Managing Self-Confidence*

Markus M. Möbius
Microsoft Research, University of Michigan and NBER

Muriel Niederle
Stanford University and NBER

Paul Niehaus
UC San Diego and NBER

Tanya S. Rosenblat
University of Michigan

August 20, 2014

Abstract

Stylized evidence suggests that people process information about their own ability in a biased manner. We provide a precise characterization of the nature and extent of these biases. We directly elicit experimental subjects' beliefs about their relative performance on an IQ quiz and track the evolution of these beliefs in response to noisy feedback. Our main result is that subjects update as if they misinterpret the information content of signals, but then process these misinterpreted signals like Bayesians. Specifically, they are *asymmetric*, over-weighting positive feedback relative to negative, and *conservative*, updating too little in response to both positive and negative feedback. These biases are substantially less pronounced in a placebo experiment where ego is not at stake, suggesting they are motivated rather than cognitive. Consistent with Bayes' rule, on the other hand, updating is *invariant* to priors (and over time) and priors are *sufficient statistics* for past information. Based on these findings, we build a model that theoretically derives the *optimal bias* of a decision-maker with ego utility and show that it naturally gives rise to both asymmetry and conservatism as *complementary* strategies in self-confidence management.

JEL Classification: C91, C93, D83
Keywords: economic experiments, Bayes' rule, asymmetric belief updating, conservatism

*We are grateful to Nageeb Ali, Roland Benabou, Gary Chamberlain, Rachel Croson, Gordon Dahl, David Eil, Glenn Ellison, Asen Ivanov, John List, Justin Rao, Al Roth, Joel Sobel, Lise Vesterlund, Roberto Weber, and participants at numerous seminars for their feedback. Aislinn Bohren and Hanzhe Zhang provided outstanding research assistance. Jenő Pál provided very helpful comments on the final draft. Niederle and Rosenblat are grateful for the hospitality of the Institute for Advanced Study where part of this paper was written. We thank the National Science Foundation, Harvard University and Wesleyan University for financial support. Niehaus acknowledges financial support from an NSF Graduate Research Fellowship.

1 Introduction

Many important economic decisions require us to assess our own abilities relative to others: we have to decide whether to choose an easy or hard major in college, whether we are talented enough to pursue a challenging career path, or whether we should compete for a sought-after promotion. A growing body of work from social psychology and economics suggests that people are not very good at forming unbiased opinions about themselves. For example, few people tend to rate themselves in the bottom half of any skill distribution (Svenson 1981, Englmaier 2006). Such mistakes can be costly – Malmendier and Tate (2005) argue that CEOs overestimate the return of their investment projects and therefore overinvest internal funds rather than return the funds to stockholders.

These facts are difficult to reconcile with the standard assumption that economic agents use Bayes' rule to process information;[1] however, there is no agreement as to what kind of alternative theory should take its place. A number of papers have proposed various models of non-Bayesian inference that depart more or less radically from Bayes' rule. For example, Rabin and Schrag (1999) and Compte and Postlewaite (2004) modify Bayes' rule slightly by allowing decision-makers to sometimes "forget" negative feedback. Such parsimonious models preserve much of the predictive power of Bayes' rule but might be too ad-hoc and restrictive to explain non-Bayesian inference in general. At the other extreme, Akerlof and Dickens (1982) and Brunnermeier and Parker (2005) allow agents to freely choose subjective beliefs. While these models can explain most patterns in the data, we can no longer take some basic facts about belief updating for granted. For example, players might sometimes react more or less strongly to the same signal, and their current beliefs might no longer be a sufficient statistic of their signal history.

In this paper, we use experiments to inform us how to generalize Bayesian updating in a disciplined way and explain the empirical facts without losing all of its predictive power. Our paper makes three contributions. First, we formulate and experimentally test a set of properties that jointly characterize Bayes' rule. Our main finding is that subjects update as if they misinterpret signal distributions in a self-serving manner, but then process these misinterpreted signals like normal Bayesians. We therefore refer to this type of updating as *biased Bayesian* updating. Second, we identify two distinct biases in subjects' interpretation of signals. On the one hand, subjects are *asymmetric* and treat positive signals as more informative than negative signals, which clearly suggests a motivated bias. On the other hand, they are also *conservative* and interpret signals as less informative than they are. Both

[1] Benoit and Dubra (2011) argue that the early evidence from social psychology based on cross-sectional data is in fact consistent with Bayesian updating. However, recent research in economics such as Eil and Rao (2011) and Burks, Carpenter, Götte and Rustichini (2013) have addressed many of those methodological shortcomings.

biases are mitigated when subjects update about an event not related to their ability, which suggests that these are motivated rather than cognitive biases. Third, we provide a theoretical justification for thinking of asymmetry and conservatism as *complementary* strategies for agents to manage their self-confidence. Specifically, we derive the updating behavior of a decision-maker who optimally interprets the informativeness of positive and negative signals to balance the demands of ego and decision-making. We show that conservatism helps the asymmetric low-skilled agent maintain a high level of confidence while limiting the risk of committing costly errors.

We use data from an experiment with 656 undergraduate students in which we track their beliefs about their performance on an IQ quiz. We focus on IQ as it is a belief domain in which decision-making and ego may conflict. We track subjects' beliefs about scoring in the top half of performers, which allows us to summarize the relevant belief distribution in a single number, the subjective probability of being in the top half. This in turn makes it feasible to elicit beliefs in an incentive compatible way using a probabilistic crossover method: we ask subjects for what value of x they would be indifferent between receiving a payoff with probability x and receiving a payoff if their score is among the top half. Unlike the quadratic scoring rule, this mechanism is robust to risk aversion (and even to non-standard preferences provided subjects prefer a higher chance of winning a fixed prize).[2] We elicit beliefs after the quiz and then repeatedly after providing subjects with informative but noisy feedback in the form of signals indicating whether they scored in the top half and which are correct with 75% probability. We then compare belief updates in response to these signals to the Bayesian benchmark. By defining the probabilistic event of interest and data generating process unambiguously, and then isolating changes in beliefs, we eliminate confounds in earlier studies that have relied on cross-sectional data (Benoit and Dubra 2011).

We find that subjects' updating rules satisfy two basic properties of Bayes' rule. Updating is *invariant* in the sense that the change in (an appropriate function of) beliefs depends only on the information received. In particular, invariance excludes *confirmatory bias* where the responsiveness to positive feedback increases with the prior (Rabin and Schrag 1999). Subjects' priors are also *sufficient statistics* in the sense that once we condition on priors at time $t-1$, events prior to $t-1$ are not predictive of beliefs at time t. Together invariance and sufficiency imply that the evolution of beliefs $\mu_t \in [0,1]$ in response to signals $\{s_t\}$ can be written as

$$f(\mu_t) - f(\mu_{t-1}) = g(s_t) \tag{1}$$

for appropriate functions f, g. To the best of our knowledge, ours is the first test of these

[2] As Schlag and van der Weele (2009) discuss, this mechanism was also described by Allen (1987) and Grether (1992) and has since been independently discovered by Karni (2009).

structural properties of Bayes' rule.

At the same time, subjects exhibit large biases when interpreting new signals: in the notation above, g differs from that predicted by Bayes' rule. Our subjects are *conservative*, revising their beliefs by only 35% as much on average as unbiased Bayesians with the same priors would. They are also *asymmetric*, revising their beliefs by 15% more on average in response to positive feedback than to negative feedback. Strikingly, subjects who received two positive and two negative signals — and thus learned nothing — ended up significantly more confident than they began.

While asymmetry clearly seems motivated, conservatism could be a purely cognitive failing. Subjects might simply misunderstand probabilities and treat a "75% correct" signal as less informative than it is.[3] To examine this issue more closely we conduct two further tests. First, we show that agents who score well on our IQ quiz – and hence are arguably cognitively more able – are as conservative (and asymmetric) as those who score poorly. Second, we conduct a placebo experiment, structurally identical to our initial experiment except that subjects report beliefs about the performance of a "robot" rather than their own performance. Belief updating in this second experiment is significantly and substantially less conservative, suggesting that ego underlies much if not all of the conservatism in our main experiment.

Interestingly, we also find that women are significantly more conservative than men (though similarly asymmetric). One implication is that high-ability women who receive the same mix of signals as high-ability men will tend to end up less confident. While only suggestive, this may help explain gender differences in other behaviors such as entry into competitive environments (Niederle and Vesterlund 2007).

Overall our data depict subjects as "biased Bayesians" who strategically misconstrue signals. This suggests that theory can develop alternatives to Bayes rule without giving up all structure; in particular, models must retain the form of Equation 1 to match our data. Our third contribution is to derive the *optimal bias* of a decision-maker who satisfies this equation but also derives utility directly from her beliefs. Unlike other models such as Compte and Postlewaite (2004), our decision-maker can choose her responsiveness to both positive and negative feedback, which allows for both conservatism and asymmetry. We show that this model is tractable and generates both conservatism and asymmetry as *complementary* strategies to manage self-confidence.

We model an agent learning about her own ability, which can be either high or low. The agent derives *instrumental utility* from making an investment decision that pays off only if her type is high, as well as direct *belief utility* from thinking she is a high type. The model is

[3]It is well-known that Bayes' rule is an imperfect positive model even when self-confidence is not at stake. We review this literature in Section 6.

agnostic as to the source of this belief utility; it could reflect any of the various mechanisms described in the literature.[4] The tension between instrumental and belief utility gives rise to an intuitive first-best: if the agent is of high ability then she would like to learn her type for sure, while if she is a low type she would like to maintain an intermediate belief which is neither too low (as that hurts her ego) nor too high (as she will make bad decisions). For example, a mediocre driver might want to think of herself as likely to be a great driver, but not so likely that she drops her car insurance.

Over time the agent receives informative signals and uses them to update her subjective beliefs. Motivated by our experimental results, we assume she does so using Bayes' rule but allow her to adopt a potentially biased interpretation of signals. For example, a driver might interpret the fact that she has not had an accident in two years as a stronger signal of her ability than is warranted. Following Brunnermeier and Parker (2005), we consider the case where the agent commits to a bias function at an initial stage that determines how she interprets the informativeness of subsequent signals.

The model generates a tight connection between the biases we observe in our experiment. Unsurprisingly, the motivated agent prefers to update asymmetrically, putting more weight on positive than negative signals. Interestingly, she also prefers to update conservatively, responding less to any signal than an unbiased Bayesian would. The intuition is as follows: asymmetry increases the agent's mean belief in her ability in the low state of the world but also increases the variance of the low-type's beliefs, and thus the likelihood of costly investment mistakes. By also updating conservatively the agent can reduce the variance of her belief distribution in the low state of the world. Finally, the agent strictly prefers not to learn her type (is information-averse) when her confidence is low as doing so would upset the careful balance between belief and decision utility. After proving these results for a specific decision problem, we also show that the same bias is approximately optimal for other problems informed by the same data, which makes it plausible that conservative and asymmetric biases arise through a process of evolution.

Our paper touches on a number of literatures, which we discuss in more detail in Section 6. The most direct empirical antecedents are papers that test the null of Bayesian updating in ego-related settings. Eil and Rao (2011) use the quadratic scoring rule to repeatedly elicit beliefs about intelligence and beauty. They find agents' posteriors are less predictable and less sensitive to signal strength after receiving negative feedback. Burks et al. (2013) obtain an alternative test by combining cross-sectional data on beliefs and ability. Both papers reject

[4]Self-confidence may directly enhance well-being (Akerlof and Dickens 1982, Caplin and Leahy 2001, Brunnermeier and Parker 2005, Kőszegi 2006), compensate for limited self-control (Brocas and Carrillo 2000, Benabou and Tirole 2002), or directly enhance performance (Compte and Postlewaite 2004).

the Bayesian null, but do not identify specific properties that hold or fail (e.g. invariance).[5]

The rest of the paper is organized as follows. Section 2 describes the details of our experimental design, and Section 3 summarizes the experimental data. Section 4 discusses econometric methods and presents results for belief updating dynamics. Section 5 unifies the experimental results theoretically. Section 6 relates the results of this paper to the existing literature and Section 7 discusses directions for future research.

2 Experimental Design and Methodology

The experiment consisted of four stages, which are explained in detail below. During the *quiz stage*, each subject completed an online IQ test. We measured each subject's belief about being among the top half of performers both before the IQ quiz and after the IQ quiz. During the *feedback stage* we repeated the following protocol four times. First, each subject received a binary signal that indicated whether the subject was among the top half of performers and was correct with 75% probability. We then measured each subject's belief about being among the top half of performers. Overall, subjects received four independent signals, and we tracked subjects' updated beliefs after each signal. In the *information purchasing stage* we gave subjects the opportunity to purchase precise information about whether her performance put her in the top half of all performers. A sub-sample of subjects were invited one month later for a *follow-up* which repeated the feedback stage but with reference to the performance of a robot rather than to their own performance.

2.1 Quiz Stage

Subjects had four minutes to answer as many questions as possible out of 30. Since the experiment was web-based and different subjects took the test at different times, we randomly assigned each subject to one of 9 different versions of the IQ test. Subjects were informed that their performance would be compared to the performance of all other students taking the same test version. The tests consisted of standard logic questions such as:

> *Question: Which one of the five choices makes the best comparison? LIVED is to DEVIL as 6323 is to (i) 2336, (ii) 6232, (iii) 3236, (iv) 3326, or (v) 6332.*
>
> *Question: A fallacious argument is (i) disturbing, (ii) valid, (iii) false, or (iv) necessary?*

[5] In other related work, Charness, Rustichini and Jeroen van de Ven (2011) find that updating about own relative performance is noisier than updating about objective events. Grossman and Owens (2010) do not find evidence of biased updating about *absolute* performance in smaller sample of 78 subjects.

A subject's final score was the number of correct answers minus the number of incorrect answers. Earnings for the quiz were the score multiplied by \$0.25. During the same period an unrelated experiment on social learning was conducted and the combined earnings of all parts of all experiments were transferred to subjects' university debit cards at the end of the study. Since earnings were variable and not itemized (and even differed across IQ tests), it would have been very difficult for subjects to infer their relative performance from earnings.

Types. We focus on subjects' learning about whether or not they scored above the median for their particular IQ quiz. Because these "types" are binary, a subject's belief about her type at any point in time is given by a single number, her subjective probability of being a high type. This will prove crucial when devising incentives to elicit beliefs, and distinguishes our work from much of the literature where only several moments of more complicated belief distributions are elicited.[6]

2.2 Feedback Stage

Signal Accuracy. Signals were independent and correct with probability 75%: if a subject was among the top half of performers, she would get a "Top" signal with probability 0.75 and a "Bottom" signal with probability 0.25. If a subject was among the bottom half of performers, she would get a Top signal with probability 0.25 and a Bottom signal with probability 0.75. To explain the accuracy of signals over the web, subjects were told that the report on their performance would be retrieved by one of two "robots" — "Wise Bob" or "Joke Bob." Each was equally likely to be chosen. Wise Bob would correctly report Top or Bottom. Joke Bob would return a random report using Top or Bottom with equal probability. We explained that this implied that the resulting report would be correct with 75% probability.

Belief elicitation. To elicit beliefs we use a *crossover* mechanism.[7] Subjects were presented with two options,

1. Receive \$3 if their score was among the top half of scores (for their quiz version).

2. Receive \$3 with probability $x \in \{0, 0.01, 0.02, ..., 0.99, 1\}$.

and asked for what value of x they would be indifferent between them. We then draw a random number $y \in \{0, 0.01, 0.02, ..., 0.99, 1\}$ and pay subjects \$3 with probability y when $y > x$ and

[6] For example, Niederle and Vesterlund (2007) elicit the mode of subjects' beliefs about their rank in groups of 4.

[7] To the best of our knowledge, ours is the first paper to implement the crossover mechanism in an experiment. After running our experiment we became aware that the same mechanism was also independently discovered by Allen (1987) and Grether (1992), and has since been proposed by Karni (2009).

otherwise pay them \$3 if their own score was among the top half.[8] We explained that subjects would maximize their probability of earning \$3 by choosing their own subjective probability of being in the top half as the threshold. We also told subjects that we would elicit beliefs several times but would implement only one choice at random for payment.

Our elicitation procedure is truth-inducing under two assumptions. First of all, subjects' preferences have to satisfy the monotonicity axiom in the sense that among lotteries that pay \$3 with probability q and \$0 with probability $1-q$, they strictly prefer those with higher q. This property holds for von-Neumann-Morgenstern preferences as well as for many non-standard models such as Prospect Theory. Monotonicity is a substantially weaker assumption than risk-neutrality which is a requirement for the widely-used quadratic scoring rule.[9] A second and more subtle condition for truth-telling is that subjects' marginal utility of earning \$3 has to be independent of whether they are in the top half. This condition is required because we explicitly allow subjects to derive direct utility from the state of the world (being in the top half). The property holds, for example, if subjects' utility from money and from being in the top half are additively separable.[10]

Moreover, the crossover mechanism does not generate perverse incentives to hedge quiz performance. Consider a subject who has predicted she will score in the top half with probability $\hat{\mu}$. Let S denote her score and F her subjective beliefs about the median score $\overline{S}$. Under quadratic scoring she will earn a piece rate of \$0.25 per point she scores and lose an amount proportional to $(I_{S\geq\overline{S}} - \hat{\mu})^2$, so her expected payoff as a function of S is

$$\$0.25 \cdot S - k \cdot \int_{\overline{S}} (I_{S\geq\overline{S}} - \hat{\mu})^2 dF(\overline{S}) \tag{2}$$

for some $k > 0$. For low values of $\hat{\mu}$ this may be *decreasing* in S, generating incentives to "hedge." In contrast, her expected payoff under the crossover mechanism is

$$\$0.25 \cdot S + \$3.00 \cdot \hat{\mu} \cdot \int_{\overline{S}} I_{S\geq\overline{S}} dF(\overline{S}), \tag{3}$$

which unambiguously increases with S. Intuitively, conditional on her own performance being the relevant one (which happens with probability $\hat{\mu}$), she always wants to do the best she can.

[8] To explain this mechanism in a simple narrative form, we told subjects that they were paired with a "robot" who had a fixed but unknown probability y between 0 and 100% of scoring among the top half of subjects. Subjects could base their chance of winning \$3 on either their own performance or their robot's, and had to indicate the threshold level of x above which they preferred to use the robot's performance.

[9] See Offerman, Sonnemans, Van de Kuilen and Wakker (2009) for an overview of the risk problem for scoring rules and a proposed risk-correction. One can of course eliminate distortions entirely by not paying subjects, but unpaid subjects tend to report inaccurate and incoherent beliefs (Grether 1992).

[10] We thank Larry Samuelson for pointing this out to us.

2.3 Follow-up Stage

We invited a random sub-sample of subjects by email to a follow-up experiment one month later. Subjects were told they had been paired with a robot who had a probability θ of being a high type. We then repeated the feedback stage of the experiment except that this time subjects received signals of the robot's ability and we tracked their beliefs about the robot being a high type.

The purpose of this follow-up was to compare subjects' processing of information about a robot's ability as opposed to their own ability. To make this comparison as effective as possible we matched experimental conditions in the follow-up as closely as possible to those in the baseline. We set the robot's initial probability of being a high type, θ, to the multiple of 5% closest to the subject's post-IQ quiz confidence. For example, if the subject had reported a confidence level of 63% after the quiz we would pair the subject with a robot that was a high type with probability $\theta = 65\%$. We then randomly picked a high or low type robot for each subject with probability θ. If the type of the robot matched the subject's type in the earlier experiment then we generated the same sequence of signals for the robot. If the types were different, we chose a new sequence of signals. In either case, signals were correctly distributed conditional on the robot's type.

3 Data

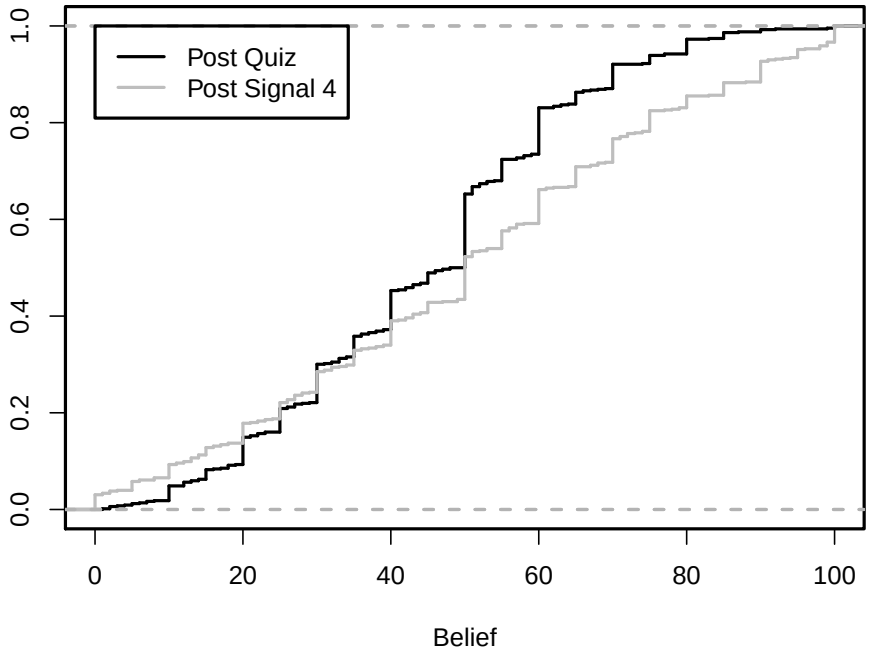
3.1 Subject Pool

The experiment was conducted in April 2005 as part of a larger sequence of experiments at a large private university with an undergraduate student body of around 6,400. A total of 2,356 students signed up in November 2004 to participate in this series of experiments by clicking a link on their home page on `www.facebook.com`, a popular social networking site.[11] These students were invited by email to participate in the belief updating study, and 1,058 of them accepted the invitation and completed the experiment online. The resulting sample is 45% male and distributed across academic years as follows: 26% seniors, 28% juniors, 30% sophomores, and 17% freshmen. Our sample includes about 33% of all sophomores, juniors, and seniors enrolled during the 2004–2005 academic year, and is thus likely to be unusually representative of the student body as a whole.

An important issue with an online experiment is how well subjects understood and were willing to follow instructions. In anticipation of this issue our software required subjects to

[11] In November 2004 more than 90% of students were members of the site and at least 60% of members logged into the site daily.

Figure 1: Belief Distributions



Empirical CDFs of subjects' beliefs after the quiz (Post Quiz) and after four rounds of feedback (Post Signal 4).

make an active choice each time they submitted a belief and allowed them to report beliefs clearly inconsistent with Bayesian updating, such as updates in the *wrong direction* and *neutral updates* (reporting the same belief as in the previous round). After each of the 4 signals, a stable proportion of about 36% of subjects reported the same belief as in the previous round.[12] About 16% of subjects did not change their beliefs at all during all four rounds of the feedback stage. In contrast, the share of subjects who updated in the wrong direction declined over time (13%, 9%, 8% and 7%), and most subjects made at most one such mistake.[13] Our primary analysis uses the restricted sample of subjects who made no updates in the wrong direction and revised their beliefs at least once. These restrictions exclude 25% and 13% of our sample, respectively, and leave us with 342 women and 314 men. While they potentially bias us against rejecting Bayes' rule, and in particular against finding evidence of conservatism, we implement them to ensure that our results are not driven by subjects who misunderstood or ignored the instructions. Our main conclusions hold on the full sample as well and we provide those estimates as robustness checks where appropriate.

To preview overall updating patterns, Figure 1 plots the empirical cumulative distribution

[12]The exact proportions were 36%, 39%, 37% and 36% for the four rounds, respectively.

[13]Overall, 19% of subjects made only one mistake, 6% made two mistake, 2% made 3 mistakes and 0.4% made 4 mistakes.

function of subjects' beliefs both directly after the quiz and after four rounds of updating. Updating yields a flatter distribution as mass shifts towards 0 (for low types) and 1 (for high types). Note that the distribution of beliefs is reasonably smooth and not merely bunched around a few focal numbers. This provides some support for the idea that the crossover elicitation method generates reasonable answers.[14]

We invited 120 subjects to participate in the follow-up stage one month later, and 78 completed this final stage of the experiment. The pattern of wrong and neutral moves was similar to the first stage of the experiment. Slightly fewer subject made neutral updates (28% of all updates) and 10% always made neutral updates. Slightly more subjects made wrong updates (22% made one mistake, 10% made two mistakes, 5% made three mistakes and 3% made 4 mistakes). The restricted sample for the follow-up has 40 subjects.

3.2 Quiz Scores

The mean score of the 656 subjects was 7.4 (s.d. 4.8), generated by 10.2 (s.d. 4.3) correct answers and 2.7 (s.d. 2.1) incorrect answers. The distribution of quiz scores (number of correct answers minus number of incorrect answers) is approximately normal, with a handful of outliers who appear to have guessed randomly. The most questions answered by a subject was 29, so the 30-question limit did not induce bunching at the top of the distribution. Table S-3 in the supplementary appendix provides further descriptive statistics broken down by gender and by quiz type. An important observation is that the 9 versions of the quiz varied substantially in difficulty, with mean scores on the easiest version (#6) fives time higher than on the hardest version (#5). Subjects who were randomly assigned to harder quiz versions were significantly less confident that they had scored in the top half after taking the quiz, presumably because they attributed some of their difficulty in solving the quiz to being a low type.[15] We will exploit this variation below, using quiz assignment as an instrument for beliefs.

4 Information Processing

We next compare subjects' observed belief updating to the Bayesian benchmark. It is easy to see that they differ starkly: if we regress subjects' logit-beliefs on those predicted by Bayes' rule, for example, we estimate a correlation of 0.57, significantly different from unity. Our goal however is to make progress by identifying the specific properties of Bayes' rule that hold or fail in the data. We therefore proceed by characterizing those properties.

[14] Hollard, Massoni and Vergnaud (2010) compare beliefs obtained using several elicitation procedures and show that using the crossover procedure results in the smoothest distribution of beliefs.

[15] Moore and Healy (2008) document a similar pattern.

As a convention, we will denote Bayesian belief at time t after receiving the t^{th} signal with μ_t and the agent's corresponding subjective (possibly non-Bayesian) belief with $\hat{\mu}_t$. For the case of binary signals (as in our experiment), we can write Bayes' rule in terms of the logistic function as

$$\text{logit}(\mu_t) = \text{logit}(\mu_{t-1}) + I(s_t = H)\lambda_H + I(s_t = L)\lambda_L \tag{4}$$

where $I(s_t = H)$ is an indicator for whether the t^{th} signal was "High", λ_H is the log likelihood ratio of a high signal, and so on. In our experiment we have $\lambda_H = -\lambda_L = \ln(3)$.

Note first that Bayes' rule satisfies *invariance* in the sense that the change in (logit) beliefs depends only on past signals. Formally, we call an updating process invariant if we can write

$$\text{logit}(\hat{\mu}_t) - \text{logit}(\hat{\mu}_{t-1}) = g_t(s_t, s_{t-1}, \ldots) \tag{5}$$

for some sequence of functions g_t that do not depend on $\hat{\mu}_{t-1}$. Next, Bayes' rule implies that the posterior $\hat{\mu}_{t-1}$ is a *sufficient statistic* for information received prior to t, so that we can write $g_t(s_t, s_{t-1}, \ldots) = g_t(s_t)$. Morever this relationship is *stable* across time, so that $g_t = g$ for all t. We think of these three properties – invariance, sufficiency and stability – as defining the core structure of Bayesian updating; they greatly reduce the potential complexity of information processing. Any updating process that satisfies them in our setting can be fully characterized by two parameters, since with binary signals $g(s_t)$ can take on at most two values. We therefore write

$$g(s_t) = I(s_t = H)\beta_H\lambda_H + I(s_t = L)\beta_L\lambda_L \tag{6}$$

The parameters β_H and β_L describe the *responsiveness* of the agent relative to a Bayesian updater, for whom $\beta_H = \beta_L = 1$.

Our empirical model nests Bayesian updating and allows us to test for the core properties of Bayesian updating (invariance, sufficiency and stability) as well measure the responsiveness to positive and negative information. The simplest version is:

$$\text{logit}(\hat{\mu}_{it}) = \delta\text{logit}(\hat{\mu}_{i,t-1}) + \beta_H I(s_{it} = H)\lambda_H + \beta_L I(s_{it} = L)\lambda_L + \epsilon_{it} \tag{7}$$

The coefficient δ equals 1 if the invariance property holds, while the coefficients β_H and β_L capture responsiveness to positive and negative information, respectively. The error term ϵ_{it} captures unsystematic errors that subject i made when updating her belief at time t. Note that we do not have to include a constant in this regression because $I(s_{it} = H) + I(s_{it} = L) = 1$. To test for stability we estimate (7) separately for each of our four rounds of updating and test whether our coefficient estimates vary across rounds. Finally, to examine whether prior beliefs are a sufficient statistic we augment the model with indicators $I(s_{i,t-\tau} = H)$ for lagged

signals on the right-hand side:

$$\text{logit}(\hat{\mu}_{it}) = \delta \text{logit}(\hat{\mu}_{i,t-1}) + \beta_H I(s_{it} = H)\lambda_H + \beta_L I(s_{it} = L)\lambda_L + \sum_{\tau=1}^{t-1} \beta_{t-\tau}[I(s_{i,t-\tau} = H)\lambda_H + I(s_{i,t-\tau} = L)\lambda_L] + \epsilon_{it} \quad (8)$$

Sufficiency predicts that the lagged coefficients $\beta_{t-\tau}$ are zero.

Identifying (7) and (8) is non-trivial because we include lagged logit-beliefs (that is, priors) as a dependent variable. If there is unobserved heterogeneity in subjects' responsiveness to information, β_L and β_H, then OLS estimation may yield upwardly biased estimates of δ due to correlation between the lagged logit-beliefs and the unobserved components $\beta_{iL} - \beta_L$ and $\beta_{iH} - \beta_H$ in the error term. Removing individual-level heterogeneity through first-differencing or fixed-effects estimation does not solve this problem but rather introduces a negative bias (Nickell 1981). In addition to these issues, there may be measurement error in self-reported logit-beliefs because subjects make mistakes or are imprecise in recording their beliefs.[16]

To address these issues we exploit the fact that subjects' random assignment to different versions of the IQ quiz generated substantial variation in their post-quiz beliefs. This allows us to construct instruments for lagged prior logit-beliefs. For each subject i we calculate the average quiz score of subjects *other* than i who took the same quiz variant to obtain a measure of the quiz difficulty level that is not correlated with subject i's own ability but highly correlated with the subject's beliefs. We report both OLS and IV estimates of Equation 7.

4.1 Invariance, Sufficiency and Stability

Table 1 presents round-by-round and pooled estimates of Equation 7.[17] Estimates in Panel A are via OLS and those in Panel B are via IV using quiz type indicators as instruments. The F-statistics reported in Panel B indicate that our instrument is strong enough to rule out weak instrument concerns (Stock and Yogo 2002).

Result 1 (Invariance)**.** *Subjects' updating behavior is invariant to their prior.*

[16] See Arellano and Honore (2001) for an overview of the issues raised in this paragraph. Instrumental variables techniques have been proposed that use lagged difference as instruments for contemporaneous ones (see, for example, Arellano and Bond (1991)); these instruments would be attractive here since the theory clearly implies that the first lag of beliefs should be a sufficient statistic for the entire preceding sequence of beliefs, but unfortunately higher-order lags have little predictive power when the autocorrelation coefficient δ is close to one, as Bayes' rule predicts.

[17] The logit function is defined only for priors and posteriors in $(0, 1)$; to balance the panel we further restrict the sample to subjects i for whom this holds for *all* rounds t. Results using the unbalanced panel, which includes another 101 subject-round observations, are essentially identical.

Table 1: Conservative and Asymmetric Belief Updating

	Regressor	Round 1	Round 2	Round 3	Round 4	All Rounds	Unrestricted
Panel A: OLS							
	δ	0.814	0.925	0.942	0.987	0.924	0.888
		$(0.030)^{***}$	$(0.015)^{***}$	$(0.023)^{***}$	$(0.022)^{***}$	$(0.011)^{***}$	$(0.014)^{***}$
	β_H	0.374	0.295	0.334	0.438	0.370	0.264
		$(0.019)^{***}$	$(0.017)^{***}$	$(0.021)^{***}$	$(0.030)^{***}$	$(0.013)^{***}$	$(0.013)^{***}$
	β_L	0.295	0.274	0.303	0.347	0.302	0.211
		$(0.025)^{***}$	$(0.020)^{***}$	$(0.022)^{***}$	$(0.024)^{***}$	$(0.012)^{***}$	$(0.011)^{***}$
	$\mathbb{P}(\beta_H = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
	$\mathbb{P}(\beta_L = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
	$\mathbb{P}(\beta_H = \beta_L)$	0.009	0.408	0.305	0.017	0.000	0.000
	N	612	612	612	612	2448	3996
	R^2	0.803	0.890	0.875	0.859	0.854	0.798
Panel B: IV							
	δ	0.955	0.882	1.103	0.924	0.963	0.977
		$(0.132)^{***}$	$(0.088)^{***}$	$(0.125)^{***}$	$(0.124)^{***}$	$(0.059)^{***}$	$(0.060)^{***}$
	β_H	0.407	0.294	0.332	0.446	0.371	0.273
		$(0.044)^{***}$	$(0.017)^{***}$	$(0.023)^{***}$	$(0.035)^{***}$	$(0.012)^{***}$	$(0.013)^{***}$
	β_L	0.254	0.283	0.273	0.362	0.294	0.174
		$(0.042)^{***}$	$(0.026)^{***}$	$(0.030)^{***}$	$(0.040)^{***}$	$(0.017)^{***}$	$(0.027)^{***}$
	$\mathbb{P}(\beta_H = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
	$\mathbb{P}(\beta_L = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
	$\mathbb{P}(\beta_H = \beta_L)$	0.056	0.725	0.089	0.053	0.001	0.004
	First Stage F-statistic	13.89	16.15	12.47	12.31	16.48	20.61
	N	612	612	612	612	2448	3996
	R^2	-	-	-	-	-	-

Notes:

1. Each column in each panel is a regression. The outcome in all regressions is the log posterior odds ratio. δ is the coefficient on the log prior odds ratio; β_H and β_L are the estimated effects of the log likelihood ratio for positive and negative signals, respectively. Bayesian updating corresponds to $\delta = \beta_H = \beta_L = 1$.
2. Estimation samples are restricted to subjects whose beliefs were always within $(0, 1)$. Columns 1-5 further restrict to subjects who updated their beliefs at least once and never in the wrong direction; Column 6 includes subjects violating this condition. Columns 1-4 examine updating in each round separately, while Columns 5-6 pool the 4 rounds of updating.
3. Estimation is via OLS in Panel A and via IV in Panel B, using the average score of other subjects who took the same (randomly assigned) quiz variety as an instrument for the log prior odds ratio.
4. Heteroskedasticity-robust standard errors in parenthesis; those in the last two columns are clustered by individual. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Invariance implies that the change in (logit) beliefs should not depend on the prior, or equivalently, that the responsiveness to positive and negative information is not a function of the prior. This implies that a coefficient $\delta = 1$ on prior logit-beliefs in Equation 7. The OLS estimate is close to but significantly less than unity; although it climbs by round, we fail to reject equality with one only in Round 4 ($p = 0.57$). These estimates may be biased upward by heterogeneity in the responsiveness coefficients, β_{iL} and β_{iH}, or may be biased downwards if subjects report beliefs with noise. The IV estimates suggest that the latter bias is more important: the pooled point estimate of 0.963 is larger and none of the estimates are significantly different from unity.

Of course, it is possible that both β_H and β_L are functions of prior logit-beliefs but that the effects cancel out to give an average estimate of $\delta = 1$. To address this possibility, Table S-5 reports estimates of an augmented version of Equation 7 that includes an interaction between the (logit) prior and the high signal $I(s_{it} = H)$. Invariance requires that the coefficient δ_H on this interaction is zero; our estimated δ_H varies in sign across rounds and is significant at the 5% level only once, in the OLS estimate for Round 1. It is small and insignificant in our pooled estimates using both OLS and by IV. All told, subjects' updating appears invariant.

Result 2 (Sufficiency)**.** *Controlling for prior beliefs, lagged information does not significantly predict posterior beliefs.*

Priors appear to be fully incorporated into posteriors – but do they fully capture what subjects have learned in the past? Table 2 reports instrumental variables estimates of Equation 8, which includes lagged signals as predictors. We can include one lag in round 2, two lags in round 3, and three lags in round 4. None of the estimated coefficients are statistically or economically significant, supporting the hypothesis that priors properly encode past information.

Result 3 (Stability)**.** *The structure of updating is largely stable across rounds.*

We test for stability by comparing the coefficients δ, β_H, and β_L across rounds. Our (preferred) IV estimates in Table 1 show some variation but without an obvious trend. Wald tests for heterogeneous coefficients are mixed; we reject the null of equality for β_H ($p < 0.01$) but not for β_L ($p = 0.24$) or for δ ($p = 0.52$). We view these results as suggestive but worth further investigation.

4.2 Conservatism and Asymmetry

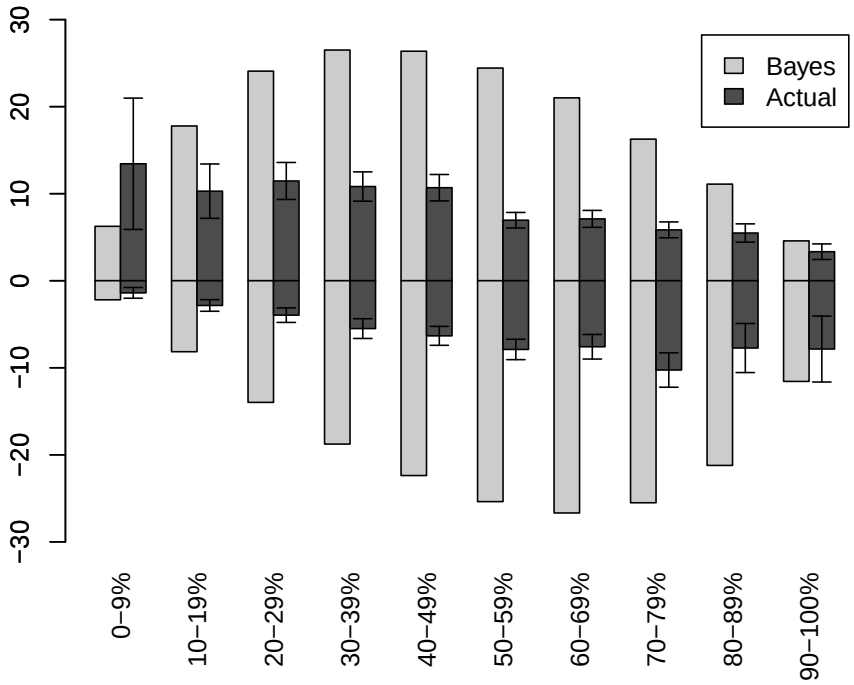
Result 4 (Conservatism)**.** *Subjects respond less to both positive and negative information than an unbiased Bayesian.*

Table 2: Priors are Sufficient for Lagged Information

Regressor	Round 2	Round 3	Round 4
δ	0.872	1.124	0.892
	$(0.100)^{***}$	$(0.158)^{***}$	$(0.152)^{***}$
β_H	0.284	0.348	0.398
	$(0.023)^{***}$	$(0.031)^{***}$	$(0.041)^{***}$
β_L	0.284	0.272	0.343
	$(0.028)^{***}$	$(0.031)^{***}$	$(0.028)^{***}$
β_{-1}	0.028	-0.027	0.045
	(0.037)	(0.051)	(0.051)
β_{-2}		-0.036	0.067
		(0.052)	(0.055)
β_{-3}			0.057
			(0.058)
N	612	612	612
R^2	-	-	-

Each column is a regression. The outcome in all regressions is the log posterior odds ratio. Estimated coefficients are those on the log prior odds ratio (δ), the log likelihood ratio for positive and negative signals (β_H and β_L), and the log likelihood ratio of the signal received τ periods earlier ($\beta_{-\tau}$). The estimation sample includes subjects whose beliefs were always within $(0, 1)$ and who updated their beliefs at least once and never in the wrong direction. Estimation is via IV using the average score of other subjects who took the same (randomly assigned) quiz variety as an instrument for the log prior odds ratio. Heteroskedasticity-robust standard errors in parenthesis. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure 2: Conservatism



Mean belief revisions broken down by decile of prior belief in being of type "Top." Responses to positive and negative signals are plotted separately in the top and bottom halves, respectively. The corresponding means that would have been observed if all subjects were unbiased Bayesians are provided for comparison. T-bars indicate 95% confidence intervals.

The OLS estimates of β_H and β_L reported in Table 1, 0.370 and 0.302, are substantially and significantly less than unity. Round-by-round estimates do not follow any obvious trend. The IV and OLS estimates are similar, suggesting there is limited bias in the latter through correlation with lagged prior beliefs.

To ensure that this result is not merely an artifact of functional form, Figure 2 presents a complementary non-parametric analysis of conservatism. The figure plots the mean belief revision in response to a Top and Bottom signal by decile of prior belief in being a top half type for each of the four observations of the 656 subjects, with the average Bayesian response plotted alongside for comparison. Belief revisions are consistently smaller than those those implied by Bayes' rule across essentially all of these categories.

Result 5 (Asymmetry)**.** *Controlling for prior beliefs, subjects respond more to positive than to negative signals.*

To quantify asymmetry we compare estimates of β_H and β_L, the responsiveness to positive and negative signals, from Table 1. The difference $\beta_H - \beta_L$ is consistently positive across all rounds and significantly different from zero in the first round, fourth round, and for the

pooled specification. While estimates of this difference in Rounds 2 and 3 are not significantly different from zero, we cannot reject the hypothesis that the estimates are equal across all four rounds ($p = 0.32$). The IV estimates are somewhat more variable but are again uniformly positive, and significantly so in Rounds 1 and 4 and in the pooled specification. The size of the difference is substantial, implying that the effect of receiving both a positive and a negative signal (that is, no information) is 26% as large as the effect of receiving only a positive signal.[18]

Figure 3 presents the analogous non-parametric analysis; it compares subjects whose prior belief was $\hat{\mu}$ and who received positive feedback with subjects whose prior belief was $1 - \hat{\mu}$ and who received negative feedback. According to Bayes' rule, the magnitude of the belief change in these situations should be identical. Instead subjects consistently respond more strongly to positive feedback across deciles of the prior. As an alternative non-parametric test we can also examine the net change in beliefs among the 224 subjects who received two positive and two negative signals. These subjects should have ended with the same beliefs as they began; instead their beliefs increased by an average of 4.8 points ($p < 0.001$).

To summarize, Bayes' rule seems to do a good job of describing the basic structure of updating, but an imperfect job predicting how subjects weigh new information. These patterns motivate the modeling approach we lay out in Section 5 below. Note also that deviations from Bayes' rule were costly within the context of the experiment. Comparing expected payoffs given observed updating (π_{actual}) to those subjects' would have earned if they updated using Bayes' rule (π_{Bayes}) or if they did not update at all ($\pi_{noupdate}$), we find that the ratio $\frac{\pi_{Bayes}-\pi_{actual}}{\pi_{Bayes}-\pi_{noupdate}}$ is 0.59. Non-Bayesian updating behavior thus cost subjects 59% of the potential gains from processing information within the experiment.

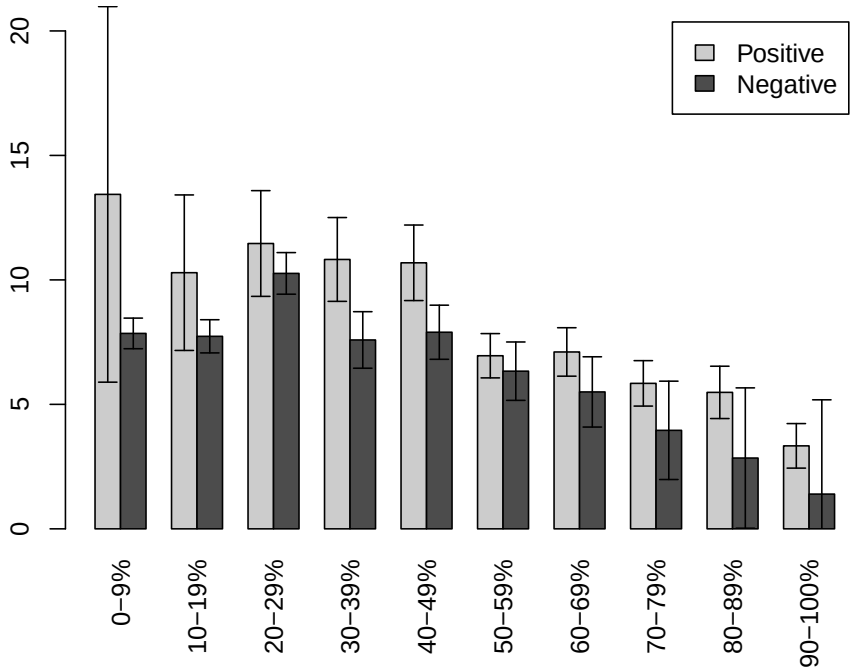
4.3 Confidence Management or Cognitive Mistakes?

Our data suggest that subjects update like Bayesians but with conservative and asymmetric biases. While asymmetry seems to reflect motivation, conservatism could plausibly be a cognitive failing. Conservatism might arise, for example, if subjects simply misinterpret the informativeness of signals and believe that the signal is only correct with 60% probability instead of 75%. Subjects might underweight signals in this way because they are used to encountering weaker ones in everyday life.

We present two pieces of evidence that suggest that conservatism is not only a cognitive error. First, we show that conservatism (and asymmetry) do not correlate with the cognitive ability of participants. Specifically, we assess whether biases are present both among high

[18]Table S-4 in the supplementary appendix shows that the results of the regression continue to hold when we pool all four rounds of observation, even when we eliminate all observations in which subjects do not change their beliefs. That is, the effect is not driven by an effect of simply not updating at all.

Figure 3: Asymmetry



Mean absolute belief revisions by decile of prior belief in being of type equal to the signal received. For example, a subject with prior belief $\hat{\mu} = 0.8$ of being in the top half who received a signal T and a subject with prior belief $\hat{\mu} = 0.2$ who received a signal B are both plotted at $x = 80\%$. T-bars indicate 95% confidence intervals.

performers (those that score in the top half) and low performers on the IQ quiz. Table 3a reports estimates of Equation 7 differentiated by ability. We find no evidence that more able (higher performing) participants update differently than less able participants: they do not differ in the way they weight their priors or in the way they incorporate positive and negative signals. This suggests that cognitive errors are not the main factor behind conservatism.

The second analysis that helps distinguish motivated behavior from a cognitive errors interpretation is to examine the results of the follow-up experiment, in which a random subset of subjects performed an updating task that was formally identical to the one in the original experiment, but which dealt with the ability of a robot rather than their own ability. For these subjects we pool the updating data from both experiments and estimate:

$$\begin{aligned} \text{logit}(\hat{\mu}^e_{it}) - \text{logit}(\hat{\mu}^e_{it}) = \beta_H \cdot I(s_{it} = H)\lambda_H + \beta_L \cdot I(s_{it} = L)\lambda_L + \\ + \beta_H^{Robot} \cdot 1(e = \text{Robot}) \cdot I(s_{it} = H)\lambda_H + \beta_L^{Robot} \cdot 1(e = \text{Robot}) \cdot I(s_{it} = L)\lambda_L + \epsilon_i^t \end{aligned} \tag{9}$$

Here, e indexes experiments (Ego or Robot), so that the interaction coefficients β_H^{Robot} and β_L^{Robot} tell us whether subjects process identical information differently across both treatments.

Table 3: Heterogeneity in Updating

(a) Heterogeneity by Ability

Regressor	OLS	IV
δ	0.918	0.966
	$(0.015)^{***}$	$(0.075)^{***}$
δ^{Able}	0.010	-0.002
	(0.022)	(0.138)
β_H	0.381	0.407
	$(0.026)^{***}$	$(0.050)^{***}$
β_L	0.317	0.296
	$(0.016)^{***}$	$(0.034)^{***}$
β_H^{Able}	-0.017	-0.048
	(0.030)	(0.054)
β_L^{Able}	-0.041	-0.011
	(0.025)	(0.049)
N	2448	2448
R^2	0.854	-

(b) Heterogeneity by Gender

Regressor	OLS	IV
δ	0.925	0.988
	$(0.015)^{***}$	$(0.103)^{***}$
δ^{Male}	-0.007	-0.047
	(0.023)	(0.125)
β_H	0.331	0.344
	$(0.017)^{***}$	$(0.031)^{***}$
β_L	0.280	0.258
	$(0.015)^{***}$	$(0.040)^{***}$
β_H^{Male}	0.080	0.063
	$(0.027)^{***}$	$(0.038)^{*}$
β_L^{Male}	0.052	0.073
	$(0.026)^{**}$	$(0.044)^{*}$
N	2448	2448
R^2	0.855	-

Each column is a separate regression. The outcome in all regressions is the log belief ratio. δ, β_H, and β_L are the estimated effects of the prior belief and log likelihood ratio for positive and negative signals, respectively. δ^j, β_H^j, and β_L^j are the differential responses attributable to being male ($j = Male$) or high ability ($j = Able$). Robust standard errors clustered by individual reported in parentheses. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Given the smaller sample available we impose $\delta = 1$ and estimate via OLS. Table 4 reports results.

Result 6. *Conservatism is significantly reduced when subjects learn about a robot's performance rather than their own performance.*

The baseline coefficients β_H and β_L are similar to their estimated values for the larger sample (see Table 1), suggesting that participation in the follow-up was not selective on updating traits. The interaction coefficients are both positive and significant — they imply that subjects are roughly twice as responsive to feedback when it concerns a robot's performance as they are when it concerns their own performance. In fact, we cannot reject the hypothesis that $\beta_H + \beta_H^{Robot} = 1$ ($p = 0.13$), though we can still reject $\beta_L + \beta_L^{Robot} = 1$ ($p = 0.004$). While conservatism does not entirely vanish, it is clearly much weaker. Interestingly, subjects are also less asymmetric in relative terms when they update about robot performance $\left(\frac{\beta_H}{\beta_L} > \frac{\beta_H + \beta_H^{Robot}}{\beta_L + \beta_L^{Robot}}\right)$. We cannot reject the hypothesis that they update symmetrically about robot performance such that $\beta_H + \beta_H^{Robot} = \beta_L + \beta_L^{Robot}$ ($p = 0.45$).

Table 4: Belief Updating: Own vs. Robot Performance

Regressor	I	II	III
β_H	0.426	0.349	0.252
	(0.087)***	(0.066)***	(0.043)***
β_L	0.330	0.241	0.161
	(0.050)***	(0.042)***	(0.033)***
β_H^{Robot}	0.362	0.227	0.058
	(0.155)**	(0.116)*	(0.081)
β_L^{Robot}	0.356	0.236	-0.006
	(0.120)***	(0.085)***	(0.089)
$\mathbb{P}(\beta_H + \beta_H^{Robot} = 1)$	0.128	0.000	0.000
$\mathbb{P}(\beta_L + \beta_L^{Robot} = 1)$	0.004	0.000	0.000
$\mathbb{P}(\beta_H = \beta_L)$	0.302	0.118	0.039
$\mathbb{P}(\beta_H + \beta_H^{Robot} = \beta_L + \beta_L^{Robot})$	0.454	0.316	0.030
N	160	248	480
R^2	0.567	0.434	0.114

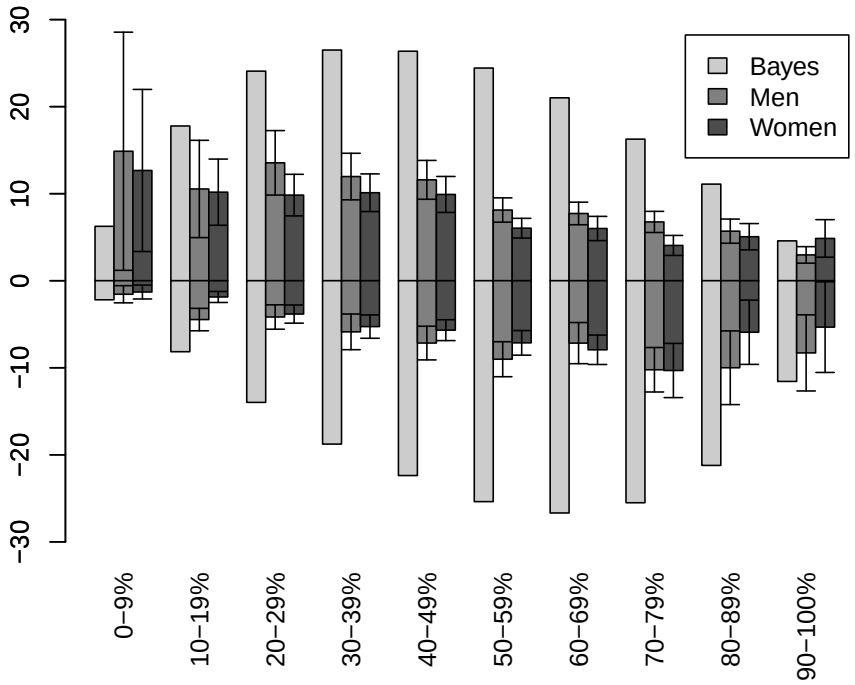
Each column is a separate regression. The outcome in all regressions is the change in the log belief ratio. β_H and β_L are the estimated effects of the log likelihood ratio for positive and negative signals, respectively. β_H^{Robot} and β_L^{Robot} are the differential response attributable to obtaining a signal about the performance of a robot as opposed to about one's own performance. Estimation samples are restricted to subjects who participated in the follow-up experiment and observed the same sequence of signals as in the main experiment. Column I includes only subjects who updated at least once in the correct direction and never in the wrong direction in both experiments. Column II adds subjects who never updated their beliefs. Column III includes all subjects. Robust standard errors clustered by individual reported in parentheses. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4.4 Gender Differences

We next turn to the question whether updating bias differs by demographic type. Heterogenous updating bias can potentially explain why different types of people can make different career choices even though they have comparable skills and preferences. Gender is a particularly relevant type: a growing body of research has shown that women tend to be less competitive than men (controlling for skill) and a large share of this difference can be explained by the relative *underconfidence* of high-ability women compared to high-ability men.

We find that women in our sample are significantly more conservative than men. This is evident in a simple graph: Figure 4 plots the mean belief revision of men and women in response to a Top and Bottom signal by decile of prior belief in being a top half type for each of the four observations of the 656 subjects, with the average Bayesian response plotted alongside for comparison. Women's belief revisions are smaller than men's in most cases and in both directions. We next test for gender effects formally by estimating our

Figure 4: Conservatism by Subject Gender



Mean belief revisions of men and women broken down by decile of prior belief in being of type "Top." Responses to positive and negative signals are plotted separately in the top and bottom halves, respectively. The corresponding means that would have been observed if all subjects were unbiased Bayesians are provided for comparison. T-bars indicate 95% confidence intervals.

empirical model (Equation 7) differentiated by gender, with results reported in Table 3b. Men are substantially less conservative than women, reacting significantly more to both positive and negative feedback and 21% more to feedback on average (23% when estimated by IV). However, we do not find significant differences in asymmetry: OLS and IV point estimates of $\frac{\beta_H+\beta_H^{Male}}{\beta_L+\beta_L^{Male}} - \frac{\beta_H}{\beta_L}$ are 0.05 and -0.10, respectively, and neither is significantly different from zero ($p = 0.64, 0.74$).

One implication of this pattern is that high-ability women who receive the same mix of signals as high-ability men will end up relatively *underconfident* due to their greater conservatism. This could help explain why, for example, women are less likely than men to select into competitive environments. Moreover, if precise feedback is less subject to updating biases than noisy feedback then the precision of feedback could itself feed into gender differences – a possibility that we leave for future investigation.

5 Optimally Biased Bayesian Updating

While subjects incorporate new signals similar to Bayesians in our experiment (in the sense that belief dynamics satisfies invariance, sufficiency and stability), they also interpret new information conservatively and asymmetrically. In this section we show that these biases arise naturally in a model that posits only invariance, sufficiency and stability.

Consider an agent who is of high type H with probability μ_0 and otherwise a low type L. The binary types reflect our experimental design where a subject is either "scoring in the top half" or not. There are T discrete time periods in each of which the agent receives a signal s_t about her ability. The agent aggregates the stream of signals up to time t into a *subjective belief* $\hat{\mu}_t$. We allow the agent's belief to differ from the *objective probability* μ_t derived using Bayes' rule. The agent balances two objectives when forming biased subjective beliefs: she wants to make good instrumental decisions, but also cares about her ego and wants to believe that she is a high type.

We first define instrumental and belief utility formally and derive the agent's *optimal* beliefs if she could choose them freely. We then derive the updating behavior of optimally biased Bayesians who manage their self-confidence. We find that agents apply both a conservative and asymmetric bias to signals. At the optimum high types learn their type quickly and with probability approaching 1 as they receive more signals. Low types, on the other hand, exhibit a "downward neutral bias": their updating biases render their logit-belief a driftless random walk, allowing them to maintain a moderate level of self-confidence even as they receive many signals. We also show that the bias function is approximately optimal even if the agent's instrumental and belief utility changes, which lets us think of the optimal bias as an evolutionary adjustment.

5.1 Utility and Optimal Beliefs

We start with instrumental utility. With equal probability, nature selects one of the T time periods as the "investment period". In this period the agent must decide whether or not to take an action that yields a positive payoff if and only if her type is high. For example, the agent might consider investing in the stock market and has to decide if she is a skilled investor, or she might consider taking a challenging major in college and has to decide whether she is smart enough. Formally, the agent can make an investment which pays 1 in the final period T if she is of high type or 0 otherwise.[19] The investment has a cost $c \in [0, 1]$ which is drawn from a well behaved continuous distribution $G \in C^2[0, 1]$ at the time of the decision. Not investing

[19]The assumption that the instrumental value of investing is realized in the last period simplifies our calculation of belief utility because the agent only learns her type in the final period and therefore manages her belief utility over all time periods $1 \leq t \leq T$.

gives a payoff of 0. The optimal decision of a Bayesian decision maker is thus to invest if and only if $c < \mu_t$. Consistent with the results of our second experiment, we assume that a biased agent behaves *as if she were a Bayesian* and invests iff $c < \hat{\mu}_t$. Hence, biasing updating is costly because it leads to worse decisions.

The agent also derives direct *belief utility* $b(\hat{\mu}_t)$ in period t from her subjective belief, where $b \in C^2[0,1]$ is a well-behaved, strictly increasing function normalized such that $b(0) = 0$. The model is agnostic over the various kinds of belief utility discussed in the literature; to capture them in a reduced-form way we make no assumptions about the shape of $b(\cdot)$ other than monotonicity.[20] The combined objective function of the agent is the sum of her average belief utility and her expected instrumental utility:

$$U(\hat{\mu}_0, .., \hat{\mu}_T) = \frac{1}{T} \sum_{t=1}^{T} \left[\underbrace{b(\hat{\mu}_t)}_{\text{belief utility}} + \underbrace{\int_0^{\hat{\mu}_t} (\mu_t - c)\, dG(c)}_{\text{instrumental utility}} \right] \tag{10}$$

When $b(\hat{\mu}) = 0$ the agent has no belief utility and behaves like a classical economic agent. Note that because payoffs are time-averaged T serves as a measure of the information-richness of the environment. In stating results we will make use of the notion of *relative time* $\tau \in [0,1]$ which we associate with absolute time $\lfloor \tau T \rfloor$.

To build intuition it will be useful to study the per-period expected utility of the low and high type agents, which we denote $L(\hat{\mu}_t)$ and $H(\hat{\mu}_t)$:

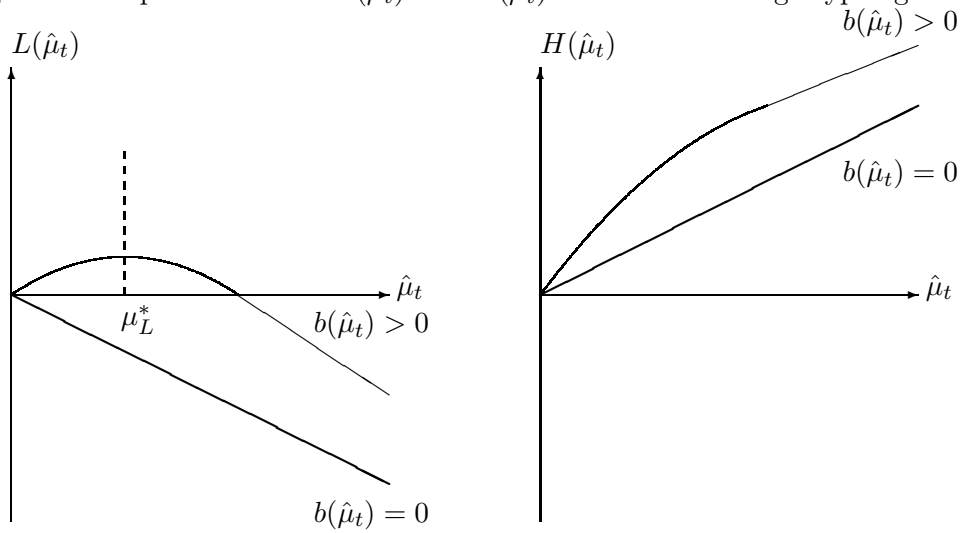
$$\begin{aligned} L(\hat{\mu}_t) &= b(\hat{\mu}_t) - \int_0^{\hat{\mu}_t} c dG(c) \\ H(\hat{\mu}_t) &= b(\hat{\mu}_t) + \int_0^{\hat{\mu}_t} (1-c) dG(c) \end{aligned} \tag{11}$$

Suppose for now that agents of low and high type could *choose* subjective beliefs μ_L^* and μ_H^* to maximize these respective expressions. As Figure 5 illustrates, the high type agent would always choose $\mu_H^* = 1$ because both her belief and instrumental utility are increasing in her subjective belief. The optimal (and possibly non-unique) μ_L^* for the low type agent depends on $b(\cdot)$, however: an agent without belief utility chooses $\mu_L^* = 0$ while an agent with ego concerns may choose $\mu_L^* > 0$. We focus on the interesting case $\mu_L^* > 0$ in which the low-type agent prefers on net to hold an inflated belief.[21] We also restrict attention to decision problems with

[20] In our model, subjective beliefs will converge for most time periods as $T \to \infty$. Other models in the literature analyze settings with few feedback periods where subjective beliefs remain noisy and hence the concavity or convexity of the belief utility function matters (see for example Kőszegi (2006)).

[21] It is not difficult to come up with conditions such that $\mu_L^* > 0$. For example, any linear belief utility function

Figure 5: Per-period utilities $L(\hat{\mu}_t)$ and $H(\hat{\mu}_t)$ of the low and high type agents



$L(1) < 0$ which implies $\mu_L^* < 1$, or in other words that the low-type agent would not want to convince herself that she was the high type. While this extreme form of bias is conceivable in situations where there are no real stakes (or belief utilities are large), it generates no interesting predictions.

5.2 Optimal Biased Bayesian Updating

Agents receive a stream of i.i.d. signals in each period t. A signal can take finitely many values which we index by k ($1 \leq k \leq K$) with distribution F_H in the high state and F_L in the low state. Let $\lambda_k = \log(F_H(k)/F_L(k))$ be the log-likelihood ratio for realization k. Every signal realization is informative such that $\lambda_k \neq 0$. Motivated by our experimental results, we assume that agents update their belief as biased Bayesians whose updating process satisfies invariance, sufficiency and stability.

Definition 1. *A biased Bayesian updating process consists of an initial subjective prior $\hat{\mu}_0$ and an updating rule*

$$logit(\hat{\mu}_{t+1}) = logit(\hat{\mu}_t) + \beta_k \lambda_k \tag{12}$$

where $\beta_k \geq 0$.

We refer to β as the *responsiveness function* and to $\tilde{\beta}_k = \beta_k / \max_k \beta_k$ as the *normalized responsiveness*.[22] Biased Bayesian updating encompasses standard Bayesian updating as a

will suffice. We know that $L(0) = 0$ and $L(1) < 0$. Moreover, for small x we have $L(x) > 0$ because G' is continuous and hence bounded and therefore $\int_0^x c dG(c) \leq \int_0^x c \max_{c\in[0,1]}(G'(c))\, dc = \frac{1}{2}(x)^2 \max_{c\in[0,1]}(G'(c))$.

[22] The normalized responsiveness is only defined for responsiveness functions which are not zero everywhere.

special case ($\hat{\mu}_0 = \mu_0$ and $\beta_k = 1$) while capturing the idea that the agent may downplay or overstate the informativeness of certain kinds of feedback. Following Brunnermeier and Parker (2005), we say that a biased Bayesian updating process is *optimal* if it maximizes expected total utility (10) among all such processes.[23] When the agent has no belief utility the optimum is, reassuringly, to be unbiased.

Proposition 1. *Let $T \geq 2$. The optimal biased Bayesian updating process for an agent without belief utility ($b(\hat{\mu}) = 0$ for all $\hat{\mu}$) is Bayes' rule: $\hat{\mu}_0 = \mu_0$ and $\beta_k = 1$ for all k.*

To characterize the case with belief utility we introduce the notions of *conservatism* and *downward neutral bias*, which is a strong form of *asymmetry*.

Definition 2. *A biased Bayesian updating process is **conservative** if the agent always responds less to new information than an unbiased Bayesian ($\max_k \beta_k < 1$). It exhibits a **downward neutral bias** (DNB) if $\sum_k F_L(k)\tilde{\beta}_k\lambda_k = 0$.*

DNB implies that the agent's expected logit-belief remains unchanged if the state is low; the agent essentially interprets the stream of information as white noise. DNB is a generalized notion of *asymmetry*: in the binary signals case, if H (L) denotes the signal with the higher (lower) log-likelihood ratio, DNB implies $\beta_H > \beta_L$.

Proposition 2. *The optimal updating process has the following features: (1) $\beta_k^T \to 0$ as $T \to \infty$ for all k so that the agent updates conservatively for large T; (2) $\sum_k F_L(k)\tilde{\beta}_k^T\lambda_k \to 0$ as $T \to \infty$ so that the agent exhibits DNB for large T; (3) if moreover the low type's optimal belief μ_L^* is unique and $L''(\mu_L^*) < 0$ then $\hat{\mu}_0^T \to \mu_L^*$; (4) for any relative time $\tau > 0$ the agent's belief converges in probability to μ_L^* in the low state and to $\mu_H^* = 1$ in the high state.*

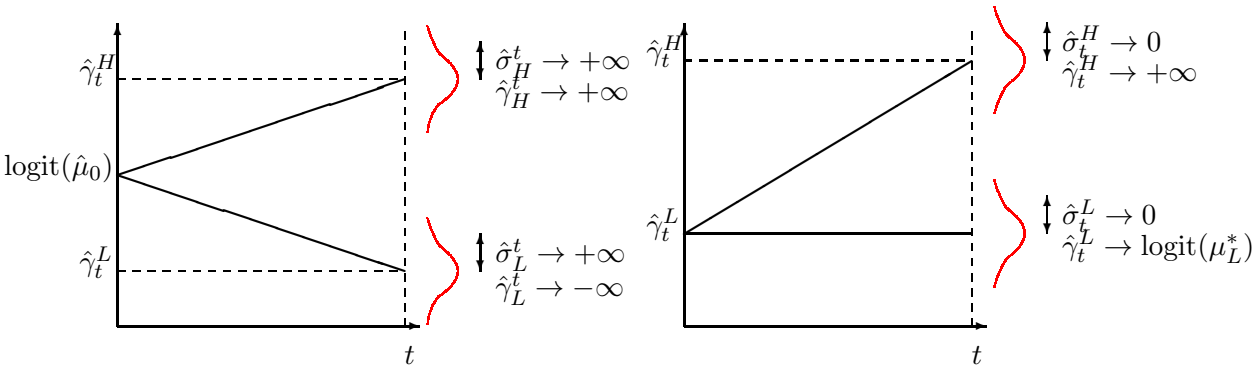
The intuition for this result can be illustrated graphically for the binary signals case. The evolution of logit-beliefs described in Equation 12 follows a random walk: in each period, the logit-belief increases by $\beta_H\lambda_H$ with probability $F_H(H)$ for the high type ($F_L(H)$ for the low type) and otherwise decreases by $\beta_L\lambda_L$. The mean logit-belief of the high type, $\hat{\gamma}_t^H$, and the variance in logit-beliefs, $\left(\hat{\sigma}_t^H\right)^2$, can hence be expressed as:

$$\begin{aligned} \hat{\gamma}_t^H &= \text{logit}(\hat{\mu}_0) + t\left[F_H(H)\beta_H\lambda_H + (1 - F_H(H))\beta_L\lambda_L\right] \\ \left(\hat{\sigma}_t^H\right)^2 &= tF_H(H)(1 - F_H(H))\left(\beta_H\lambda_H - \beta_L\lambda_L\right)^2 \end{aligned} \tag{13}$$

We can derive analogous expressions $\hat{\gamma}_t^L$ and $\left(\hat{\sigma}_t^L\right)^2$ for the mean and variance of the low type's logit-belief by replacing the probability $F_H(H)$ with $F_L(H)$. The left panel of Figure 6 shows

[23] Existence is guaranteed since (a) expected utility is continuous in $\hat{\mu}_0 \in (0,1)$ and β_k; (b) using the logic of proposition 2, one can show that there are $\epsilon > 0$ and $M > 0$ such it is never optimal to choose $\hat{\mu}_0 < \epsilon$, $\hat{\mu}_0 > 1-\epsilon$ or $\beta_k > M$. Hence, the optimal parameters live in a compact Euclidean metric space.

Figure 6: Evolution of logit-beliefs of an unbiased Bayesian (left panel) and an optimally biased Bayesian (right panel)

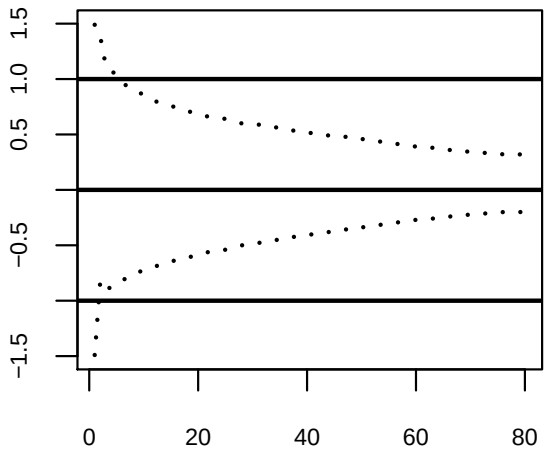


the mean logit belief of the high type (increasing solid line) and low type (decreasing solid line) when the agent is an unbiased Bayesian. Note that the mean logit beliefs of both types converge to $+\infty$ and $-\infty$ at rate t while the standard deviation increases only at rate $\sqrt{t}$. Therefore, beliefs converge to either 1 or 0 in probability.

The biased Bayesian would prefer keep her beliefs close to either 1 (in the high state) and $\mu_L^* > 0$ (in the low state). By choosing an initial belief close to her optimal low-type's belief μ_L^* and by becoming asymmetric ($\beta_H/\beta_L \uparrow$) she can slow the rate at which the low type's logit-belief drifts to $-\infty$, or even eliminate this drift alltogether by choosing a DNB. The right panel of Figure 6 illustrates this idea. Asymmetry alone is insufficient, however, without conservatism: unless the agent also reduces her responsiveness to information the variance of the low type's logit-beliefs will make it impossible to keep logit-beliefs close to μ_L^*. Although the agent's mean logit-belief in the low state stays close to μ_L^*, her realized logit-belief will typically be either very small or very large. Since $L(0) = 0$ and $L(1) < 0$ this is costly; the low-type agent would in fact be worse off than under unbiased Bayesian updating. Conservatism addresses this problem by keeping the low-type agent's beliefs close to μ_L^* in probability. The proof of Proposition 2 formalizes this intuition: it shows that any updating process that is not both conservative and downward-neutral biased must do strictly worse than a process that is, and that an optimal updating process allows the agent to closely approximate her "first best" payoffs by keeping her belief bounded away from zero at μ_L^* in the low state while still learning her type rapidly in the high state.

While Proposition 2 characterizes optimal behavior for large T, we can also characterize the finite-T case numerically. Figure 7 shows the optimal updating policy over the range $1 \leq T \leq 80$ for a binary signals example with a uniform cost distribution, an objective prior of $\mu_0 = \frac{1}{2}$ and belief utility $b(\hat{\mu}) = \frac{1}{4}\hat{\mu}$. These parameters satisfy the long-term learning

Figure 7: Numerical optima for finite T and binary signals



Plots optimal responsiveness to positive and negative signals (β_H and $-\beta_L$) for the unbiased (solid lines) and optimally biased (dotted lines) cases over $1 \leq T \leq 80$. The remaining parameters are fixed at $\mu_0 = 0.5$, $c \sim U[0,1]$, $b(\hat{\mu}) = \frac{1}{4}\hat{\mu}$, $p = 0.75$, $q = 0.25$

condition $L(1) < 0$ and imply $\mu_L^* = \frac{1}{4}$: the agent would like to maintain a confidence level of 25% in the low state. As in our experiment signals are accurate with probability 0.75. The agent is optimally asymmetric over the entire range, conservative for $T > 8$, and increasingly conservative as T increases.

5.3 Robustness of Biased Bayesian Updating

The optimal updating rule β^T that we characterized in Proposition 2 depends on the specific decision problem (summarized by per-period utilities $L(\hat{\mu})$ and $H(\hat{\mu})$). However, we can show that this dependence is weak in the following sense: if the agent faces a new decision problem $(\tilde{L}, \tilde{H})$ and continues to use the old updating rule β^T, then she can do almost as well as when she uses the new optimal updating rule $\tilde{\beta}^T$.

Proposition 3. *Fix a signal distribution (F_H, F_L). Consider two decision problems (L, H) and $(\tilde{L}, \tilde{H})$ with optimal updating rules β^T and $\tilde{\beta}^T$, respectively. Assume, that the agent uses the updating rule β^T for the latter problem. Then the agent's combined utility and subjective belief at any relative time τ converge in probability to the first-best values as $T \to \infty$.*

The result implies that the agent can do very well by applying a uniform updating bias (independent of the decision problem) and by choosing an initial subjective prior close to the low-type's optimal belief. This observation allows for the possibility of an evolutionary process

in which Nature selects an updating rule for a generic decision problem which the agent then applies to different specific problems throughout life.

6 Discussion

In this section we discuss how our results on updating fit into the existing literature on biased beliefs in psychology and economics, organized by three broad themes. First, researchers have documented *cognitive* updating biases that arise even when ego is not at stake. Second, psychologists have argued that additional biases may be *motivated* by desires to hold certain beliefs, such as high self-confidence. Third, economic theory has recently begun exploring motives and mechanisms that give rise to motivated biases.

6.1 Cognitive Limitations

Work on cognitive limits dates back at least as far as the 1960s, when psychologists began testing Bayes' rule as a positive model in ego-neutral settings (see Slovic and Lichtenstein (1971), Fischhoff and Beyth-Marom (1983), and Rabin (1998) for reviews). A prototypical experiment involved showing subjects two urns containing 50% and 75% red balls, respectively, and then asking them to predict from which urn a sample of balls was drawn. The main conclusion from these studies was that subjects did not appropriately weight old versus new information. Early studies found conservative updating, a consensus which was then upset by Kahneman and Tversky's (1973) discovery of the "base rate fallacy," seen as "the antithesis of conservatism" (Fischhoff and Beyth-Marom 1983, 248–249). Recently Massey and Wu (2005) reconcile these results by generating both conservative and anti-conservative updating within a single experiment: their subjects underweight signals with high likelihood ratios, but overweight signals with low likelihood ratios. We complement this literature by showing that subjects are *more* conservative when updating about their own ability than about a robot's.

6.2 Motivated Biases

Our experimental results are most closely related to work on motivated biases, and in particular to a series of concepts from psychology.

Attribution bias. Social psychologists have argued that people exhibit self-serving "attribution biases," or tendencies to take credit for good outcomes and deny blame for bad ones. Attribution bias could be a mechanism for asymmetry in our experiment: subjects may attribute positive signals to performance and negative signals to noise. Yet the existing evidence on this link is ambiguous for two reasons. First, attribution and updating are distinct processes:

one can make attributions without updating or vice versa.[24] Second, psychologists themselves have argued that attribution bias studies "seem readily interpreted in information-processing terms" (Miller and Ross 1975, p. 224) either because the data-generating processes were not clearly defined (Wetzel 1982) or because key outcome variables were not objectively defined or elicited incentive-compatibly.[25] Relative to this literature our contribution is to (1) clearly define the probabilistic event (scoring in the top half) and outcome variables (subjective beliefs about the probability of that event) of interest, and (2) explicitly inform subjects about the conditional likelihood of observing different signals.[26]

Confirmatory bias. This literature argues that people over-weight information that confirms their prior views (see Rabin and Schrag (1999) for a review). We do not find evidence of confirmatory bias in our data: asymmetry is not obviously more pronounced among subjects with a more optimistic prior (Figure 3), and we cannot reject the property of invariance, which rules out confirmatory bias.

Overconfidence. Asymmetric updating induces *overconfidence*, in the sense that individuals will over-estimate their probability of succeeding at a task compared to the forecast of an unbiased Bayesian who began with the *same* prior and observed the *same* stream of signals. We emphasize this definition to contrast it with others frequently used in the literature. Findings that more than $x\%$ of a population believe that they are in the top $x\%$ in terms of some desirable trait are commonly taken as evidence of irrational overconfidence, but Zábojník (2004), Van den Steen (2004), Santos-Pinto and Sobel (2005), and Benoit and Dubra (2011) have all illustrated how such results can obtain under unbiased Bayesian information processing in cross-sectional data. Burks et al. (2013) address this criticism by studying the joint distribution of beliefs and actual ability, but do not directly measure updating. Lastly, complementary work by (Mayraz 2011) suggests that subjects tend to choose prior beliefs that are beneficial to them ("wishful thinking"), a distinct mechanism for overconfidence.

Overconfidence and Forecasting. In the finance literature, overconfidence is often used to describe an agent who is too confident in the *precision* of her forecast. For example, a stock

[24] To illustrate, consider the prototypical experimental paradigm in which subjects taught a student and then attributed the student's subsequent performance either to their teaching or to other factors. A common finding is that subjects attribute poor performances to lack of student effort, while taking credit for good performances. This is clearly consistent with the fixed beliefs that (a) student effort and teacher ability are complementary and (b) the teacher is capable.

[25] For example, Wolosin, Sherman and Till (1973) had subjects place 100 metal washers on three wooden dowels according to the degree to which they felt that they, their partner, and the situation were "responsible" for the outcome. Santos-Pinto and Sobel (2005) show that if agents disagree over the interpretation of concepts like "responsibility," this can generate positive self-image on average, and conclude that "there is a parsimonious way to organize the findings that does not depend on assuming that individuals process information irrationally..." (p. 1387).

[26] Of course, limiting ambiguity makes our test for asymmetry both unconfounded and relatively stringent, since it may be precisely in the interpretation of ambiguous concepts that agents are most biased.

analyst forming beliefs about the future returns on an asset may have overly tight posteriors if she over-estimates her ability as a forecaster (e.g. De Long, Shleifer and Waldmann (1991)). We view this alterative definition of overconfidence as another manifestation of the same underlying idea, that people overestimate their own abilities. Higher-skilled analysts presumably have a lower variance in their forecasts - hence an analyst with ego utility might adopt the belief that her forecasts are more accurate than they actually are, and recommend a non-diversified portfolio as a result.

6.3 Theoretical Models of Updating Biases

Our results lend support to recent theoretical work in behavioral economics that examines *motives* for elevating one's self-confidence. One is simply direct ego utility: people like to think well of themselves (e.g. Akerlof and Dickens (1982) and Kőszegi (2006)). Alternatively, people may derive higher *anticipatory* utility when they believe their future will be bright (Caplin and Leahy 2001, Brunnermeier and Parker 2005). Self-confidence may also help limit self-control problems or make it easier to credibly impress others (Carrillo and Mariotti 2000, Benabou and Tirole 2002).[27] Finally, confidence may directly enhance performance (Compte and Postlewaite 2004). Our results are broadly consistent with any of these motives and we remain agnostic among them; in this spirit, we incorporate demand for self-confidence into our model in a reduced-form way (Section 5).

Our emphasis in the paper is rather on the *mechanisms* that agents use to bias their beliefs. The theoretical literature has examined three broad cases. In one, theorists retain Bayes' rule in order to focus on imperfect memory (Mullainathan 2002, Benabou and Tirole 2002, Wilson 2003, Gennaioli and Shleifer 2010). Our experiment does not speak directly to this idea as it was intentionally designed to minimize forgetfulness, compressing updating into a short time period and reminding subjects of the full history of signals at each update. Subjects may well forget more over longer periods. A second category examines selective information acquisition; for example, Kőszegi (2006) studies a model in which agents with ego utility avoid information. While we focus here on updating biases in the feedback stage of our experiment – where information was costless and unavoidable – we do also find some evidence of information aversion. For brevity we discuss these results in Appendix S-1. Finally, our data most directly support a third group of papers in which agents directly manipulate their beliefs (Akerlof and Dickens 1982) or their interpretations of signals Brunnermeier and Parker (2005).

A distinct strand of behavioral theory questions the very idea of a single, monolithic "belief." Work on *ambiguity aversion* in particular assumes that agents hold multiple priors and use the most pessimistic to assess any given situation (Gilboa and Schmeidler 1989). Generally

[27] See also Burks et al. (2013) for empirical support for this hypothesis.

speaking, such models can generate updating patterns inconsistent with Bayes' rule. Because we study a binary event, however, it is still true that if a subject updates a family of priors by applying Bayes' rule to each then the most pessimistic prior must yield the most pessimistic posterior, and vice versa. Bayes' rule should still be satisfied in our data even if our subjects are ambiguity averse. In other words, our data are neither inconsistent with ambiguity aversion nor implied by it.

7 Conclusion

We use a large-scale experiment to open the black box of belief updating in a setting where ego is at stake. While we can soundly reject the hypothesis that agents use Bayesian updating, we do find empirical support for three core properties of Bayes' rule, namely invariance, sufficiency and stability. Subjects' differ from Bayes' rule in the way they interpret signals; they do so with pronounced conservative and asymmetric biases. The facts that these biases are equally prevalent among more and less able subjects and are mitigated in a placebo treatment both suggest that they arise from subjects' desire to protect their ego, rather than cognitive errors. Subjects' valuations for information are also biased, as a substantial minority – and low-confidence subjects in particular – are averse to obtaining informative feedback.

Taken together, the experimental data suggest a disciplined way for theorists to relax Bayes' rule, preserving the core properties of invariance, sufficiency and stability while allowing for biased interpretations. We pursue this approach in the second half of the paper. We find that conservatism, asymmetry, and an aversion to information all emerge naturally as optimal and complementary biases. These findings provide a potential explanation for our empirical results and illustrate how they can be incorporated into tractable, refutable theories.

References

Akerlof, George A. and William T. Dickens, "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, 1982, *72 (3)*, 307–319.

Allen, Franklin, "Discovering personal probabilities when utility functions are unknown," *Management Science*, 1987, *33* (4), 542–544.

Arellano, Manuel and Bo Honore, "Panel data models: some recent developments," in J.J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics*, Vol. 5 of *Handbook of Econometrics*, Elsevier, 2001, chapter 53, pp. 3229–3296.

___ **and Stephen Bond**, "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, April 1991, *58* (2), 277–97.

Benabou, Roland and Jean Tirole, "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics*, 2002, *117 (3)*, 871–915.

Benoit, JeanPierre and Juan Dubra, "Apparent Overconfidence," *Econometrica*, 09 2011, *79* (5), 1591–1625.

Brocas, Isabelle and Juan D. Carrillo, "The value of information when preferences are dynamically inconsistent," *European Economic Review*, 2000, *44*, 1104–1115.

Brunnermeier, Markus K. and Jonathan A. Parker, "Optimal Expectations," *American Economic Review*, September 2005, *95* (4), 1092–1118.

Burks, Stephen V., Jeffrey P. Carpenter, Lorenz Götte, and Aldo Rustichini, "Overconfidence and Social Signalling," *Review of Economic Studies*, 2013, *80* (3), 949–983.

Caplin, Andrew and John Leahy, "Psychological Expected Utility Theory And Anticipatory Feelings," *The Quarterly Journal of Economics*, February 2001, *116* (1), 55–79.

Carrillo, Juan D. and Thomas Mariotti, "Strategic Ignorance as a Self-Disciplining Device," *Review of Economic Studies*, 2000, *67*, 529–544.

Charness, Gary, Aldo Rustichini, and Jeroen van de Ven, "Overconfidence, self-esteem, and strategic deterrence," Technical Report, U.C. Santa Barbara 2011.

Compte, Olivier and Andrew Postlewaite, "Confidence-Enhanced Performance," *American Economic Review*, December 2004, *94* (5), 1536–1557.

Eil, David and Justin M. Rao, "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 2011, *3* (2), 114–38.

Eliaz, Kfir and Andrew Schotter, "Paying for Confidence: an Experimental Study of the Demand for Non-Instrumental Information," *Games and Economic Behavior*, November 2010, *70* (2), 304–324.

Englmaier, Florian, "A Brief Survey on Overconfidence," in D. Satish, ed., *Behavioral Finance – an Introduction*, ICFAI University Press, 2006.

Fischhoff, Baruch and Ruth Beyth-Marom, "Hypothesis Evaluation from a Bayesian Perspective," *Psychological Review*, 1983, *90 (3)*, 239–260.

Gennaioli, Nicola and Andrei Shleifer, "What Comes to Mind," *Quarterly Journal of Economics*, November 2010, pp. 1399–1433.

Gilboa, Itzhak and David Schmeidler, "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, April 1989, *18* (2), 141–153.

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

Grossman, Zachary and David Owens, "An Unlucky Feeling: Overconfidence and Noisy Feedback," Technical Report, UC Santa Barbara 2010.

Hollard, Guillaume, Sebastien Massoni, and Jean-Christophe Vergnaud, "Comparing three elicitation rules: the case of confidence in own performance," Technical Report, Universite Paris June 2010.

Kahneman, Daniel and Amos Tversky, "On the Psychology of Prediction," *Psychological Review*, 1973, *80* (4), 237–251.

Karni, Edi, "A Mechanism for Eliciting Probabilities," *Econometrica*, 03 2009, *77* (2), 603–606.

Kőszegi, Botond, "Ego Utility, Overconfidence, and Task Choice," *Joural of the European Economic Association*, 2006, *4* (4), 673–707.

Long, J Bradford De, Andrei Shleifer, and Robert Waldmann, "The Survival of Noise Traders in Financial Markets," *The Journal of Business*, January 1991, *64* (1), 1–19.

Malmendier, Ulrike and Geoffrey Tate, "CEO Overconfidence and Corporate Investment," *The Journal of Finance*, 2005, *60* (6), 2661–2700.

Massey, Cade and George Wu, "Detecting Regime Shifts: the Causes of Under- and Overreaction," *Management Science*, 2005, *51* (6), 932–947.

Mayraz, Guy, "Wishful Thinking," Technical Report, Paris School of Economics 2011.

Miller, Dale and Michael Ross, "Self-Serving Biases in the Attribution of Causality: Fact or Fiction?," *Psychology Bulletin*, 1975, *82* (2), 213–225.

Moore, Don A. and Paul J. Healy, "The Trouble With Overconfidence," *Psychological Review*, April 2008, *115* (2), 502517.

Mullainathan, Sendhil, "A Memory-Based Model Of Bounded Rationality," *The Quarterly Journal of Economics*, August 2002, *117* (3), 735–774.

Nickell, Stephen J, "Biases in Dynamic Models with Fixed Effects," *Econometrica*, November 1981, *49* (6), 1417–1426.

Niederle, Muriel and Lise Vesterlund, "Do Women Shy Away from Competition? Do Men Compete Too Much?," *The Quarterly Journal of Economics*, August 2007, *122* (3), 1067–1101.

Offerman, Theo, Joep Sonnemans, Gijs Van de Kuilen, and Peter Wakker, "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes," *The Review of Economic Studies*, October 2009, *76* (29), 1461–1489.

Rabin, Matthew, "Psychology and Economics," *Journal of Economic Literature*, March 1998, *36* (1), 11–46.

___ and Joel Schrag, "First Impressions Matter: A Model Of Confirmatory Bias," *The Quarterly Journal of Economics*, February 1999, *114* (1), 37–82.

Santos-Pinto, Luis and Joel Sobel, "A Model of Positive Self-Image in Subjective Assessments," *American Economic Review*, December 2005, *95* (5), 1386–1402.

Schlag, Karl and Joel van der Weele, "Eliciting Probabilities, Means, Medians, Variances and Covariances without assuming Risk Neutrality," Technical Report, Universitat Pompeu Fabr October 2009.

Slovic, Paul and Sarah Lichtenstein, "Comparison of Bayesian and Regression Approaches

to the Study of Information Processing in Judgment," *Organizational Behavior and Human Performance*, 1971, *6*, 649–744.

Stein, Charles, "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1972, p. 583602.

Stock, James H. and Motohiro Yogo, "Testing for Weak Instruments in Linear IV Regression," NBER Technical Working Papers 0284, National Bureau of Economic Research, Inc November 2002.

Svenson, Ola, "Are We All Less Risky and More Skillful Than Our Fellow Drivers?," *Acta Psychologica*, 1981, *47*, 143–148.

Van den Steen, Eric, "Rational Overoptimism (and Other Biases)," *American Economic Review*, September 2004, *94* (4), 1141–1151.

Wetzel, Christopher, "Self-Serving Biases in Attribution: a Bayesian Analysis," *Journal of Personality and Social Psychology*, 1982, *43* (2), 197–209.

Wilson, Andrea, "Bounded Memory and Biases in Information Processing," NajEcon Working Paper Reviews, www.najecon.org April 2003.

Wolosin, Robert J., Steven Sherman, and Amnon Till, "Effects of Cooperation and Competition on Responsibility Attribution After Success and Failure," *Journal of Experimental Social Psychology*, 1973, *9*, 220–235.

Zábojník, Ján, "A model of rational bias in self-assessments," *Economic Theory*, January 2004, *23* (2), 259–282.

A Proofs

A.1 Proof of Proposition 1

When $b(\hat{\mu}) = 0$ for all $\hat{\mu}$, the objective function in (10) is maximized if and only if for any possible history of signals at any time $t \leq T$ and associated Bayesian belief μ_t the following holds: $\hat{\mu}^t > c$ iff $\mu^t > c$. Since the cost distribution is continuous and positive, this implies $\hat{\mu}^t = \mu^t$ for any signal history that generates the objective Bayesian posterior μ^t. Because all signal realizations are informative (and hence occur with positive probability) we obtain for $t = 1$ already K linear equations of the form $\text{logit}(\hat{\mu}^0) + \beta_k \lambda_k = \text{logit}(\mu^0) + \lambda_k$, one for each signal realization. As we have $K+1$ unknowns we can use any of the signal realizations at time $t = 2$ – e.g. two consecutive $k = 1$ realizations – to uniquely pin down $\beta_k = 1$ and $\hat{\mu}^0 = \mu^0$.

A.2 Auxiliary Approximation Lemma

For our proofs, we will frequently exploit that logit beliefs in our model are sums of independent random variables. While these variables are i.i.d. their distribution generally depends on T (because the responsiveness function changes with T), so we cannot use the standard central limit theorem. Instead we use Stein's (1972) method to bound the approximation error of the central limit theorem in our framework.

Consider the random variable Y defined over the realizations k of a single signal:

$$Y(k) = \hat{\beta}_k \lambda_k \text{ with probability } F_L(k) \tag{14}$$

where $\hat{\beta}_k \leq 1$ is the normalized responsiveness (which implies that for at least one realization we have $\hat{\beta}_k = 1$). The following lemma will be useful:

Lemma 1. *Consider any normalized responsiveness function. Let* $k^* = \arg\min_k |\lambda_k|$. *We then have* $Var(Y) \geq F_L(k^*)\left(1 - F_L(k^*)\right)\lambda_{k^*}$.

Proof: The variance of Y is minimized over all normalized responsiveness functions if $\beta_{k^*} = 1$ and $\beta_k = 0$ for all $k \neq k^*$. This reduces Y to a simple Bernoulli random variable and the result follows.

We define two new constants:

$$\begin{aligned} M_L &= 5\left(\frac{\max_k \lambda_k}{\sqrt{F_L(k^*)\left(1 - F_L(k^*)\right)\lambda_{k^*}}}\right)^3 \\ M_H &= 5\left(\frac{\max_k \lambda_k}{\sqrt{F_H(k^*)\left(1 - F_H(k^*)\right)\lambda_{k^*}}}\right)^3 \end{aligned}$$

We can now prove the following approximation for subjectve beliefs:

Lemma 2. *Let $\epsilon > 0$ and $-\infty \leq a < b \leq \infty$. The random variable $W = \frac{logit(\hat{\mu}_{\lfloor \tau T \rfloor}) - \hat{\gamma}^L_{\lfloor \tau T \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T \rfloor}}$ satisfies:*

$$Prob(a \leq W \leq b | L) \leq \Phi(b + 2\epsilon) - \Phi(a - 2\epsilon) + \frac{M_L}{\epsilon\sqrt{\tau T}}$$

where Φ is the cdf of the normal distribution $N(0,1)$. An analogous result holds for beliefs in the high state where M_L is replaced by M_H.

Note, that the upper bound depends only on ϵ, τT and the distribution of the signal distribution but (importantly) *not* on the particular responsiveness function.

Proof: WLOG we focus on low-state beliefs only. We define the function h:[28]

$$h(x) = \begin{cases} 0 & \text{if } x < a - 2\epsilon \\ \frac{1}{2\epsilon^2}(x - a + 2\epsilon)^2 & \text{if } a - 2\epsilon \leq x < a - \epsilon \\ 1 - \frac{1}{2\epsilon^2}(x - a)^2 & \text{if } a - \epsilon \leq x < b \\ 1 & \text{if } a \leq x < b \\ 1 - \frac{1}{2\epsilon^2}(x - b)^2 & \text{if } b \leq x < b + \epsilon \\ \frac{1}{2\epsilon^2}(x - b - 2\epsilon)^2 & \text{if } b + \epsilon \leq x < b + 2\epsilon \\ 0 & \text{if } b + 2\epsilon \leq x \end{cases}$$

This function approximates the indicator function that takes value 1 on the interval $[a, b]$ such that h is bounded above by the indicator function on the interval $[a - 2\epsilon, b + 2\epsilon]$, bounded below by the indicator function on $[a, b]$ and bounded derivative $|h'(x)| \leq \frac{1}{\epsilon}$. Now we use Stein's inequality to establish

$$|\mathbb{E}[h(W)] - \mathbb{E}[h(Z)]| \leq \frac{\max_x h'(x) 5E|X_i|^3}{\sqrt{\tau T}}$$

where $Z \sim N(0,1)$ and X_i are i.i.d. random variables of the form $X = \frac{Y - E(Y)}{\sqrt{Var(Y)}}$. Thus

$$Prob(a \leq W \leq b) \leq \mathbb{E}[h(W)] \leq \mathbb{E}[h(Z)] + \frac{(\max_x h'(x))\, 5E|X_i|^3}{\sqrt{\tau T}}$$

and the result of the lemma then follows.

A.3 Uniform Downward-Neutral Bias

We define a particular responsiveness function which we call the *uniform downward neutral bias* that approximates the utility of the unrestricted agent who can freely choose her beliefs in both states of the world. This will be useful to prove proposition 2 where we show that non-conservative responsiveness functions or those which do not satisfy the DNB property cannot be optimal because they cannot approximate the utility of the unrestricted agent.

For a given signal distribution, we partition the set of possible realizations into an "Up-set" $U = \{k | \lambda_k > 0\}$ and a "Down-set" $D = \{k | \lambda_k < 0\}$. We fix a constant $\frac{1}{2} < \theta < 1$. For each T

[28] For $a = -\infty$ ($b = \infty$) we adapt the definition naturally and let $h(x) = 1$ for $x < b$ ($x > a$).

we define the following biased Bayesian updating process:

$$\begin{aligned} \hat{\mu}_0^T &= \mu_L^* \\ \beta_k &= \begin{cases} T^{-\theta} & \text{for } k \in U \\ T^{-\theta} \underbrace{\dfrac{\sum_{k\in U} F_L(k)\lambda_k}{-\sum_{k\in D} F_L(k)\lambda_k}}_{\kappa} & \text{for } k \in D \end{cases} \end{aligned} \tag{15}$$

Note, that $0 < \kappa < 1$ because the unbiased agent's expected change in logit-beliefs in the low state has to be negative (hence, $\sum_{k\in U} F_L(k)\lambda_k + \sum_{k\in D} F_L(k)\lambda_k < 0$). We can derive the mean and variance of logit-beliefs at relative time τ in both states:

$$\begin{aligned} \hat{\gamma}_{\tau T}^H &= \text{logit}(\mu_L^*) + \tau T^{1-\theta} \underbrace{\left(\sum_{k\in U} F_H(k)\lambda_k + \kappa \sum_{k\in D} F_H(k)\lambda_k \right)}_{\Gamma_H} \\ \hat{\gamma}_{\tau T}^L &= \text{logit}(\mu_L^*) \\ \left(\hat{\sigma}_{\tau T}^H\right)^2 &= \tau T^{1-2\theta} \underbrace{\left(\sum_{k\in U} F_H(k)\lambda_k^2 + \kappa^2 \sum_{k\in D} F_H(k)\lambda_k^2 - \Gamma_H^2 \right)}_{\Sigma_H > 0} \\ \left(\hat{\sigma}_{\tau T}^L\right)^2 &= \tau T^{1-2\theta} \underbrace{\left(\sum_{k\in U} F_L(k)\lambda_k^2 + \kappa^2 \sum_{k\in D} F_L(k)\lambda_k^2 \right)}_{\Sigma_L > 0} \end{aligned} \tag{16}$$

Note, that $\Gamma_H > 0$ because the unbiased agent's expected change in logit-beliefs in the high state is strictly positive (hence, $\sum_{k\in U} F_H(k)\lambda_k + \sum_{k\in D} F_H(k)\lambda_k > 0$) and $\kappa < 1$. We call this particular updating process the uniform downward-neutral bias (uniform DNB) because a uniform bias factor is applied to up and down signal realizations, respectively, and logit-beliefs for the low type follow a random walk without drift.

Lemma 3. *Assume a biased Bayesian with uniform DNB. At any relative time $\tau > 0$, the agent's high state belief converges in probability to 1 while the agent's low state belief converges in probability to μ_L^*. The total utility (10) of the agent converges to the total utility of an unrestricted agent with belief μ_L^* in the low state and belief 1 in the high state.*

Figure 6 illustrates the intuition for the lemma. In the high state, the agent's logit-belief at relative time τ is of order $\tau T^{1-\theta}$ according to (16). This expression converges to infinity. In the low state, the agent's logit-belief behaves like a driftless random walk whose standard deviation is of order $\sqrt{\tau} T^{\frac{1}{2}-\theta}$, which converges to 0.

To formalize this argument, we first show that for any lower bound m the probability that

the high type's logit-belief lies above m at relative time τ converges to 1 as $T \to \infty$:

$$\begin{aligned} P(\text{logit}(\hat{\mu}_{\lfloor \tau T \rfloor}) < m | H) &= P\left(\frac{\text{logit}(\mu_{\lfloor \tau T \rfloor}) - \hat{\gamma}^H_{\lfloor \tau T \rfloor}}{\hat{\sigma}^H_{\tau T}} < \frac{m - \hat{\gamma}^H_{\lfloor \tau T \rfloor}}{\sqrt{\tau} T^{\frac{1}{2}-\theta} \sqrt{\Sigma_H}} | H \right) \\ &\leq \Phi\left(\frac{m - \hat{\gamma}^H_{\lfloor \tau T \rfloor}}{\sqrt{\tau} T^{\frac{1}{2}-\theta} \sqrt{\Sigma_H}} + 2\epsilon \right) + \frac{M_H}{\epsilon \sqrt{\tau T}} \end{aligned}$$

For the last inequality we use our approximation lemma 2 with $a = -\infty$ and any $\epsilon > 0$. We now exploit the fact that $\frac{m - \hat{\gamma}^H_{\lfloor \tau T \rfloor}}{\sqrt{\tau} T^{\frac{1}{2}-\theta} \sqrt{\Sigma_H}} \to -\infty$, which holds since $\hat{\gamma}^H_{\lfloor \tau T \rfloor} \to \infty$ and the numerator is of order $O(\tau T^{1-\theta})$ while the denominator is only of order $O(\sqrt{\tau} T^{\frac{1}{2}-\theta})$.

We next show that for any $\epsilon' > 0$ the probability that the low type's belief stays within an ϵ'-neighborhood around $\text{logit}(\mu^*_L)$ converges to 1 in probability as $T \to \infty$. Note, that the expected logit-belief at any relative time τ is $\text{logit}(\mu^*_L)$ under the uniform DNB:

$$\begin{aligned} & P(|\text{logit}(\hat{\mu}^{\lfloor \tau T \rfloor}) - \text{logit}(\mu^*_L)| > \epsilon' | L) = \\ = \; & P\left(\frac{\text{logit}(\hat{\mu}^{\lfloor \tau T \rfloor}) - \text{logit}(\mu^*_L)}{\hat{\sigma}^L_{\tau T}} < -\frac{\epsilon'}{\hat{\sigma}^L_{\tau T}} | L \right) + P\left(\frac{\text{logit}(\hat{\mu}^{\lfloor \tau T \rfloor}) - \text{logit}(\mu^*_L)}{\hat{\sigma}^L_{\tau T}} > \frac{\epsilon'}{\hat{\sigma}^L_{\tau T}} | L \right) \\ \leq \; & \Phi\left(\frac{-\epsilon'}{\sqrt{\tau} T^{\frac{1}{2}-\theta} \sqrt{\Sigma_L}} + 2\epsilon \right) + 1 - \Phi\left(\frac{\epsilon'}{\sqrt{\tau} T^{\frac{1}{2}-\theta} \sqrt{\Sigma_L}} - 2\epsilon \right) + \frac{2M_L}{\epsilon \sqrt{\tau T}} \end{aligned}$$

For the last inequality we fix any $\epsilon > 0$ and use our approximation lemma 2 twice. We can make this upper bound as small as we want for sufficiently high T since $\theta > \frac{1}{2}$.

Also note that we can obtain a uniform upper bound for all relative time by setting $\tau = 1$ on the RHS. Since the cost distribution is atomless, it follows that the expected utility of the low type agent converges to the utility of the unconstrained low type with constant belief μ^*_L.

A.4 Proof of Proposition 2

Step 1: Conservatism We first show conservatism (claim 1 of the proposition) through proof by contradiction. The intuition for conservatism is as follows: assume the agent's responsiveness does not converge to 0. There will be some realization k and a sequence (T^j), such that $|\beta^{T^j}_k| > \delta > 0$ for some $\delta > 0$. We will show that the agent's total utility in the low state converges to at most 0 as $T^j \to \infty$. According to lemma 3 an agent with uniform DNB would do strictly better: hence the agent cannot be optimally biased.

We start by bounding the probability that subjective beliefs fall within the interval $[\epsilon', 1-\epsilon']$

in the low state:

$$\begin{aligned} & P(\epsilon' < \hat{\mu}_{\lfloor \tau T^j \rfloor} < 1-\epsilon' | L) \\ = \; & P\left(\frac{\text{logit}(\epsilon') - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} < \frac{\text{logit}(\hat{\mu}^{\lfloor \tau T \rfloor}) - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} < \frac{\text{logit}(1-\epsilon') - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} | L \right) \\ \leq \; & \Phi\left(\frac{\text{logit}(1-\epsilon') - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} + 2\epsilon \right) - \Phi\left(\frac{\text{logit}(\epsilon') - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} - 2\epsilon \right) + \frac{M_L}{\epsilon\sqrt{\tau T}} \end{aligned}$$

For the last inequality we fix any $\epsilon > 0$ and use our approximation lemma 2. We next replicate the proof of lemma 1 to show:

$$\hat{\sigma}_{\tau T^j}^L \geq \sqrt{\tau T^j} \underbrace{\sqrt{F_L(k^*)\left(1 - F_L(k^*)\right)\lambda_{k^*}\delta}}_{M' > 0}$$

We can therefore simplify the upper bound:

$$P(\epsilon' < \hat{\mu}_{\lfloor \tau T^j \rfloor} < 1-\epsilon' | L) \leq \frac{1}{\sqrt{2\pi}} \left(\frac{\text{logit}(1-\epsilon') - \text{logit}(\epsilon')}{\sqrt{\tau T^j} M'} + 4\epsilon \right) + \frac{M_L}{\epsilon\sqrt{\tau T^j}} = M''\epsilon + \frac{M'''(\epsilon, \epsilon')}{\sqrt{\tau T^j}}$$

Now fix a relative time τ^*. We can bound the total utility of the low type above by $\tau^* b(1) + (1-\tau^*)K$ where

$$K = \underbrace{\left(M''\epsilon + \frac{M'''(\epsilon,\epsilon')}{\sqrt{\tau T^j}} \right) b(1)}_{\substack{\text{Bound on expected utility from posterior} \\ \text{falling within } [\epsilon', 1-\epsilon'] \\ \text{after relative time } \tau^*}} + \underbrace{b(\epsilon')}_{\substack{\text{Bound on expected util-} \\ \text{ity from posteriors be-} \\ \text{low } \epsilon' \text{ after relative time} \\ \tau^*}} + \underbrace{A\left[b(1) - \int_0^{1-\epsilon'} c dG(c) \right]}_{\substack{\text{Bound on expected utility} \\ \text{from posteriors above } 1-\epsilon' \text{ af-} \\ \text{ter relative time } \tau^* \text{ (probabil-} \\ \text{ity A)}}}$$

Due to the fact that the cost distribution is non-atomic, the last term is negative for sufficiently small ϵ' as $L(1) < 0$. Next, choose first τ^* and ϵ' and then T^* to make $\tau^* b(1)$ and the first two terms of K as small as desired for all $T^j > T^*$. Therefore, the low type's utility cannot be bounded away from 0 and the biased Bayesian does not do strictly better than an unbiased Bayesian for large T^j.

Step 2: DNB The proof of claim 2 of the proposition proceeds in 2 sub-steps. (A) We first show that for any constant $M > 0$ we have $\max_k \beta_k^T > \frac{M}{T}$ for any sufficiently large T. (B) Next, if optimal updating does not exhibit DNB for large T then the mean logit low-type belief converges either to plus or minus infinity. In both cases, the biased agent's utility will be strictly lower than under the uniform DNB.

We start with part A. Assume this claim is wrong. Then, we can find some M and a sub-sequence T^j such that $\max_k \beta_k^{T^j} < \frac{M}{T^j}$. This implies that mean logit-belief in the high state at any relative time τ is bounded above by $M^* = M \max_k \lambda_k$. But since belief utility is strictly increasing, her utility will be strictly lower than the utility of the unrestricted agent,

and therefore also strictly lower than for the agent with uniform DNB for any large enough T. This is a contradiction since we assumed that the responsiveness function is optimal.

Next consider claim B. Assume that $\sum_k F_L(k)\hat{\beta}_k^T\lambda_k$ does not converge to 0. Then there is some $\epsilon > 0$ and a sub-sequence T^j such that $|\sum_k F_L(k)\hat{\beta}_k^{T^j}\lambda_k| > \epsilon$. For any constant M, this implies $|\sum_k F_L(k)\beta_k^{T^j}\lambda_k| > \frac{M\epsilon}{T^j}$ as long as T^j is sufficiently big. Hence, the mean logit-belief of the low type converges either to $-\infty$ or $+\infty$.

We fix $\tau^* < 1$ and look at the case $\hat{\gamma}^L_{\lfloor\tau^* T^j\rfloor} \to -\infty$ first. Take a constant $B < \text{logit}(\mu_L^*)$. We use our approximation lemma 2 (for some $\epsilon > 0$):

$$\begin{aligned}
P(\text{logit}(\hat{\mu}_{\lfloor\tau^* T^j\rfloor}) > B|L) &= P\left(\frac{\text{logit}(\hat{\mu}_{\lfloor\tau^* T^j\rfloor}) - \hat{\gamma}^L_{\lfloor\tau^* T^j\rfloor}}{\hat{\sigma}^L_{\tau^* T^j}} > \frac{B - \hat{\gamma}^L_{\lfloor\tau^* T^j\rfloor}}{\hat{\sigma}^L_{\tau^* T^j}}|L\right) \\
&\leq 1 - \Phi\left(\frac{B - \hat{\gamma}^L_{\lfloor\tau T^j\rfloor}}{\hat{\sigma}^L_{\tau^* T^j}} - 2\epsilon\right) + \frac{M_L}{\epsilon\sqrt{\tau^* T^j}} \\
&\leq 1 - \Phi(-2\epsilon) + \frac{M_L}{\epsilon\sqrt{\tau^* T^j}} \\
&\leq \frac{2}{3} \quad \text{for } \epsilon \text{ small enough and large enough } T^j
\end{aligned}$$

Hence, the probability of the low-type's logit-belief being below B for relative times $\tau > \tau^*$ is at least $\frac{1}{3}$. Hence, the low-type's utility is strictly lower than for an agent with unrestricted beliefs. This is a contradiction since we assumed that the responsiveness function is optimal. We can arrive at a similar contradiction for the case $\hat{\gamma}^L_{\lfloor\tau^* T^j\rfloor} \to \infty$.

Step 3: Initial Beliefs

We prove claims 3 and 4 of proposition 2 in 3 sub-steps. (A) We define an upper envelope function $U(x)$ for $L(x)$. (B) We show that $\hat{\sigma}^L_{\tau T} \to 0$ as $T \to \infty$, which is a strong form of conservatism. (C) We show that this implies claims (3) and (4) of proposition 2.

We start with part A. Using Taylor's theorem we can write

$$L(x) = L(\mu_L^*) + \frac{1}{2}L''(y)(x - \mu_L^*)^2 \tag{17}$$

for some $y \in [x, \mu_L^*]$. Note that L'' is continuous and hence strictly negative in an ϵ-neighborhood of μ_L^*, since $L''(\mu_L^*) < 0$. We can assume that $L''(y) \leq -A$ for some $A > 0$ in that neighborhood. We can now define the upper envelope function $U(x)$ for $L(x)$ as follows:

$$U(x) = \begin{cases} L(\mu_L^*) - \frac{A}{2}(\mu_L^* - \epsilon)^2 & \text{for } x \leq \mu_L^* - \epsilon \\ L(\mu_L^*) - \frac{A}{2}(x - \mu_L^*)^2 & \text{for } \mu_L^* - \epsilon \leq x \leq \mu_L^* + \epsilon \\ L(\mu_L^*) - \frac{A}{2}(\mu_L^* + \epsilon)^2 & \text{for } x \geq \mu_L^* + \epsilon \end{cases} \tag{18}$$

This upper envelope will lie above $L(x)$ in the ϵ-neighborhood. We can refine the upper envelope function such that the upper envelope function dominates $L(x)$ on the interval $[0, 1]$ by considering the following set M that includes all local maxima outside the ϵ-neighborhood:

$$M = \left\{x|L'(x) = 0\right\} \setminus [\mu_L^* - \epsilon, \mu_L^* + \epsilon]$$

Denote the supremum of the $L(M)$ with m^*. Due to the Bolzano-Weierstrass theorem, there is a sequence $(x^j) \subset M$ such that $L(x^j)$ converges to m^*. Due to continuity, there is a subsequence $(x^{j'})$ of (x^j) and a $\tilde{x}$ such that $x^{j'} \to \tilde{x}$ and $L(x^{j'}) \to m^*$ and $L(\tilde{x}) = m^*$. If $m^* \geq L(\mu_L^*)$ then we get a contradiction because we assumed that the maximum at μ_L^* is unique. Hence, $m^* < L(\mu_L^*)$. Therefore, we can simply make the ϵ-neighborhood of the upper-envelope function small enough such that it always lies above m^*. This will ensure that the upper envelope function dominates L on the interval $[0, 1]$.[29]

For part B, assume that $\hat{\sigma}_T^L$ does not converge to 0 as $T \to \infty$. Then there is a subsequence (T^j) and some $\delta > 0$ such that $\hat{\sigma}_{T^j}^L > \delta$. Let $\delta' < \frac{\delta\sqrt{2\pi}}{4}$ and $\tau^* < 1$. We use our approximation lemma 2 (for some $\epsilon > 0$ and any $\tau > \tau^*$):

$$
\begin{aligned}
& P(|\text{logit}(\hat{\mu}_{\lfloor \tau T^j \rfloor}) - \text{logit}(\mu_L^*)| < \delta' | L) \\
= \;& P\left(\frac{\text{logit}(\mu_L^*) - \delta' - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T^j \rfloor}} < \frac{\text{logit}(\hat{\gamma}^L_{\lfloor \tau T^j \rfloor}) - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T^j \rfloor}} < \frac{\text{logit}(\mu_L^*) + \delta' - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T^j \rfloor}} | L \right) \\
\leq \;& \Phi\left(\frac{\text{logit}(\mu_L^*) + \delta' - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T^j \rfloor}} + 2\epsilon \right) - \Phi\left(\frac{\text{logit}(\mu_L^*) - \delta' - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T^j \rfloor}} - 2\epsilon \right) + \frac{M_L}{\epsilon\sqrt{\tau T^j}} \\
\leq \;& \frac{1}{\sqrt{2\pi}} \left(\frac{2\delta'}{\hat{\sigma}^L_{T^j}} + 4\epsilon \right) + \frac{M_L}{\epsilon\sqrt{\tau T^j}} \\
\leq \;& \frac{1}{2} + \frac{4\epsilon}{\sqrt{2\pi}} + \frac{M_L}{\epsilon\sqrt{\tau^* T^j}} \\
\leq \;& \frac{2}{3} \quad \text{for } \epsilon \text{ small enough and large enough } T^j
\end{aligned}
$$

Hence, the probability that subjective beliefs fall outside the interval $[\text{logit}^{-1}(\text{logit}(\mu_L^* - \delta')), \text{logit}^{-1}(\text{logit}(\mu_L^* + \delta'))]$ for $\tau > \tau^*$ is at least 1/3. The utility of the low-type agent using the upper-envelope function $U(x)$ accumulated over time $\tau > \tau^*$ is always strictly worse than the utility of the agent with a uniform DNB who can maintain beliefs arbitrarily closely to the optimal μ_L^*. Since her actual utility is even lower, we can strictly improve the agent's utility by using a uniform DNB. This is a contradiction since we assumed that the responsiveness function is optimal. Hence we proved $\hat{\sigma}_T^L \to 0$.

It follows that $\hat{\mu}_0^T \to \mu_L^*$. Otherwise, there would be a δ-neighborhood of μ_L^* and a subsequence (T^j) such that the initial prior $\hat{\mu}_0^{T^j}$ falls outside that interval. Combined with part A, this would imply that the agent's utility is strictly lower than under the uniform DNB along this sequence for large T^j which is a contradiction.

Combining part A with claim (3) of the proposition we immediately get convergence of low-type beliefs at any relative time τ to μ_L^*. Part A of step 2 also establishes that high-type mean-logit beliefs converge to $+\infty$. It is easy to see that $\hat{\sigma}_T^L \to 0$ implies $\hat{\sigma}_T^H \to 0$. Using lemma 2 then establishes that high-type beliefs converge to 1 in probability at any relative time $\tau > 0$.

[29] If there are finitely many local maxima, then the argument simplifies to m^* being the second-highest maximum.

A.5 Proof of Proposition 3

We have established in step 3 of the proof of proposition 2 that $\hat{\sigma}_T^L \to 0$. Using lemma 2 we can show that the probability that the low-type's beliefs remain in an interval around the new optimal low-type beliefs converges to 1 for any relative time τ. High-type belief convergence to 1 at all relative times is not affected by choosing a different prior.

A.6 Proof of Proposition 4

We know that high-type beliefs converge to 1 while low type beliefs stay close to μ_L^*. We also know that $\hat{\sigma}_T^L \to 0$ and $\hat{\sigma}_T^H \to 0$ and that there are constants $m_1, m_2 > 0$ such that $m_1 < \hat{\sigma}_T^L / \hat{\sigma}_T^H < m_2$. Hence, the probability at relative time τ that the agent is a low type provided that $\hat{\mu}_{\lfloor \tau T^j \rfloor} < 1$ converges to 1. Therefore, learning one's type decreases the agent's total utility to 0 with probability approaching 1 as $T \to \infty$ and destroys belief utility $(1-\tau)b(\hat{\mu}_\tau)$ (since low type logit-beliefs follow a driftless random walk with vanishing variance).