. In particular, the origination fee is decreasing in the degree of bias and share of biased entrepreneurs, while the interest rate is increasing. As either parameter increases, the lender leverages the optimistic entrepreneur's unanticipated willingness to borrow large amounts at high interest rates by offering a contract that is cheaper to enter but has a higher interest rate. This reduces investment by the unbiased entrepreneurs relative to their level in the optimal contract for their type. We formally present these comparative statics in Appendix C.

7 Dynamics

In order to consider dynamics, we extend the definition of an updating rule to specify a posterior belief for each possible signal realization and prior belief $p \in \Delta(\Omega)$, $h(z, p)$, and similarly for a forecast $\hat{\rho}(x,p)$. In this case, the analysis from Sections 4 and 5 pins down the misspecified model(s) that represent the updating rule and/or forecast at each prior.

7.1 Prior-Independent Representations.

An important question in this expanded framework is whether there exists a representation that is independent of the prior. We explore this question for updating rules below and present an analogous analysis for forecasts in Appendix D.3.

Definition 10 (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 subjective model representing the updating rule does not vary with the prior belief about the likelihood of each state. This makes it a conceptually appealing property for biases in which an agent is inherently Bayesian but has a mistaken understanding of the information generating process that is independent of her current worldview. For example, biases such as overreaction and optimism are not inherently linked to the agent's prior belief. In contrast, the property is conceptually at odds with biases that directly depend on the agent's current worldview in the sense that this worldview influences her perception of information. For example, confirmation bias is inherently linked to the agent's prior belief about the state, and therefore, is naturally represented by a model that varies with the prior. As we will show below, the property is also at odds with some biases in which an agent is non-Bayesian, as representing such biases in a Bayesian framework can require prior-dependence (e.g. Epstein et al. (2008)).

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_i/p_j 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 therefore, can form a prior-independent representation.

Proposition 6 (Prior-Independent Representation). *Fix an updating rule $h(z,p))$ such that $p \in S(h(\cdot,p))$ for 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 at prior p also represents h at all other priors $p' \in \Delta(\Omega)$.*[26] This property 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 canonical parameterizations of common biases have prior-independent representations. For example, the parameterization of overreaction in Example 4 and the parameterization of partisan bias in Bohren and Hauser (2021) both have prior-independent representations (see Appendix F.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.

Many biases are also naturally parameterized in a way that only admits prior-dependent representations. For example, the direction of confirmation bias and the magnitude of base rate neglect depend on the prior. Therefore, updating rules that only admit prior-dependent representations are essential for capturing the essence of these biases (see Page 8 for examples of such updating rules). While less obvious, the linear parameterization of over/underreaction in Epstein et al. (2008) (see Example 2) and the posterior parameterization of partisan bias in Example 3 only admit prior-dependent representations (see Appendix F.1). In the former, even though the over/underreaction parameter is independent of the prior, the additivity of the non-Bayesian updating rule with respect to the prior and the Bayesian posterior differs structurally from the multiplicative form of Bayes rule with respect to the prior and signal likelihoods, and therefore, can only be represented by Bayesian updating in a misspecified model when the model varies with the prior belief. 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 capturing the bias is independent of the prior. Similarly, the misspecified causal models from Spiegler (2020) only admit prior-dependent representations.[27]

Even when a prior-independent representation exists for a given updating rule, the unique model that represents a forecast-updating rule pair may not be prior-independent due to the dependence of the forecast on the prior. This brings us to the following result,

[26] Whenever an updating rule h can be represented by at least two models at some prior p, then trivially a prior-dependent representation exists even when a prior-independent representation also 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 $P \subset \Delta(\Omega)$ and select the other model to represent h at the remaining priors $\Delta(\Omega) \setminus P$.

[27] 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 belief (Bohren and Hauser 2021; Frick et al. 2020).

which establishes a desirable property for the naive consistent forecast.

Proposition 7. *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 in which a sequence of signals is observed, it turns out that 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 posterior belief in period $t+1$, in a dynamic setting one can also define the period-t forecast of the posterior belief in any future period $\tau > t$. It turns out that the representation of the naive consistent forecast and an updating rule that has a prior-independent representation induces period-t forecasts over posterior beliefs in any future period $\tau > t$ that are equal to the period-t forecast of beliefs in period τ induced by the correctly specified model.

7.2 Time Inconsistency.

Time inconsistency is a key property of many dynamic behavioral models. We next discuss how a prior-dependent representation is a natural way to allow for dynamic inconsistency.

Consider a dynamic setting in which a state ω is drawn at the beginning of the game. An agent observes a sequence of signals drawn independently from μ_i when the realized state is ω_i. Suppose the agent's updating rule and forecast is represented by prior-independent model $(\hat{\mu}_i)_{\omega_i \in \Omega}$, and the agent accurately anticipates that she will use this updating rule and forecast in all periods. In contrast to many dynamic behavioral models, this leads to behavior that is dynamically consistent: the optimal action an agent chooses following any signal realization is the same regardless of whether she commits to an action strategy before the signal is realized or selects an action after the signal is realized.

While dynamic consistency is desirable in certain settings, dynamic 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, representing such biases requires a misspecified model that can exhibit dynamic inconsistency. That is, the misspecified model an agent believes they will use in future periods must differ from the (potentially) misspecified model they actually use to form beliefs in future periods. A prior-dependent representation is a natural way to allow for this. For example, suppose the forecast and updating rule at prior p are represented by model $(\hat{\mu}_i(\cdot;p))_{\omega_i \in \Omega}$ and the agent believes she will use the forecast and updating rule induced by this model in all future periods. In contrast, the agent's actual updating rule and forecast has a prior-dependent representation denoted by family of models $((\hat{\mu}_i(\cdot;p))_{\omega_i\in\Omega})_{p\in\Delta(\Omega)}$. This can lead to dynamically inconsistent behavior: since the agent's model of how to interpret information changes with her belief but she does not anticipate this, the agent may wish to deviate from her ex-ante action strategy after observing the signal and updating her belief.

Prior-dependent representations do not always lead to dynamic inconsistency. When the agent accurately anticipates how her model varies with the prior, a prior-dependent representation can capture an agent who is time consistent. For example, if the true model varies with the prior as in active and social learning environments, then the unique representation of an agent who is Bayesian and has an accurate forecast will be prior-dependent. Alternatively, a biased agent who is sophisticated about her bias will accurately predict how her future updating rule and forecast will vary with her future belief, and therefore, exhibit time consistency.

8 Conclusion

We develop a representation that links updating rules to misspecified models. We show that any misspecified model can be represented through an update rule and a plausible forecast and vice-versa under mild conditions. This provides a natural tool for expressing a misspecified model and it's implications on decision making entirely in terms of the two important biases it induces; the prospective and retrospective bias. In addition, this provides a natural way to complete an update rule through the construction of a forecast. We identify paths to complete an update rule – the introspection-proof model and the naive consistent forecast – and provide necessary and sufficient conditions for these to exist. These results allow us to embed well-documented information processing biases into economic decision problems where the update rule on its own would have been insufficient. This decomposition also highlights the importance of eliciting more than the agent's updating rule in experimental work in order to get a complete picture of how the economically relevant ways an agent reasons about information.

A Proofs from Section 4

Proof of Lemma 1. (If:) Let $F \equiv \{x : x_i = \int_{\mathcal{Z}} h(z)_i \, d\hat{\mu},\ \hat{\mu} \in \Delta^*(\mathcal{Z})\}$. 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. $\sum_{j=1}^{K} \lambda_j a_j = x$, $\lambda_j > 0$, $\sum \lambda_j = 1$. Fix $\varepsilon \in (0, \min\{\lambda_j\})$, 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 $|\int_{\mathcal{Z}} h(z)_i d\hat{\mu} - x_i| \leq \varepsilon$, 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 $m_i = \int_{\mathcal{Z}} h(z)_i \, d\mu$, the expected value of the misspecified posterior under the true unconditional distribution, which exists, lies in F, and has non-zero Radon-Nikodym derivative ν-a.e.. Since 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. $q = \int_{\mathcal{Z}} h(z)_i \, d\gamma$. Consider the compound lottery where with probability $\frac{1}{1+\varepsilon}$ the signal z is drawn from γ and with complementary probability is it is drawn from μ. Call this measure $\hat{\mu}$. Then $\int_{\mathcal{Z}} h(z)_i \, d\hat{\mu} = p_i$. 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 in the under μ so it must occur with positive probability under $\hat{\mu}$, which is a contradiction. Therefore, by part 3 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 $m_i = \int_{\mathcal{Z}} h(z)_i d\hat{\mu}$. 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, and 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 since any neighborhood of that point occurs with positive probability $v \cdot m = \int v \cdot s \, d\hat{\rho}_{\hat{\mu}} < v \cdot m$ by the full support assumption, which 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 $\hat{\mu} = \hat{\rho} \circ g^{-1}$ be the pushforward measure. By change of variables formula

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

so by Lemma 2, a misspecified model with unconditional signal distribution $\hat{\mu}$ exists and induces update rule $g(z)$. This misspecified model has forecast $\hat{\rho}$ by construction.

(Only If:) Fix a misspecified model $(\hat{\mu}_i)_{i=1}^N$. Let $h(z)$ be the update 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

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

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

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

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

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

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

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

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—in this case, applying this condition 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 pin down the entire set of conditional signal distributions or violate Bayes-plausibility, and therefore, cannot be part of a misspecified model that represents the updating rule.

Lemma 3 also 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 uniquely select a misspecified model.

Proof of Lemma 3. **Part 1:** First suppose $h(z)$ can be represented by a misspecified model with perceived unconditional signal distribution $\hat{\mu}$. It follows from standard argument that beliefs must be a martingale, so if $h(z)$ describes posteriors then

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

Now suppose that $\hat{\mu}$ is a measure where

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

Define conditional distributions

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

for all $Z \in \mathcal{F}$. These are probability distributions. It remains to show this induces the correct belief. Since $\hat{\mu}_i$ is absolutely continuous with respect to μ, Bayes rule and the properties of the Radon-Nikodym derivative imply that

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

so these distributions induce the correct posteriors. Finally any family of misspecified models must solve

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

$\hat{\mu}$-a.s. so the conditional distributions are unique.

Part 2. Suppose $h(z)$ can be represented by a misspecified model with conditional signal distribution $\hat{\mu}_j$. Then, by standard argument, for any i the likelihood ratios

$h(z)_i/h(z)_j$ must be martingales with respect to $\hat{\mu}_j$ so

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

Now suppose that $\hat{\mu}_j$ is a measure that satisfies

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

for update rule h and all i. Define the misspecified model

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

This is a misspecified model that induces update rule $h(z)$. Finally any family of misspecified models must solve

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

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

Proof of Theorem 1. (If:) 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

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

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 measurable Z, define

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

For any Borel set X, note that

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

Therefore, for any $Z \in \mathcal{P}$, this agrees with $\hat{\rho}$ which implies $\hat{\mu}_i(\mathcal{Z}) = \hat{\rho}_i(\Delta(\Omega)) = 1$ by the assumption that $\hat{\rho}$ is a forecast. Therefore $\hat{\mu}_i$ is a probability measure that induces the specified forecast.

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 $(\mathcal{Z}, \mathcal{F})$. Moreover, $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ 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. Uniqueness for sets in $\mathcal{P}$ follows from Lemma 3 applied to the transformed signal spaces where signals are posteriors.

(Only If:) The forecast must be plausible by Lemma 2. Suppose that there exists a Borel 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.

This result follows from Lemma 3: an unconditional measure is enough to uniquely identify the misspecified model that represents a given updating rule. Similarly, it is immediate that a forecast and a misspecified model consistent with that forecast uniquely identify an updating rule and that an updating rule and a misspecified model that induces that updating rule uniquely identify a forecast. □

B Proofs from Section 5

Proof of Proposition 1. Suppose $h(z)$ is an updating rule such that

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

then by Lemma 1 there exists a misspecified model $(\hat{\mu}_\omega)_{\omega \in \Omega}$ that induces unconditional distribution μ over $\mathcal{Z}$ and is represented by updating rule $h(z)$. By the proof of Lemma 1,

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

describes a family of misspecified models that induce the desired distribution and updating rule. Moreover, as argued before any family of misspecified models must solve

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

so there is at most one Radon-Nikodym derivative that solves this equation, and thus the misspecified models are unique.

Now suppose that $(\hat{\mu}_i)_{\omega_i \in \Omega}$ describes a family of introspection proof misspecified models that are represented by updating rule $h(z)$ and have unconditional distribution μ. By the above logic, the Radon-Nikodym derivative $\frac{d\hat{\mu}^i}{d\mu} = \frac{1}{p_i} h(z)_i$. This implies that

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

so the desired condition holds. □

Proof of Proposition 2. (If:) The existence of a naive-consistent forecast follows immediately from Theorem 1. For any X such that $Z = h^{-1}(X)$, note that $\hat{\mu}_i(Z) = \hat{\rho}_i(X) = \mu_i(h_B^{-1}(X)) = \mu_i(h_B^{-1}(h(Z)))$ by construction of $\hat{\rho}_i$ 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 $\hat{\rho}(X) > 0$. □

C Proofs from Section 6

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, $(h(g, F) + h(b, F))/2 = \frac{1-q\gamma}{2(1-q)}$. 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, $\gamma \equiv E(h(z_m;p)|M) = (h(g,M)+h(b,M))/2$. 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. The optimal origination fee satisfies the participation constraint with equality, $c = E_{\hat{\rho}}[x^2]/(1+r)$. Plugging this into the lender's profit simplifies the problem to

$$\max_{r \geq 0} E_{\hat{\rho}}[x^2]/(1+r) + rE[h(z)^2]/(1+r)^2. \tag{23}$$

Taking the first order condition and setting it equal to zero yields Eq. (18). □

Proof of Corollary 2. A correctly specified entrepreneur uses update rule h_B and forecast ρ_B. Note that $V_{h_B} = V_{\rho_B}$ and $m_{h_B} = 1/2$. From Eq. (18), this implies that the optimal interest rate is zero, $r^*(h_B, \rho_B) = 0$. When the entrepreneur uses updating rule h and forecast ρ_h, she correctly anticipates her posterior beliefs. Therefore, $V_h = V_{\rho_h}$ and $m_h = m_{\rho_h} = 1/2$. This is the only forecast that can be jointly represented by a misspecified model with h. □

Proof of Proposition 6. (If:) Fix an interior prior $p \in \Delta(\Omega)$. By Lemma 1, there exists a misspecified model $(\hat{\mu}_i)_{i=1}^n$ that represents $h(z,p)$ at p. Therefore, by Bayes rule, for ν-almost all z

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

So the condition from observation 1 implies that

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

which is exactly the condition $h(z,p')$ must satisfy to be induced by $(\hat{\mu}_i)_{i=1}^n$ at p'.

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

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

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 7. Fix a prior p and let $(\hat{\mu}_i)_{i=1}^N$ be the essentially unique representation of $h(z,p)$ and the naive consistent forecast given by **??** at prior p.

It follows from Proposition 6 that this induces $h(z,p)$ at every prior, as for any p' the likelihood ratio of the heuristic 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 at p', as for any Borel X

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

so the forecast induced by $(\hat{\mu}_i)_{i=1}^N$ is the ρ^{NCF}. □

Analysis from Section 6.3. Since the lender cannot screen types, he offers a contract that specifies an origination fee $c \in \mathbb{R}$ and a borrowing rate $r \in \mathbb{R}$ to maximize his expected revenue subject to the constraint that both types choose to open a line of credit,

$$\max_{c,r\in\mathbb{R}} c + r(\alpha E_{\rho_B}[I^*(x;r)] + (1-\alpha)E_{\rho_h}[I^*(x;r)]) \tag{24}$$
$$\text{s.t. } E_{\rho_B}[x^2]/(1+r) \geq c.$$

The expectation in the constraint is taken with respect to the accurate forecast ρ_B since both types of entrepreneurs use this forecast.

Although both types have the same ex-ante prediction of future beliefs when deciding whether to open the line of credit, when deciding how much to borrow, the biased entrepreneur overestimates her future returns and borrows more than the unbiased entrepreneur. Further, for a given signal, the amount borrowed is increasing in the entrepreneur's bias. This leads to the following result.

Proposition 8.

1. *Fixing the level of bias γ, the optimal origination fee is decreasing and the optimal interest rate is increasing in the share of biased types $1-\alpha$.*

2. *Fixing the frequency of biased entrepreneurs α, the optimal origination fee is decreasing and the optimal interest rate is increasing in the level of bias $1-\gamma$.*

As the bias increases (γ decreases), the optimal interest rate increases and the lender profits from the increased optimism of the entrepreneur. The origination fee has an inverse relationship with the level of bias: a higher bias leads to a lower fee. This is the result of naive consistency: since all entrepreneurs form the same expectations over their future borrowing behavior ex-ante, the only way for the lender to offer a contract with a higher interest rate is to lower the upfront fee.

Proof. The lender chooses c to satisfy the constraint $E[h_B(z)^2]/(1+r) \geq c$ with equality. Therefore we can rewrite Eq. (23) as

$$\max_r E[h_B(z)^2]/(1+r) + r(\alpha E[h_B(z)^2]/(1+r)^2 + (1-\alpha)E[h_B(z)^{2\gamma}]/(1+r)^2)$$

This has first order condition

$$\frac{-E(h_B(z)^2)}{(1+r)^2} - \frac{2r(\alpha E(h_B(z)^2)}{(1+r)^3} + \frac{(1-\alpha)E(h_B(z)^{2\gamma})}{(1+r)^3} + \frac{\alpha E(h_B(z)^2)}{(1+r)^2} + \frac{(1-\alpha)E(h_B(z)^{2\gamma})}{(1+r)^2} = 0.$$

Solving this expression for r leads to

$$r^*(\alpha,\gamma) = \frac{(1-\alpha)(E(h_B(z)^{2\gamma}) - E(h_B(z)^2))}{(1+\alpha)E(h_B(z)^2) + (1-\alpha)E(h_B(z)^{2\gamma})}$$

The optimal origination fee is $c^*(\alpha,\gamma) = E[h_B(z)^2]/(1+r^*)$. If all entrepreneurs were unbiased ($\alpha = 1$), then the lender would want to offer an interest rate of $r^*(1,\gamma) = 0$, as the entrepreneur invests optimally. This results in an origination fee of $c^*(1,\gamma) = E[h_B(z)^2]$. Otherwise, the lender offers a lower origination fee and a higher interest rate. □

D Extensions

D.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, (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 11. *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. Using this, we can define another class of misspecified models:

Definition 12. *Given an updating rule $h(z)$, 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 it solves*

$$\begin{aligned} &\inf_{\hat{\mu}} \sup_{T^n} \liminf -\frac{\ln Pr(T_n = 0)}{n} \\ &s.t.\ \hat{P}r(T_n = 1) \geq 1 - \alpha \text{ for all } n \\ &\hat{\mu} \text{ represents } h(z). \end{aligned}$$

That is, given any hypothesis test that rejects the misspecified model with probability less than α, the α-introspection-proof model minimizes the worst-case probability of rejection under the true distribution as n grows large.

This has a natural connection to a well-studied problem in statistics. We are interested in the probabilities of type 1 and 2 error for a hypothesis test where:

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

By the Chernoff-Stein lemma, for any $\hat{\mu}$, the solution to the inner optimization problem is $D(\hat{\mu}||\mu)$, where D is KL-divergence, so we can reframe this problem as

$$\begin{aligned} &\min D(\hat{\mu}||\mu) \\ &\text{s.t. } \int h_i(z)\frac{d\hat{\mu}}{d\nu}d\nu = p_i \text{ for all i.} \end{aligned}$$

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

Theorem 2. *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 $h(\cdot)$, the α-introspection-proof misspecified model is given by:*

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

where $\lambda \in \mathbb{R}^n_+$ solves $\int (p_i - h_i(z))e^{\lambda \cdot h(z)}d\mu = 0$ for each i.[28] *This model has KL-divergence $p \cdot \lambda - \log \psi_h(\lambda)$ from the truth.*

The updating rule $h(z)$ 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 $x \in \Delta(\Omega)$

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

where λ is the λ from above and $\rho_h = \mu \circ h^{-1}$.

D.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 13. *A family of misspecified models $(\hat{\mu}_i)$ representing updating rule $h(z)$ is* ω_i-Introspection-proof Model Relative to ω_j *if for all measurable A*

$$\hat{\mu}_j(A) = \mu_i(A)$$

This restriction requires there to exist some state ω_j where the observed frequencies

[28] Since $h(z)$ 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 convex hull condition is satisfied.

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.

Theorem 3. *Fix an updating rule $h(z)$. This can be represented by an ω_i-introspection-proof misspecified model relative to ω_j, $(\hat{\mu}^k)_{j=1}^N$ if and only if for all $k \in \{1, 2, \ldots N\}$*

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

Moreover, if this representation exists, for any k and any measurable A

$$\hat{\mu}^k(A) = \int_A \frac{p_j}{p_k} \frac{h(z)_k}{h(z)_j} \, d\mu_i.$$

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 it seems very similar to the original introspection-proof condition, this condition is in fact, much less restrictive.

D.3 Prior-Independent Representations of Forecasts

Observation 1. *A forecast has a prior-independent representation if $\hat{\rho}(x', p') = \hat{\rho}(x, p)$ for x' such that $x'_i/x_1 = \frac{p'_1}{p'_i}\frac{p_i}{p_1}\frac{x_i}{x_1}$.*

E Comparison to Blackwell's Order

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 $\mathcal{Z} = \{z_1, z_2, \ldots z_K\}$ 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 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.

F Additional Examples

F.1 Examples of Updating Rules with Prior-Independent and Prior-Dependent Representations

In this section we show that the parameterization of overreaction in Example 4 and the parameterization of partisan bias in Bohren and Hauser (2021) satisfy the condition in Proposition 6, and therefore, have a prior-independent representation. We also show that the parameterization of over/underreaction in Epstein et al. (2008) (see Example 2) and the parameterization of partisan bias in Example 3 do not satisfy the condition in Proposition 6, and therefore, do not have a prior-independent representation.

In Example 4,

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

It is straightforward to see that it is possible to factor the prior out of this updating rule.

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, $z = \frac{d\mu_1}{d\nu}/(\frac{d\mu_2}{d\nu} + \frac{d\mu_1}{d\nu})$, with support $\mathcal{Z} \subset [0,1]$. Consider updating rule $h(z,p)_1/h(z,p)_2 = p_1 z^{\alpha}/(1-p_1)(1-z^{\alpha})$, where $\alpha \in (0,\infty)$ is the partisan bias parameter. Again it is straightforward to see that it is possible to factor the prior out of this updating rule.

In contrast, the model of over/underreaction in Example 2 does not satisfy the condition in Proposition 6, as

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

clearly depends on the prior. Similarly, in the model of partisan bias in Example 3,

$$\frac{p_2}{p_1}\frac{h(z,p)_1}{h(z,p)_2} = \frac{p_2}{p_1}\left(\frac{h_B(p,z)_1^2}{1 - h_B(p,z)_1^2}\right)$$

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

F.2 Linear Under- and Overreaction

Fix a correctly specified model $(\mu_i)_{\omega_i \in \Omega}$, and consider the updating rule for under- and overreaction defined by Epstein et al. (2008):

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

for some $\alpha \in (-\infty, 1]$. We can use Lemma 3 to find misspecified models that represent this updating rule. For instance, consider a misspecified model with an unconditional measure that is equal to the true unconditional measure, $\hat{\mu} = \mu$. Then $\hat{\mu}$ satisfies Eq. (21) as $\int_{\mathcal{Z}} h_B(z)\, d\hat{\mu} = \int_{\mathcal{Z}} h_B(z)\, d\mu = p$ by standard argument, and therefore, $\int_{\mathcal{Z}}(\alpha h_B(z) + (1-\alpha)p)\, d\hat{\mu} = p$. Given this unconditional distribution, the state-contingent distribution in state ω_i is given by:

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

In other words, it is completely pinned down by the true unconditional measure $\hat{\mu}$, the Bayesian updating updating rule h_B, and the under- or overreaction parameter α.

This representation is not unique. Suppose for instance that $|\Omega| = 2$, $\mathcal{Z} = [0,1]$, $p = 1/2$, μ is the uniform distribution over $\mathcal{Z}$ and $|h_B(z)_1 - \frac{1}{2}|$ is symmetric about $z = 1/2$. Then the distribution with pdf $f(z) = 3/2 - 6(z-1/2)^2$ also satisfies $\int_{\mathcal{Z}} h_B(z)f(z)dz = 1/2$, and therefore, $\int (\alpha h_B(z) + (1-\alpha)/2)\, f(z)dz = 1/2$. While in the first case, the agent correctly anticipates the frequencies of different signals but underreacts to them, in this case, the agent underestimates the frequency of "extreme" signal realizations which, given that $h_B(z)$ is monotone, means that in addition to underreacting to the signal, the agent also anticipates that she'll observe signal realizations which, on average, are 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. (2021): "Robust Model Misspecification and Paradigm Shift," Mimeo.

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, A. (2016): "Informational Herding with Model Misspecification," *Journal of Economic Theory*, 222–247.

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

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

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

de Clippel, G. and X. Zhang (2019): "Non-bayesian persuasion," .

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

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 (2019): "Asymptotic Behavior of Bayesian Learners with Misspecified Models," .

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

Frick, M., R. Iijima, and Y. Ishii (2020): "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 (2020): "Limits Points of Endogenous Misspecified Learning," .

——— (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," .

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

He, K. (2020): "Mislearning from Censored Data: The Gambler's Fallacy in Optimal-Stopping Problems," .

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.

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

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

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 (2020): "Cheap Talk with Prior-biased Inferences," .

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.

Mailath, G. J. and L. Samuelson (2019): "The Wisdom of a Confused Crowd : Model-Based Inference," .

Molavi, P. (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.

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.

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.

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