Base-Rate Neglect: Foundations and Implications[*]

Dan Benjamin and Aaron Bodoh-Creed[†]and Matthew Rabin

[Please See Corresponding Author's Website for Latest Version]

July 19, 2019

Abstract

We explore the implications of a model in which people underweight the use of their prior beliefs when updating based on new information. Our model is an extension, clarification, and "completion" of previous formalizations of base-rate neglect. Although base-rate neglect can cause beliefs to overreact to new information, beliefs are too moderate on average, and a person may even weaken her beliefs in a hypothesis following supportive evidence. Under a natural interpretation of how base-rate neglect plays out in dynamic settings, a person's beliefs will reflect the most recent signals and not converge to certainty even given abundant information. We also demonstrate that BRN can generate effects similar to other well-known biases such as the hot-hand fallacy, prediction momentum, and adaptive expectations. We examine implications of the model in settings of individual forecasting and information gathering, as well studying how rational firms would engage in persuasion and reputation-building for BRN investors and consumers. We also explore many of the features, such as intrinsic framing effects, that make BRN difficult to formulaically incorporate into economic analysis, but which also have important implications.

[*] For helpful comments, we thank seminar participants at Berkeley, Harvard, LSE, Queen's University, Stanford, and WZB Berlin, and conference participants at BEAM, SITE, and the 29th International Conference on Game Theory at Stony Brook. We are grateful to Kevin He for valuable research assistance.

[†] Corresponding Author. E-mail: acreed@berkeley.edu WWW: http://faculty.haas.berkeley.edu/acreed

1 Introduction

This paper explores a model capturing one common interpretation of the psychology literature on *base-rate neglect* (BRN): when revising beliefs in light of new information, people tend to underuse their previous information relative to the requirements of Bayes' law. We build on previous formulations, but clarify and flesh out the implications of this interpretation of BRN-as-downweighting-priors approach, especially exploring how the short- and long-run beliefs contrast with Bayesian beliefs. We identify intrinsic framing effects and other features of the model, and explore the implications of specifying assumptions related to BRN that have previously been unspecified. We show how the two key features of BRN—(1) a tendency towards moderate beliefs and (2) excessive movement of those beliefs in the short run, which generate persistent fluctuation of beliefs that never converge towards certainty in the long run—have implications in a number of economic settings (such as prediction-making, sequential sampling, persuasion, and reputation building), that are different from (and robust to) the implications of other biases in reasoning.

In Section 2 we present a simple formulation of BRN. Given priors $p(\theta)$ that the hypothesis θ (drawn from a set of hypotheses Θ) is true, a Bayesian agent—whom we call Tommy—forms his posterior after observing signal s according to Bayes's Rule. Here and throughout the paper, denoting correct beliefs over any domain by $p(\cdot)$, Bayes's Rule is $p(\theta|s) = \frac{p(s|\theta)p(\theta)}{\Sigma_{\theta'}p(s|\theta')p(\theta')}$. An agent who suffers from base-rate neglect—whom we call Saki—uses likelihood information exactly as Tommy does, but underweights her prior beliefs: $p_\alpha(\theta|s) = \frac{p(s|\theta)p(\theta)^\alpha}{\Sigma_{\theta'}p(s|\theta')p(\theta')^\alpha}$, where $\alpha \in [0,1)$. If $\alpha = 0$, Saki completely ignores her prior when she updates. If $\alpha \in (0,1)$, Saki under-uses, but does not completely neglect, her priors. This formulation has been employed previously to empirically measure BRN by estimating the value of α.[1]

Along with discussing some of the evidence for BRN, Section 2 explores some of the meaning and implications of the bias. Although Saki's beliefs tend to fluctuate more than a Bayesian's, she will in fact hold overly moderate beliefs on average. The evidence suggests that people give less weight to a base rate even after processing (what turns out to be) non-diagnostic information. In those cases, doing so will necessarily moderate her beliefs. The most striking form of Saki's propensity towards moderation is a generalization of this feature: if Saki's prior belief strongly favors one hypothesis and she observes sufficiently weak evidence *in favor* of that hypothesis, she will update to believe the hypothesis *less* strongly. Along with the seminal evidence that non-diagnostic information can moderate beliefs, the only studies on BRN we know of that include concrete probability assessments and treatments where base rates and new information both lean in the same direction, Griffin and Tversky (1992) Study 2 and Bar-Hillel (1980) Problem 4, provide support for this extreme moderation effect.

[1] Following Grether (1980) and Grether (1992), several papers have analyzed belief evolution in updating experiements by linearly regressing $\ln\left(\frac{p(\theta|s)}{p(\theta'|s)}\right)$ on $\ln\left(\frac{p(s|\theta)}{p(s|\theta')}\right)$ and $\ln\left(\frac{p(\theta)}{p(\theta')}\right)$, where the estimated coefficient on the second term is an estimate of α.

Section 2 also explores some challenging features of the model. As with other non-Bayesian models, the effects of BRN inherently depend on how Saki divides up all possibilities into hypotheses. For example, consider a scenario where Saki is a manager assessing two of her employees, Heidi and Tarso, updating her beliefs about two hypotheses: (1) Heidi is at least as effective as Tarso, versus (2) Heidi is less effective than Tarso. Compare this to a second scenario, where Saki thinks about three hypotheses: (1) Heidi is at least as effective, versus (dividing the second hypothesis in two) (2a) Heidi is less effective but competent, versus (2b) Heidi is less effective and incompetent. Upon updating after receiving the same information, Saki will be more doubtful of Heidi's relative effectiveness in the second scenario than the first. More generally, comparing across different ways Saki might divide the world into hypotheses, Saki will exhibit subadditivity of the form Tversky and Koehler (1994) highlighted in a different context: an event is viewed as more likely if it is framed in terms of disjoint sub-events. Framing can also generate violations of conjunction of the form highlighted in Kahneman and Tversky (1983): after a particularly bad performance by Heidi, Saki might put more weight on Heidi being *both* less effective than Tarso *and* incompetent in the second scenario than she would put on Heidi being less effective than Tarso in the first scenario. Such a violation of the conjunction principle in probability—that a sub-event can't be more likely than the broader event—occurs whenever a signal is very likely conditional on an a priori unlikely event, but Saki neglects the low prior probability of the unlikely event.

In Section 3, we examine BRN when an agent observes a sequence of informative signals over time. We assume that each time Saki receives a signal her updated beliefs become her priors when interpreting the next signal.[2] This means that new information is fully weighted when first received, but is subsequently downweighted when it becomes part of the prior beliefs as further signals are received. Moreover, each signal is increasingly neglected as more information arrives, and the signal eventually has a negligible influence on beliefs after many newer signals have been observed. To lay out such dynamics in simplest form, Section 3 then studies an environment where Saki receives an infinite stream of signals that are conditionally independent. In this case, the influence of a signal on Saki's current belief is exponentially declining in the number of intervening signals. This generates a recency effect whereby early signals eventually have no influence on her beliefs. In effect, Saki's beliefs after any number of signals are always determined by the informational equivalent of a finite number of signals. Whereas Tommy eventually learns the correct hypothesis almost surely in this setting, Saki's beliefs perpetually vacillate (due to the recency effect), and she will never be confident about which hypothesis is correct (due to the effective finiteness of her information). Moreover, although the long-run *frequency* of different beliefs de-

[2] While BRN has most often been framed and illustrated as neglect specifically of "base rates" in the sense of the statistical proportion of a population that meets a description—our model assumes that faced with new evidence *any* prior beliefs are downweighted, regardless of whether these prior beliefs happen to be base rates. As we discuss in Section 2, the evidence from experiments in which subjects observe a sequence of signals and report updated beliefs after each signal (in the tradition of Grether (1992)) provides support for the assumption that during any phase of updating priors are downweighted even when the source of those priors are previous signals.

pends on the truth, the *range* of beliefs she returns to does not: Saki will occasionally return to believing in each hypothesis strongly when she happens to get a string of signals supporting that belief.

At the end of Section 3 we specify an aspect of our model necessary to make it complete. BRN as it has been previously formulated specifies Saki's beliefs *retrospectively* after she sees evidence, but it does not specify what she thinks *prospectively* about what her beliefs will be when she gets additional information. Retrospective and prospective beliefs are intrinsically equivalent for a Bayesian. But as discussed and modeled in Benjamin, Rabin, and Raymond (2016) in the context of a different error, the the two need not be consistent given cognitive errors such as BRN.[3] Without much empirically to go on, but in line with the underlying psychology of prospective beliefs in similar types of non-Bayesian models, we assume that Saki thinks that her future updating will be Bayesian.

We explore some of the consequences of BRN in Sections 4-7, and illustrate how the properties discussed above have economic implications. Although Sections 5 and 6 explore how a rational economic actor can profitably take advantage of Saki's errors, in Section 4 we flesh out how BRN affects the beliefs of individuals who are trying to predict outcomes under uncertainty. Imagine an economic actor is trying to figure something out over time, such as an investor attempting to discover patterns in stock prices or an unemployed worker inferring her likely future prospects from interviews and offers she gets. Suppose that when she begins, Saki entertains some set of theories of the world, each of which we think of as a potentially true data generating process, meaning each theory is a plausible explanation of an infinite series of signals she observes. So, for instance, there could be an autocorrelative process that Saki is uncertain about—she might think there is positive, negative, or no autocorrelation in stock prices for instance. In addition, we focus on situations where the observed signals are sufficient for Tommy to eventually learn the dynamic structure.

Consider situations where Saki and Tommy begin with priors about all the possible theories of the world and where they put positive weight on the actual true theory of the world. Tommy will eventually learn the truth, so that his predictions settle down to the appropriate prediction given the recent past. If there is positive autocorrelation, Tommy's predictions will exhibit positive autocorrelation. If the world is i.i.d., Tommy's predictions will be independent of recent events. In contrast, each observation causes Saki to move her beliefs towards the theory that best explains the recent past, so her predictions are a composite of her ever-changing beliefs about the theories'

[3] Besides Benjamin, Rabin, and Raymond (2016), the only paper we are familiar with that explores the relationship between prospective and retrospective beliefs is He and Xiao (2017). They explore the implications of imposing consistency between the two, and show how some features of well-known biases can and (mostly) cannot be captured in such consistent models. Models in which agents are fully Bayesian but have a wrong model (e.g., Barberis, Shleifer, and Vishny (1998); Rabin (2002); Rabin and Vayanos (2010)) don't mention prospective beliefs separately for the same reason Bayesian models do not: it is implicit that the updating is consistent. Models where people misread some of their signals, as in Rabin and Schrag (1999), implicitly assume that people will prospectively anticipate the true way they will analyze future signals conditional on how they read (or misread) those signals.

likelihoods and what each theory predicts about the world. The recency effect implies that Saki's beliefs will often exhibit prediction momentum, meaning Saki predicts the future will resemble the recent past. For example, suppose that Saki believes that a sports team's performance is i.i.d. across games, but she is trying to figure out the team's permanent ability. Saki will believe that recent good performances will continue *not* because she believes teams vacillate between episodes of good performance and bad performance (per the "hot hand" interpretation of Camerer (1989)), but because her beliefs vacillate between thinking the team is *permanently* good or *permanently* bad. In a similar vein, in a canonical normal-normal updating setting, the recency effect generates patterns of beliefs similar to adaptive expectations when the economy"s structure is stable, but Saki is responsive to endogenous changes in the economy because she understands the impact of these changes on what she observes.

Section 4 also provides a new perspective on a large literature on "quasi-Bayesian models." In this growing genre of models, it is assumed that people dogmatically believe in a mis-specified model of the world, in the sense that they believe that the true model of the world is impossible. For example, Barberis, Shleifer, and Vishny (1998) models investors who dogmatically and incorrectly believe that a firm's performance is determined by random switches between two stochastic processes. Because a Bayesian would (quite generally) eventually learn the correct model if he believed that model were possible, these models rely on the agent's dogmatic dismissal of the model in order for the error to persist.[4] BRN provides a simple alternative explanation for a failure to learn the true model of the world: BRN causes Saki's beliefs to perpetually vacillate, based on the most recent observations, about what theory best explains the world. Saki is entertaining the true model, and in fact always believes it is possible, and often believes it is very likely—but (per the core dynamic features of BRN) she never converges to certainty, and she always fluctuates based on her most recent signals.

Although prospective beliefs do not play a role in most of our applications in this paper, they are crucial in any economic situation involving decision-makers who endogenously decide what information to gather or pay attention to. Section 5 studies a sequential-sampling model with costly signals. If Saki is prospectively Bayesian and believes that she will process these signals as per Bayes's rule, then she may be caught in a "learning trap" wherein she is unable to become sufficiently confident to make a choice. As a result, while lowering the cost of the signals is obviously good for Tommy, it can actually make Saki worse off by inducing this learning trap. We also explore some of the implications of different assumptions about Saki's prospective beliefs in

[4] The assumption that agents do not even entertain the true data generating process as a possibility is extreme, of course. Such models can be viewed as a good approximation to capture beliefs of agents with strong priors up until the very long horizon. But not considering all of the infinite number of ways the world can be might in fact be more realistic, and Gagnon-Bartsch, Rabin, and Schwartzstein (2018) argue that (subjectively rational) inattention might lead people not to notice they are wrong even in the very long run. Our assumption that agents might forever vascillate in their belief about the way the world works without realizing something is amiss can similarly be seen as an extreme way to capture beliefs short of the very long run. But here it may be even less clear that somebody would notice that she is making an error.

this context.

In Sections 6 and 7 we explore some economic implications of BRN in cases where a rational economic actor interacts with somebody known to suffer from BRN. In Section 6, we consider a "persuader" who can influence an audience's beliefs by choosing whether or not to reveal a signal. A revealed signal is verifiable, but the existence, absence, and nature of an unrevealed signal is not verifiable. If the audience is Bayesian, the unique sequential equilibrium is for information to be revealed if and only if it moves the audience's beliefs in a direction the persuader wants. In equilibrium, the audience deduces that the absence of a revealed signal implies either a bad signal or no information. We analyze the contrasting implications of BRN. We assume that, despite neglecting base rates, Saki is strategically sophisticated: she understands how the motives and available actions of would-be persuaders may influence what messages they wish to send—or whether they wish to send information at all. We show that—despite this sophistication about what silence may mean—a persuader who is happy with status-quo beliefs may prefer to not reveal even favorable information to Saki so as to prevent the status-quo beliefs from being neglected. On the flipside, if Saki's prior is unfavorable to the persuader, then the persuader might be willing to reveal even a bad signal because (again due to the moderation effect) the signal muddies Saki's beliefs.

In Section 7 we examine the implications of BRN for reputation-building. We consider the prototypical setting studied by Fudenberg and Levine (1992): with high probability, a long-run, patient firm is a "strategic" player that decides each period whether to "shirk" or "work" on unobservably high quality that period; and with low probability, the firm is a "committed type" that automatically works each period. Each period, a new consumer decides whether or not to buy without knowing the current quality, but after observing the history of product quality, which is informative about whether the firm is strategic or committed. Fudenberg and Levine (1992) showed that, if the consumers are Bayesians, then a patient, strategic firm gets almost the same utility as a committed firm. Cripps, Mailath, and Samuelson (2004) in turn showed that in equilibrium, the firm will work with high probability initially, but as time goes on, strategic firms are found out and begin to shirk consistently. We show that because Saki's beliefs permanently vacillate, then consumers that neglect base rates will never become confident of the firm's type, and the firm's reputation will never be completely and permanently destroyed.

In Section 8 we discuss the relationship between BRN and other types of biases. While there can be interesting interactions between BRN and biases such as the Law of Small Numbers (LSN) (Rabin (2002)), Nonbelief in the Law of Large Numbers (NBLLN) (Benjamin, Rabin, and Raymond (2016)), limited memory, and motivated cognition, the most urgent comparison is with models of confirmatory bias. As modeled in Rabin and Schrag (1999), confirmatory bias formalizes the psychology whereby people tend to misread evidence as supporting their currently held beliefs. "Currently held" is interpreted in their model as the person's beliefs about which of two hypotheses is more likely. As such, confirmatory bias leads to people being *over*-influenced by priors entering a period. We propose three factors that can identify the separate effects of BRN and confirmation bias.

Section 8.3 discusses the link between attempts to estimate the down-weighting of priors from BRN and efforts to identify whether experimental subjects misinterpret signals when updating beliefs. Using data presented in Griffin and Tversky (1992), we argue that experimental subjects appear to underweight both base-rates and signals on average, which means that it is crucial to account for both of these effects in any attempt to estimate the magnitude of BRN or signal misweighting. If the base-rate and the signal point in opposite directions, then a failure to account for BRN could yield estimates that suggest that subject's *over*-weight signals when in fact the signals are *under*-weighted. Symmetrically, if the base-rate and signal point in the same direction, then failing to account for BRN would suggest subjects underweight signals more than they actually do.

In Section 9 we discuss some more speculative extensions of the model. Section 10 discusses areas for future experimental tests, theoretical challenges, and potential applications.

2 Base-Rate Neglect

In this section, we review the motivating evidence for BRN and relate it to the basic formulation of BRN for a single act of updating priors. We turn to the dynamic model in Section 3.

2.1 Evidence and Basic Model

The abundant experimental literature on BRN includes three main types of experiments. In the paper that launched the literature on BRN, Kahneman and Tversky (1973) introduced the first type. They asked subjects to assign a probability to the event that a person is an engineer rather than a lawyer based on the following description:

> Jack is a 45 year old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.

Before being given this description, one group of subjects was told that the description was randomly drawn from 100 possible descriptions consisting of 30 engineers and 70 lawyers. A second group was provided the same description but told the population consists of 70 engineers and 30 lawyers. According to Bayes' Rule, the posterior odds ratio that Jack is an engineer versus a lawyer is given this description is

$$\frac{p(\text{engineer}|\text{description})}{p(\text{lawyer}|\text{description})} = \frac{p(\text{description}|\text{engineer})p(\text{engineer})}{p(\text{description}|\text{lawyer})p(\text{lawyer})}. \tag{1}$$

The base rates, $p(\text{engineer})$ and $p(\text{lawyer})$, are objective and provided to the subjects, but the informativeness of the descriptions, $p(\text{description}|\text{engineer})$ and $p(\text{description}|\text{lawyer})$, were left to the subjects' judgment. One elegant feature of this experiment is that it does not rely on these judgments: irrespective of participants' assessment of those likelihoods, an unambiguous prediction of Bayes' Rule is that the ratio of equation 1 across the base-rate conditions would be $\frac{70/30}{30/70} \approx 5.4$ times higher if the description were known to be drawn from the population that is 70% engineers relative to the population that is 30% engineers.[5] Contrary to this, the mean probability that Jack is an engineer offered by subjects, averaged across this description and four similar others, was 55% when engineers were 70% of the population and 50% when they were 30%, yielding a ratio of only $\frac{55/45}{50/50} = 1.2$—implying that subjects' posteriors are too insensitive to the base rate.

A number of psychological mechanisms were proffered for BRN in the lawyer-engineer problem. Kahneman and Tversky (1973) argued that subjects' posteriors are based on their judgments of whether the description is "representative" of a lawyer or engineer. Nisbett, Borgida, Crandall, and Reed (1976) argued that the descriptions are weighted more heavily because they are "vivid, salient, and concrete," while the base rates are "remote, pallid, and abstract." Bar-Hillel (1980) argued that neither of these provides a sufficient explanation of BRN because it also occurs in experiments in which both the base rates and likelihood information are abstract statistics. An example of this second type of experiment is the Cab Problem (originally due to Kahneman and Tversky (1972a)), in which subjects are asked:

> Two cab companies operate in a given city, the Blue and the Green (according to the color of cab they run). Eighty-five percent of the cabs in the city are Blue, and the remaining 15% are Green.
>
> A cab was involved in a hit-and-run accident at night.
>
> A witness later identified the cab as a Green cab.
>
> The court tested the witness' ability to distinguish between Blue and Green cabs under nighttime visibility conditions. It found that the witness was able to identify each color correctly about 80% of the time, but confused it with the other color about 20% of the time.
>
> What do you think are the chances that the errant cab was indeed Green, as the witness claimed?

The correct answer is $\frac{(0.8)(0.15)}{(0.8)(0.15)+(0.2)(0.85)} \approx 41\%$. Bar-Hillel found that only 10% of subjects gave answers close to the correct answer. The modal answer, which was given by 36% of subjects, was 80%—an answer that exactly mirrors the information from the signal and thus reflects complete

[5] Since this prediction is independent of how the subjects interpret the description, it is also clear that violations of this prediction are produced by incorrect utilization of base-rate information.

base-rate neglect. (Roughly 5% of subjects gave an answer close to 15%, reflecting complete "signal neglect.") On the basis of this problem and many variants, Bar-Hillel provided evidence against a number of potential explanations of BRN and argued that BRN is due to subjects' perception that the base rate is less relevant than the likelihood information.

Most of the subsequent literature has used the first or second type of experiment to study factors that increase or decrease the extent of base-rate neglect. Although Kahneman and Tversky (1973), Bar-Hillel (1980), and some other studies found high numbers of people who completely neglected base rates, Koehler (1996) reviews the literature and concludes that base rates are almost always used to some extent, and that their underuse varies by elicitation mode. For example, reviews by Koehler (1996) and Barbey and Sloman (2007) concluded that framing updating problems in terms of frequencies rather than probabilities weakens BRN, but the frequency frame does not *eliminate* it. Goodie and Fantino (1999) found that BRN can be reduced, but not eliminated, by extensive training with explicit feedback. In a large, representative sample of the German population, Dohmen, Falk, Huffman, Marklein, and Sunde (2009) found a large extent of BRN. Ganguly, Kagel, and Moser (2000) studied BRN with monetary incentives in market settings, doing so using both a contextualized vignette (the second type of BRN experiment) and an abstract experiment of the third type discussed below. Both methods induced BRN, with the contextualized setting producing significantly more.

Updating problems in many field settings resemble the first or second type of experiment, and researchers have documented the pervasivness of BRN in a variety of settings including courts' judgments in trials (Tribe (1971)), doctors' diagnoses of patients (Eddy (1982)), and psychologists' interpretations of diagnostic tests (Meehl and Rosen (1955)). Eide (2011) found a similar degree of BRN in the Cab Problem among law students as that typical of undergraduate samples. In response to realistic hypothetical scenarios, Kennedy, Willis, and D. Faust (1997) found that school psychologists were more confident but less accurate in assessing learning disability when base-rate information was supplemented with scores from a diagnostic screening.

The third type of experiment, often called a "bookbag-and-poker-chip experiment," involves abstract objects (e.g., balls drawn from urns) and is often conducted with incentives in a controlled laboratory setting. Such experiments were pioneered by Phillips and Edwards (1966), introduced into experimental economics by Grether (1980), and integrated into asset-market experiments starting with Camerer (1987). Griffin and Tversky (1992) Study 2 provides especially detailed evidence on BRN in an incentivized experiment. Because it illustrates some key issues that we identify below, we refer back to it several times throughout the paper. They considered inference from samples of 10 spins of what is either a 60% "heads" coin or 40% "heads" coin, explaining to subjects that a coin that is fair when tossed could nonetheless be biased when spun. The experimental design and results are shown in Table 1, which we have constructed from Griffin and Tversky's Table 2. The base rate of a 60% heads bias was varied from 0.10 to 0.90 (the "$p(H)$" column), and the strength of the evidence presented to the subjects was varied from 5 heads out of 10 spins to 9 out of 10 (the "# Heads (s)" column). After being told the prior probability and the data, the subjects were asked their "confidence" in percentage terms that the coin is biased in favor of heads. The median response for each experimental condition is shown in the "Median" column of the table; we treat these median responses as though they were offered by a single respondent providing her posterior beliefs in the different conditions. For comparison, the Bayes column shows the correct, Bayesian posteriors.

We return in Section 8.3.1 to a discussion of how focusing solely on the mis-weighting of signals may have lead researchers to attribute distorted posterior beliefs to overweighting of the signal.[6] But as noted above, Kahneman and Tversky (1973)'s elegant design reveals the underweighting of base rates in an absolute sense, regardless of any misweighting of signals. Moreover, following Grether (1980), a number of studies have used experiments with variation in both base rates and signal strength to separately identify BRN from misweighting of signals and have generally found that *both* are underweighted (for a review, see Benjamin (2018)). Table 1 illustrates the intuition for how these different weightings can be separately identified, and provides evidence for the general underweighting. In all of the rows with a prior of 0.50, BRN cannot have an effect on posterior beliefs, so deviations from Bayesian updating entirely reflect misweighting of signals. In the 0.5 prior condition, the median report posterior belief following the observation of 5 out of 10 heads was (unsurprisingly) 50%. But the median subject's posterior was less certain than the Bayesian posterior when there were 6, 7, 8 or 9 heads out of 10, which implies underweighting of signals. Analogously, in all of the rows with 5 heads out of 10, the signal is uninformative, so that deviations from Bayesian updating entirely reflects BRN. Although the median subject correctly reported posteriors equal to the prior when the prior was 0.33 or 0.50, the median subject's posterior was less certain than the prior when the prior equaled 0.10, 0.67, or 0.90, implicating BRN.

This logic can be extended to obtain estimates of our model's parameter α in all of the conditions where the prior is not equal to 0.50, while taking misweighting of signals into account. We assume that in the 0.50-prior conditions, the median subject's posteriors $\pi(H|s)$ are equal to the (mistaken) beliefs $\pi(s|H)$ about the likelihood of $s = 5, 6, 7, 8, 9$ heads out of 10 conditional on a heads-biased coin; and since $\pi(T|s) = 1 - \pi(H|s)$, we set the (mistaken) beliefs $\pi(s|T)$ equal to $1 - \pi(H|s)$. Then, using the median subject's posteriors $\pi(H|s)$ in all of the non-0.50-prior conditions, we calculate the implied value of α from the following updating rule

$$\pi(H|s) = \frac{\pi(s|H)p(H)^{\alpha}}{\pi(s|H)p(H)^{\alpha} + \pi(s|T)p(T)^{\alpha}}. \tag{2}$$

The column "Actual α" shows the results from this calculation. While these estimates should not be taken too seriously (e.g., we are using median responses and do not have standard errors), it is

[6] The analysis of Section 8.3.1 uses the final two columns of Table 1.

$p(H)$	# Heads (s)	Bayes	Median	Actual α	(Mis)est. α	$P(s\mid h)/P(s\mid T)$	$P_U(s\mid H)/P_U(s\mid T)$
0.10	5	0.10	0.23	0.56	0.56	1.00	2.61
0.10	6	0.20	0.45	0.28	0.46	2.25	7.36
0.10	7	0.36	0.60	0.20	0.55	5.06	13.5
0.10	8	0.55	0.80	0.00	0.48	11.4	36.0
0.10	9	0.74	0.85	0.21	0.69	25.6	51.0
0.33	5	0.33	0.33	1.02	1.02	1.00	0.99
0.33	6	0.53	0.50	0.58	1.17	2.25	2.00
0.33	7	0.72	0.57	0.82	1.93	5.06	2.65
0.33	8	0.85	0.77	0.26	1.77	11.4	6.70
0.33	9	0.93	0.90	0.00	1.51	25.6	18.0
0.50	5	0.50	0.50	n/a	n/a	1.00	1.00
0.50	6	0.69	0.60	n/a	n/a	2.25	1.5
0.50	7	0.84	0.70	n/a	n/a	5.06	2.33
0.50	8	0.92	0.80	n/a	n/a	11.4	4.00
0.50	9	0.96	0.90	n/a	n/a	25.6	9.00
0.67	5	0.67	0.55	0.29	0.29	1.00	0.61
0.67	6	0.82	0.65	0.31	-0.28	2.25	0.93
0.67	7	0.91	0.71	0.07	-1.05	5.06	1.22
0.67	8	0.96	0.83	0.24	-1.27	11.4	2.36
0.67	9	0.98	0.90	0.00	-1.51	25.6	4.50
0.90	5	0.90	0.60	0.18	0.18	1.00	0.17
0.90	6	0.95	0.70	0.20	0.02	2.25	0.26
0.90	7	0.98	0.85	0.40	0.05	5.06	0.63
0.90	8	0.99	0.93	0.51	0.04	11.4	1.37
0.90	9	0.996	0.99	0.90	0.43	25.6	7.30

Table 1: Estimating α from Griffin and Tversky (1992)

noteworthy that all but one of the inferred values of α fall within $[0, 1]$, indicating that the results from most of the conditions of Griffin and Tversky (1992) provide evidence of BRN.

As mentioned above, researchers sometimes draw conclusions about misweighting signals without accounting for BRN—and conversely, some of the literature on BRN does not take into account misweighting of signals. The column "(Mis)est. α" of Table 1 shows how estimates of BRN can be off if misweighting of signals is not accounted for. The estimates in this column are the implied values of α from the updating rule $p_\alpha(H|s) = \frac{p(s|H)p(H)^\alpha}{p(s|H)p(H)^\alpha + p(s|T)p(T)^\alpha}$, where the difference from equation 2 is that the true likelihoods, $p(s|H)$ and $p(s|T)$, are used in the equation instead of estimates of subjects' mistaken likelihoods. When the prior is 0.33, the signal and prior point in opposite directions, so underweighting the signal is misattributed to *overweighting* the prior (i.e., $\alpha > 1$). Even more perversely, when the prior is 0.67, the signal and prior point in the same direction, so underweighting the signal is misinterpreted as evidence for treating the prior as favoring the tails-biased coin ($\alpha < 0$)!

From a meta-analysis of bookbag-and-poker-chip experiments from 14 papers in which subjects are provided with a single sample of data, Benjamin (2018) estimated α using equation 2 above, where the (mistaken) beliefs $\pi(s|H)$ and $\pi(s|T)$ are estimated from experiments with 50-50 priors. Benjamin found $\widehat{\alpha} = 0.61$ (SE = 0.07). When restricting to the subsample of 6 papers with incentivized experiments, the point estimate is *smaller*: $\widehat{\alpha} = 0.43$ (SE = 0.09).

As mentioned in the Introduction, our model of BRN is based on an empirical regression specification introduced by Grether (1992), Experiments II and III. Grether conducted an incentivized, bookbag-and-poker-chip experiment in which subjects were given an initial prior, presented with a sequence of samples, and asked to report their posterior belief after each sample. For notational consistency, we denote the two urns as H and T and the t^{th} sample as s_t. Using this notation, Grether estimated the panel regression

$$\ln\left(\frac{\pi_{it}(H|s_t, s_{t-1}, ..., s_1)}{\pi_{it}(T|s_t, s_{t-1}, ..., s_1)}\right) = \beta_0 + \beta_1 \ln\left(\frac{p_t(s_t|H)}{p_t(s_t|T)}\right) + \beta_2 \ln\left(\frac{\pi_{it}(H|s_{t-1}, s_{t-2}, ..., s_1)}{\pi_{it}(T|s_{t-1}, s_{t-2}, ..., s_1)}\right) + \epsilon_{it}, \quad (3)$$

where $\pi_{it}(H|s_{t-1}, s_{t-2}, ..., s_1)$ and $\pi_{it}(T|s_{t-1}, s_{t-2}, ..., s_1)$ are individual i's posteriors after having observed samples $s_1, s_2, ..., s_{t-1}$. These posteriors are assumed to become the priors when updating upon observing s_t, and $\pi_{i0}(H)$ and $\pi_{i0}(T)$ are the initial priors (which are the same for all individuals). The coefficient β_1 captures misweighting of signals, while the coefficient β_2 captures misweighting of priors and corresponds to our parameter α. Benjamin (2018) meta-analyzed results from eight experiments following the design of Grether (1992). All but one of the β_2 estimates were in $(0, 1)$ (the exception is Grether (1992) Experiment II, which found $\widehat{\beta}_2 > 1$). The inverse-variance-weighted mean estimates of β_1 and β_2 from the meta-analysis were 0.53 (SE = 0.01) and 0.88 (SE = 0.01), consistent with underweighting of both signals and underweighting of priors.

We end our review of the literature by highlighting that while the most of the evidence on BRN focuses on settings where priors are objective and provided to subjects (i.e., the priors correspond to base rates in the population), our model and applications assume that BRN applies more generally when updating, including updating from subjective priors.[7] Some support for this is provided by sequential-sample bookbag-and-poker-chip experiments that build on Grether (1992), which provide subjects with some initial priors and then estimates BRN from the subjects' updates. The subjects' updates and the resulting estimates are based not only on their initial priors but also (and mostly) from their subsequent priors. This would seem to rule out the possibility that only objective base rates are neglected (while subjective prior beliefs are given full weight).[8] Regardless of the stand one takes on the difference between base rates and prior beliefs, the characterization of BRN updating from a single signal presented in this section continues to hold for one-shot updating experiments that provide base rates to the subjects.

2.2 Immoderate Movement and Moderate Posteriors

Many "textbook" examples of BRN are like the Cab Problem: they focus on how people update following a signal with a likelihood ratio in the opposite direction of the priors. Another iconic example, based on Eddy (1982), is about medical diagnosis. A person is getting a test for a disease: if she has the disease, the test will be positive 90% of the time; if she doesn't have the disease, it will be negative 90% of the time. Assume that 5% of the population being tested have the disease. If the test is positive, what is the probability of disease? The right answer is around 32%—the low base rate of the disease means that the positive test result is probably a false positive. In her extreme form ($\alpha = 0$), by contrast, Saki will give an answer of 90%. In these textbook examples, Saki will think the wrong hypothesis most likely, and have an extreme opinion about it too.

But, of course, signals generally point in the *same* direction as the prior. If most people do not have a disease, then test results will be negative 86% ($= \frac{5}{100} \cdot \frac{1}{10} + \frac{95}{100} \cdot \frac{9}{10}$) of the time. When the

[7] One might interpret our model as predicting that in the lawyer-engineer problem, it should matter whether the description of Jack is presented to subjects before or after the base rates. In particular, our model might be interpreted as predicting that if the description comes first, then it would be treated as the subjects' priors and the description—rather than the base rates—would be downweighted when updating. We do not view such evidence as a clear test of our assumption, though, for two reasons: (i) reversing the order of presentation may not cause people to reverse what is the prior and what is the signal since the base rates (a probability distribution) are more naturally treated as priors than the description (a sample realization); and (ii) information from signals is *also* underweighted (see Sections 2.1 and 8.3), so we expect the standard tests for BRN to also show underweighting of base rates in the reversed presentation. In an early literature review, Borgida and Brekke (1981) concluded that "there appears to be no direct relationship between base rate utilization and order of presentation" (p. 73). Two recent papers containing multiple experiments consistently found that subjects underuse base rates by more when the base rates are presented before the description (Krosnick, Li, and Lehman (1990); Chun and Kruglanski (2006)). For example, Krosnick, Li, and Lehman (1990), Experiment 1, replicated the result of Kahneman and Tversky (1973) when the base rates were presented first, with subjects' mean probability of engineer being 53% regardless of whether the base rate of engineers was 70% or 30%. When the description of Jack was presented first, subjects' mean probability of engineer was 57% when the base rate was 70% and 39% when the base rate was 30%.

[8] It should be noted, however, that the analysis of these experiments relies on the assumption that posterior beliefs today become (potentially neglected) priors when the beliefs are next updated; we discuss the evidence related to this assumption in Section 3.

test result is negative, Tommy would believe there is $\frac{.95\times.9}{.95\times.9+.05\times.1} > 99\%$ chance he is disease-free following a negative result, while Saki thinks there is only a 90% chance. In other words, in the more common case wherein the evidence supports prior beliefs, Saki's beliefs are *less* extreme than Tommy's. We refer to the fact that Saki typically holds less extreme beliefs than Tommy as the *moderation effect*. In fact, using a common measure of uncertainty over binary questions, $p(1-p)$, for which 0 is maximal certainty and 0.25 is maximal uncertainty, Augenblick and Rabin (2017) show that the reduction in uncertainty in one-shot updating from correct priors is on average lower for Saki than for Tommy. In this sense, BRN leads people to be less certain on average than they should be. There is a simple intuition for this: when a person is using some information correctly but under-using other information, on average she'll be less certain than she should be. (In Section 3, we show how the dynamic extension of BRN manifests this moderation effect: Saki's beliefs as she receives an infinite flow of signals will be forever bounded away from certainty.)

The fact that Saki's beliefs exhibit the moderation effect, making her (on average) underconfident about the truth relative to Tommy, should not be conflated with how much she is changing her mind. When the correctly used information is new (the signal) and the under-used information is old (the prior), her beliefs will "move too much" in response to new information. To get a handle on what it means for beliefs to "move too much," consider the metric $(p_{t+1}-p_t)^2$, where p_t denotes the probability an agent believes a hypothesis (vs. the complement) is true in period t. Augenblick and Rabin (2017) show that the expected value of $(p_{t+1}-p_t)^2$ is always greater for Saki than for Tommy when updating from correct priors. Returning to the medical testing example, in the 14% of the time where the medical test gives a positive result, Tommy's beliefs move up 0.27, from 5% to 32%. In the 86% of the time she gets a negative result her beliefs move down 0.045, from 5% to less than 1%. Saki's beliefs, on the other hand, move up 0.85, from 5% to 90%, following positive results and move up 0.05, from 5% to 10%, following negative results. Saki's beliefs are moving around more.

A second sense in which Saki's beliefs tend to be too moderate seems likely to be less economically important, but is more dramatic and more surprising: when starting out believing in a hypothesis, Saki may believe in it *less* following weak supportive evidence, which we call the *extreme moderation effect*. This was seen in the thought experiment wherein we turned the textbook medical-test example around: as noted, (extreme) Saki who started out thinking there was only a 5% chance of having the disease and got a *negative* test result would believe that she now has the disease with 10% chance. Her belief that she is disease-free is weaker despite the reassuring medical test.

The extreme moderation effect can be framed in two ways, as described in Proposition 1.

Proposition 1 *Fix any $\alpha < 1$. Consider two hypotheses θ and θ'. Then:*

1. *For all signals s where $1 < \frac{p(s|\theta)}{p(s|\theta')} < \infty$, there exist prior beliefs $p(\theta), p(\theta') < 1$ such that $\frac{p_\alpha(\theta|s)}{p_\alpha(\theta'|s)} < \frac{p(\theta)}{p(\theta')}$.*

2. *For all* $p(\theta) > p(\theta')$, *there exists* $z > 1$ *such that for all signals* s *where* $\frac{p(s|\theta)}{p(s|\theta')} < z$, $\frac{p_\alpha(\theta|s)}{p_\alpha(\theta'|s)} < \frac{p(\theta)}{p(\theta')}$.

Part (1) says that for any signal favoring θ that does not perfectly distinguish θ and θ', there exists a (strong enough) prior on θ such that the signal will make Saki believe less in θ. Part (2) states that if Saki believes the hypothesis θ is more likely to be true, then weak-enough signals in favor of that belief will make Saki less confident that θ is true.

While the extreme moderation effect is surprising, Griffin and Tversky (1992) Study 2 provides some evidence for it (albeit not highlighted by Griffin and Tversky). Consider again the study's design and results, shown in Table 1. In two conditions, subjects were given a base rate of 90% that a coin is heads-biased and were told that 6 or 7 of the 10 spins resulted in heads. The median subject assigned an average posterior of *less* than 90% in these conditions. This represents a puzzle for the Bayesian model since the sample is evidence in favor of the hypothesis that the coin was biased towards heads, so the subject should be more confident the coin is biased (i.e., put a probability of greater than 90% on this hypothesis) rather than less.

Bar-Hillel (1980) also provides some evidence of the extreme moderation effect, although unfortunately she only reports subjects' modal beliefs. In one version of her Problem 4, subjects were told that 8 out of 10 urns contain 75% blue beads and 25% red beads, while 2 out of 10 contain the reversed proportions. Four beads are drawn from a randomly selected urn, and three are blue. This is again a case where both the prior and the signal favor the blue-majority urn. While the correct Bayesian posterior in favor of the blue-majority urn is 97%, the modal reported posterior (reported by 14 out of 54 subjects) was 75%—lower than the base rate of 80%.

Further support for the extreme moderation effect comes from cases where people's beliefs become more moderate following uninformative data. For example, Griffin and Tversky (1992) Study 2 also has a condition with a base rate of 90% and 5 heads out of the 10 spins. For a Bayesian, beliefs should be unaffected by uninformative data. The median subject's posterior, however, fell to 60%.

2.3 Hypothesis Dependence

As with all non-Bayesian models we are familiar with, the effects of BRN inherently depend on the set of hypotheses considered. This hypothesis dependence obviously interferes with ease of applying the model and establishing general implications of BRN. But it is an empirical reality that updating is subject to these effects, and for all their problematic features, in essence they should be a feature of the model.

The most prominent empirical form of hypothesis-dependence is at the core of *support theory* (Tversky and Koehler (1994)), namely that the probability a person attributes to an "event" is lower than the total probability that person would ascribe to the disjoint subevents that comprise that event. For example, Fischhoff, Slovic, and Lichtenstein (1978) asked subjects to assess the probability of various reasons why a car might fail to start. The subjects assigned a probability of 0.22 to the residual hypothesis of "The cause of failure is something other than the battery, the fuel system, or the engine," but the probability increased to 0.44 once the residual hypothesis was unpacked into subevents (e.g., "The cause of failure is the ignition system.")[9]

Most of the evidence for subadditivity, as in the car example, focuses on settings with no obvious connection to updating. We find that, due to BRN, subadditivity is also induced (or reinforced) by the process of updating. Given a set of (disjoint) hypotheses A, B, and C, Saki could consider the set of hypotheses $\Theta_1 = \{A, B, C\}$ or the set $\Theta_2 = \{A, B \cup C\}$. Saki does not try to distinguish between B and C under Θ_2, whereas she does under Θ_1. Tommy would have the same beliefs about the likelihood of $B \cup C$ under either Θ_1 or Θ_2.

Let $p_\alpha^{\Theta_i}(\theta|s), \theta \in \Theta_i$ denote Saki's posterior beliefs about θ after observing s while considering Θ_i. BRN causes Saki to (partially) neglect the rarity of B and C under Θ_1 relative to the greater prior probability of the union of these hypotheses under Θ_2. Proposition 2 proves that this results in subadditivity.

Proposition 2 $p_\alpha^{\Theta_2}(B|s) + p_\alpha^{\Theta_2}(C|s) > p_\alpha^{\Theta_1}(B \cup C|s)$ *when* $\alpha < 1$.

Proposition 2 should be interpreted as an interpersonal comparative static relating the beliefs of two Sakis considering different sets of hypotheses. Saki's beliefs are never subadditive when considering a single, fixed set of hypotheses.

BRN can also induce another well known empirical regularity: violations of the conjunction principle. Conjunction violations occur when a decision maker places a strictly higher probability on a hypothesis A than on $B \supset A$, thereby violating the axioms of probability theory. For example, Kahneman and Tversky (1983) offered subjects the following description:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

In one of their experiments, Kahneman and Tversky then asked subjects whether it is more likely that "Linda is a bank teller" or "Linda is a bank teller and is active in the feminist movement." 85% of their respondents selected the second statement as more likely, in violation of the conjunction principle. Kahneman and Tversky (1983) associated violations of the conjunction principle with the representativeness heuristic.

[9] See Benjamin (2018) for a review of the large body of evidence in favor of such "subadditivity" and a discussion of psychological mechanisms that contribute to it. The hypothesis-dependence of beliefs has also been the focus of decision-theoretic treatments of non-Bayesian reasoning (e.g., Ahn and Ergin (2010) and Blume, Easley, and Halpern (2009)).

The experiment of Kahneman and Tversky (1983) found intrapersonal conjunction fallacies in the sense that the subjects were asked their beliefs after being informed of all of the hypotheses that they would be required to evaluate (and without being provided information). BRN can create conjunction fallacies when we compare Saki's responses when updating under different sets of hypotheses, which is analogous to an experiment comparing the beliefs of two Sakis entertaining different sets of hypotheses. Saki's beliefs would never exhibit a conjunction fallacy when considering a single, fixed set of hypotheses.

Proposition 3 *Suppose the state space contains events $A \subset B \subset \Omega$, and consider the sets of hypotheses $\Theta_1 = \{A, \Omega - A\}$ and $\Theta_2 = \{B, \Omega - B\}$. Suppose that there is some signal s such that $p(s|A) > p(s|B)$ and $p(s|\Omega - B) \geq p(s|\Omega - A)$. For $\alpha \in [0, 1]$ sufficiently small, $p^{\Theta_1}_{\alpha}(A|s) > p^{\Theta_2}_{\alpha}(B|s)$.*

BRN also predicts some unfamiliar types of hypothesis dependence, which could serve as out-of-sample tests of the formulation of BRN we explore. To illustrate, suppose that two workers, Heidi and Tarso, have just been hired. They will be evaluated based on who performed better each month—no ties are announced. Each month, we observe either h (Heidi is better) or t (Tarso is better). Relative performance each month is a "toss up"–there is a 50-50 chance each month, no matter what happened before, that Heidi will be deemed better that month. Suppose Saki is wondering about the hypothesis "What is the probability that Heidi outperforms Tarso the first 5 months?" Saki and Tommy will think initially that there is a $\frac{1}{32}$ probability that the hypothesis is true. Moreover, if she observes that Tarso outperforms Heidi in Month 1 ($s_1 = t$), Saki (like Tommy) would place probability 0 on that hypothesis.

But what happens if she observes Heidi outperform Tarso in Month 1 ($s_1 = h$)? Suppose Saki continues to consider the hypothesis, "What is the probability that Heidi outperforms Tarso the first 5 months?"and the complement of that hypothesis. Saki and Tommy would try to integrate the signal into their prior beliefs about the probability of Heidi winning a 5 month sweep (i.e., the hypothesis above). If Heidi sweeps Tarso, she would for sure outperform Tarso in month 1. If Heidi does not sweep the first 5 months, there is a $\frac{15}{31}$ chance that she outperforms Tarso in the first month. Tommy, applying Bayes Rule, finds that Heidi will sweep with probability $(\frac{31}{31}\frac{1}{32})/(\frac{31}{31}\frac{1}{32} + \frac{15}{31}\frac{31}{32}) = \frac{1}{16}$. In contrast, if Saki completely neglects her prior beliefs ($\alpha = 0$), Saki would reach the conclusion that there is a $(\frac{31}{31})/(\frac{31}{31} + \frac{15}{31}) = \frac{31}{46} \approx .674$ chance that Heidi would sweep. Because Saki is about twice as likely to see Heidi win month 1 if she will sweep than if she won't, Saki thinks there is a 2/3 chance of a sweep.[10]

After observing Heidi outperform Tarso on Month 1, one could reframe the question facing Saki and Tommy as the (different) hypothesis "What is the probability that Heidi outperforms Tarso the next 4 months?"and the complementary hypothesis. The structure of the problem provides a

[10] This prediction is, of course, extreme; if $\alpha = 0.5$, which is within the range of values found in experiments, then Saki will update to think the probability of a sweep is $(\frac{31}{31}\sqrt{\frac{1}{32}})/(\frac{31}{31}\sqrt{\frac{1}{32}} + \frac{15}{31}\sqrt{\frac{31}{32}}) = 0.270$.

natural base rate $\left(\frac{1}{16}\right)$ for the hypothesis, and since no signals have been observed to inform the likelihood of this hypothesis, Saki and Tommy agree that the probability the hypothesis is true is $\frac{1}{16}$. Tommy's beliefs do not depend on the framing, while Saki's beliefs change after the reframing.

2.4 What is a Signal?

For Tommy, receiving an uninformative realization of a signal is equivalent to receiving no signal at all. For Saki, however, since BRN is triggered by the act of updating, there is a difference: receiving no signal leads to no updating, while receiving an uninformative realization of a signal causes updating and hence moderation of beliefs.

While this prediction may seem exotic, we already discussed some evidence for it from Griffin and Tversky (1992) Study 2. It was also found in the original experiment on BRN in Kahneman and Tversky (1973). After showing BRN from the Jack description (see Section 2.1), Kahneman and Tversky conducted the same experiment two more times. In one experiment, subjects were given "no information whatsoever about a person chosen at random from the sample." With no evidence to combine with the base rate, subjects did not neglect base rates at all. The median probability assigned to an engineer was 70% in the 70% base-rate group and 30% in the 30% base-rate group. In the other experiment, subjects were given the following description, intended to be completely uninformative:

> Dick is a 30 year old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

In both groups, the median probability assigned to Dick being an engineer was 50%, consistent with the subjects interpreting the description as uninformative and ignoring the base rates.

As with results in the lawyer-engineer problem using informative descriptions, when BRN is found from uninformative descriptions, the base rates are often not *completely* ignored. Moreover, while some experiments have replicated BRN from uninformative descriptions, others have found no BRN at all (see Koehler (1996) Table 1). Sometimes BRN and no BRN have been found across different groups of subjects or stimuli within the same experiment (Zukier and Pepitone (1984) and Gigerenzer, Hell, and Blank (1988), Experiment 1).

Although limited, there is also evidence from bookbag-and-poker-chip experiments. Troutman and Shanteau (1977) drew beads from one of two boxes that were equally likely to be chosen, showed subjects a cup containing the beads, and elicited subjects' posteriors. The numbers of red, white, and blue beads were 70, 30, and 50 in Box A and 30, 70, and 50 in Box B. Troutman and Shanteau (1977) studied the effects on subjects' posteriors of three kinds of uninformative data sets: "null samples," which were cups containing no beads at all; "irrelevant samples," containing two blue beads; and "mixed samples" of one red and one white bead, and found that all three types of uninformative signal realizations caused moderation of beliefs. Likewise, in studying how beliefs are affected by mixed samples (but not null or irrelevant samples), Shanteau (1975) found that the effects of initial evidence favoring one of the hypotheses were moderated. However, in an experiment closely modeled on Troutman and Shanteau (1977), Labella and Koehler (2004) studied the effects of irrelevant and mixed (but not null) samples and did not replicate their results. Labella and Koehler (2004) found that beliefs did not change following irrelevant samples and became *more* extreme following mixed samples.[11] Although mixed, we view the evidence overall as supportive.

The distinction between what is an uninformative signal realization vs. no signal would seem to suggest that when Saki sees something that she codes as a signal, even if she in the end concludes that the signal realization is uninformative, per the model she updates her beliefs and discounts her prior beliefs in the process. To complete our model, we assume that Saki codes any random variable with a distribution that depends on θ as a signal. Formally, s_t is considered a signal if there exists hypotheses θ and θ' such that $p(s_t|\theta, s_{t-1}, ..., s_1) \neq p(s_t|\theta', s_{t-1}, ..., s_1)$, where $(s_{t-1}, ..., s_1)$ denote previously observed signals. In our applications, we will evade (rather than solve) the issue by defining what is and is not a signal.

3 The Dynamics of Updating

In this section we flesh out the dynamic model in a setting where Saki updates her beliefs repeatedly as she observes a series of signals. Periods are indexed by $t \in \{1, 2, ...\}$ and, as in Section 2, hypotheses are denoted $\theta \in \Theta$ with prior belief $p(\theta)$. In each period t the agent may observe a signal $s_t \in S$.

In order to extend the model to a dynamic setting, the key additional assumption—the major modeling gambit of the paper—is that a person who sequentially processes information will treat updated posteriors as the priors for further updating. Although we know of no direct tests of this assumption and discuss alternatives later, we think it is natural and all econometric analyses of BRN from experiments (reviewed in Section 2.1) make it. Moreover, while the priors are "base rates" (proportions in the population that have some characteristic) in most of the evidence for BRN, our model assumes that people also downweight the priors that have resulted from earlier updating. The estimates from regression equation 3, which indicate that priors are underweighted in sequential-updating experiments, provide some support for that assumption.[12]

In period $t = 1$, Saki receives a signal s_1 with conditional likelihood $p(s_1|\theta)$ and updates her

[11] They intepreted the latter result as consistent with confirmatory bias; we discuss the tension between BRN and confirmatory bias in Section 8.3.3.

[12] This evidence does not rule out that there may be additional under-use of base rates beyond the under-use of other priors.

beliefs to $p_\alpha(\theta|s_1) = \frac{p(s_1|\theta)p(\theta)^\alpha}{\Sigma_{\theta'}p(s_1|\theta')p(\theta')^\alpha}$. When period $t = 2$ starts, Saki repeats this process with the posterior formed at the close of period $t = 1$ serving as the prior belief in $t = 2$. Therefore, if Saki receives a signal s_2, Saki forms a belief about the relative likelihood of hypotheses θ and $\widetilde{\theta}$ equal to

$$\frac{p_\alpha(\theta|s_1, s_2)}{p_\alpha(\widetilde{\theta}|s_1, s_2)} = \frac{p(s_2|\theta, s_1)}{p(s_2|\widetilde{\theta}, s_1)} \left(\frac{p(s_1|\theta)}{p(s_1|\widetilde{\theta})}\right)^\alpha \left(\frac{p(\theta)}{p(\widetilde{\theta})}\right)^{\alpha^2}. \tag{4}$$

Note that $p(s_2|\theta, s_1)$ indicates the interpretation of s_2 can depend on s_1 as well as θ. This posterior is then used in period $t = 3$ as the prior belief. Using this procedure we can recursively define Saki's posteriors after observing a sequence of signals observed in successive periods. Beliefs after updating in period t have the form

$$\frac{p_\alpha(\theta|s_1, s_2, ..., s_t)}{p_\alpha(\widetilde{\theta}|s_1, s_2, ..., s_t)} = \left(\frac{p(\theta)}{p(\widetilde{\theta})}\right)^{\alpha^t} \prod_{\tau=1}^{t} \left(\frac{p(s_\tau|\theta, s_1, s_2, ..., s_{\tau-1})}{p(s_\tau|\widetilde{\theta}, s_1, s_2, ..., s_{\tau-1})}\right)^{\alpha^{(t-\tau)}}.$$

At period t the informational content of the previous signals (e.g., $p(s_{t-1}|\theta, s_1, .., s_{t-2})$) is neglected, but the previous signals provide context for interpreting the current signal through the conditioning.

Although Saki downweights past signals, we emphasize that this is different than forgetting them (an issue we return to in Section 8.2). As an example to illustrate the distinction, suppose Saki sequentially surveys five employees at a firm to learn about the hypothesis "Do at least half of the employees agree with the manager's strategy?" Assume that each employee truthfully reveals whether he or she agrees and Saki updates her prior (and neglects past information) after she interviews each employee. Once Saki observes the third employee agree with the manager, Saki immediately becomes certain that the hypothesis is true. This requires that Saki remember all of her previous employee interviews (even if she neglects the information revealed in her updating) and understand how the previous interviews affects the informativeness of the current one.

Although we find the assumption that Saki's posterior beliefs become her prior beliefs for further updating intuitive, we can think of two alternatives. One alternative hypothesis, which Benjamin, Rabin, and Raymond (2016) call "pooling," is that the current signal is pooled with all past signals, which is then treated as a single sample, and people update from initial priors in response to the pooled sample. Pooling predicts that posteriors from sequential samples should be equal to posteriors from a simultaneous sample that contains the same set of signals. The two papers that have directly tested this hypothesis have found evidence against it (Shu and Wu (2003), Study 3, and Kraemer and Weber (2004)).[13] We find it difficult to interpret what exactly

[13] For example, Shu and Wu found that subjects ended up with different posteriors when they updated their beliefs after each of 10 signals than when they updated after observing the same signals in groups of two or groups of five. In a review of the literature, Benjamin (2018) compared findings across papers and concluded that there is additional evidence against the pooling hypothesis.

pooling would mean, unless there is some reason to think that beliefs we see at the moment we first peer in at Saki—or experimental participants—are their original priors before getting any information. Presumably the beliefs people hold are the result of previous instances of updating, and as such it must be implicit in the formulation that the downweighting must occur on the beliefs at the beginning of an act of updating, not the beginning of time. That all being said, in Section 9.2 we offer a formulation of our model where hypothetical original priors are correctly weighted in updating and the information contained in signals is neglected as successive signals are observed. Almost all of our findings continue to hold in this alternative model.

An alternative possibility is that Saki collects information by passively collecting the information when not making decisions, and only uses this information to update beliefs when using the information to make a choice. The "clump" of information observed between decisions would then be treated as a single aggregate signal that is used to update prior beliefs.[14] The formal results we establish in the paper continue to hold, but in cases where a great many informative signals were observed between decisions, Saki's beliefs would be closer to Tommy's because the pooled signals would have a more extreme total likelihood.

We now focus on a setting that will allow us to most simply illustrate some of its basic features. Consider a person observing signals that are i.i.d. conditional on her hypotheses, such as when the hypothesis is the bias of a coin and the signals are flips (or spins) of that coin. If the signals are independent, then we can use the simpler formula

$$\frac{p_\alpha(\theta|(s_\tau)_{\tau=1}^t)}{p_\alpha(\widetilde{\theta}|(s_\tau)_{\tau=1}^t)} = \left(\frac{p(\theta)}{p(\widetilde{\theta})}\right)^{\alpha^t} \prod_{\tau=1}^{t} \left(\frac{p(s_\tau|\theta)}{p(s_\tau|\widetilde{\theta})}\right)^{\alpha^{(t-\tau)}}. \tag{5}$$

It is often notationally convenient to write our formulas in log-likelihoods, letting

$$L(\theta,\widetilde{\theta}|(s_\tau)_{\tau=1}^t) = \ln\frac{p_\alpha(\theta|(s_\tau)_{\tau=1}^t)}{p_\alpha(\widetilde{\theta}|(s_\tau)_{\tau=1}^t)} = \sum_{\tau=1}^{t} \alpha^{t-\tau} l_\tau(\theta,\widetilde{\theta}) + \alpha^t l_0(\theta,\widetilde{\theta}) \tag{6}$$

where

$$l_\tau(\theta,\widetilde{\theta}) = \ln\frac{p(s_\tau|\theta)}{p(s_\tau|\widetilde{\theta})}, \quad l_0(\theta,\widetilde{\theta}) = \ln\frac{p(\theta)}{p(\widetilde{\theta})}.$$

Note that equation 6 predicts a long-run form of the moderation effect. If Saki observes a long sequence of signals that happen to be uninformative, equation 6 implies that Saki's beliefs will converge towards a uniform distribution over the set of hypotheses she entertains.

While Tommy's beliefs are unaffected by the order in which signals are observed, equation 6 implies that Saki's beliefs exhibit a recency bias—she draws stronger inferences from signals observed recently relative to signals observed in the more distant past. BRN predicts that, as for Tommy, Saki's beliefs will eventually be uninfluenced by her priors. Whereas this happens for

[14] This modeling issue is similar to the issue of clumping discussed in Benjamin, Rabin, and Raymond (2016).

Tommy because he has settled near the truth and is no longer updating, with Saki the original priors stop mattering even though she is continuing to update.

Equation 6 also implies that so long as the signals have bounded informativeness, Saki will never become confident about the true hypothesis describing the world regardless of how much information she has received. Suppose that the informativeness of the realizations of the signals, $l_\tau(\theta, \widetilde{\theta})$, are bounded above and below by $\overline{L}$ and $-\overline{L}$ for some $\overline{L} < \infty$. Using equation 6 we have for any infinite sequence of signals $(s_\tau)_{\tau=-\infty}^{t}$ that

$$\ln \frac{p_\alpha(\theta|(s_\tau)_{\tau=-\infty}^{t})}{p_\alpha(\widetilde{\theta}|(s_\tau)_{\tau=-\infty}^{t})} = \sum_{\tau=-\infty}^{t} \alpha^{t-\tau} l_\tau(\theta, \widetilde{\theta}) \leq \overline{L} \sum_{\tau=-\infty}^{t} \alpha^{(t-\tau)} = \frac{\overline{L}}{1-\alpha}. \tag{7}$$

Since the log-likelihood ratio cannot diverge to infinity, it must be the case that Saki's beliefs are always bounded away from certainty. In our model, Saki's beliefs not only do not converge to the truth, Saki's beliefs fail to converge entirely. In effect, Saki will never become confident in any particular hypothesis. In contrast, under mild identification assumptions, Tommy's beliefs almost surely converge to the true hypothesis after receiving an infinite number of such signals.

As long as Saki continues to receive signals, her beliefs will continue to change. Treating Saki's beliefs as a random variable, equation 6 implies that average log likelihood ratio she places on two hypotheses θ and $\widetilde{\theta}$ if the hypothesis $\widetilde{\theta}$ is true has a mean of

$$\mathbb{E}\left[L(\theta, \widetilde{\theta}|(s_\tau)_{\tau=-\infty}^{\infty}, \alpha) \middle| \widetilde{\theta} \right] = \mathbb{E}\left[\ln \frac{p_\alpha(\theta|(s_\tau)_{\tau=-\infty}^{\infty})}{p_\alpha(\widetilde{\theta}|(s_\tau)_{\tau=-\infty}^{t})} \middle| \widetilde{\theta} \right] = \frac{1}{1-\alpha} \mathbb{E}\left[\ln \frac{p(s|\theta)}{p(s|\widetilde{\theta})} \middle| \widetilde{\theta} \right]. \tag{8}$$

The conditioning on $\widetilde{\theta}$ reflects the fact that the distribution of beliefs is determined by the distribution of signals, which is in turn determined by the true hypothesis, $\widetilde{\theta}$. In the limit as $\alpha \to 1$, the log-likelihood diverges as would be the case for Tommy. It is also straightforward to show that the "likelihood ratio martingale" is not, in fact, a martingale under BRN, which means that we have to use non-standard techniques to characterize the evolution of Saki's beliefs. In situations where the distribution of l_τ has full support, one can often prove that Saki's beliefs have an ergodic distribution with nontrivial support (see Section 7).

Although it does not matter for most of the applications from Sections 4, 6, or 7, in most economic setting a person's anticipation of her future beliefs matters. For example, if a person making an error in statistical reasoning is choosing whether to acquire more information, making predictions about the person's behavior requires assumptions about how the person thinks they will process information in the future. The implicit assumption in any Bayesian model is that people think that they will process information correctly in the future. And for many types of errors people make, there is every reason to believe people correctly anticipate how they will use information in the future. For example, quasi-Bayesian models posit agents that dogmatically believe the true model of the world is impossible, but the agents process information in a Bayesian fashion at all points in time—and know that they do so. But for cognitive biases of various sorts, it is less likely that beliefs about future processing will match actual processing. As with any error in reasoning, there is reason to doubt that there will be 'sophistication' about the error, while continuing to make the error. Although we know of no evidence on prospective beliefs about the role of base rates, to complete our theory of base-rate neglect, we assume that Saki believes she will be a Bayesian when she observes future information.

However, suppose Saki actively plans how she will gather and use future information, which naturally requires answering questions such "How much will I know after each signal?" If she believes that she will be Bayesian in the future, then her predictions of her own future beliefs conditional on future signals will be incorrect. However, the tension between her predicted beliefs and her actual beliefs under the influence of BRN might cause her to adopt, or at least move towards, Bayesian posteriors. We refer to this as *pre-emptive Bayesianism* since the act of forward-looking planning pre-empts Saki's natural tendency to neglect base rates. This notion that attention to base rates at a moment in time might be "grandfathered" in by past attention to base rates in the contingency that has arisen seems plausible. On the other hand, if she is confident that she will use Bayes' rule for future updating, there may be no particular reason for her to later attend to the fact that she does not (see Gagnon-Bartsch, Rabin, and Schwartzstein (2018)). We view this as an interesting target for future experimental investigation.

Finally, it is worth discussing the difference between Saki and a Bayesian updater that believes the true hypothesis describing the world might be changing. Both should discount information learned long ago and both should find it impossible to learn the true hypothesis unless signals are very strong. The key difference between the two is that the rate at which Saki discounts the past is dictated by the rate at which information arrives, whereas a Bayesian will neglect the past more when the rate (and potentially the magnitude) of hypothesis changes is large. Although it is easy to generate examples where the two rates are correlated, there are also many plausible models where they are not. This is particularly true in cases where the information gathering rate is endogenously determined (e.g., see Section 5).

4 Theory Updating and Forecasting

In this section we discuss how Saki forecasts forthcoming events when she is using recent signals to both (1) update her beliefs over hypotheses about what that signal-generating process is and (2) predict forthcoming events *given* a hypothesis about the signal-generating process. Our analysis has two themes. First, since Saki's beliefs vacillate between hypotheses about the underlying theory of the world, her beliefs often mimic "quasi-Bayesian" models wherein agents dogmatically believe in a mis-specified theory of the world (e.g., autocorrelation in i.i.d. environments, adaptive expectations in normal-normal updating problems). Second, we show that Saki's beliefs exhibit *prediction momentum* (i.e., she can appear to project the recent past into the future) even when she dogmatically believes that there is no autocorrelation. For most of this section we focus on the $\alpha = 0$ case to cleanly demonstrate the effects we are studying.

Papers beginning with Barberis, Shleifer, and Vishny (1998) have looked at quasi-Bayesian models where the true signal-generating process, θ^*, is not amongst those considered possible (i.e., $\theta^* \notin \Theta$). They then investigate what Bayesian updating implies given the real process is θ^*, and typically these models find that mispredictions based on a false theory can last forever.[15] The key difference in our setup is that Saki does entertain the true hypothesis (i.e., $\theta^* \in \Theta$), but BRN prevents her from learning what it is. Moreover, Saki may never believe θ^* is the most likely element of Θ. We do not intend this as a criticism of the previous literature as we do not believe that the dogmatism assumed in quasi-Bayesian models is intended as an extreme assumption. Moreover, a quasi-Bayesian agent will not learn his model of the world is incorrect if the agent has no reason to attend to the erroneous aspects of his model's predictions (see Gagnon-Bartsch, Rabin, and Schwartzstein (2018) for a formal treatment). Similarly, there may be no reason for Saki to realize her beliefs are perpetually vacillating if she has no reason to attend to that fact.

Previous research has shown that economic actors seem to believe in autocorrelation between events, which can be interpreted as a quasi-Bayesian model if no autocorrelation is actually present. Gilovich, Vallone, and Tversky (1985) document that most basketball fans believe player performance is autocorrelated within a game and refer to this autocorrelation as a "hot hand."[16] For example, their survey evidence shows 91% of fans believe that "a player has a better chance of making a shot after having just made his last two or three shots than he does after having just missed his last two or three shots." Camerer (1989) observed in a betting market that bettors act

[15] Examples of papers that attempt to provide a general model that be integrated into different applications include Rabin (2002) and Rabin and Vayanos (2010) on the gambler's and hot-hand fallacies, Benjamin, Rabin, and Raymond (2016) on the non-belief in the law of large numbers, Esponda and Pouzo (2016) on strategic interactions with mis-specified models, and Spiegler (2016) on biases in causal reasoning. Confirmation bias, interpreting signals so that they conform with one's prior beliefs, has been modeled by Rabin and Schrag (1999) and Fryer, Harms, and Jackson (2018). Models of coarse or categorical thinking include Mullainathan (2000), Fryer and Jackson (2008), Jehiel (2005), Jehiel and Koessler (2008), Mullainathan, Schwartzstein, and Shleifer (2008), and Eyster and Piccione (2013). Inspired by effects such as the "Winner's Curse," there are a number of papers that model the failure to make inferences from the actions of others (Eyster and Rabin (2005); Esponda (2008); Madarász (2012)), and these models have been repeatedly applied in social learning settings (DeMarzo, Vayanos, and Zwiebel (2003); Eyster and Rabin (2010, 2014); Bohren (2016); Gagnon-Bartsch and Rabin (2017)). Models that assume false beliefs about others' strategic reasoning or information include Camerer, Ho, and Chong (2004) and Crawford and Iriberri (2007). Misspecified models have also been considered in specific applications, such as firms learning about demand (Kirman (1975); Nyarko (1991)) as well as macroeconomic forecasting (Sargent (1993); Evans and Honkapohja (2001)). Loewenstein, O'Donoghue, and Rabin (2003) model projection bias, which assumes one can have mis-specified beliefs about future preferences, which is thematically related to our model's assumption that Saki has mis-specified beliefs about her future updating. (This footnote is adapted from Gagnon-Bartsch, Rabin, and Schwartzstein (2018)).

[16] Gilovich, Vallone, and Tversky (1985) argued that there was no autocorrelation and characterized this belief as a fallacy. Miller and Sanjurjo (2018) argue that autocorrelation actually is present. We do not intend to take a stand on whether autocorrelation of this form exists, we merely propose a theory for why individuals might falsely believe it exists if it does not.

as if each team's performance is autocorrelated across games. This belief in autocorrelation is not limited to sporting events. Greenwood and Shleifer (2014) argue that investors seem to believe that market returns are positively autocorrelated, but actual market returns are negatively correlated with the investors' expectations. Our first goal is to highlight that much of the existing stylized evidence could be generated by BRN, and one can test between BRN and a belief in a hot hand using the longer run implications of the theories.

Although the primary inspiration for this material is economic forecasting, the basic intuition for how BRN could generate an *apparent* belief in autocorrelation is seen straightforwardly in coin-flipping settings. Let h_τ denote a coin flip of heads in period τ, t_τ denote a flip of tails in period τ, and $f_\tau \in \{h, t\}$ denote a generic flip in period τ. When the period is not relevant, we drop the subscript. Θ denotes the set of $N > 1$ hypotheses the agent entertains about the bias of the coin.[17] A hypothesis is $\theta \in \Theta$ where $\theta \equiv p(h|\theta)$ where there is no autocorrelation. After observing a flip, Saki's posteriors are

$$p_{\alpha=0}(\theta_i|h) = \frac{\theta_i}{\sum_{\theta_j\in\Theta}\theta_j} \text{ and } p_{\alpha=0}(\theta_i|t) = \frac{1-\theta_i}{\sum_{\theta_j\in\Theta}1-\theta_j}.$$

Her prediction for the next flip is

$$p_{\alpha=0}(h|h) = \frac{\sum_{\theta_j\in\Theta}(\theta_j)^2}{\sum_{\theta_j\in\Theta}\theta_j} \text{ and } p_{\alpha=0}(h|t) = \frac{\sum_{\theta_j\in\Theta}\theta_j(1-\theta_j)}{\sum_{\theta_j\in\Theta}1-\theta_j}. \tag{9}$$

So, for instance, if $\Theta = \{0.75, 0.5, 0.25\}$, then Saki's prediction will be $p_{\alpha=0}(h_{\tau+1}|h_\tau) = \frac{7}{12}$ and $p_{\alpha=0}(h_{\tau+1}|t_\tau) = \frac{5}{12}$.

We now provide a formal definition of prediction momentum. Let $\mathbb{E}_{\tau,\alpha}$ denote Saki's expectations in period τ, where $\mathbb{E}_{\tau,\alpha=1}$ represents the expectations of Tommy or a Bayesian observer.

Definition 1 *Beliefs exhibit* ***Prediction Momentum*** *if for $\varepsilon > 0$ and $\tau \to \infty$ we have*

$$\mathbb{E}_{\tau,\alpha=1}[p_\alpha(h_{\tau+1}|h_\tau, f_{\tau-1}..., f_1)|h_\tau, \theta^*] - \mathbb{E}_{\tau,\alpha=1}[p_\alpha(h_{\tau+1}|t_\tau, f_{\tau-1}..., f_1)|t_\tau, \theta^*] > \epsilon.$$

Saki's beliefs exhibit more prediction momentum than θ^* if as $\tau \to \infty$

$$\mathbb{E}_{\tau,\alpha=1}[p_\alpha(h_{\tau+1}|h_\tau, f_{\tau-1}..., f_1)|h_\tau, \theta^*] - \mathbb{E}_{\tau,\alpha=1}[p_\alpha(h_{\tau+1}|t_\tau, f_{\tau-1}..., f_1)|t_\tau, \theta^*] > p(h_{\tau+1}|h_\tau, \theta^*) - p(h_{\tau+1}|t_\tau, \theta^*)$$

Note that the definition describes prediction momentum in terms of a Bayesian outside observer (i.e., $\mathbb{E}_{\tau,\alpha=1}$) about Saki's beliefs in period τ conditional on the *observer* only having seen f_τ (although the outside observer also knows θ^*), whereas Saki's beliefs condition on all past information (i.e., $p_\alpha(h_{\tau+1}|t_\tau, f_{\tau-1}..., f_1)$). Note that, in the limit as $\tau \to \infty$, Tommy would learn

[17] If Saki and Tommy only entertained one theory (i.e., $\Theta = \{\theta^*\}$), then there is nothing to learn and their predictions would be identical.

θ^* and not exhibit prediction momentum in our i.i.d. model. Finally, our definition of more prediction momentum than θ^* defines prediction momentum relative to the true autocorrelation of the underlying signal-generating process, which is the degree of prediction momentum Tommy expects asymptotically.

One can show that Saki's beliefs exhibit prediction momentum in this coin-flipping setting.[18]

Proposition 4 *Saki's beliefs exhibit prediction momentum and prediction momentum relative to θ^* if Θ has at least two elements.*

Mapping this into the basketball setting of Gilovich, Vallone, and Tversky (1985), suppose that the true probability that a team will win (h) or lose (t) each game is i.i.d. and given by the unknown parameter $\theta^* \in \Theta$. The fan is certain that the outcomes of any two games are i.i.d., but does not know θ^*. Equation 9 implies that if the fan is Saki with $\alpha = 0$, then the fan believes it is most likely that the maximal element of Θ is true (i.e., the team has high ability) after the team wins a game. Symmetrically, Saki believes it is most likely that the minimal element of Θ is true (i.e., the team has low ability) after the team loses a game. In other words, the pattern of the fan's predictions exhibit prediction momentum in that the probability the fan places on the team winning the next game is higher following previous wins than following previous losses. In addition (and unlike belief in a hot hand), Saki thinks that the team will be more successful in the long run after observing a win because Saki is optimistic about the team's permanent underlying ability. To differentiate BRN and belief in a hot hand, one would need to measure the beliefs of fans about long-run statistics (i.e., the performance of the team at the end of the season). In contrast, any prediction momentum exhibited by Tommy will fade as he learns θ^*.

Of what (should be) more interest than sports, various patterns of investor mispredictions have been the focus of recent research (e.g., extrapolative expectations over market returns: Beshears et al. (2013); Greenwood and Shleifer (2014); Landier, Ma, and Thesmar (2017)), and these mispredictions (to the extent they do not fade with learning) represent a form of prediction momentum. These findings are sometimes framed as individuals placing too much weight on recent experience (e.g., Beshears et al. 2013), but Proposition 4 shows that they can also be generated by BRN and a correct interpretation of the informativeness of the signal.[19] If BRN is driving these

[18] If $\alpha > 0$, the belief Saki holds will depend on the full sequence of flips so far. Suppose Saki observes a sequence of flips $(f_1, ..., f_\tau)$. Then we can define the "headsness" of the sequence as $F_h = \sum_{i=0}^{\tau-1} \alpha^{\tau-i}\mathbf{1}\{f_{i+1} = h\}$, which describes the degree to which recent flips have been discounted by α. Saki assigns a likelihood ratio to two possible hypotheses $\theta, \widetilde{\theta} \in \Theta$ equal to $\frac{p_\alpha(\theta|(f_1,...,f_\tau))}{p_\alpha(\widetilde{\theta}|(f_1,...,f_\tau))} = \left(\frac{\theta}{\widetilde{\theta}}\right)^{F_h}\left(\frac{1-\theta}{1-\widetilde{\theta}}\right)^{1-F_h}$, and Saki will assign a higher probability of the next flip being heads if she has recently seen heads rather than tails, which is the essence of prediction momentum. If Θ is symmetric around .5, and $\alpha \leq \frac{1}{2}$, Saki will always think heads is more likely than tails if the last two flips have been heads. For such symmetric Θ and all $\alpha < 1$, there will exist some number of recent heads such that Saki predicts heads is more likely than tails no matter what came before. In contrast, Tommy's beliefs equally weight all of the prior flips in making inferences about θ^*, which implies his beliefs do not exhibit momentum.

[19] Stated more strongly (and speculatively), extrapolative expectations might even result from BRN and an *underweighted* signal. Sections 2.1 and 8.3.1 argue underweighting signals is typical in (some) canonical updating experiments.

$(f_\tau, f_{\tau-1})$	**Pos. Auto.**	**Fair**	**Neg. Auto**
(h,h)	0.60	0.33	0.07
(h,t)	0.07	0.33	0.60
(h,h)	0.07	0.33	0.60
(t,t)	0.60	0.33	0.07

Table 2: **Saki's Conditional Beliefs**

extrapolative expectations (and not a belief in autocorrelated returns), then investors expect significantly improved returns over the long-run after observing a few quarters of high returns, which would then have a significant impact on the investors' view of the net present value of market returns.[20] On the other hand, if investors believe that market returns are autocorrelated, then investors' beliefs about the market's long-run performance will be changed only slightly after a few quarters of high returns, and the effect on the perceived net present value of market returns will be muted accordingly. Again, one must ellicit beliefs about long-run market performance to tease out the differences between BRN and short-run autocorrelation.

Now we extend our analysis to the case where Saki entertains hypotheses that the coin is autocorrelated. Hypotheses have the form $\theta = (r, s) \in \Theta$, $r \neq 1 - s$, where $p(h|h, \theta) = r$ and $p(t|t, \theta) = s$. Since autocorrelation is possible, Saki uses the realization of $f_{\tau-1}$ to interpret f_τ, meaning that making a prediction about $f_{\tau+1}$ requires conditioning on $f_{\tau-1}$ and f_τ even if $\alpha = 0$. We can describe Saki's beliefs after seeing two successive flips as

$$p_{\alpha=0}(\theta_i = (r_i, s_i)|h_\tau, h_{\tau-1}) = \frac{r_i}{\sum_{\theta_j=(r_j,s_j)\in\Theta} r_j} \tag{10}$$

$$p_{\alpha=0}(\theta_i = (r_i, s_i)|t_\tau, h_{\tau-1}) = \frac{1 - r_i}{\sum_{\theta_j=(r_j,s_j)\in\Theta} 1 - r_j} \tag{11}$$

$$p_{\alpha=0}(\theta_i = (r_i, s_i)|h_\tau, t_{\tau-1}) = \frac{1 - s_i}{\sum_{\theta_j=(r_j,s_j)\in\Theta} 1 - s_j} \tag{12}$$

$$p_{\alpha=0}(\theta_i = (r_i, s_i)|t_\tau, t_{\tau-1}) = \frac{s_i}{\sum_{\theta_j=(r_j,s_j)\in\Theta} s_j} \tag{13}$$

For example, after observing (h, h), Saki places the highest probability on the hypothesis with the highest autocorrelation of heads. After observing (t, h), Saki places the highest probability on the hypothesis with the highest probability of reversal.

When the theories are $\Theta = \{(0.9, 0.9), (0.5, 0.5), (0.1, 0.1)\}$, which correspond to positively autocorrelated, independent, and negatively autocorrelated coins, Saki's beliefs following any pair of successive flips are described in Table 2. Saki places the highest probability on whichever coin

[20] These shifts in beliefs may not be reflected in prices if (for example) arbitrageurs can trade against the optimism of the typical investor.

best explains her last two observations. If the coin is actually uncorrelated, Saki never believes that the true model of the world is most likely. Instead, Saki's beliefs oscillate between believing the coin is positively or negatively autocorrelated based on which better reflects the most recent flips.

Tommy would, under weak conditions, learn θ^* and not exhibit prediction momentum relative to θ^*. However, BRN presents several interesting issues that do not arise in a Bayesian model. For example, Saki's prediction $p_{\alpha=0}(h_{\tau+1}|, h_\tau, h_{\tau-1})$ depends on the set of theories in Θ (rather than their probabilities). The hypothesis dependence of Saki's beliefs implies that even if she observes $(h_\tau, h_{\tau-1})$ and entertains a theory θ_i that allows for positive autocorrelation (i.e., $r_i > 0.5$), if most $\theta_j \in \Theta$ do not (i.e., $r_j \leq 0.5$), then Saki's predictions will not exhibit prediction momentum.

Therefore, we focus on the case where Θ has a symmetric structure in the sense that if $\theta_i = (r_i, s_i) \in \Theta$, then there exists $\theta_j = (1 - r_i, 1 - s_i) \in \Theta$ and $p(\theta_i) = p(\theta_j)$. θ_i and θ_j are symmetric theories in that θ_i places the same probability on $h_{\tau+1}$ given h_τ as θ_j places on $t_{\tau+1}$ given h_τ (and a similar relation holds conditional on t_τ). The symmetric structure implies that ex ante neither Tommy nor Saki expects the coin to exhibit autocorrelation, which serves as a useful benchmark.

An additional complication of our analysis of Saki's beliefs is that since the observer's expectation conditions only on the most recent flip, we must account for the relative probability of having observed $h_{\tau-1}$ or $t_{\tau-1}$ conditional on having observed h_τ. These probabilities are determined by the true theory of the world, θ^*, and can be computed by treating the evolution of Saki's beliefs as a Markov process and computing the long-run distribution over the different possible realizations of $(f_\tau, f_{\tau-1})$.

Proposition 5 *Suppose the true theory of the world is $\theta^* = (r^*, s^*)$. From the perspective of a Bayesian observer, Saki's expected predictions are:*

$$\mathbb{E}_{\tau,\alpha=1}[p_{\alpha=0}(h_{\tau+1}|h_\tau, f_{\tau-1})|h_\tau, \theta^*] = \sum_{\theta_i=(r_i,s_i)} \frac{r_i^2}{\sum_{\theta_j=(r_j,s_j)} r_j} r^* + \frac{r_i(1-s_i)}{\sum_{\theta_i=(r_i,s_i)} 1 - s_j}(1 - r^*)$$

$$\mathbb{E}_{\tau,\alpha=1}[p_{\alpha=0}(h_{\tau+1}|t_\tau, f_{\tau-1})|t_\tau, \theta^*] = \sum_{\theta_i=(r_i,s_i)} \frac{r_i(1-r_i)}{\sum_{\theta_j=(r_j,s_j)}(1 - r_j)}(1 - s^*) + \frac{r_i s_i}{\sum_{\theta_i=(r_i,s_i)} s_j} s^*$$

Saki's beliefs exhibit more prediction momentum than θ^* if

$$\mathbb{E}_{\tau,\alpha=1}[p_{\alpha=0}(h_{\tau+1}|h_\tau, f_{\tau-1})|h_\tau, \theta^*] - \mathbb{E}_{\tau,\alpha=1}[p_{\alpha=0}(h_{\tau+1}|t_\tau, f_{\tau-1})|t_\tau, \theta^*] > r^* + s^* - 1$$

Finally, the prediction momentum caused by BRN in a canonical normal-normal updating model naturally gives rise to forecasts reminiscent of adaptive-expectations models, another case where BRN can generate belief dynamics similar to a classical non-Bayesian model. While the adaptive-expectations model has often provided a useful description of belief dynamics (for a review, see Fuster, Laibson, and Mendel (2010)), the classical models could not say much about how individuals respond to endogenous structural changes in the economy since these models typically violated the rational expectations hypothesis. Saki's beliefs exhibit patterns similar to adaptive expectations when the economy's structure is stable, but Saki is responsive to endogenous changes in the economy because she understands the impact of these changes on what she observes.

Suppose s_τ is a time series of economic interest with true stochastic process $s_\tau = \theta + \varepsilon_\tau$, where ε_τ is normally distributed with mean 0 and precision ρ_ε. The agent does not know θ and infers it from the observed time series s_τ, starting from a prior for θ that is normally distributed with mean μ_0 and precision ρ_0. Saki's expectations take a particularly simple functional form:

$$\mathbb{E}_{\tau,\alpha}\theta = \frac{\alpha^\tau \rho_0 \mu_0 + \sum_{i=1}^{\tau} \alpha^{\tau-i} \rho_\varepsilon s_i}{\alpha^\tau \rho_0 + \sum_{i=1}^{\tau} \alpha^{t-i} \rho_\varepsilon}.$$

Rewriting this equation recursively and taking the limit as $\tau \to \infty$, we find

$$\mathbb{E}_{\tau,\alpha}\theta - \mathbb{E}_{\tau-1,\alpha}\theta \to (1-\alpha)(s_t - \mathbb{E}_{\tau-1,\alpha}\theta), \tag{14}$$

Our framework does not suffer from some of the key drawbacks of the early adaptive-expectations models. Prior models specified a fixed rule for expectations formation—i.e., they *assumed* adaptive expectations—and were therefore not sensitive to changes to the true signal-generating process. These models were subject to the "Lucas critique," namely that agents' forecasting rules should be (but are not in these classic models) sensitive to changes in policy that influence the data generating process. In contrast, ours is a model of biased *updating* relative to Bayesian updating. It therefore provides a framework for mapping the data-generating process into a rule for expectations formation, and this mapping takes policy changes into account.

As a simple example, suppose that on date $\widetilde{\tau}$, the data-generating process changes from $s_\tau = \theta + \varepsilon_\tau$ to $s_\tau = \gamma + \theta + \varepsilon_\tau$, where $\gamma \neq 0$ is known. Written recursively, Saki's updating rule for $\tau > \widetilde{\tau}$ would be

$$\mathbb{E}_\tau\theta - \mathbb{E}_{\tau-1}\theta \to (1-\alpha)(s_\tau - \gamma - \widehat{\theta}_{\tau-1}). \tag{15}$$

While the agent's expectation of θ would continue to overweight recent signals both before and after date $\widetilde{\tau}$, the forecast rule itself would immediately adjust to appropriately account for the change in the data-generating process. We suspect that most practitioners would have identified equation 15 as the natural change to equation 14 to account for such a simple change in policy. This analysis provides a well founded reason for what might otherwise be an ad hoc choice. Our theory may also be able to identify the most reasonable way of accounting for policy changes when there are multiple ad hoc ways for accounting for the change.

5 Learning Traps

Consider situations where Saki is collecting costly, informative signals in advance of a decision. Suppose $\theta \in \{U, D\}$ and each period Saki can either take an action $a \in \{U, D\}$ or pay for a signal with cost c. If Saki takes an action, her payoff (not including signal costs) is $u(\theta, a) = 1$ if $\theta = a$, and 0 otherwise. Saki's total payoff after gathering N signals and choosing a is $u(a, \theta) - Nc$. We argue that if Saki naively believes she will use future information as a Bayesian would, the boundedness of Saki's beliefs puts her at risk of being caught in learning traps.

Standard results on optimal sequential sampling (Wald (1947)) show that the optimal plan for Tommy is defined in terms of two cutoffs $\rho_U, \rho_D \in (0, 1)$. Tommy chooses $a = u$ in period t if $p_{\alpha=1}(U|s_1, ..., s_t) \geq \rho_U$ and $a = d$ if $p_{\alpha=1}(U|s_1, ..., s_t) \leq \rho_D$. The solution has two crucial properties. First, as c drops, ρ_U increases and ρ_D decreases, meaning that Tommy has to be more confident to make a decision. Second, Tommy's posteriors eventually enter one of these two regions almost surely. A lower value of c is obviously good for Tommy—he spends less per signal, gathers more information, and is more confident in his choice.

Having a lower value of c is not obviously better for Saki. If Saki is naive about her future updating and believes she will process information using Bayes' rule, she adopts the same sequential sampling plan as Tommy. As noted in Section 3, it is possible that Saki's beliefs, $p_\alpha(U|s_1, ..., s_t)$, are bounded within $[\rho_D, \rho_U]$ regardless of how many signals are observed. If this happens, Saki will pay for signals each period without ever making a decision. In other words, Saki is caught in a learning trap.[21]

In contrast, if Saki realized that she would process future information using BRN updating and be caught in a learning trap, then one of two predictions seems natural. First, she might simply refuse to gather signals knowing that they would never be informative enough for her to make a choice. Second, she could choose to collect several signals simultaneously (if possible), the basic idea being that she would in effect get a single very informative signal that might be sufficient to inform her choices. In either case, the behavior of a prospective Bayesian and a prospective BRN version of Saki are qualitatively quite different.

In many settings, the sequential sampling problem refers to signals that are external to the agent and are gathered through an effortful process. However, there are other situations where it is plausible to think of the signal generation process as being conducted through introspection. When considering which of two options to choose, the decider needs to learn about the features of the different choices and also introspect about her own preferences over these features and the trade-offs entailed in choosing one option over the other. For example, a potential apartment renter must introspect about her own preferences over trade-offs between apartment size, location,

[21] In addition, if the range of Saki's beliefs allow $p_\alpha(U|s_1, ..., s_t) > \rho_U$ and not $p_\alpha(U|s_1, ..., s_t) < \rho_D$, then Saki will almost surely choose $a = U$ regardless of the true hypothesis. A symmetric result holds if Saki's beliefs can drop below ρ_D and can't exceed ρ_U.

and rental price offered by different housing options even after all of the "hard facts" about the apartments are known.

If the decision maker is Tommy, he will introspect until the benefits of further introspection are outweighed by the costs. If the introspection costs are small, then Tommy will be able to eliminate almost all of the uncertainty and will be confident about his preferences and the resulting choice. On the other hand, if Saki is naive about her future updating, she will never be able to become confident about her own tastes and will be indecisive. Moreover, if the signals generated by introspection are weak, then it could be that the moderation effect causes introspection to erode any ex ante confidence she had in the merits of one choice over the other.

6 Persuasion

We now show that base-rate neglect has striking implications in a simple persuasion environment. We assume that a strategic persuader can choose to either reveal or withhold a signal about the hypothesis from an audience, but the persuader cannot reveal a false signal or credibly reveal that a signal has not been received. A key difference between the analysis in this section and in earlier sections is that instead of receiving exogenous signals, here the receiver—although she may be Saki—fully understands that any signal she observes is filtered by the strategy of the persuader. That is, in equilibrium, the audience takes into account that the persuader chooses to reveal signals selectively.

The hypotheses are that $\theta = G$ (good for the persuader) or $\theta = B$ (bad). Neither the persuader nor the audience knows the correct hypothesis. The persuader gets a private signal $s \in \{g, b, \varnothing\}$, where g is more likely than b in hypothesis G, b is more likely than g in hypothesis B, and $\varnothing$ is uninformative. In particular, when $\theta = G$ we have $p(g|G) = m$, $p(b|G) = n$, and $p(\varnothing|G) = r = 1 - m - n$ where $m > n > 0$. When $\theta = B$ we have $p(g|B) = n$, $p(b|B) = m$, and $p(\varnothing|B) = r = 1 - m - n$. The symmetry of the probabilities across hypotheses does not matter for the results below, but it simplifies exposition. The persuader decides whether or not to reveal s to the audience. Signals g and b are verifiable (thus are *not* cheap talk), but $\varnothing$ is not verifiable, so the persuader can choose to conceal a signal by claiming $\varnothing$. The persuader's strategy is $\sigma(s)$, and $\sigma(s)$ equals either $\varnothing$ or the observed signal s.

In many applications, a persuader would attempt to persuade the audience to take an action such as buying a product. To simplify the analysis, however, we assume that the audience's only role is to update beliefs (and not to take any action). We model the persuader's utility U as depending directly on the audience's beliefs: $U = p_\alpha(G|\sigma(s))$, where $p_\alpha(G|\sigma(s))$ is the probability the audience puts on $\theta = G$ at the end of the game.

Intrinsic to BRN, there is a distinction for Saki between receiving an uninformative signal and not receiving a signal at all. Therefore, a key question is: how does Saki react when the persuader does not reveal a signal? In an equilibrium in which observing $\varnothing$ is informative, Saki updates using that information. But in an equilibrium in which $\varnothing$ is completely uninformative—$\sigma(s) = \varnothing$ for all signals s—Saki does not treat the observation of $\varnothing$ as a signal and does not update. We view this model of Saki as capturing BRN while retaining the standard assumption of strategic sophistication.

In a *silent equilibrium* the persuader adopts the strategy $\sigma(g) = \sigma(b) = \varnothing$, observing $\varnothing$ is uninformative, and so Saki does not update in equilibrium. In contrast, in a *bragging equilibrium*, the persuader adopts the strategy of always releasing g and withholding b: $\sigma(g) = g$ and $\sigma(b) = \varnothing$. In this equilibrium, observing $\varnothing$ is informative about the hypothesis, and so Saki updates her beliefs as follows:

$$\begin{aligned}\frac{p_\alpha(G|\sigma(s) = \varnothing)}{1 - p_\alpha(G|\sigma(s) = \varnothing)} &= \frac{p(s = \varnothing|\theta = G) + p(s = b|\theta = G)}{p(s = \varnothing|\theta = B) + p(s = b|\theta = B)} \left(\frac{p(G)}{1 - p(G)}\right)^\alpha \\ &= \frac{r + n}{r + m} \left(\frac{p(G)}{1 - p(G)}\right)^\alpha .\end{aligned}$$

The fact that $\frac{r+n}{r+m} < 1$ implies that when Saki observes $\varnothing$, she treats it as a signal that $\theta = B$.

When the audience is Tommy, the bragging equilibrium is the only equilibrium.

Proposition 6 *If $\alpha = 1$ (i.e., the audience is Tommy), then the bragging equilibrium is the unique sequential equilibrium.*

In contrast, when Saki is the audience, a silent equilibrium can occur. If Saki places sufficient prior probability on $\theta = G$, then the persuader will find that releasing any signal, even a g signal, will reduce the posterior probability Saki places on $\theta = G$. This is a consequence of the moderation effect, and the possibility of a silent equilibrium is a distinctive prediction of BRN.

Proposition 7 *A silent sequential equilibrium exists if and only if $\alpha < 1$ and $p(G) \geq \frac{m^{1/(1-\alpha)}}{m^{1/(1-\alpha)} + n^{1/(1-\alpha)}}$*

Note that as the informativeness of a signal (i.e., $\frac{m}{n}$) increases, the threshold prior belief approaches 1. Consequently, for any prior belief Saki might have, if $\frac{m}{n}$ is large enough, the silent equilibrium disappears.

In an earlier draft, we explored an extension of this static model to the case of two competing persuaders, one supporting $\theta = G$ (i.e., his payoff equals $p_\alpha(G|\sigma(s))$) and a second supporting $\theta = B$ (i.e., his payoff equals $1 - p_\alpha(G|\sigma(s))$). We assume that neither persuader has private information about the correct hypothesis, both observe the same signal, and both have the option as to whether to release this signal. In this case, there is no equilibrium in which both persuaders are silent. To see this, recall that silence can occur because the persuader does not want to moderate Saki's beliefs. However, in the presence of competing persuaders, if one persuader would prefer not to moderate Saki's beliefs, the other persuader has a strict incentive to do so. Therefore every signal will be revealed by one of the competing persuaders.

Our persuasion results were cast in a static model to highlight the basic effects of BRN and to avoid the unrelated assumptions needed to flesh out a dynamic model. In a dynamic context, the bounded confidence of Saki's beliefs suggests that if she is continually talked to by the persuader, her beliefs will never converge. In effect, questions are always up for discussion and nothing ever gets settled. If there is only one persuader, he has an incentive to communicate with Saki until her beliefs are as favorable as her bounded confidence affords, and then to cease communication. If there are two risk-neutral competing persuaders, we conjecture that at least one persuader will reveal each signal and Saki's beliefs will never stop moving.

7 Reputation

Reputation is a key theme in economics because, among other things, it provides a mechanism for an individual, government, or firm to be perceived as truthworthy by another party that will only have a one-shot interaction with the agent. Reputation has been captured by modeling a long-run player (LRP) who interacts with an infinite sequence of short-run players (SRPs). The outcome of each interaction is a random variable that is a influenced by whether the LRP undertakes costly investment. The LRP is either a "strategic type"($\theta = S$) that maximizes expected discounted profit or a "committed type"($\theta = C$) that always invests. The outcomes of previous interactions are public information and allow the SRPs to learn about the LRP's type. The LRP's reputation is reflected in the likelihood that SRPs place on the LRP choosing to invest, and the SRPs are more willing to interact with LRPs that have a good reputation. This last fact means strategic LRPs may invest to prevent the SRPs from learning that $\theta = S$. For example, the SRPs may be consumers, each of whom is sequentially deciding whether to buy a car from the LRP, Toyota. Consumers are more willing to buy if Toyota has a reputation for producing high-quality cars, but producing high-quality cars requires that Toyota make costly, unobservable (to consumers) R&D investments.

Fudenberg and Levine (1992) showed that if the LRP is sufficiently patient and the SRPs are Bayesian, then there exists an equilibrium where LRPs of type $\theta = S$ earn a payoff that is arbitrarily close to the (higher) payoff that LRPs of type $\theta = C$ earn. This implies that (at least for a long time) the LRP succeeds in maintaining a good reputation (i.e., the SRPs expect the LRP to invest), but the main theorem of Fudenberg and Levine (1992) is about payoffs and remains silent regarding the dynamic path of beliefs. Cripps, Mailath, and Samuelson (2004) fill this gap by showing that the strategic type of LRP is found out in the long run, and the SRPs eventually stop buying. Intuitively, due to the imperfect monitoring, any single outcome has a small effect on the SRPs' beliefs (e.g., a low-quality car can be attributed to bad luck). Even though the LRP's shirking is rare, the Tommys will eventually infer—from the accumulation of weak evidence—that the LRP is the strategic type. At that point, play will collapse to the stage-game Nash equilibrium in which the LRP shirks and the SRPs do not buy. In short, while the LRP initially maintains a good reputation, it is exploited and is ultimately transient.

We assume throughout this section that Saki is sophisticated in that she has correct beliefs regarding the LRP's strategy and uses this information when forming her beliefs. As the Sakis update based on observed outcomes in each period in equilibrium, old information is increasingly underused and thus does not accumulate sufficiently for the SRPs to become confident that the LRP is the strategic type. The limited impact of past outcomes implies that the LRPs' past good and bad behavior has limited effect on the current SRPs. This provides the LRP with the opportunity to build and exploit his reputation repeatedly over time, resulting in a fluctuating long-run reputation.[22]

While the possibility of maintaining a long-run reputation follows from the ergodicity of the SRPs' beliefs (and the concomitant failure of these beliefs to converge to the truth), the SRPs' base-rate neglect further implies that their belief about the likelihood that the LRP is the committed type is bounded in the long run. If the LRP must have a reputation that is "too good"—i.e., above the long-run upper bound on Saki's belief that $\theta = C$—in order for a Saki SRP to trust him enough to buy, then the SRPs will never buy and the LRP has no incentive to try to maintain a reputation. Symmetrically, if a Saki SRP would trust the LRP enough to buy even at the lower bound of her belief that $\theta = C$, then Saki will always buy and the LRP again has no incentive to maintain a reputation.

7.1 Model Setup and Tommy Results

We consider a LRP with discount factor $\delta \in (0, 1)$ who plays a simultaneous move stage game against a sequence of SRPs. The periods of the game are indexed $t = 1, ..., T \leq \infty$, and a different SRP plays in every period. The strategic LRP chooses whether to take the action "invest" ($a_t^{\text{LRP}} = 1$) at cost $c > 0$ or the action "shirk" ($a_t^{\text{LRP}} = 0$) at a cost of 0; the committed type mechanically chooses to invest every period. The prior probability that the LRP is committed is $\phi_0 = p(\theta = C) \in (0, 1)$, but we focus on what happens when the LRP is in fact strategic. Output quality is i.i.d. and drawn from distribution $y_t \sim f(\cdot|a_t^{\text{LRP}})$, which depends on the LRP's period-t action. We denote the period-t SRP's belief by $\phi_t = p_\alpha(\theta = C|y_1, ..., y_{t-1})$.

Simultaneous with the LRP's action and after having observed $y_1, ..., y_{t-1}$, the period-t SRP chooses whether to take action "buy" ($a_t^{\text{SRP}} = 1$), in which case the SRP's payoff is $u^{\text{SRP}}\left(y_t, a_t^{\text{SRP}} = 1\right)$, which is an increasing and bounded function of the quality of output y_t, or the action "refuse" ($a_t^{\text{SRP}} = 0$), which yields payoff $u^{\text{SRP}}\left(y_t, a_t^{\text{SRP}} = 0\right) \equiv 0$. If the LRP shirks, his stage-game payoff

[22] Liu and Skrzypacz (2011) develop a reputation model in which the SRPs (who are Tommys) observe only a finite set of records regarding the past behavior of the LRP. Unlike in our model, the equilibrium of their model features predictable cycles: after the LRP exploits his good reputation, the current generation of SRPs perfectly identify that he is the strategic type, but then the SRPs "collude" with the LRP in the process of rebuilding his reputation (after which the LRP exploits the next generation of SRPs).

is 1 or 0, depending on whether the SRP buys or refuses; if the LRP invests, his stage-game payoff is $1 - c$ or $-c$, respectively, $c \in (0, 1)$. We assume that

$$E\left[u^{\text{SRP}}\left(y_t, a_t^{\text{SRP}} = 1\right) \middle| a_t^{\text{LRP}} = 1\right] > 0 > E\left[u^{\text{SRP}}\left(y_t, a_t^{\text{SRP}} = 1\right) \middle| a_t^{\text{LRP}} = 0\right],$$

which implies that buying the product is optimal if and only if the SRP is sufficiently confident that the LRP will invest. The LRP's lifetime utility is $(1-\delta)\sum_{t=0}^{\infty}\delta^t\left[a_t^{\text{SRP}} - c1\left\{a_t^{\text{LRP}} = 1\right\}\right]$.

While it is standard to assume that y_t has finite support, we instead assume a continuous support because doing so greatly facilitates proving results about ergodicity. Let

$$l_\beta(y_t) \equiv \ln\frac{f(y_t|a_t^{\text{LRP}} = 1)}{\beta f(y_t|a_t^{\text{LRP}} = 1) + (1-\beta)\, f(y_t|a_t^{\text{LRP}} = 0)} \tag{16}$$

denote the informativeness of quality level y_t when the strategic type plays a mixed action with probability β of investing. The numerator of equation 16 is the probability that y_t is observed when the LRP is committed to $a_t = 1$, and the denominator is the same probability when a strategic agent chooses $a_t = 1$ with probability β.

When the SRPs are Sakis, her beliefs about the LRP's type will fluctuate without reaching certainty even in the long run. To facilitate proving a strong version of this result—ergodicity of the SRPs' beliefs—we make the following technical assumption:

Assumption (A) *For any $\beta \in [0, 1)$, the distribution of $l_\beta(y_t)$ has full support on $\mathbb{R}$.*

Assumption (A) implies that as long as the LRP shirks with probability less than 1, it is possible that an arbitrarily informative realization of y_t could occur. This assumption ensures that as long as the LRP invests with positive probability, it is possible that the SRPs' beliefs could move from any belief to any other belief with a single realization of y_t.[23]

We also make the following regularity assumptions on l_β:

Assumption 1 *For any $\beta \in [0, 1]$ we have:*[24]

1. $E[\|l_\beta\|] < \infty$ *and* $E\left[l_\beta^2\right] < \infty$.
2. *For all x and all β we have* $E[l_\beta \mid l_\beta > x] < \infty$

The first point of Assumption 1 implies that the first and second moments of $l_\beta(y_t)$ exist, which is relatively innocuous. The second point implies that the expected value of large deviations is

[23] As an alternative to Assumption (A), we could instead assume that the (endogenous) belief process to be aperiodic. This alternative would similarly facilitate proving ergodicity but would allow for beliefs to have bounded support. We use Assumption (A) because it is a condition on exogenous variables.

[24] The assumption requires bounded moments for *any* mixing probability $\beta \in [0, 1]$, and hence it can can be checked independently of the LRP's strategy. While we could instead have stated the assumption as more primitive conditions on the moments of $f(y_t|a)$, we prefer the formulation in Assumption 1 because it makes the economic meaning clearer.

well-defined, which we use to argue that the LRP's belief process is well-behaved even after very rare outcomes.[25]

Since each SRP lives only for one period, the SRPs' equilibrium strategy is straightforward to characterize: the period-t SRP buys if and only if $E_t\left[u^{\text{SRP}}\left(y_t, a_t^{\text{SRP}} = 1\right)\right] > 0$. This condition is satisfied if and only if the SRP believes that the probability that the LRP invests is at least μ^* for some $\mu^* \in (0, 1)$.

7.2 Fluctuating Long-Run Reputations

We focus on Markov equilibria, in which the LRPs' strategy $\sigma(\phi_t)$ specifies his mixing probability as a function of the payoff-relevant information—namely, the current SRP's belief. The Markov equilibrium restriction is appealing in this context because the payoff-irrelevant aspect of the history refer to the beliefs and actions of different SRPs. Proposition 8 states that Saki's beliefs about the LRP's type do not converge in the long run.

Proposition 8 *Let Assumptions (A) and 1 hold. For any $\alpha < 1$ and in any equilibrium, the process ϕ_t is ergodic with non-degenerate support.*

To prove that Saki's beliefs are ergodic, we treat ϕ_t as a Harris chain and prove the process is aperiodic, admits an invariant measure, is Harris recurrent, and has uniformly bounded, finite expected return times as long as ϕ_t starts in some compact set. Aperiodicity follows from the fact that ϕ_t can jump to any point in $(0, 1)$ for a sufficiently strong signal. The existence of an invariant measure follows from the fact that Saki's beliefs will constantly drift back towards some (compact) set in $(0, 1)$ as information is neglected.

Harris recurrence is more difficult to prove since we must show that we can choose some compact set $C \subset (0, 1)$ of posterior beliefs such that ϕ_t eventually returns to C almost surely regardless of the current value of ϕ_t. We choose C so that it contains an open neighborhood of 0.5. We define $N(\phi)$ as the number of periods required for the process to drift from ϕ to within a neighborhood of 0.5 (i.e., C) if a sequence of uninformative signals are observed. Of course, Saki observes informative signals during these $N(\phi)$ periods since $\sigma(\phi_t) < 1$. We treat these accumulated signals as a "jump" that is then decayed away. For ϕ_t to remain away from 0.5, these "jumps" must always be of sufficiently large magnitude. Since even an infinite sequence of signals has in expectation a finite influence on beliefs due to the gradual neglect of past observations, we can provide bounds on the mean and variance of the effect of any sequence of signals on Saki's beliefs. Our proof establishes that because of these bounds, ϕ_t almost surely returns to C. A similar logic provides a uniform, finite bound on the expected return times to C so long as ϕ_t starts within C. Therefore ϕ_t is Harris recurrent.

[25] We use this assumption in the proof of Proposition 8, but only at a particular x. Unfortunately, the definition of this x uses a great deal of notation, so for expositional ease we make the stronger assumption here.

7.3 Always Buying or Never Buying

Since Saki's beliefs are bounded by the strength of the signals, it is possible that the LRP's action cannot affect his own reputation enough to influence the SRPs' action. That is, the SRP may always buy or never buy *regardless* of the history of previous outcomes. But in that case, the LRP has no incentive to ever invest. In such an equilibrium—in stark contrast to the classic (Tommy) result that a sufficiently patient LRP will always try to establish a good reputation at the beginning of the game—here the LRP *always shirks.*

To show this formally, we replace Assumption (A)—that output quality can be arbitrarily informative—with the assumption that the informativeness of output quality is bounded:

Assumption (A') *There is some l^* such that $l_\beta(y_t) \in [-l^*, l^*]$ for all $\beta \in [0,1]$.*

As per the logic of equation 7 from Section 3, Assumption (A') implies that the long-run beliefs of the Sakis are bounded.

$$\ln \frac{\phi_t}{1-\phi_t} \in \left[\frac{-l^*}{1-\alpha}, \frac{l^*}{1-\alpha}\right].$$

Recall that the SRP buys if and only if she believes that the probability that the LRP invests is at least μ^* for some $\mu^* \in (0,1)$.

Proposition 9 *Let Assumption (A′) hold. For any $\delta \in (0,1)$, the following hold:*

1. *If $\mu^* > 0$ is sufficiently small, then in equilibrium the SRPs always buy and the strategic LRP always shirks.*

2. *If $\mu^* < 1$ is sufficiently large and $T < \infty$, then in equilibrium the SRPs always refuse to buy and the strategic LRP always shirks.*

The first part of the proposition describes the equilibrium when μ^* is sufficiently low. In this case, the SRP is willing to buy even if a strategic LRP does not invest because the lower bound on ϕ_t is higher than μ^*. The second part of the proposition shows that the same behavior on the part of the LRP also obtains when μ^* is sufficiently high. This result is not obvious since one might imagine that in equilibrium the strategic LRP invests with sufficiently high probability that, even though the SRP is quite confident that the LRP is strategic, that it is still optimal for the SRP to buy. We exploit the finite T by using backward induction. In the final period, a strategic LRP optimally chooses not to invest. If $\ln\left(\frac{\mu^*}{1-\mu^*}\right) > \frac{l^*}{1-\alpha}$, it is impossible for the SRP to believe that it is sufficiently likely that the LRP is committed to warrant buying in period T. Therefore, the SRP will not buy in period T. In period $T-1$, the LRP knows that his action cannot influence the period-T behavior of the SRP. Therefore the LRP does not invest, but as above, this implies that the SRP will not buy in period $T-1$. By backward induction, the LRP never invests and the SRP never buys in any period. It remains an open question whether this outcome is the unique subgame-perfect equilibrium in the infinite-horizon game.

8 Base-Rate Neglect and Other Biases

In this section we discuss how BRN relates to, differs from, interacts with, and contradicts other biases. We also discuss how some of these other biases might be mis-identified in particular contexts. In the process we revisit alternative interpretations of Kahneman and Tversky (1973).

8.1 Other Renditions of Representativeness

In Kahneman and Tversky (1972b), Kahneman and Tversky (1973), and elsewhere, Kahneman and Tversky proposed that cognition is influenced by the "representativeness heuristic." Although the representativeness heuristic has been criticized for lacking a sharp definition (e.g., Gigerenzer and Hoffrage (1995)), a common informal definition is that individuals "predict the outcome that appears most representative of the evidence" (Kahneman and Tversky (1973)). Although representativeness was suggested by Kahneman and Tversky (1973) as the source of base-rate neglect, the basic intuition can be infused into economics in many different ways. As we understand its meaning as a heuristic, representativeness is a central part of the most familiar of all updating models: Bayes's rule. The heuristic turns into a bias when, one way or another, people over- or underuse particular elements embedded in the likelihood function. BRN's relationship to representativeness is indirect: downweighting a base rate induces a *relative* over-use of likelihood information.

Gennaioli and Shleifer (2010) develop a model of "local thinking" based on representativeness, in which (in the extreme case) each hypothesis is treated as equivalent to the most representative state in that hypothesis. A state is representative of a hypothesis to the extent it is more likely under the hypothesis than under the complement of the hypothesis. A local thinker uses a version of Bayes's rule where the probability of hypotheses are replaced with the probability of the states that represent them. When one models the beliefs of a local thinker in the context of a coin-flipping example, the local thinker often uses Bayes's rule, without neglecting her priors. For example, Griffin and Tversky (1992) ask subjects whether a coin has 60% bias in favor of heads or a 60% bias in favor of tails based on the outcome of 10 coin spins. Under a standard definition of a state space, each state defines the bias of the coin and the 10 outcomes of the spins. This means there is nothing to represent under the model and the local thinker is Bayesian. If one instead defines each state to be the bias of the coin and sequences of $N > 10$ coin spins of which 10 are revealed, then the most representative state for a bias in favor of heads (tails) given the 10 revealed spins is a sequence of spins with all heads (tails) for the $N - 10$ unobserved bins. When one uses these states in an extreme local thinker's version of Bayes's rule and the probabilities under each hypothesis are symmetric, the probabilities of the unobserved coin spins cancel out and you are left with Bayes rule.

8.2 Limited Memory

Models that incorporate forgetting often generate recency biases similar to those predicted by our theory of base-rate neglect. For example, both Mullainathan (2002) and Bodoh-Creed (2018) predict that the average effect of observing information will slowly dissipate as it is forgotten, since data learned long ago are more likely to be forgotten than data learned recently. We refer to a forgetful agent as Mattie.

Saki and Mattie can be differentiated based on how their beliefs evolve in settings with correlated signals. When Saki infers from a sequence of correlated signals, she fully accounts for how the previously observed signals affect the probability of observing a signal today under different hypotheses. This occurs despite the fact that Saki has neglected the information that caused her to change her beliefs when she first observed these past signals. In contrast, if a signal observed previously has been forgotten, it could not be used to interpret the meaning of a signal observed today. Mattie must either completely account for a signal observed in the past (because he has remembered it) or completely neglect a signal in the past (because the signal has been forgotten).[26] This in turn highlights that base-rate neglect is not about *forgetting* past signals, nor even partially forgetting them.

Recall the example from Section 3 wherein Saki surveys the five employees of a firm to determine whether at least three of the employees agree with the manager's strategy. Assume the probability each employee agrees with the manager is 0.5. Suppose that each employee reveals whether he or she agrees, and Saki updates her prior (and neglects past information) after she interviews each employee. Let H denote the hypothesis that at least three employees agree with the manager's strategy, and let $s_i = a$ denote that the i^{th} interview was with an employee that agreed. If Saki, with $\alpha = 0$, interviews three employees that all agree with the manager, her beliefs are

$$\begin{aligned}
p_{\alpha=0}(H|s_1 = a) &= \frac{0.6875(0.5)^\alpha}{0.6875(0.5)^\alpha + 0.3125(0.5)^\alpha} = 0.6875 \\
p_{\alpha=0}(H|s_1 = s_2 = a) &= \frac{0.875(0.6875)^\alpha}{0.875(0.6875)^\alpha + 0.125(0.3125)^\alpha} = 0.875 \\
p_{\alpha=0}(H|s_1 = s_2 = s_3 = a) &= 1
\end{aligned}$$

Saki does neglect what she learns in the prior period. For example, Tommy would believe $p(E|s_1 = s_2 = a) = 0.939$. However, once Saki observes the third employee agree with the manager, Saki is certain that H is true. This requires that Saki remember all of her previous employee interviews. In contrast, Mattie might not agree with Saki if she had forgotten that one of the employees agreed

[26] Relatedly, the difference between BRN and limited memory can also be observed in the magnitude of the effect of repetitions. Suppose that a signal repeated to Saki is fortified as per the second model of Section 9.1. If Saki has $\alpha \in (0, 1)$, then the fortified signal eliminates the effect of partial neglect. A signal repeated to Mattie has either no effect (because she never forgot it) or has the full effect of the signal (because it was forgotten and only now remembered).

with the manager's strategy.

8.3 Misinterpreting Signals

8.3.1 Underweighting of Unambiguous Signals

There are theories that predict that agents underweight the information in signals they receive. In a review and meta-analysis of the extensive literature on bookbag-and-poker-chip experiments, Benjamin (2018) concludes that, by and large, people underinfer from signals. While BRN and underweighting signals seem as if they are opposite biases—BRN involves underweighting priors and underweighting signals involves underweighting likelihood information—as discussed in Sections 2.1, we believe that both of these effects typically operate simultaneously.

We now return attention to Table 1 of Section 2. As we argued in Section 2, understanding how experimental subjects update their beliefs requires controlling for both base-rate neglect and misinterpretations of signals. Suppose that our research interest had been estimating the subjective informativeness of the data given to the subjects under the assumption that base-rates are correctly used by the subjects. The column labeled "$P_U(s|H)/P_U(s|T)$" is the subjective informativeness the signal must have had if the subjects held the belief in column "Median" and were using base-rates correctly. For comparison, the signals' true informativeness is displayed in the column labeled "$P(s|h)/P(s|T)$."

If one assumes that base-rates are used correctly, then estimates of the signals subjective informativeness will be incorrect. Suppose the signal and the prior beliefs point in different directions. If a researcher fails to control for BRN, the estimates of the signals' informativeness would be biased upwards since the neglect of the prior would be attributed to the signal strength (i.e., $P_U(d|H)/P_U(d|T) > P(d|h)/P(d|T)$). We see this in the 10% prior probability condition in Table 1. We do not see the same effect in the 30% prior probability condition, but only because the underweighting of the signals' informativeness is stronger than the effects of BRN. In all of the cases where the signals and priors point in the same direction, we find that a researcher that fails to control for BRN would come to believe that the subjects underweight their signals (which is right, but for the wrong reasons).

8.3.2 Non-belief in the Law of Large Numbers

A well-documented bias that results in signal underweighting is non-belief in the law of large numbers (NBLLN): people believe that even in very large random samples, the sample mean might depart significantly from the overall population mean (Kahneman and Tversky (1972b)).[27] Benjamin, Rabin, and Raymond (2016) provide a model of NBLLN with a focus on coin flipping

[27] In contrast, some papers predict overweighting of the evidence (e.g., Bushong and Gagnon-Bartsch (2018)), and the literature on extrapolative expectations is sometimes interpreted as overweighting of signals (see discussion in Section 4).

experiments. The signal S_N observed in a given period is a set of N realizations of a binomial random variable. The agent's goal is to learn about an underlying state of the world $\theta \in \Theta$. If an NBLLN agent observes signal S_N in a given period, her beliefs change as if the signal were generated in a two step process. First, a "subjective rate" $\beta \in (0,1)$ is drawn from a distribution $p^{\psi}(\beta|\theta)$ that has full support over $(0,1)$. Second, the probability of observing N_a realizations of a and N_b realizations of b is determined by the binomial distribution with rate β. Letting $p_{S_N}(S_N|\beta)$ denote the binomial probability mass function for S_N given rate β, the agent believes that the likelihood of signal S_N conditional on state θ is

$$p^{\psi}(S_N|\theta) = \int_{\beta\in[0,1]} p_{S_N}(S_N|\beta)p^{\psi}(\beta|\theta)d\beta$$

Intuitively, observing many signals in the current period is very informative about β, but the NBLLN agent only learns about θ to the extent that a single draw from $p^{\psi}(\beta|\theta)$ is informative.

We combine the effects of NBLLN with BRN by assuming that the agent's interpretation of a signal, $p^{\psi}(S_N|\theta)$, is determined by NBLLN, but the agent's use of her prior is determined by BRN. Formally, the agent's beliefs obey

$$p^{\psi}_{\alpha}(\theta|s) = \frac{p^{\psi}(S_N|\theta)p(\theta)^{\alpha}}{\sum_{\theta'\in\Theta} p^{\psi}(S_N|\theta)p(\theta')^{\alpha}}.$$

The combination of these two effects means that the agent neglects both the information she collects in the current period and her prior knowledge. Proposition 10 reveals that the combination of NBLLN and BRN prevents an agent from ever learning the state of the world regardless of how precise her signal is (due to NBLLN) or how many signals she observes over time (due to BRN).

Proposition 10 *For any θ_1, $\theta_2 \in \Theta$ and prior $p(\theta_1), p(\ \theta_2) \in (0,1)$, we have that as $N \to \infty$*

$$\frac{p^{\psi}_{\alpha}(\theta_1|S_N)}{p^{\psi}_{\alpha}(\theta_2|S_N)} \underset{a.s}{\rightarrow} \frac{p^{\psi}(\beta|\theta_1)}{p^{\psi}(\beta|\theta_2)}\left(\frac{p(\theta_1)}{p(\theta_2)}\right)^{\alpha}$$

8.3.3 Confirmatory Bias and Motivated Cognition

There is a substantial psychological literature on "confirmatory bias," which argues that people tend to misread ambiguous evidence as supportive of their current theories. When boiled down to the essential predictions and formalizations, this is the *opposite* of BRN. To take a classical context in which confirmatory bias is posited, consider a teacher evaluating the talent of a student over the course of a school year. Research and psychological theory suggest that, insofar as the teacher is making subjective judgments about the student's performance, he will be inclined to read ambiguous evidence as supporting his earlier impressions. Although modeled differently in Rabin and Schrag (1999), Augenblick and Rabin (2016) argue confirmatory bias can be usefully modeled using our functional form with $\alpha > 1$. Insofar as this bias is the opposite of BRN, it would seem to be a significant problem with the literature (and those of us modeling the biases) that two compelling factors in psychology can lead systematically in opposite directions.

In trying to tease apart the potentially conflicting influences of BRN and confirmation bias, we rely on three factors. First, the psychology of confirmatory bias suggests that it is not merely that subjects attend too much to prior information, as the $\alpha > 1$ model would suggest. Instead, subjects interpret new information so that it confirms their prior beliefs. This means that subjects with different priors will ascribe different meanings to new pieces of information in a systematic fashion, whereas simply setting $\alpha > 1$ yields a model where the agents have identical interpretations of each signal (and the interpretations are therefore independent of their prior beliefs). Second, the psychology literature suggests that confirmatory bias will be more prevalent when the signal is ambiguous and, stated informally, amenable to multiple interpretations that allow individuals to pick the interpretation that conforms to his or her prior beliefs. This requisite ambiguity suggests a natural bound on the applicability of confirmatory bias that does not apply to BRN. Third, and most speculatively, we believe that subjects exhibiting confirmation bias will not be aware that a subject with differing prior beliefs would interpret the signal differently. Although we are not aware of evidence of this final effect, it would be expected if the experimental subjects are not aware of the biases in their own belief formation process. Again, modeling confirmation bias with $\alpha > 1$ fails to capture this third effect since all subjects would, in fact, agree about the meaning of each signal.

We use the term motivated cognition to refer to misinterpretations of the signal that are caused by an agent's preferences (e.g., Brunnermeier and Parker (2005)). For example, agents could interpret signals as evidence that their preferred states are more likely. This form of misinterpretation is distinguished from confirmation bias by the fact that agents with different preferences (and possibly the same prior beliefs) will interpret the information in different ways. Unlike confirmation bias, it is hard to provide a simple model that captures these effects in a form that is the "opposite" of BRN. However, many of the comments applied to confirmation bias apply to motivated cognition. For example, some degree of ambiguity in the meaning of the signal is required so that alternative interpretations are feasible. Moreover, we speculate that agents subject to motivated cognition are unaware that other agents hold different interpretations of the signal.

8.3.4 The Relative Importance of Base-Rate Neglect and Misinterpretation of Signals

As mentioned above, we believe that misinterpretations of signals and BRN occur together, and the evidence discussed in Section 8.3.1 supports that view, at least in the case of unambiguous signals. There is a question of the relative magnitude of the effects. In particular, one might worry that the misinterpretations generate signals that are so strong that the effects of BRN (e.g., moderation effect, perpetually fluctuating beliefs) are simply irrelevant. These core predictions of BRN will hold even if signals are overweighted, underweighted, or misinterpreted, so long as signals observed in the distant past have relatively little influence on Saki's beliefs relative to a signal observed today. Therefore, the relative importance of BRN and misweighted signals is determined by whether Saki collects multiple, imperfectly informative signals (and BRN takes effect) or Saki stops updating her beliefs after collecting a few signals (and the effect of neglect is relatively small).

If a single strong signal were observed and (mis)interpreted as an extremely informative breakthrough, then the agent might never choose to collect other signals, as they are very likely to confirm what she already knows, and the effects of BRN will be muted. We concede that it is entirely possible that some signals are sufficiently ambiguous that there is enough room to interpret the information as a very strong signal when (in fact) the signal is much weaker. To the extent the bias is driven by motivated cognition (or confirmation bias interpreted as a preference for confirming information), it would appear natural (and maybe necessary) that the theory include a tension between the benefit of biasing our beliefs to encourage optimism and the cost in terms of suboptimal decisions due to these biased beliefs (e.g., Brunnermeier and Parker (2005)). Such a tension would prevent wild misinterpretations of signals unless the impact on the agent's utility were slight.

This suggests agents would gather multiple signals to learn about the world. Since the starkest effects of BRN occur when a sequence of signals is observed, we suspect the effects of BRN would be significant given the relatively low values of α observed in experimental subjects. There are surely some instances where most of an agent's knowledge is represented by a "breakthrough" that forecloses the need for further learning. However, the gradual process of learning through the acquisition of many moderately informative signals over time is also common (and perhaps even typical). We do not dispute the importance of misinterpreted signals, but we believe BRN is also usually important as well given the high degree of BRN observed (i.e., low values of α) in experimental subjects.

9 Model Extensions

9.1 Repeated (and Fortified) Signals

Suppose that Saki receives independent signals s_1 and s_2 in successive periods. She then receives s_3, which she knows ex ante is a repetition of s_1, and is thus not informative about the hypotheses conditional on the rest of the information she has received. What is the effect (if any) on her beliefs? Although we know of no evidence related to this in the psychology literature, there are at least two possible approaches to applying our model. First (and taking our model literally), Saki does not update her beliefs since s_3 is entirely determined by the already known s_1 and, therefore, is not a signal as we have defined them in Section 2.4. Therefore, Saki would think the relative probability of two hypotheses θ and $\widetilde{\theta}$ is:

$$\frac{p_\alpha(\theta|s_1,s_2,s_3)}{p_\alpha(\widetilde{\theta}|s_1,s_2,s_3)} = \frac{p(s_2|\theta)}{p(s_2|\widetilde{\theta})}\left(\frac{p(s_1|\theta)}{p(s_1|\widetilde{\theta})}\right)^{\alpha}\left(\frac{p(\theta)}{p(\widetilde{\theta})}\right)^{\alpha^2}$$

A possible extension of our model would be to assume that s_3 draws Saki's attention to the realization of s_1, which means s_1 would be *fortified* and given full Bayesian weight in the updating process. Note that the signal is not being double counted by Saki—the fortification has made it just as informative for Saki as it would be for Tommy. Therefore, observing s_3 does not cause Saki to neglect her previous observations, so

$$\frac{p_\alpha(\theta|s_1,s_2,s_3)}{p_\alpha(\widetilde{\theta}|s_1,s_2,s_3)} = \frac{p(s_2|\theta)}{p(s_2|\widetilde{\theta})}\frac{p(s_1|\theta)}{p(s_1|\widetilde{\theta})}\left(\frac{p(\theta)}{p(\widetilde{\theta})}\right)^{\alpha^2}$$

This, of course, sounds very passive in that Saki's attention is directed exogenously. This is a deep problem with any non-Bayesian model of signal updating.

Another application of these ideas is to situations where Saki is presented with a sufficient statistic on which she could base her beliefs. For example, in the coin flipping settings of Section 4, an extreme Saki with $\alpha = 0$ is unable to learn the bias of a coin even in an i.i.d. environment. It would seem reasonable that if the fraction of previously observed "head"s outcomes is provided to Saki each period, then she might very well learn the bias of the coin via the sufficient statistic. The effect is, as if, the outcome of every prior flip had been fortified. Of course, such a sufficient statistic is useless for Tommy since his beliefs converge to the truth after observing many flips. For such an effect to be operative in Saki, there must exist a sufficient statistic and, importantly, Saki must choose to rely on that statistic in lieu of the beliefs she has formed by sequentially updating based on each successive flip. If Saki has no reason to think her beliefs are not accurate, it is not clear why she would not simply ignore the sufficient statistic (see Gagnon-Bartsch, Rabin, and Schwartzstein (2018) for a formal treatment).

9.2 Unneglected Prior Beliefs and Persistent Theories

We now consider an extension of our model wherein the agent neglects the signals she receives in prior periods when a new signal is observed, but she always gives her $t = 0$ prior beliefs full weight. We refer to this agent as Peggy. This modification of our model is motivated by the idea that agents may maintain a disposition towards certain beliefs that does not get neglected as new information arrives. The goal of this section is to lay out our model of Peggy and point out how our predictions about the evolution of the beliefs of Saki and Peggy differ.

Suppose that Peggy has prior beliefs $p(\theta)$ and observes signals s_1 and s_2 in periods 1 and 2 respectively. After s_1 is observed, her beliefs about hypothesis θ following the signal are

$$p_\alpha(\theta|s_1) = \frac{p(\theta|s_1)p(\theta)}{\sum_{\theta'} p(\theta'|s_1)p(\theta')}.$$

These are precisely the beliefs that Tommy would hold. Once s_2 is observed, Peggy discounts the information in her first signal and her beliefs have the form

$$p_\alpha(\theta|s_1) = \frac{p(\theta|s_2)\left(p(\theta|s_1)\right)^\alpha p(\theta)}{\sum_{\theta'} p(\theta'|s_2)\left(p(\theta'|s_1)\right)^\alpha p(\theta')}.$$

Each time Peggy observes a new signal, she futher neglects the signals she observed previously, but her priors are never neglected.

The qualitative differences between the evolution of Saki's and Peggy's beliefs turns on whether the neglect of the prior belief in particular (as opposed to neglect of the signals) is important. For example, Peggy may not exhibit conjunction violations (Proposition 3) as these can require almost complete neglect of prior beliefs. In addition, if Saki observes a sequence of uninformative signals, then her beliefs will converge to a uniform distribution over the hypotheses (i.e., a fully neglected prior). In contrast, the beliefs of Peggy will approach her priors if provided the same sequence of signals. We also find that her beliefs continuously evolve and never converge, but Peggy's prior now affects the distribution of these vacillations. However, most of our results do not qualitatively depend on neglect of the prior beliefs specifically and apply to Peggy as well. Please see Appendix B for a complete discussion

Our model of Peggy is not necessarily inconsistent with the typical experimental designs used to test base-rate neglect (e.g., Kahneman and Tversky (1973); Griffin and Tversky (1992)). Suppose that before the experiment the subjects have a nondogmatic prior about the base-rate that will be used in the experimental design. When the experimenters reveal the base-rate in the experiment, the subjects treat this as a perfect signal of the true base-rate. After receiving this perfect signal, the initial prior belief over the set of possible base-rates is irrelevant—it does not matter if Peggy neglects her priors or not once the perfect signal is received.

10 Conclusion

This paper takes the evidence from existing experiments to build out a fully fleshed model of dynamic base-rate neglect. We conclude by highlighting some predictions that merit testing and identifying promising applications.

Some basic features of BRN are under-tested and merit increased focus. For example, the moderation effect and the distinction between an uninformative signal and no signal have tentative support from previous experiments, but these issues deserve to be the focus of future experimental work. Other predictions, such as the hypothesis-dependence of Section 2.3, have not (to our knowledge) been directly tested before in an updating context and deserve examination.

Perhaps the most promising target for experimental testing is our assumption that yesterday's posteriors become today's priors. Although this assumption is common within the existing literature (e.g., Grether (1992)), we are not aware of any work that directly and explicitly tests this assumption against others that we discuss in Section 3. In our setting, it makes a big difference whether Saki perceives two pieces of information to arrive at the same (and hence neither is discounted) or whether the first piece of information is fully incorporated into her priors and thus discounted when the second piece of information arrives. We would not be surprised if people update their priors when faced with a new decision, which suggests that the definition of a period is simply the length of time between decisions. We would also not be surprised if people update their priors if a sufficient amount of time has passed even if they have not faced a decision to which a signal is relevant. In any case, the exact mechanism that prompts updating is an important empirical question in our model (as well as many other non-Bayesian models).

Among the missing evidence is the study of how individuals prospectively view their own inference. The application to sequential sampling in Section 5 exposed some important implications of whether individuals are aware of the biases in their own updating. Although we would be surprised if individuals were fully cognizant of their own biases, this is a conjecture that merits testing. A related idea is whether people are aware that others suffer from base-rate neglect. Our applications addressed these issues to some extent. For example, the persuader's behavior in Section 6 depends heavily on whether the persuader believes the audience is Tommy or Saki.

A useful theoretical application of our model would be to study how well BRN can be used as a replacement for the assumption of dogmatic, incorrect beliefs about the correct model of the world. For example, would the conclusions of Barberis, Shleifer, and Vishny (1998) change if the agents thought it was possible that asset prices follow a random walk, but were subject to BRN and were thus unable to learn this? More generally, can we identify features of a model or the role mis-specified beliefs play in the analysis that indicate when BRN can mimic the effects of the mis-specification?

One of the signature features of BRN that we identify is the failure of Saki's beliefs to converge—in other words, Saki fails to learn. Since effective learning is such a foundational feature of many economic models, the failure to do so suggests both novel economic effects and room for welfare-enhancing interventions. Our application to reputation, for example, suggests that since Saki will fail to differentiate between strategic firms and those committed to producing high quality, it may be futile to expect the behavior of firms to be disciplined by reputational concerns alone, and this could motivate interventions such as quality certification or regulation. More abstractly, the failure to learn raises questions about whether Saki can attain beliefs close to rational expectations via learning and whether learning can be used to motivate game-theoretic equilibrium notions.

More interesting questions arise when we consider applications that involve more elaborate multi-agent interaction than our applications have assumed. Again, these applications will have to make assumptions about what agents believe about their own updating and about the updating of other agents. These choices raise questions without obvious answers. For example, is it possible for Saki to realize that Tommy does not update as she does? If she is aware of this, is it possible for Saki to simultaneously realize Tommy does not neglect his prior beliefs while she does? If she does realize that Tommy updates differently, can this cause Saki to form Bayesian posteriors as per our (speculative) discussion of pre-emptive Bayesianism? And of course there are further questions about what Saki believes Tommy believes about Saki's belief updating process. While these questions may be amenable to being resolved through theoretical inquiry, they might also be persuasively answered through experimental investigation.

References

Ahn, D., and H. Ergin (2010): "Framing Contingencies," *Econometrica*, 78(2), 655–695.

Augenblick, N., and M. Rabin (2017): "Belief Movement, Uncertainty Reduction, and Rational Updating," Working Paper.

Bar-Hillel, M. (1980): "The Base-Rate Fallacy in Probability Judgments," *Acta Psychologica*, 44, 211–233.

Barberis, N., A. Shleifer, and R. Vishny (1998): "A Model of Investor Sentiment," *Journal of Financial Economics*, 49, 307–343.

Barbey, A., and S. Sloman (2007): "Base-rate respect: From ecological rationality to dual processes," *Memory & Cognition*, 30(3), 241–254.

Benjamin, D. (2018): "Errors in Probabilistic Reasoning and Judgmental Biases," Working Paper.

Benjamin, D., M. Rabin, and C. Raymond (2016): "A Model of Non-Belief in the Law of Large Numbers," *Journal of the European Economic Association*, 14(2), 515–544.

Beshears, J., et al. (2013): "What Goes Up Must Come Down? Experimental Evidence on Intuitive Forecasting," *American Economic Review: Papers & Proceedings*, 103(3), 570–574.

Blume, L., D. Easley, and J. Halpern (2009): "Constructive Decision Theory," Working Paper.

Bodoh-Creed, A. (2018): "Mood, Memory, and the Evaluation of Asset Prices," forthcoming.

Bohren, A. (2016): "Informational Herding with Model Misspecification," *Journal of Economic Theory*, 163, 222–247.

Borgida, E., and N. Brekke (1981): “The Base Rate Fallacy in Attribution and Prediction,” in *New Directions in Attribution Research*, ed. by J. Harvey, W. Ickes, and R. Kidd, vol. 3. Erlbaum, Hillsdale, NJ.

Brunnermeier, M., and J. Parker (2005): “Optimal Expectations,” *The American Economic Review*, 95(4), 1092–1118.

Bushong, B., and T. Gagnon-Bartsch (2018): “Learning with Misattribution of Reference Dependence,” Working Paper.

Camerer, C. (1987): “Do Biases in Probability Judgements Matter in Markets? Experimental Evidence,” *American Economic Review*, 77, 981–997.

Camerer, C. (1989): “Does the Basketball Market Believe in the ‘Hot Hand’?,” *American Economic Review*, 79(5), 1257–1261.

Camerer, C., T. Ho, and J. Chong (2004): “A Cognitive Hierarchy Model of Games,” *The Quarterly Journal of Economics*, 119(3), 861–898.

Chun, W., and A. Kruglanski (2006): “The Role of Task Demands and Processing Resources in the Use of Base-Rate and Individuating Information,” *The Journal of Personality and Social Psychology*, 91(2), 205–217.

Crawford, V., and N. Iriberri (2007): “Level-K Auctions: Can a Non-equilibrium Model of Strategic Thinking Explain the Winner’s Curse and Overbidding in Private-Value Auctions?,” *Econometrica*, 75(6), 1721–1770.

Cripps, M., G. Mailath, and L. Samuelson (2004): “Imperfect Monitoring and Impermanent Reputations,” *Econometrica*, 72, 407–432.

DeMarzo, P., D. Vayanos, and J. Zwiebel (2003): “Persuasion Bias, Social Influence, and Uni-Dimensional Opinions,” *The Quarterly Journal of Economics*, 118, 909–968.

Dohmen, T., A. Falk, D. Huffman, F. Marklein, and U. Sunde (2009): “The Non-Use of Bayes Rule: Representative Evidence on Bounded Rationality,” Working Paper.

Eddy, D. (1982): “Probabilistic reasoning in clinical medicine: Problems and opportunities,” in *Judgment under uncertainty: Heuristics and biases*, ed. by D. Kahneman, P. Slovic, and A. Tversky, pp. 249–267. Cambridge University Press, Cambridge.

Eide, E. (2011): “Two tests of the base rate neglect among law students,” Working Paper.

Esponda, I. (2008): “Behavioral Equilibrium in Economies with Adverse Selection,” *American Economic Review*, 98(4), 1269–1291.

Esponda, I., and D. Pouzo (2016): “Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models,” *Econometrica*, 84(3), 1093–1130.

Evans, G., and S. Honkapohja (2001): *Learning and Expectations in Macroeconomics*. Princeton University Press, Princeton, NJ.

Eyster, E., and M. Piccione (2013): "An Approach to Asset Pricing Under Incomplete and Diverse Perceptions," *Econometrica*, 81, 1483–1506.

Eyster, E., and M. Rabin (2005): "Cursed Equilibrium," *Econometrica*, 73(5), 1623–1672.

——— (2010): "Naive Herding in Rich-Information Settings," *American Economic Journal: Microeconomics*, 2(4), 221–243.

——— (2014): "Extensive Imitation is Irrational and Harmful," *The Quarterly Journal of Economics*, 129(4), 1861–1898.

Fischhoff, K., P. Slovic, and S. Lichtenstein (1978): "Fault Trees: Sensitivity of estimated failure to problem representation," *Journal of Experimental Psychology: Human Perception and Performance*, 4, 330–344.

Fryer, R., P. Harms, and M. Jackson (2018): "Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization," *Journal of the European Economic Association*, forthcoming.

Fryer, R., and M. Jackson (2008): "A Categorical Model of Cognition and Biased Decision-Making," *B. E. Journal of Theoretical Economics*, 8, 1–44.

Fudenberg, D., and D. Levine (1992): "Maintaining a Reputation when Strategies are Imperfectly Observed," *The Review of Economic Studies*, 59, 561–579.

Fuster, A., D. Laibson, and B. Mendel (2010): "Natural Expectations and Macroeconomic Fluctuations," *Journal of Economic Perspectives*, 24(4), 67–84.

Gagnon-Bartsch, T., and M. Rabin (2017): "Naive Social Learning, Mislearning, and Unlearning," Working Paper.

Gagnon-Bartsch, T., M. Rabin, and J. Schwartzstein (2018): "Channeled Attention and Stable Errors," Working Paper.

Ganguly, A., J. Kagel, and D. Moser (2000): "Do Asset Market Prices Reflect Traders' Judgment Biases," *Journal of Risk and Uncertainty*, 20(3), 219–245.

Gennaioli, N., and A. Shleifer (2010): "What Comes to Mind," *The Quarterly Journal of Economics*, 125(4), 1399–1433.

Gigerenzer, G., W. Hell, and H. Blank (1988): "Presentation and Content: The Use of Base Rates as a Continuous Variable," *Journal of Experimental Psychology: Human Perception and Performance*, 14(3), 513–525.

Gigerenzer, G., and U. Hoffrage (1995): "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats," *Psychological Review*, 102(4), 684–704.

Gilovich, T., R. Vallone, and A. Tversky (1985): "The Hot Hand in Basketball: On the Misperception of Random Sequences," *Cognitive Psychology,*, 17, 295–314.

Goodie, A. S., and E. Fantino (1999): "What does and does not alleviate base-rate neglect under direct experience," *Journal of Behavioral Decision Making*, 12(4), 307–335.

Greenwood, R., and A. Shleifer (2014): "Expectations of Returns and Expected Returns," *The Review of Financial Studies*, 27(3), 714–746.

Grether, D. (1980): "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *The Quarterly Journal of Economics*, 95, 537–557.

——— (1992): "Testing bayes rule and the representativeness heuristic: Some experimental evidence," *Journal of Economic Behavior and Organization*, 17(1), 31–57.

Griffin, D., and A. Tversky (1992): "The Weighing of Evidence and the Determinants of Confidence," *Cognitive Psychology*, 24, 411–435.

He, X. D., and D. Xiao (2017): "Processing Consistency in Non-Bayesian Inference," Working Paper.

Jehiel, P. (2005): "Analogy-Based Expectation Equilibrium," *Journal of Economic Theory*, 123, 81–104.

Jehiel, P., and F. Koessler (2008): "Revisiting games of incomplete information with analogy-based expectations," *Games and Economic Behavior*, 62(2), 533–557.

Kahneman, D., and A. Tversky (1972a): *On Prediction and Judgment.* Oregon Research Institute Monograph.

——— (1972b): "Subjective Probability: A Judgement of Representativeness," *Cognitive Psychology*, 3, 430–454.

——— (1973): "On the Psychology of Prediction," *Psychological Review*, 80, 237–251.

——— (1983): "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment," *Psychological Review*, 90, 293–315.

Kennedy, M., W. Willis, and D. D. Faust (1997): "The Base-Rate Fallacy in School Psychology," *Journal of Psychoeducational Assessment*, 15(4), 292–307.

Kirman, A. (1975): "Learning by Firms About Demand Conditions," in *Adaptive Economic Models*, ed. by R. Day, and T. Groves, pp. 137–156. Academic Press.

Koehler, J. (1996): "The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges," *Behavioral and Brain Sciences*, 19, 1–53.

Kraemer, C., and M. Weber (2004): "How Do People Take into Account Weight, Strength, and Quality of Segregated vs. Aggregated Data? Experimental Evidence," *The Journal of Risk and Uncertainty*, 29(2), 113–142.

Krosnick, J., F. Li, and D. Lehman (1990): "Conversational Conventions, Order of Information Acquisition, and the Effect of Base Rates and Individuating Information on Social Judgments," *Journal of Personality and Social Psychology*, 59(6), 1140–1152.

Labella, C., and D. Koehler (2004): "Dilution and confirmation of probability judgments based on nondiagnostic evidence," *Memory & Cognition*, 32(7), 1076–1089.

Landier, A., Y. Ma, and D. Thesmar (2017): “New Experimental Evidence on Expectations Formation,” Working Paper.

Liu, Q., and A. Skrzypacz (2011): “Limited Records and Reputation,” Working Paper.

Loewenstein, G., T. O’Donoghue, and M. Rabin (2003): “Projection Bias in Prediction Future Utility,” *The Quarterly Journal of Economics*, pp. 1209–1248.

Madarász, K. (2012): “Information Projection: Model and Applications,” *The Review of Economic Studies*, 79, 961–985.

Meehl, P. E., and A. Rosen (1955): “Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores,” *Psychological Bulletin*, 52(3), 194–216.

Meyn, S., and R. Tweedie (1993): *Markov Chains and Stochastic Stability.* Springer-Verlag, New York.

Miller, J., and A. Sanjurjo (2018): “Surprised by the Hot Hand Fallacy? A Truth in the Law of Small Numbers,” *Econometrica*, 86, 2019–2047.

Mullainathan, S. (2000): “Thinking Through Categories,” Working Paper.

——— (2002): “A Memory-Based Model of Bounded Rationality,” *The Quarterly Journal of Economics*, 117(3), 735–774.

Mullainathan, S., J. Schwartzstein, and A. Shleifer (2008): “Coarse Thinking and Persuasion,” *The Quarterly Journal of Economics*, 123, 577–619.

Nisbett, R., E. Borgida, R. Crandall, and H. Reed (1976): “Popular induction: Information is not necessarily informative,” in *Cognition and Social Behavior*, ed. by J. Carroll, and J. Payne. Erlbaum, Hillsdale, NJ.

Nyarko, Y. (1991): “Learning in Misspecified Models and the Possibility of Cycles,” *Journal of Economic Theory*, 55(2), 416–427.

Phillips, L., and W. Edwards (1966): “Conservatism in a Simple Probability Inference Task,” *Journal of Experimental Psychology*, 72(3), 346–354.

Rabin, M. (2002): “Inference by Believers in the Law of Small Numbers,” *The Quarterly Journal of Economics,*, 117, 775–816.

Rabin, M., and J. Schrag (1999): “First Impressions Matter: A Model of Confirmatory Bias,” *The Quarterly Journal of Economics*, 114(1), 37–82.

Rabin, M., and D. Vayanos (2010): “The Gambler’s and Hot-Hand Fallacies: Theory and Applications,” *The Review of Economic Studies*, 77, 730–778.

Sargent, T. (1993): *Bounded Rationality in Macroeconomics.* Oxford University Press, Oxford.

Shanteau, J. (1975): “Averaging versus multiplying combination rules of inference judgment,” *Acta Psychologica*, 39(1), 83–89.

Shu, S., and G. Wu (2003): “Belief Bracketing: Can Partitioning Information Change Consumer Judgments?,” Working Paper.

Spiegler, R. (2016): “Bayesian Networks and Boundedly Rational Expectations,” *The Quarterly Journal of Economics*, 131, 1243–1290.

Tribe, L. H. (1971): “Trial by Mathematics: Precision and Ritual in the Legal Process,” *Harvard Law Review*, 84(6), 1329–1393.

Troutman, C., and J. Shanteau (1977): “Inferences Based on Nondiagnostic Information,” *Organizational Behavior and Human Performance*, 19(1), 43–55.

Tversky, A., and D. Koehler (1994): “Support Theory: A nonextensional representation of subjective probability,” *Psychological Review*, 101(4), 547–567.

Wald, A. (1947): *Sequential Analysis.* Wiley, New York.

Zukier, H., and A. Pepitone (1984): “Social Roles and Strategies in Prediction: Some Determinants of the Use of Base-Rate Information,” *Journal of Personality and Social Psychology*, 47(2), 349–360.

A Proofs

Proposition 1 *Choose any $\alpha < 1$. Consider two hypothesis θ and θ'. Then:*

1. *For all signals s where $1 < \frac{p(s|\theta)}{p(s|\theta')} < \infty$, there exist prior beliefs $p(\theta), p(\theta') < 1$ such that $\frac{p_\alpha(\theta|s)}{p_\alpha(\theta'|s)} < \frac{p(\theta)}{p(\theta')}$.*
2. *For all $p(\theta) > p(\theta')$, there exists $z > 1$ such that for all signals s where $\frac{p(s|\theta)}{p(s|\theta')} < z$, $\frac{p_\alpha(\theta|s)}{p_\alpha(\theta'|s)} < \frac{p(\theta)}{p(\theta')}$.*

Proof. (1): Writing down the respective updating formula, we have

$$\frac{p_\alpha(\theta|s)}{p_\alpha(\theta'|s)} = \frac{p(s|\theta)}{p(s|\theta')}\left(\frac{p(\theta)}{p(\theta')}\right)^{\alpha}. \tag{17}$$

Substituting in any $\frac{p(\theta)}{p(\theta')} > \left(\frac{p(s|\theta)}{p(s|\theta')}\right)^{1/(1-\alpha)}$, we find that our claim holds.

(2): Using equation 17, if one lets $z = \left(\frac{p(\theta)}{p(\theta')}\right)^{1-\alpha}$, then our claim holds. Since $p(\theta) > p(\theta')$, we have $z > 1$. ■

Proposition 2 *For all $\alpha < 1$, $p_\alpha^{\Theta_2}(B|s) + p_\alpha^{\Theta_2}(C|s) > p_\alpha^{\Theta_1}(B \cup C|s)$.*

Proof. Our claim is equivalent to $p_\alpha^{\Theta_1}(A|s) > p_\alpha^{\Theta_2}(A|s)$, which we can write

$$\frac{p(s|A)p(A)^\alpha}{p(s|A)p(A)^\alpha + p(s|B \cup C)p(B \cup C)^\alpha} > \frac{p(s|A)p(A)^\alpha}{p(s|A)p(A)^\alpha + p(s|B)p(B)^\alpha + p(s|C)p(C)^\alpha}.$$

Simplifying, we find that this is equivalent to

$$p(s|B \cup C)p(B \cup C)^\alpha < p(s|B)p(B)^\alpha + p(s|C)p(C)^\alpha.$$

Further simplifying yields

$$p(s|B)p(B)^\alpha \left[\left(\frac{p(B)}{p(B \cup C)}\right)^{1-\alpha} - 1\right] < p(s|C)p(C)^\alpha \left[1 - \left(\frac{p(C)}{p(B \cup C)}\right)^{1-\alpha}\right].$$

This inequality holds if $\alpha < 1$, proving our claim. ■

Proposition 3 *Suppose the state space contains events $A \subset B \subset \Omega$, and consider the sets of hypotheses $\Theta_1 = \{A, \Omega - A\}$ and $\Theta_2 = \{B, \Omega - B\}$. Suppose that there is some signal s such that $p(s|A) > p(s|B)$ and $p(s|\Omega - B) \geq p(s|\Omega - A)$. For $\alpha \in [0, 1]$ sufficiently small, $p_\alpha^{\Theta_1}(A|s) > p_\alpha^{\Theta_2}(B|s)$.*

Proof. Proving our claim is equivalent to

$$\frac{p(s|A)p(A)^\alpha}{p(s|A)p(A)^\alpha + p(s|\Omega - A)p(\Omega - A)^\alpha} \geq \frac{p(s|B)p(B)^\alpha}{p(s|B)p(B)^\alpha + p(s|\Omega - B)p(\Omega - B)^\alpha},$$

which in turn implies

$$\frac{p(s|A)}{p(s|A) + p(s|\Omega - A)\left[\frac{p(\Omega-A)}{p(A)}\right]^\alpha} \geq \frac{p(s|B)}{p(s|B) + p(s|\Omega - B)\left[\frac{p(\Omega-B)}{p(B)}\right]^\alpha}.$$

Since $p(s|A) \geq p(s|B)$ and $p(s|\Omega - A) \leq p(s|\Omega - B)$, the final equality holds if $\alpha \in [0, 1]$ is sufficiently small. ■

Proposition 4 *If Θ has at least two elements, Saki's beliefs exhibit prediction momentum and prediction momentum relative to θ^* .*

Proof. First note that both kinds of prediction momentum are equivalent since $p(h_{\tau+1}|h_\tau, \theta^*) - p(h_{\tau+1}|t_\tau, \theta^*) = 0$. In this setting, it suffices to show that $p_{\alpha=0}(h|h) > p_{\alpha=0}(h|t)$ since there is no need to condition on signals observed before period t to predict Saki's beliefs. From equation 9 we know $p_{\alpha=0}(h|h) > p_{\alpha=0}(h|t)$ if and only if

$$\frac{\sum_{\theta_j \in \Theta} (\theta_j)^2}{\sum_{\theta_j \in \Theta} \theta_j} > \frac{\sum_{\theta_j \in \Theta} \theta_j(1 - \theta_j)}{\sum_{\theta_j \in \Theta} 1 - \theta_j}$$

Simplifying this yields $N \sum_{\theta_j \in \Theta} (\theta_j)^2 - (\sum_{\theta_j \in \Theta} \theta_j)^2$, which is strictly positive if Θ has at least two elements. ■

Proposition 5 *Suppose the true theory of the world is $\theta^* = (r^*, s^*)$. From the perspective of a Bayesian observer, Saki's expected predictions are:*

$$\mathbb{E}_{\tau,\alpha=1}[p_{\alpha=0}(h_{\tau+1}|h_\tau, f_{\tau-1})|h_\tau, \theta^*] = \sum_{\theta_i=(r_i,s_i)} \frac{r_i^2}{\sum_{\theta_j=(r_j,s_j)} r_j} r^* + \frac{r_i(1-s_i)}{\sum_{\theta_i=(r_i,s_i)} 1-s_j}(1-r^*)$$

$$\mathbb{E}_{\tau,\alpha=1}[p_{\alpha=0}(h_{\tau+1}|t_\tau, f_{\tau-1})|t_\tau, \theta^*] = \sum_{\theta_i=(r_i,s_i)} \frac{r_i(1-r_i)}{\sum_{\theta_j=(r_j,s_j)}(1-r_j)}(1-s^*) + \frac{r_i s_i}{\sum_{\theta_i=(r_i,s_i)} s_j} s^*$$

Saki's beliefs exhibit more prediction momentum than θ^* if

$$\mathbb{E}_{\tau,\alpha=1}[p_{\alpha=0}(h_{\tau+1}|h_\tau, f_{\tau-1})|h_\tau, \theta^*] - \mathbb{E}_{\tau,\alpha=1}[p_{\alpha=0}(h_{\tau+1}|t_\tau, f_{\tau-1})|t_\tau, \theta^*] > r^* + s^* - 1$$

Proof. Suppose the theory of the world generating the sequence of coin flips is $\theta^* = (r^*, s^*)$. By treating the sequence of coin flips as a Markov process, we find that the probability that h_τ is preceded by $h_{\tau-1}$ is r^*. We can write the observers expectations of Saki's beliefs condition on the theory of the world θ^* as

$$\mathbb{E}_{\alpha=1}[p_\alpha(h_{\tau+1}|h_\tau)|\theta^*] = \sum_{(\theta_i, f_{\tau-1})} p(h_{\tau+1}|h_\tau, \theta_i, f_{\tau-1}) p_\alpha(\theta_i|h_\tau, f_{\tau-1}) p(f_{\tau-1}|h_\tau|\theta^*)$$

Note that only the final term is condition on θ^* as this is the term that describes the evolution of the coin flips. Substituting in terms yields

$$\mathbb{E}_{\alpha=1}[p_\alpha(h_{\tau+1}|h_\tau)|\theta^*] = \sum_{(\theta_i, f_{\tau-1})} r_i \left[\frac{r_i}{\sum_{(\theta_i,h_{\tau-1})} r_i} r^* + \frac{1-s_i}{\sum_{(\theta_i,h_{\tau-1})} 1-s_i}(1-r^*)\right]$$

Simplifying we find

$$\mathbb{E}_{\alpha=1}[p_\alpha(h_{\tau+1}|h_\tau)|\theta^*] = \sum_{(\theta_i, f_{\tau-1})} \left[\frac{r_i^2}{\sum_{(\theta_i,h_{\tau-1})} r_i} r^* + \frac{r_i(1-s_i)}{\sum_{(\theta_i,h_{\tau-1})} 1-s_i}(1-r^*)\right]$$

A similar argument yields the formula for conditioning on t_τ.

Combining these results with the definition of θ^* yields our final result regarding prediction momentum with respect to θ^*. ■

Proposition 6 *If $\alpha = 1$ (i.e., the audience is Tommy), then the bragging equilibrium is the unique sequential equilibrium.*

Proof. Consider any sequential equilibrium $\widetilde{\sigma}$. Since we are focusing on sequential equilibria, the persuadee's beliefs following a deviation must be consistent with a sequence of fully mixed strategies (and the associated Bayesian beliefs) with the limit of the sequence of strategies being $\widetilde{\sigma}$. Note that since the persuader does not know whether her hypothesis is true or not, none of the strategies in this sequence can be conditioned on θ. In other words, if the i^{th} strategy in the sequence releases g with probability $\widetilde{\sigma}^i(g)$ and b with probability $\widetilde{\sigma}^i(b)$, the associated Bayesian beliefs are:

$$\begin{aligned}
\frac{p_{\alpha=1}(G|\widetilde{\sigma}^i(s))}{1-p_{\alpha=1}(G|\widetilde{\sigma}^i(s))} &= \frac{p(g \text{ revealed}|\theta=G,\widetilde{\sigma}^i(s))}{p(g \text{ revealed}|\theta=B,\widetilde{\sigma}^i(s))}\left(\frac{p(G)}{1-p(G)}\right) \\
&= \frac{\widetilde{\sigma}^i(g)m}{\widetilde{\sigma}^i(g)n}\left(\frac{p(G)}{1-p(G)}\right) \\
&= \frac{m}{n}\left(\frac{p(G)}{1-p(G)}\right) > \frac{p(G)}{1-p(G)}.
\end{aligned}$$

Therefore, if $\widetilde{\sigma}(g)=\varnothing$, the off-path beliefs following the release of a g signal are

$$\frac{p_{\alpha=1}(G|g \text{ revealed})}{1-p_{\alpha=1}(G|g \text{ revealed})} = \frac{m}{n} = \left(\frac{p(G)}{1-p(G)}\right) > \frac{p(G)}{1-p(G)}$$

Since the persuader's utility is increasing in $p_{\alpha=1}(G|\widetilde{\sigma}(s))$, he has a strict incentive to deviate to revealing g.

To determine whether withholding a b signal is optimal, we must compute the audience's beliefs following the release of $\varnothing$ by the persuader, and these beliefs are:

$$\begin{aligned}
\frac{p_{\alpha=1}(G|\varnothing \text{ revealed})}{1-p_{\alpha=1}(G|\varnothing \text{ revealed})} &= \frac{m(1-\widetilde{\sigma}^i(g))+n(1-\widetilde{\sigma}^i(b))+r}{n(1-\widetilde{\sigma}^i(g))+m(1-\widetilde{\sigma}^i(b))+r}\left(\frac{p(G)}{1-p(G)}\right) \\
&= \frac{n(1-\widetilde{\sigma}^i(b))+r}{m(1-\widetilde{\sigma}^i(b))+r}\left(\frac{p(G)}{1-p(G)}\right) \\
&\geq \frac{n}{m}\left(\frac{p(G)}{1-p(G)}\right)
\end{aligned}$$

with a strict inequality when $\widetilde{\sigma}^i(b)>0$. Therefore $\sigma(b)=0$ is optimal, and our claim is proven. ■

Proposition 7 *A silent sequential equilibrium exists if and only if* $\alpha<1$ *and*

$$p(G) \geq \frac{m^{1/(1-\alpha)}}{m^{1/(1-\alpha)}+n^{1/(1-\alpha)}}. \tag{18}$$

Proof. Using the notation of Proposition 6, in a silent equilibrium $\sigma(s)$ the off-path beliefs following the revelation of a signal have the form

$$\begin{aligned}
\frac{p_\alpha(G|\sigma(s)=g)}{1-p_\alpha(G|\sigma(s)=g)} &= \frac{p(g \text{ revealed}|\theta=G,\sigma(s))}{p(g \text{ revealed}|\theta=B,\sigma(s))}\left(\frac{p(G)}{1-p(G)}\right)^\alpha = \frac{m}{n}\left(\frac{p(G)}{1-p(G)}\right)^\alpha \\
\frac{p_\alpha(G|\sigma(s)=b)}{1-p_\alpha(G|\sigma(s)=b)} &= \frac{p(b \text{ revealed}|\theta=G,\sigma(s))}{p(b \text{ revealed}|\theta=B,\sigma(s))}\left(\frac{p(G)}{1-p(G)}\right)^\alpha = \frac{n}{m}\left(\frac{p(G)}{1-p(G)}\right)^\alpha.
\end{aligned}$$

If the persuader deviates from a silent equilibrium strategy and releases a g signal, then the resulting posteriors are

$$\frac{p_\alpha(G|\sigma(s)=g)}{1-p_\alpha(G|\sigma(s)=g)} = \frac{m}{n}\left(\frac{p(G)}{1-p(G)}\right)^\alpha \leq \frac{p(G)}{1-p(G)}$$

Since revealing g is the most favorable signal that the persuader can reveal and this signal causes the persuadees beliefs to moderate, then it is not optimal for the persuader to deviate from the silent equilibrium strategy.

Now suppose a silent equilibrium exists and equation 18 does not hold. Then as per the algebra above, releasing a g signal must result in

$$\frac{p_\alpha(G|\sigma(s)=g)}{1-p_\alpha(G|\sigma(s)=g)} > \frac{p(G)}{1-p(G)}$$

which is a profitable deviation for the persuader. Therefore equation 18 is necesary for a silent equilibrium to exist. ■

Proposition 8 *Let Assumptions 1 and (A) hold. For any $\alpha < 1$, in any equilibrium, the process ϕ_t is ergodic with non-degenerate support.*

Proof. Recall that the LRP's strategy can be described by a Markov process that takes the the beliefs of the SRP, ϕ_t, as the state variable. We denote the probability the LRP "works" as $\sigma(\phi_t)$. Furthermore, $\sigma(\phi_t) < 1$ in all periods. On the contrary suppose $\sigma(\phi_t) = 1$ for some ϕ_t. Since $\sigma(\phi_t) = 1$, the SRP would not treat y_t as a signal since it is (by design) uninformative. However, if the SRP does not infer from y_t, it is an optimal deviation for the LRP to "shirk." We conclude that in equilibrium it must be the case that $\sigma(\phi_t) < 1$.

We will prove that the log likelihood process, $L_t = \ln \frac{\phi_t}{1-\phi_t}$, is ergodic, which implies ϕ_t is ergodic. Since L_t is a monotone function of ϕ_t, we often use L_t as the argument of $\sigma(\circ)$. We use the following notation

$$\overline{m} = \max_{\beta\in[0,1]} E\left\|l_\beta\right\| \text{ and } \gamma^2 = \max_\beta Var(l_\beta)$$

Where convenient we supress the dependence of the period t shock on the state and the LRP's strategy by letting l_t denote $l_{\sigma(L_t)}(y_t)$ and l denote a generic likelihood shock. Given L_t, we can write $L_{t+\tau}$ as

$$L_{t+\tau} = \alpha^\tau L_t + S_\tau \text{ where } S_\tau = \sum_{i=1}^{\tau} \alpha^{\tau-i} l_{t+i}$$

Now we prove that L_t is ergodic. Throughout we refer the reader to Meyn and Tweedie (1993) for standard definitions regarding Harris chains. We use Theorem 13.01.1, statement (iii), of Meyn and Tweedie (1993) to prove ergodicity. This theorem requires that we show the L_t process satisfies the following properties, where C, defined later, refers to a compact subset of the real numbers.

- Strong aperiodicity and irreducibility, which follows from the full support of l_t
- Existence of an invariant measure
- Finite expected return times for points in C to C
- Harris recurrence

We now proceed with proving these properties hold.

Lemma 1 *L_t possesses an invariant measure.*

Proof. To prove that the chain is recurrent, we use Theorem 8.0.2, statement (ii), of Meyn and Tweedie (1993). Using the notation of this theorem, let $V(l) = \|l\|$ be the drift function and define the set

$$C = \left[\frac{-1}{1-\alpha}\,\overline{m}, \frac{1}{1-\alpha}\,\overline{m}\right]$$

Given this definition, for any $l \notin C$ and $l > 0$ we have (using the fact that $\|l\| > \frac{1}{1-\alpha}\,\overline{m}$)

$$\Delta V(l) \leq \|\alpha l\| + E\left[\left\|l_{\sigma(l)}(y)\right\|\right] - \|l\| = E\left[\left\|l_{\sigma(l)}(y)\right\| - (1-\alpha)\,\|l\| \leq \overline{m} - (1-\alpha)\,\|l\|$$

Since $\Delta V(l) \leq 0$ for all $l \notin C$, Theorem 8.0.2, statement (ii), implies L_t is recurrent. Theorem 10.0.1 of Meyn and Tweedie (1993) then implies there exists an invariant measure for L_t ■

Lemma 2 *There exists a compact set C such that $E_L\left[\tau_C\right] < \infty$ for $L \in C$.*[28]

Proof. From Assumption 1 we know that $E[l_t^2]$ is bounded. Therefore, for some $\gamma > 0$ we can write

$$Var(S_N) \leq \sum_{i,j=0}^{\infty} \alpha^{i+j} Cov(l_{t+N-i}, l_{t+N-j}) \leq \frac{\gamma^2}{(1-\alpha)^2}$$

From the Chebyshev inequality we can write for any N that

$$p(S_N \notin [-\overline{m} - k, \overline{m} + k]) \leq \frac{\gamma^2}{k(1-\alpha)^2} \tag{19}$$

Choose $k = \frac{(1-\alpha)^2}{2\gamma^2}$, so we have $p(S_N \notin [-\overline{m} - k, \overline{m} + k]) \leq \frac{1}{2}$.

Let $C = [-\overline{m} - k - \lambda, \overline{m} + k + \lambda]$ for some small $\lambda > 0$. Note that the set C is petite since l_t has full support for all t and C is compact. The core idea of our proof is to define $N(L)$ as the number of periods required for an initial state L to decay to within $[-\lambda, \lambda]$ if a sequence of uninformative signals is observed. (Of course, informative signals must be observed over these $N(L_t)$ periods in equilibrium.) Equation 19 implies that there is a 0.5 probability that these $N(L_t)$ signals yield $L_{t+N(L_t)} \in C$. In this event, we have a return time $\tau_C \leq N(L_t)$. In the complementary event, we provide a finite upper bound W on the expected number of periods required for L_t to hit $[-\lambda, \lambda]$ if a sequence of uninformative signals is observed. Again, there is less than a 0.5 chance that the signals observed in these periods push L outside of C. Repeating this argument, it is easy to see that $E_L\left[\tau_c\right] \leq 0.5N(L) + 2W$.

Let $N(L)$ be defined by

$$N(L) = \lceil log(\|L\|/\lambda)/\log(1/\alpha)\rceil$$

where $\lceil r\rceil$ denotes the smallest greater integer. This definition implies

$$\alpha^{N(L)} * L \leq \lambda$$

We then have that if $L_t = L$ that

$$L_{t+N(L)} = \alpha^{N(L)} * L + S_{N(L)} < \lambda + S_{N(L)}$$

From equation 19 we have $p(L_{t+N(L)} \in C) > \frac{1}{2}$.

[28] $E_L[\tau_C]$ refers to the expected hitting time of a set C from initial point L.

However, suppose $\left\|S_{N(L)}\right\| > \overline{m}+k$. Then we need to compute the expected time until the next probability $\frac{1}{2}$ event of entering C (i.e. $N(L_{t+N(L)})$ periods pass). The expected length is

$$\begin{aligned}
E[N(L_{t+N(L)})|\left\|S_{N(L)}\right\| &> \overline{m}+k] < \frac{E[\log(\left\|L_{t+N(L)}\right\|/\lambda)|\left\|S_{N(L)}\right\| > \overline{m}+k]}{\log(1/\alpha)}+1 \\
&= \frac{E[\log(\left\|\alpha^{N(L)}L+S_{N(L)}\right\|/\lambda)|\left\|S_{N(L)}\right\| > \overline{m}+k]}{\log(1/\alpha)}+1 \\
&\le \frac{E[\log(\alpha^{N(L)}\left\|L\right\|+\left\|S_{N(L)}\right\|/\lambda)|\left\|S_{N(L)}\right\| > \overline{m}+k]}{\log(1/\alpha)}+1 \\
&< \frac{E[\log(1+\left\|S_{N(L)}\right\|/\lambda)|\left\|S_{N(L)}\right\| > \overline{m}+k]}{\log(1/\alpha)}+1
\end{aligned}$$

Using the concavity of the logarithm function, we can write

$$\begin{aligned}
E[N(L_{t+N(L)})|\left\|S_{N(L)}\right\| &< \frac{E[\log(\left\|S_{N(lL)}\right\|/\lambda)|\left\|S_{N(L)}\right\| > \overline{m}+k]}{\log(1/\alpha)}+1+\frac{1}{\log(1/\alpha)} \\
&= \frac{E[\log\left\|S_{N(L)}\right\| \mid \left\|S_{N(L)}\right\| > \overline{m}+k]-\log\lambda}{\log(1/\alpha)}+1+\frac{1}{\log(1/\alpha)}
\end{aligned}$$

Now we apply Jensen's inequality to obtain

$$\begin{aligned}
E[N(L_{t+N(L)})|\left\|S_{N(L)}\right\| &\le \frac{\log E[\left\|S_{N(L)}\right\| \mid \left\|S_{N(L)}\right\| > \overline{m}+k]-\log\lambda}{\log(1/\alpha)}+1+\frac{1}{\log(1/\alpha)} \\
&\le \frac{\log\sum_{i=0}^{N(L)}\alpha^i E[\left\|l_{N(L)-i}\right\| \mid \left\|S_{N(L)}\right\| > \overline{m}+k]-\log\lambda}{\log(1/\alpha)}+1+\frac{1}{\log(1/\alpha)}
\end{aligned}$$

From Assumption 1 there exists $u<\infty$ such that

$$E[\left\|l_{N(L)-i}\right\| \mid \left\||S_{N(L)}\right\| > \overline{m}+k] \le E[l_{N(L)-i}| \text{ For all } 0\le i\le N(l),\ l_{N(L)-i} > \overline{m}+k] \le u$$

Using this fact, we have

$$\begin{aligned}
E[N(L)|\left\|S_{N(L)}\right\| &> \overline{m}+k] \le \frac{\log\left(u\sum_{i=0}^{N(L)}\alpha^i\right)-\log\lambda}{\log(1/\alpha)}+1 \qquad (20) \\
&< \frac{\log\left(\frac{u}{1-\alpha}\right)-\log\lambda}{\log(1/\alpha)}+1
\end{aligned}$$

where we denote this final quantity W and note that $W<\infty$. In other words, starting at L_t, there is a 50% chance that after $N(L_t)$ periods the chain has returned to C (i.e., if $\left\|S_{N(L)}\right\| \le \overline{m}+k$) and a 50% chance it has not (i.e., if $\left\|S_{N(L)}\right\| > \overline{m}+k$). In the later case, there is a 50% chance that after W periods the chain has returned to C (i.e.,if $\left\|S_W\right\| \le \overline{m}+k$) and a 50% chance it has not (i.e.,if $\left\|S_W\right\| > \overline{m}+k$). We can recursively use this argument since the length W is independent of L_t or $N(L_t)$. This then yields:

$$E_L[\tau_c] \leq \frac{1}{2}N(L) + \frac{1}{2}\sum_{i=1}^{\infty}\left(\frac{1}{2}\right)^i iW = \frac{1}{2}N(L) + 2W \tag{21}$$

Since $N(L) < \log((\overline{m} + k + \lambda)/\lambda)/\log(1/\alpha)$ for all $L \in C$, equation 21 yields $E_L[\tau_c] < \infty$ for all $L \in C$.

■

Lemma 3 *L_t is Harris recurrent.*

Proof. To prove Harris recurrence, we use theorem 9.1.7, statement (ii), of Meyn and Tweedie (1993). This theorem requires us to show that there exists a compact set C such that $p(L_{t+\tau} \in C$ for some $\tau \geq 0) = 1$ for any $L \in \mathbb{R}$. We borrow a great deal of the notation, including the definition of the set C, from Lemma 2.

The argument used in Lemma 2 proves our claim for $L \in C$ since $E_L[\tau_c] < \infty$ implies $p(L_{t+\tau} \in C$ for some $\tau \geq 0) = 1$. Consider $L \notin C$, and note that after $N(L)$ uninformative signals we have $L \in [-\lambda, \lambda] \subset C$. Equation 21 continues to bound $E_L[\tau_c]$ for $L \notin C$, so $p(L_{t+\tau} \in C$ for some $\tau \geq 0) = 1$ for any $L \in \mathbb{R}$.[29] Therefore, we conclude that L_t is Harris Recurrent. ■

■

Proposition 9 *Let Assumption (A′) hold. For any $\delta \in (0, 1)$, the following hold:*

1. *If $\mu^* > 0$ sufficiently small, then in equilibrium the SRPs always buy and the strategic LRP always shirks.*

2. *If $\mu^* < 1$ sufficiently large and $T < \infty$, then in equilibrium the SRPs always refuse to buy and the strategic LRP always shirks.*

Proof. Consider our first claim. Define γ as follows:

$$\gamma = min\{\log\left(\frac{\phi_0}{1-\phi_0}\right), 0\} - \frac{l^*}{1-\alpha}$$

From equation 6 we have

$$\log\left(\frac{\phi_t}{1-\phi_t}\right) \geq \alpha^t \log\left(\frac{\phi_0}{1-\phi_0}\right) - \frac{l^*}{1-\alpha} \geq \gamma$$

which in turn means that

$$\phi_t \geq \frac{e^\gamma}{1+e^\gamma}$$

If the LRP always shirks, then the probability that the LRP invests in any given period is ϕ_t. Therefore, if $\mu^* \leq \frac{e^\gamma}{1+e^\gamma}$, then the SRP always purchases and the optimal choice for a strategic LRP is to shirk.

[29] We required a uniform bound on $E_L[\tau_c]$ in Lemma 2. We can prove such a uniform bound for $L \in C$, but not for $L \notin C$.

Now consider our second claim. Define χ as follows:

$$\chi = max\{\log\left(\frac{\phi_0}{1-\phi_0}\right), 0\} + \frac{l^*}{1-\alpha}$$

From equation 6 we have

$$\log\left(\frac{\phi_t}{1-\phi_t}\right) \leq \alpha^t \log\left(\frac{\phi_0}{1-\phi_0}\right) + \frac{l^*}{1-\alpha} \leq \chi$$

which in turn means that

$$\phi_t \leq \frac{e^\chi}{1+e^\chi}$$

If the LRP shirks, then the probability that the LRP invests in any given period is ϕ_t. Therefore, if $\mu^* \geq \frac{e^\chi}{1+e^\chi}$, then the SRP will not purchase (again, assuming the LRP shirks in equilibrium).

Consider period T. Since this is the last period, it is clearly optimal for the LRP to shirk. Given this, if $\mu^* \geq \frac{e^\chi}{1+e^\chi}$, then it is optimal for the SRP to not purchase. Now consider period $T-1$. Since the SRP will not purchase in period T, it is optimal for the LRP to shirk in period $T-1$, and thus it is optimal for the SRP to not purchase given $\mu^* \geq \frac{e^\chi}{1+e^\chi}$. Continuing this backward induction, we find that in equilibrium it must be the case that the LRP shirks in every period and the SRP never purchases.

■

Proposition 10 *For any θ_1, $\theta_2 \in \Theta$ and prior $p(\theta_1), p(\ \theta_2) \in (0,1)$, we have that as $N \to \infty$*

$$\frac{p^\psi_\alpha(\theta_1|S_N)}{p^\psi_\alpha(\theta_2|S_N)} \underset{a.s}{\rightarrow} \frac{p^\psi(\beta|\theta_1)}{p^\psi(\beta|\theta_2)} \left(\frac{p(\theta_1)}{p(\theta_2)}\right)^\alpha$$

Proof. Let β denote the asymptotic frequency of a in S_N as $N \to \infty$. From the Law of Large Numbers, we have

$$\frac{\int_{\beta\in[0,1]} p_{S_N}(S_N|\beta)p^\psi(\beta|\theta_1)d\beta}{\int_{\beta\in[0,1]} p_{S_N}(S_N|\beta)p^\psi(\beta|\theta_2)d\beta} \underset{a.s}{\rightarrow} \frac{p^\psi(\beta|\theta_1)}{p^\psi(\beta|\theta_2)}$$

Combining this our formula for $p^\psi_\alpha(\theta|S_N)$ yields the desired result. ■

B Unneglected Prior Beliefs and Persistent Theories

We now provide a more complete analysis of how Peggy—an agent that always gives her $t=0$ prior beliefs full weight, but neglects the signals she observed in the past as newer signals arrive—does and does not differ from Saki.

For the most part, the evolution of Peggy's beliefs will be qualitatively similar to those of Saki described in Section 3, but replacing "flat" priors with her maintained priors each period. For example, in the case of i.i.d. signals we can describe Peggy's beliefs using

$$\frac{p_\alpha(\theta|(s_\tau)_{\tau=1}^t)}{p_\alpha(\widetilde{\theta}|(s_\tau)_{\tau=1}^t)} = \frac{p(\theta)}{p(\widetilde{\theta})} \prod_{\tau=1}^{t} \left(\frac{p(s_\tau|\theta)}{p(s_\tau|\widetilde{\theta})}\right)^{\alpha^{(t-\tau)}}.$$

In analogy with equation 6, we can describe the beliefs of Peggy using log-likelihoods:

$$\ln \frac{p_\alpha(\theta|(s_\tau)_{\tau=1}^t)}{p_\alpha(\widetilde{\theta}|(s_\tau)_{\tau=1}^t)} = \sum_{\tau=1}^t \alpha^{t-\tau} * l_\tau(\theta,\widetilde{\theta}) + l_0(\theta,\widetilde{\theta})$$

The bound on the range of Peggy's beliefs can be formulated as

$$l_0(\theta,\widetilde{\theta}) - \frac{1}{1-\alpha}\underline{L} \leq \ln \frac{p_\alpha(\theta|(s_\tau)_{\tau=1}^t)}{p_\alpha(\widetilde{\theta}|(s_\tau)_{\tau=1}^t)} \leq l_0(\theta,\widetilde{\theta}) + \frac{1}{1-\alpha}\overline{L} \tag{22}$$

where $\underline{L} = \min_{\theta,\theta'} l_\tau(\theta,\widetilde{\theta})$ and $\overline{L} = \max_{\theta,\theta'} l_\tau(\theta,\widetilde{\theta})$. It is clear that Peggy's beliefs are always bounded away from certainty and fail to converge.

It is also possible for a weak signal to have a moderating effect for Peggy just as in the case of Saki. However, Proposition 1 does not hold as stated for Peggy since Peggy's posterior matches a Bayesian's after a single signal has been collected. Instead, the following modified proposition holds[30]:

Proposition 1' *Choose any $\alpha < 1$. Consider two hypothesis θ and θ' and suppose signal s_1 has been observed. Then:*

1. *For all signals s_2 where $1 < \frac{p(s_2|\theta)}{p(s_2|\theta')} < \infty$, there exists $p(\theta|s_1), p(\theta'|s_1) \in (0,1)$ such that $\frac{p_\alpha(\theta|s_1,s_2)}{p_\alpha(\theta'|s_1,s_2)} < \frac{p(\theta|s_1)}{p(\theta'|s_1)}$.*
2. *For all $p(\theta|s_1) > p(\theta'|s_1)$, there exists $z > 1$ such that for all signals s_2 where $\frac{p(s_2|\theta)}{p(s_2|\theta')} < z$, $\frac{p_\alpha(\theta|s_1,s_2)}{p_\alpha(\theta'|s_1,s_2)} < \frac{p(\theta|s_1)}{p(\theta'|s_1)}$.*

Proposition 2 on subadditivity will continue to hold for Peggy as it required only an arbitrarily small degree of neglect of her prior beliefs. So long as Peggy has observed previous signals, the neglect of these prior signals will generate subadditivity. Proposition 3 on the conjunction fallacy for events $A \subset B$ may not hold as it requires that Saki neglect the fact that $p(A) < p(B)$, and nearly complete neglect may be required if A is much less likely than B. If the prior beliefs are not neglected at all, this condition can fail to hold.

The analysis of Sections 4 and 5 turned on the nonconvergence of beliefs, which is still true for Peggie. However, the formulas must change to account for the persistent prior belief. For example, the description of when Saki's beliefs exhibit more prediction momentum than θ^* provided by Proposition 5 would need to include terms reflecting the prior beliefs.

Much of the analysis of the persuasion problem in Section 6 continues to hold if we make slight modifications to our model. One would need to include at least three potential signals: a very favorable signal (g), a moderately favorable signal (m), and an unfavorable signal (b). Formally, we require

$$\frac{p(g|\theta=G)}{p(g|\theta=B)} > \frac{p(m|\theta=G)}{p(m|\theta=B)} > 1 > \frac{p(b|\theta=G)}{p(b|\theta=B)}.$$

Since Peggy does not discount her prior beliefs, in a one period example there is no reason for the persuader not to reveal a g or m signal. Now consider a persuader that has the option of revealing

[30] The claims are stated to maximize ease of comparison with Proposition 1. The assumptions could be reframed in terms of whether $p(s_1|\theta)/p(s_1|\theta')$ was sufficiently large relative to $p(s_2|\theta)/p(s_2|\theta')$. We omit the proof, which is a straightforward modification of the proof of Proposition 1.

two successive signals. Again, it is straightforward to show that the persuader has an incentive to reveal any g or m signal in the first period. In the second period, it is also straightforward to show that any g signal is released since the only cost of such a revelation is to cause Peggy to (potentially) discount a weakly less favorable signal observed in period 1. On the other hand, if the persuader released a g signal in the first period, it may be in the persuader's interest to withold an m signal. This can occur if the neglect of the first g signal reduces the probability Peggy places on $\theta = G$ more than than the m signal increases it.

The first result in our reputation application, that the short-run player's never learn whether or not the long-run player is strategic (Proposition 8), hinged on two properties of Saki's beliefs. First, when provided a sequence of uninformative signals, Saki neglects everything she has learned and her beliefs approach a uniform distribution. Even though Peggy's beliefs would approach her prior in the same situation, this is immaterial for our proof.[31] The second feature of our model that makes the proof possible is that an infinite sequence of signals has the same effective informational content for Saki as a finite sequence of signals would have for Tommy. Equation 22 shows that this is also true for Peggy, and so only slight modifications of our proof have to be made. Proposition 9 continues to hold as written since it is formulated in terms of the probability that the short-run player places on the long-run player invests. However, the formula for this probability must now account for the effect of the prior beliefs on the bounds on Peggy's beliefs as per equation 22.

[31] Formally, the set C defined in the proof would have to center on Peggy's prior belief that the long-run player is strategic rather than around $\phi_t = 0.5$.