ABSTRACT

Title of Dissertation: ESSAYS ON INFORMATION AND NON-BAYESIAN BELIEFS

ZHENXUN LIU
Doctor of Philosophy, 2024

Dissertation Directed by: Professor Emel Filiz-Ozbay
Department of Economics

In this dissertation, I present a comprehensive discussion of a class of biases within the realm of probabilistic reasoning, namely confirmation bias (encompassing or closely related to commonly seen terms in the literature such as motivated reasoning and wishful thinking).

The dissertation consists of three main chapters. In Chapter 1, I propose a new and improved belief updating model that can accommodate both motivated and unmotivated confirmation bias. The model improves upon existing models in its ability to explain data better, and its applicability to settings beyond binary-state spaces. I characterize the model with three intuitive axioms. In two extended applications, I show that the model establishes a link between confirmation bias and several well-known phenomena, such as the significance of first impressions, the polarization of beliefs, and the perseverance of inaccurate beliefs.

In Chapter 2, I turn to the experimental elicitation of motivated and unmotivated confirmation bias. Previous experiments have provided evidence for motivated and unmotivated confirmation bias individually, but never discussed the possibility that the two can occur together in depth.

This chapter presents one of the first experiments that examines both forms of confirmation bias together. Subjects were asked to update their beliefs regarding both politically contextualized questions and neutral questions. Subjects exhibited both motivated and unmotivated confirmation bias, but there was also significant heterogeneity among them. Notably, motivated confirmation bias is significantly stronger in later rounds of the experimental tasks, which may be correlated with the shorter response times in the later rounds.

In Chapter 3, which is joint work with Emel Filiz-Ozbay, we discuss wishful thinking (motivated confirmation bias) within a major application. In a rational inattention setting where consumers acquire information on the good's quality before making purchasing decisions, we examine the implications of the presence of consumers with wishful thinking. These biased consumers are unaware of their bias, and weigh any good news about the product quality more heavily than a Bayesian consumer. The firm, which aims to increase the volume of sales, can strategically constrain the accuracy of the information that consumers can acquire. We show that in the presence of biased consumers, a firm would find it profitable to constrain information acquisition unless the prior belief on the quality of the product is too low. We characterize the conditions under which the entry of a competitor firm can effectively alleviate this type of exploitation. Our findings shed light on the incentives of review platforms for bombarding wishful consumers with low quality product reviews and limit consumers' ability to identify to reviews with informative contents.

ESSAYS ON INFORMATION AND NON-BAYESIAN BELIEFS

by

Zhenxun Liu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Professor Emel Filiz-Ozbay, Chair/Advisor
Professor Erkut Ozbay
Professor Yusufcan Masatlioglu
Professor Daniel Reck
Professor Neslihan Uler

Acknowledgments

I owe my gratitude to all the people who have supported me throughout this journey. I would not have enjoyed it as much as I have without them. This has been an experience I will remember forever.

First and foremost, I would like to thank my advisor, Professor Emel Filiz-Ozbay, for the amount of supervision, guidance, and support she has provided throughout the past years. For a large part of the journey, she has had more confidence in me than myself. It has been a pleasure discussing my research and collaborating with her.

I am equally grateful for my partner, Yi Li, who has been patient and caring through the most depressing and trying moments. Although we were on the opposite sides of the globe for the majority of the journey, she has always been there to support me.

I would also like to thank Professor Erkut Ozbay, Professor Yusufcan Masatlioglu, Professor Daniel Reck, and Professor Neslihan Uler not only for agreeing to serve on my dissertation committee, but also for their comments and guidance regarding my research.

A lot has happened over the years, and the pandemic has been especially difficult for everyone. I would also like to thank all my cohort who have persevered and supported each other along the way. I have learned a great deal from them. Most notably, Elif Osun has provided a lot of guidance regarding experimental design and implementation. Many others were also supportive regarding research (and venting about the difficult PhD life), including but not limited to Keaton

Ellis, Ece Yegane, and Ming Fang.

Finally, I would like to thank my family and friends for their financial and mental support.

One journey has come to an end, and another journey is set to begin. I am uncertain where this new journey leads, but the joy and relief of conquering the current one is for certain. As an atheist, I do not thank God, but as an idealist who leans towards determinism, I would like to thank destiny.

Table of Contents

List of Tables

List of Figures

Chapter 1: A Model of Confirmation Bias

1.1 Introduction

The belief updating processes in humans generally do not adhere to the Bayes rule, as numerous studies have shown. One of the most prominent biases affecting this process is confirmation bias, where individuals tend to interpret new information in a way that confirms their existing hypothesis. In this paper, I propose a new model of confirmation bias, and provide the axioms that characterize the model. The model improves upon the existing literature in three main ways. First, it accommodates both "motivated" and "unmotivated" confirmation bias (as explained later), allowing further analyses on their differences and similarities. Second, it extends the discussion on confirmation bias to richer state spaces, while existing discussions are confined to binary states. Third, it not only exhibits desirable asymptotic properties as discussed by earlier models, but also well behaves when there are few signals, making it suitable for estimation of experimental data with finite signals.

Confirmation bias has been observed to impact behavior across various domains, such as scientific hypothesis testing (1), gambling (2), finance (3), and political ideologies (4, 5). Despite an extensive body of research on confirmation bias in psychology (see [6], for a comprehensive review), theoretical discussions in economics have been limited so far (7, 8, 9, 10). Economists often interpret confirmation bias as asymmetric belief updating arising from skewed *prior* beliefs.

This mainly captures what [6] referred to as "unmotivated" confirmation bias (henceforth UCB), where the "existing hypothesis" an individual is testing is the most probable one, but neglects what he called "motivated" confirmation bias (henceforth MCB), where individuals have tangible reasons (monetary, ego related, etc.) to defend their existing hypothesis. This is often referred to as "wishful thinking", where the individual *prefers* the hypothesis to be true.[1] That is to say, confirmation bias may originate from not only skewed prior beliefs, but also preferences, as both can influence which hypothesis the individual wants to test.

The existing literature has hardly considered both MCB and UCB simultaneously, even though many economic problems under uncertainty involve both the decision maker's preferences and beliefs. Consider a principal trying to assess an agent's intellectual capabilities, for instance a teacher or employer assessing a student or employee. How would the principal react to a piece of negative evidence regarding the agent's intelligence? If the agent is a total stranger to the principal, there are no preferences nor any specific prior probabilities to go by, and the assessment would be relatively objective. If the agent has a reputation for her intelligence, such as a well known researcher, the principal's prior belief might make her hesitant about fully accepting the negative evidence. On top of that, if the two are good friends or collaborators in work, the principal would be happy that the agent is smart, and both prior beliefs and personal preferences would cause the principal to downplay the negative evidence. In this scenario, only considering UCB or MCB would result in an underestimation of the magnitude of the bias. Conversely, if the agent is a friend with a below-average reputation regarding her abilities, the principal's UCB and MCB will act against each other. Only considering UCB or MCB will give contradicting predic-

[1]This is not to say that motivated confirmation bias is deliberate. To quote Nickerson (1998, pp 175): "... confirmation bias connotes a less explicit, less consciously one-sided case-building process." Neither is it the case that motivated confirmation bias will always dominate unmotivated confirmation bias, as one might interpret from the experimental results of [11].

tions in regards to the direction of the bias. Importantly, [12] provides evidence that individuals can exhibit both UCB and MCB simultaneously.

This paper is the first to examine both forms of confirmation bias in the same framework. The model features an updating process that generalizes Bayesian updating, where the posterior likelihood ratio between two states is the Bayesian posterior likelihood ratio multiplied by the bias ratio.

$$\frac{\mu^{CB}(\omega|s)}{\mu^{CB}(\omega'|s)} = \frac{\mu^{B}(\omega|s)f(\mu_0(\omega), v(\omega))}{\mu^{B}(\omega'|s)f(\mu_0(\omega'), v(\omega'))}$$

The bias $f(\mu_0(\omega), v(\omega))$ increases in both prior beliefs μ_0 and preferences v over the states, representing UCB and MCB, respectively. The model can be characterized by three simple axioms defined on observable individual beliefs and the signal structure. In particular, one axiom characterizes the general structure of the belief updating process, while the other two are necessary and sufficient conditions for the properties of $f(\cdot)$, which produce UCB and MCB, respectively. Several other papers characterize a belief updating process with axioms (e.g. 8, 9, 10, 13, 14, 15, 16, 17, 18). The process described here can be viewed as an intersection of the updating processes in [18] and [10], where the former is a model of wishful thinking, and the latter can accommodate UCB. A crucial difference is that their models are constructed within the [19] framework, while my model is independent of choice.

Additionally, previous discussions on (unmotivated) confirmation bias are limited to binary state spaces (e.g. 7, 8, 9, 10), while this model expands the analysis of confirmation bias to a general finite state space. This is critical since many relevant applications cannot be easily reduced to binary states. Examples range from weather to political inclination. Applicability to multi-state situations allows me to propose a novel but natural criterion within the framework of

this model to identify the states towards which an individual is biased.

Finally, the model proposed here provides a better explanation for empirical data. For instance, [7] assumes that a biased individual has an exogenous probability of misinterpreting a dis-confirmatory signal as a confirmatory one. However, experimental evidence indicates that people generally update their beliefs in the correct direction, though with asymmetry in the magnitude. In contrast, on top of [12], models similar to mine have shown to work well in experimental data analyses in [20] and [21].

The model has intuitive implications for various applications: 1) the order in which signals are received matters, as earlier signals carry more weight than later ones; 2) the same signal source can result in polarized beliefs, depending on the differences in people's prior beliefs and preferences; 3) an individual's belief in an incorrect state may converge to 1. I discuss how UCB and MCB relate to these implications differently. My model provides theoretical support for the empirical evidence presented in previous research, such as [22], [5], and [23].

In what follows, I will briefly discuss the related literature, then describe the model in detail, and provide the characterization results. Finally, I will present the interesting and intuitive implications that can be derived from the model in applications.

Relation to the Literature

Theoretical work:

[7] proposed a model of (unmotivated) confirmation bias where agents misinterpret signals that dis-confirm their beliefs as signals that confirm their beliefs with a certain exogenous probability. [24] and [18] proposed similar models for (motivated) confirmation bias, with the latter providing its axiomatization within the [19] framework. [25] differs in its ability to address the

"opposite" of MCB, a type of pessimistic behavior where the person's updated beliefs are more biased towards worse states. [10] proposed a model of inertial updating, where the individual minimizes the distance between the prior and posterior, and can accommodate UCB as a special case.

My model relates to other non-Bayesian belief updating models, such as the two-period model in [8], where individuals face a cost for not succumbing to temptation to update incorrectly.[2] [9] involves ambiguity and has the agent form plausible priors and update beliefs in two steps, maximizing the event's likelihood and choosing the least different prior from a benchmark. [13] aimed to explain base-rate neglect, conjunction fallacy, and disjunction fallacy, while another paper focused on updating beliefs with qualitative information (14). To address the issue that Bayesian individuals cannot update belief upon realization of zero probability events, [26] proposed an axiomatic model that permits belief updating when facing zero probability events by having the individual retroactively choose priors that rationalize observations.

[27] presented a model where the individual has anticipatory utility, forms optimal beliefs about future states, and makes consumption choices to maximize total utility. Their model is inherently not a belief updating model, and cannot explain UCB. Additionally, optimal expectations and confirmation bias have different intuitive appeal: optimal expectation describes the behavior of setting a positive mindset before receiving information, while confirmation bias describes the behavior of biased interpretation of information at the time of receiving it.

Experimental work:[3]

[2]The paper's discussion on confirmation bias is limited to specific signal structures. In particular, the discussion relied on symmetric signal structures, i.e. $\pi(a|A) = \pi(b|B) = \theta$, given there are two signals $\{a, b\}$ and two states $\{A, B\}$.

[3]There is a much richer body of evidence documented mainly by the psychology literature, and others such as the political science and finance literature (e.g. 1, 2, 3, 4, 5, 28, 29, 30, 31). Both [6] and [32] are comprehensive reviews on the literature.

This paper relates to the experimental literature on confirmation bias, where experiments typically provide a clear signal structure and the experimenters compare subjects' updating process with a Bayesian benchmark. [33] conducted an experiment with the classic balls and urns setting, discovering evidence of UCB in the form of increased conservatism when the signal disconfirms prior beliefs. [34] observed that subjects weigh their first signal more than their second one, implying the existence of UCB. [35] observed UCB in people's beliefs about their own abilities. [11] had participants assess and update their beliefs about their IQ or appearance based on binary comparisons with other subjects, and observed MCB that biased beliefs in the direction of "higher IQ" and "more attractive". [36] also investigated belief updating about cognitive abilities and observed similar results. [21] noted strong MCB in politically charged questions, despite the use of uninformative signals. [37] observed MCB regarding bad events, where subjects are more inclined to believe losses will not occur.[4] These experiments further highlight the need for a theoretical model that accommodates both UCB and MCB.

Motivated reasoning:

My model relates confirmation bias with motivated reasoning, where the latter topic has attracted more attention over the past years (e.g. 11, 24, 33, 34, 36, 38). The economics literature has largely discussed the concepts of confirmation bias and motivated reasoning either in isolation of each other or as distinct and competing theories. However, the two concepts can, in fact, be considered two parts of the same information processing mechanism. [32] argued that it is possible that the process of motivated reasoning involves the motivation creating a direc-

[4]Mixed results exist, with [33] and [34] observing UCB but not MCB, despite both experiments having monetary incentives to induce motivation. [38] found that context impacts MCB, and pure monetary incentives were ineffective in generating MCB-induced belief distortion. While my proposed model does not distinguish between contexts, it offers a framework for future research to explore the relation between the two forms of confirmation bias and how context influences each of them.

tional hypothesis, after which the individual performs hypothesis confirming tests. While the two concepts have their nuanced differences, one may view motivated reasoning and motivated confirmation bias as behaviorally indifferentiable under most economic contexts.

1.2 Model

The basic setup of the model involves the following:

- Ω is the finite set of all possible states, $\omega \in \Omega$.
- $v : \Omega \to \mathbb{R}$ denotes an individual's valuation, or preferences, over each state.[5]
- $\mu \in M = int(\Delta(\Omega))$ is a probability distribution over Ω and $\mu(\omega)$ denotes an individual's belief over the state ω.
- A signal s is drawn from the finite set of signals, S, according to a family of conditional probabilities $\{\pi(s|\omega)\}_{\omega\in\Omega}$, with $\pi(s|\omega) \in (0,1)\ \forall s, \omega$.

The individual obtains information to learn about the underlying state of the world. To model confirmation bias, define a positive valued weight function: $F_0(\omega) = f(\mu_0(\omega), v(\omega))$, with $f : (0,1) \times \mathbb{R} \to \mathbb{R}_+$ being a continuous function of the prior belief $\mu_0(\cdot)$ and the valuation $v(\cdot)$ over the state.

When the biased individual receives a signal s, her posterior belief is distorted by confirmation bias in the following way:

$$\mu^{CB}(\omega|s) = \frac{\mu_0(\omega)\pi(s|\omega)F_0(\omega)}{\sum_{\omega'} \mu_0(\omega')\pi(s|\omega')F_0(\omega')} \tag{1.1}$$

[5]These preferences could originate from consumption or material payoff associated with each state, or from more abstract emotional gains such as satisfaction from being intelligent.

Or equivalently, for all $\omega, \omega' \in \Omega$,

$$\frac{\mu^{CB}(\omega|s)}{\mu^{CB}(\omega'|s)} = \frac{\mu_0(\omega)\pi(s|\omega)F_0(\omega)}{\mu_0(\omega')\pi(s|\omega')F_0(\omega')} = \frac{\mu^{B}(\omega|s)F_0(\omega)}{\mu^{B}(\omega'|s)F_0(\omega')} \tag{1.2}$$

This process is a generalization of Bayesian updating - when F_0 is a constant function, the updating process becomes Bayesian. Compared to that of a Bayesian individual, the biased individual's beliefs are distorted by F_0.

$F_0(\omega)$ should generally be regarded as an aggregation function of the individual's attitude towards a state under the influence of confirmation bias. The function has the following monotonicity properties:

Property (Monotonicity)**.**

1) Monotonicity in Prior:

$$\mu_0(\omega) \geq \mu_0'(\omega) \Rightarrow f(\mu_0(\omega), v(\omega)) \geq f(\mu_0'(\omega), v(\omega))$$

2) Monotonicity in Value:

$$v(\omega) \geq v'(\omega) \Rightarrow f(\mu_0(\omega), v(\omega)) \geq f(\mu_0(\omega), v'(\omega))$$

That is, the higher the prior probability or valuation, the more favorable a state is. The distortion of beliefs is directed towards the most "favorable" states, or the hypotheses the individual wants to test the most, as determined by $F_0(\omega)$. UCB and MCB are induced by the monotonicity of $f(\mu_0(\omega), v(\omega))$ in each of its arguments, respectively.

1.2.1 Basic Properties

Confirmation bias effectively shifts the correct Bayesian posterior beliefs in a direction, and with a magnitude, that depends on both the individual's prior and valuation of the states. The following example illustrates this updating process in detail.

Example 1. Depicted in the simplex shown in Figure 1.1 is a three-state situation, with states $\{A, B, C\}$ represented by the respective vertices. Each interior point represents a probability distribution, and its distance to each edge represents the probability of the state (as represented by a vertex) opposite from the edge. Consider an individual with a prior belief of $\mu_0 = (\mu_0(A), \mu_0(B), \mu_0(C)) = (0.5, 0.25, 0.25)$ and valuations $v = (v(A), v(B), v(C)) = (0.25, 0, 1)$. That is to say, state A is the most probable state by prior beliefs, while state C is the most preferred state. MCB would motivate the individual towards believing in state C, while UCB would mainly drive the individual to believe in state A. The figure shows the individual's beliefs change after receiving a signal s such that $\pi(s|A) = \pi(s|B) = 0.6$ and $\pi(s|C) = 0.3$, and how the combined effect of MCB and UCB would differ from just one of the two.

The Bayesian posterior is denoted μ^B, and represents a belief shift towards states A and B. First, consider the case without MCB and assume $F_0(\omega) = \mu_0(\omega)$. The biased posterior belief due to UCB alone is denoted μ^{CB}, which is a shift away from the Bayesian posterior directed straight towards A. Alternatively, assume $F_0(\omega) = \mu_0(\omega) + v(\omega)$, i.e. there are both UCB and MCB, and denote the resulting posterior belief as μ'^{CB}. Since state C has the highest valuation, μ'^{CB} is much closer to C than μ^{CB}, indicating a large increase in the probability placed on state C.

The model can accommodate a smooth resolution of the conflict between the motivated

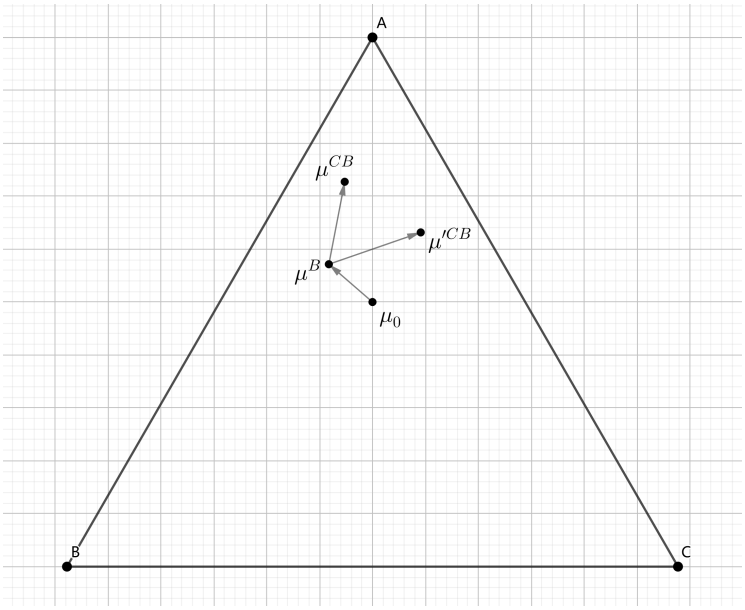


Figure 1.1: Figurative example of UCB and MCB

and the unmotivated components of confirmation bias. That is, the stronger of the two does not necessarily cancel out the other, and the direction of the belief shift acknowledges both the motivated and unmotivated confirmation bias. This is dependent on the structure of $F_0(\omega)$, though, and may change if F_0 adopts a different functional form. For example, if $F_0(\omega) = max\{k(\mu_0(\omega)), h(v(\omega))\}$, then the stronger bias of the two dominates, or crowds out, the other.

It is worthy to note that one can obtain (1.2) from the maximization of the following:

$$\mu^{CB} = \underset{\mu \in M}{argmax}\{E_\mu[g(\mu_0(\omega), v(\omega))] - D_{KL}(\mu||\mu^B)\} \tag{1.3}$$

Here, g increases in both $\mu_0(\omega)$ and $v(\omega)$, $\mu^B(\omega|s) = \frac{\mu_0(\omega)\pi(s|\omega)}{\sum_{\omega'}\mu_0(\omega')\pi(s|\omega')}$ is the Bayesian posterior belief, and $D_{KL}(\mu||\mu^B) = \sum \mu(\omega) ln(\frac{\mu(\omega)}{\mu^B(\omega)})$ is the Kullback-Leibler divergence.[6] Specifically,

[6]Observe that $\mu^B(\omega) \in (0,1)\ \forall\omega$ is required for all states, else the KL divergence is undefined. Having the prior taken from M and the conditional probabilities of signals confined within $(0,1)$, as stated in the basic setup, guarantee that $\mu^B \in M$. Consequently, the biased posterior beliefs will also remain within M.

the solution to this optimization problem is, for any $\omega, \omega' \in \Omega$,

$$\frac{\mu^{CB}(\omega|s)}{\mu^{CB}(\omega'|s)} = \frac{\mu^{B}(\omega|s)e^{g(\mu_0(\omega),v(\omega))}}{\mu^{B}(\omega'|s)e^{g(\mu_0(\omega'),v(\omega'))}}$$

Defining $g(\mu_0(\omega), v(\omega)) = ln(f(\mu_0(\omega), v(\omega)))$ will produce the formulation in (1.2) exactly.

Understanding the model as an optimization problem can allow certain interpretations. For example, the optimization problem can be viewed in light of the discussion in [4] as people, under a political context, having "accuracy goals" and "partisan goals" when they seek and interpret information. Alternatively, the function g can be interpreted as a utility function akin to the psychological preferences described in [39] and [40], where beliefs have both "instrumental" and "psychological" values.[7] This form of preference encompasses many situations that involve the effect of emotions on a person's welfare. While the logic behind motivated confirmation bias is intuitive, there has yet been a satisfactory theory explaining what drives unmotivated confirmation bias. This certain interpretation can serve as a candidate explanatory theory.

Conveniently, the model also allows two alternative ways to interpret the mechanisms of the bias. The bias as interpreted as a shift on the Bayesian posterior can be equivalently interpreted as a shift on the prior belief, after which Bayesian updating is applied. This can be observed from (1.2), as one can combine the distortion F_0 with μ_0 first to obtain a set of distorted priors. Figure 1.2 depicts this property, building upon Example 1. μ_0^* is the prior belief such that a Bayesian individual endowed with μ_0^* would obtain μ^{CB} as her posterior belief upon observing signal s. The vectors $\mu^{CB} - \mu^B$ and $\mu_0^* - \mu_0$ can both be considered as the shift due to confirmation bias, and the vectors $\mu^{CB} - \mu_0^*$ and $\mu^B - \mu_0$ are both belief shifts due to Bayesian updating upon

[7] Instrumental value refers to utilization of beliefs to guide decisions under uncertainty. Psychological value refers to beliefs entering the utility function directly.

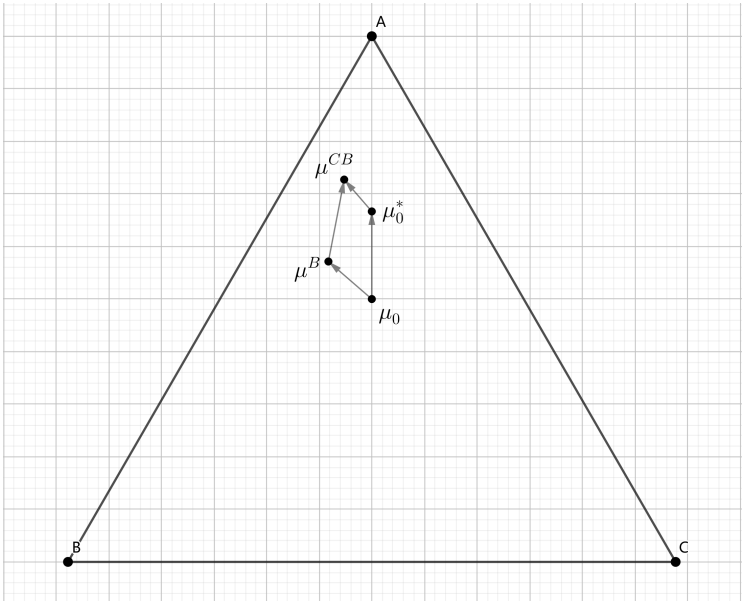


Figure 1.2: Equivalence to Bayesian updating with distorted prior.

observing s.

Alternatively, one may also realize that there should be a signal s^* with conditional probabilities $\pi^*(s^*|\cdot)$ such that a Bayesian individual endowed with prior μ_0 would arrive directly at μ^{CB} after observing s^*. This is also readily observable from (1.2) as one can apply F_0 to π instead of μ_0. Figure 1.3 depicts this property, where the vector $\mu^{CB} - \mu_0$ denotes the Bayesian updating process given such a signal s^*. The conditional probabilities $\pi^*(s^*|\cdot)$ can be inferred in a similar fashion to how μ_0^* is inferred.[8]

By the same logic, (1.3) can be equated to solving either of two similar problems, one where the solution is the distorted prior and one where the solution is the distorted signal, and then finding the Bayesian posterior belief given the solution. These observations are summarized in the proposition below.

[8]Here, it should be noted that the distortion as applied to signal probabilities are the same for all signals $s \in S$: $\frac{\pi(s|\omega)}{\pi(s|\omega')} e^{F_0(\omega)-F_0(\omega')}$. This results in $\sum_s \pi(s|\omega)e^{F_0(\omega)} \neq \sum_s \pi(s|\omega')e^{F_0(\omega')}$, and the sums cannot be normalized to 1 for all states simultaneously. Hence, strictly speaking, the distorted probabilities do not constitute probability measures.

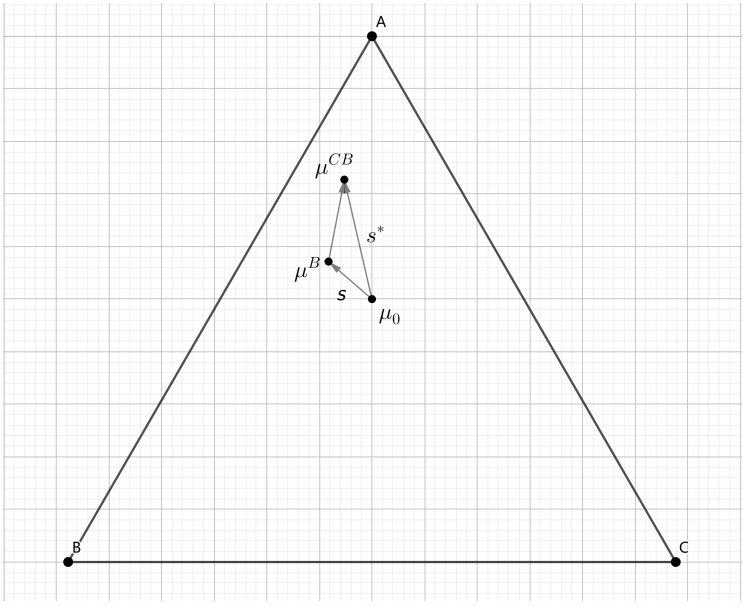


Figure 1.3: Equivalence to Bayesian updating with distorted signal.

Proposition 1. *The following are equivalent:*

- *Solving for (1.3) given signal s and prior belief μ_0.*

- *Finding the Bayesian posterior given signal s and distorted prior belief μ_0^*, where*

$$\mu_0^* = \underset{\mu \in M}{argmax}\{E_\mu[g(\mu_0(\omega), v(\omega))] - D_{KL}(\mu||\mu_0)\}$$

- *Finding the Bayesian posterior given distorted signal s^* and prior belief μ_0, where*

$$\pi^*(s^*|\cdot) = \underset{\hat{\pi}(s|\cdot)\in(0,1)^K}{argmax}\{\sum_\omega \hat{\pi}(s|\omega)g(\mu_0(\omega), v(\omega)) - D_{KL}(\hat{\pi}(s|\cdot)||\pi(s|\cdot))\}$$

with $|\Omega| = K$ and $\pi(s|\cdot) = (\pi(s|\omega_1), \pi(s|\omega_2), ..., \pi(s|\omega_K))$.

To summarize, it is equivalent to apply the distortion induced by confirmation bias to the Bayesian posterior, the prior, or the signal probabilities. With this equivalence property, the

model can be easily applied to certain scenarios such as information design settings where many existing theoretical results regarding Bayesian individuals have been established.

1.2.2 Characterization

Next, I will provide an axiomatic characterization of the confirmation bias model presented earlier.[9] As a starting point, define an updating process that generalizes Bayesian updating by allowing a multiplicative bias term. I will call this, in short, the BUME representation.

Definition 1. An individual's beliefs have a *Bayesian Updating with Multiplicative Error (BUME)* representation if, given the prior belief μ_0, for each state ω_i there exists an induced error ϵ_i such that posterior beliefs can be expressed as

$$\mu(\omega_i|s) = \frac{\epsilon_i \mu_0(\omega_i)\pi(s|\omega_i)}{\sum_j \epsilon_j \mu_0(\omega_j)\pi(s|\omega_j)}$$

Note that when ϵ_i is constant for all i, BUME coincides with Bayesian updating. Broadly speaking, this form of belief updating may be used to describe biases other than confirmation bias, depending on the properties ϵ has. The updating process as described in my model corresponds to $\epsilon_i = F_0(\omega_i)$ that increases in the prior belief and valuation over state ω_i. Though BUME is a generalization, such a belief updating process may still be obtained via (1.3). Regardless of what ϵ represents, BUME has the following important characteristic.

[9]The discussion will be centered around the inference of the unobservable confirmation bias, with individual beliefs and the signal structure observable. That is, the axioms will be constructed to be testable within typical belief updating experiments.

Axiom 1. (IIA - States) Given μ_0, for any two states ω_i, ω_j and two signals s and s',

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)}\frac{\mu(\omega_j|s')}{\mu(\omega_i|s')} = \frac{\pi(s|\omega_i)}{\pi(s|\omega_j)}\frac{\pi(s'|\omega_j)}{\pi(s'|\omega_i)}$$

Put simply, this axiom implies that the individual understands the meaning of the signals. The axiom relates the posterior ratios with the signal probability ratios, and states that when the signal changes, the posterior belief ratio should change by the same proportion as the signal probability ratio. This axiom is named IIA - States since one may also interpret it as information on other states have no effect on the posterior ratios between ω and ω'. In this regard, the axiom can be likened to the IIA axioms in [41]. The two IIA axioms, IIA (Consequences) and IIA (Actions), of Matějka and McKay (2015) describe how the relative probabilities of choosing two actions stay the same as long as payoffs stay the same across states or actions. In the same spirit, A1*, which is a necessary condition of Axiom 1, describes how the posterior probability ratio between two states stays the same for different signal realizations as long as the signals' conditional probability ratio is preserved.

A1*. Given μ_0, for any two states ω_i, ω_j and two signals s and s',

$$\frac{\pi(s|\omega_i)}{\pi(s|\omega_j)} = \frac{\pi(s'|\omega_i)}{\pi(s'|\omega_j)} \Rightarrow \frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} = \frac{\mu(\omega_i|s')}{\mu(\omega_j|s')}$$

Axiom 1 is a more general version of A1* where $\frac{\pi(s|\omega_i)}{\pi(s|\omega_j)} = \frac{\pi(s'|\omega_i)}{\pi(s'|\omega_j)}$ is not required. Instead, Axiom 1 imposes that the differences in posterior ratios can be fully explained by the differences in the signals' conditional probabilities. As stated in the next result, Axiom 1 is not only necessary, but also sufficient for BUME.

Proposition 2. *An individual's beliefs satisfy Axiom 1 if and only if they have a BUME representation. The induced errors* $\epsilon_i = \epsilon(\omega_i, \mu_0)$ *are uniquely identified up to a linear transformation.*

The proof is provided in Appendix A. The intuition behind this is that only a multiplicative bias can be completely cancelled when calculating the ratio between posterior ratios. Since Bayesian updating easily satisfies Axiom 1, this result also confirms that the BUME representation is a generalization of Bayesian updating. As the model proposed earlier is a specific type of BUME, where the ϵ's are non-decreasing in the priors, μ_0, and valuations, v, I will present two axioms that will give rise to ϵ's that exhibit such monotonic properties - one for monotonicity in prior and one for monotonicity in valuation. Correspondingly, the two axioms will represent belief updating with UCB and MCB, respectively.

Recall that UCB refers to confirmation bias resulting from a skewed prior belief. An individual with UCB would have higher posterior beliefs over states which they held higher prior beliefs. More specifically, given two prior beliefs and two signals such that the respective Bayesian posteriors should be equal, any difference in the observed posterior beliefs of the individual should be due to the difference in prior. Axiom 2 describes exactly this, and will be the only additional axiom needed to characterize UCB.

Axiom 2. (UCB) For any μ_0, μ_0' and any ω_i, ω_j such that $\mu_0(\omega_i) \geq \mu_0'(\omega_i)$ and $\mu_0(\omega_j) \leq \mu_0'(\omega_j)$,

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} \geq \frac{\mu'(\omega_i|s')}{\mu'(\omega_j|s')}$$

for all s, s' such that

$$\frac{\pi(s|\omega_i)}{\pi(s|\omega_j)} \frac{\pi(s'|\omega_j)}{\pi(s'|\omega_i)} = \frac{\mu_0(\omega_j)}{\mu_0(\omega_i)} \frac{\mu_0'(\omega_i)}{\mu_0'(\omega_j)}$$

I denote $\mu'(\cdot|s)$ as the posterior obtained from the prior μ'_0. Note that the condition on the signals implies that the respective Bayesian posterior belief ratios should be equalized for both signals. Based on BUME, I may define a more specific representation as follows.

Definition 2. An individual's beliefs have a BUME-UCB representation if they have a BUME representation, and for any state ω_i and any two different priors μ_0, μ'_0 such that $\mu_0(\omega_i) \geq \mu'_0(\omega_i)$, the respective induced errors satisfy $\epsilon_i = \epsilon(\omega_i, \mu_0(\omega_i)) \geq \epsilon(\omega_i, \mu'_0(\omega_i)) = \epsilon'_i$.

BUME-UCB is a BUME representation where the induced ϵ's increase in the prior beliefs. The addition of Axiom 2 is enough to obtain BUME-UCB.

Proposition 3. *An individual's beliefs satisfy Axioms 1 and 2 if and only if they have a BUME-UCB representation.*

The above proposition states that Axioms 1 and 2 fully characterize belief updating with unmotivated confirmation bias. Axiom 2 and the definition of BUME-UCB as presented above gives a set of ϵ's that are weakly increasing in prior beliefs. Note that with Axiom 2, ϵ_i only depends on the prior over ω_i, but not other states. One can simply change the weak inequalities in Axiom 2 to strict inequalities to obtain ϵ's that are strictly increasing in prior beliefs.

Next, I will characterize belief updating with motivated confirmation bias. Recall that MCB refers to confirmation bias resulting from personal preferences. An individual with MCB would have higher posterior beliefs over states which they prefer more than others. If the prior beliefs over two states are the same, then any differences in posterior beliefs unaccounted for by the signals should be due to the differences in the preferences. However, preferences are not observable. Therefore, the following axiom is instead a condition that allows us to infer preference orderings among states, in a standard revealed preference argument. That is, it states

that such differences in posterior belief unaccounted for by the signal should be preserved for all priors that weigh the two states equally.

Axiom 3. (MCB) If $\forall s$ and for some μ_0 such that $\mu_0(\omega_i) = \mu_0(\omega_j)$,

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} \geq \frac{\pi(s|\omega_i)}{\pi(s|\omega_j)}$$

then, $\forall s$ and $\forall \mu_0$ such that $\mu_0(\omega_i) = \mu_0(\omega_j)$,

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} \geq \frac{\pi(s|\omega_i)}{\pi(s|\omega_j)}$$

This axiom imposes that when prior beliefs are the same, then any discrepancy between the posterior likelihood ratio and the signal probability ratio needs to be preserved.[10] For comparison, a Bayesian individual should always satisfy this axiom with equality. Based on this, the following statement then defines an updating process distorted by MCB. Since an individual who exhibits MCB would be more biased towards preferred states, a BUME-MCB representation is a BUME representation where the ordering over induced errors can be rationalized by a preference relation.

Definition 3. An individual's beliefs have a BUME-MCB representation if they have a BUME representation, and there exists a rational preference relation $\succeq_P$ such that $\omega_i \succeq_P \omega_j$ if and only if the respective induced errors satisfy $\epsilon_i \geq \epsilon_j$ for some μ_0 such that $\mu_0(\omega_i) = \mu_0(\omega_j)$.

In the same spirit as the classical revealed preference, it is then not surprising that the

[10]The idea behind this axiom differs from the wishful thinking axiom in [18], which is based on the SEU framework. The axiom states if an endowment act g provides higher payoff in all states within event A compared to event B, then the individual's beliefs of event A will be biased upward relative to B. The endowment is analogous to innate preferences v in my model, but are observable. This observability allows for this straightforward way of describing the bias.

addition of Axiom 3 is necessary and sufficient for the BUME-MCB representation.

Proposition 4. *An individual's beliefs satisfy Axioms 1 and 3 if and only if they have a BUME-MCB representation.*

In sum, beliefs that satisfy Axioms 1 and 3 have a BUME representation where the induced errors $\epsilon(\omega_i, \mu_0(\omega_i))$ have an implied ordering that can be rationalized by a preference ordering. In the definition, ω_i being revealed preferred to ω_j only requires there to be one prior such that the induced errors satisfy $\epsilon(\omega_i, \mu_0(\omega_i)) \geq \epsilon(\omega_j, \mu_0(\omega_j))$. Axiom 3, analogous to revealed preference, makes this relation consistent across all priors beliefs that weigh the two states equally. Consequently, this allows the ordering over induced errors to be consistent with a rational preference relation. Contrary to some of the previous models where the updating process is inferred from choice or actions, the opposite is done here, where preferences are inferred from observed belief updating.

To conclude this discussion, it is natural that individual beliefs would satisfy Axioms 1-3 if and only if they allow both BUME-UCB and BUME-MCB representations, or as a whole, a BUME-CB representation.

Definition 4. An individual's beliefs have a BUME-CB representation if they allow both a BUME-UCB and a BUME-MCB representation.

Corollary 1. *An individual's beliefs satisfy Axioms 1-3 if and only if they have a BUME-CB representation.*

1.2.3 Comparison with Other Models

In this section, I compare my model with the closely related models in the literature.

Rabin and Schrag, 1999 (henceforth RS)

The RS model is one of the first models for (unmotivated) confirmation bias. The essence of the RS model is that individuals misinterpret dis-confirmatory signals as confirmatory ones with a certain probability p. What constitutes a confirmatory or dis-confirmatory signal is determined by prior beliefs. Since the prior belief is skewed towards state L, l is the confirmatory signal while h is the dis-confirmatory signal.

The RS model does not accommodate MCB. Moreover, one can easily check that the RS model does not satisfy Axiom 1 or 2 as presented in the previous section, and thus describe essentially different belief updating processes.

Since the RS model predicts that there is a probability p that dis-confirmatory signals will be misinterpreted, it might not fit empirical data very well. Most experiments feature belief updating in response to no more than 6 consecutive signals, while such drastic denial of dis-confirmatory information does not seem to be the general observation (e.g. 11, 12, 33, 34, 38).[11]

It is also difficult to generalize the RS model to settings with more than two states. In the RS model, a signal s is confirmatory if and only if $\mu_0(\omega) > \mu_0(\omega')$ and the Bayesian posterior $\mu^B(\omega|s) > \mu_0(\omega|s)$, for $\omega, \omega' \in \Omega$. In the two-state two-signal setting, this is very straightforward. When one expands the state and signal spaces, there is the immediate question of which state an individual is biased towards and which signal is confirmatory.[12]

Wishful Thinking Model (henceforth WT)

[18] characterized a model of wishful thinking within the broader SEU framework proposed in [19]. Under the WT model, the individual has a "default" act g, or endowment, which is

[11] Even if people were found to be so biased as to update beliefs in opposite directions, my model can also accommodate this with large enough F_0 for certain states.

[12] In Section 1.3, I propose a natural extended definition of confirmation bias under the framework of my model that is compatible with multiple states.

realized no matter what the state is. States associated with higher endowment are more attractive, acting as the motivation for wishful thinking. The representation of the belief updating process is very similar to that proposed in this paper:

$$\frac{\mu_1(\omega|s)}{\mu_1(\omega'|s)} = \frac{\mu^B(\omega|s)}{\mu^B(\omega'|s)} \frac{\delta(u(g(\omega)))}{\delta(u(g(\omega')))}$$

Here, δ is an increasing distortion function, $g(\omega)$ is the outcome of act g in state ω, and u is a utility function. The updating process of the WT model is essentially a BUME model, but the distortion function δ does not take the prior beliefs as an argument. Hence, unlike my model, the WT model does not accommodate UCB.

[25] and [24] both proposed similar models of wishful thinking. The former is also based on the SEU framework, but additionally allows pessimism, where the individual's beliefs are biased towards worse states, and extensively discusses "logit distortions", which are similar to the exponential biases obtained from solving (1.3). The latter has the individual jointly choose action and beliefs to maximize expected utility subject to a cost, which is comparable to (1.3), but with action instead of prior beliefs as arguments for the utility function.[13]

Note that these models satisfy the following stronger version of Axiom 2.

[13] Specifically, the individual solves

$$\mu^{CL}(\omega) = argmax_{\mu,a}\{\mu \cdot u(a,\omega) - \gamma D_{KL}(\mu||\bar{\mu})\}$$

where $\bar{\mu}$ is an exogenous belief that is interpreted as the mainstream belief in a society. Viewed in a belief updating context, this mainstream belief could be the Bayesian posterior. The authors also discuss an alternative formulation where the individual chooses signal structures instead of beliefs.

A2*. For any μ_0, μ_0', and any ω_i, ω_j,

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} = \frac{\mu'(\omega_i|s')}{\mu'(\omega_j|s')}$$

for all s, s' such that

$$\frac{\pi(s|\omega_i)}{\pi(s|\omega_j)}\frac{\pi(s'|\omega_j)}{\pi(s'|\omega_i)} = \frac{\mu_0(\omega_j)}{\mu_0(\omega_i)}\frac{\mu_0'(\omega_i)}{\mu_0'(\omega_j)}$$

This condition actually describes an updating process without UCB, as opposed to the (weak) UCB as described in Axiom 2. Additionally, while [25] does not necessarily satisfy Axiom 3, the other two models satisfy both Axioms 1 and 3.

In contrast to my model, the lack of UCB in these wishful thinking models makes them insensitive to signal ordering. According to [16], these updating models can also be considered as "Virtual Bayesian" updating, a type of generalization of Bayesian updating (discussed more in the Appendix).

Inertial Updating (henceforth, IU)

[10] proposed a model of Inertial Updating, which is also built upon the SEU framework and describes a individual who tries to minimize the distance between her final posterior belief and her prior belief. In other words, beliefs are sticky, and hence the name "inertial updating". They characterize both Bayesian updating and biased updating in this framework, with the latter including biases such as conservatism and (unmotivated) confirmation bias. This is, however, reliant on the assumption that the bias distorts the prior beliefs of the individual, and the individual updates beliefs from the distorted prior. Recall Proposition 1, and one can immediately notice how the IU model can describe UCB in the same way as my model.

The IU model cannot explain MCB. One of its axioms prescribes that if an individual's prior assigns the same probability to two states, she will continue to do so after observing an event that contains both states.[14] However, this directly contradicts the intuition of MCB, where if there is an innate preference over states, posterior probability ratios will not stay the same (as can be observed in 21). The authors attempt to explain wishful thinking using another concept called "Mixed Bayesians", that is the individual's beliefs are a convex combination of the original one and another belief:

$$\mu_E = \frac{\mu_0(E)}{\mu_0(E) + \rho(E)}\mu^B(\mu_0, E) + \frac{\rho(E)}{\mu_0(E) + \rho(E)}\mu^B(\rho, E)$$

Here, μ_0 is the original prior, ρ is the other prior it is "mixed" with, E is the event, μ^B is the Bayesian posterior given the event and prior, and μ_E is the posterior belief given event E. If one were to explain MCB, ρ should be connected to preferences, which requires representing preferences with a belief. Another issue is that, in order to generate a bias towards a certain preferred state, ρ needs to change along with μ_0. This is because ρ always needs to be more skewed towards the preferred state than μ_0 is in order to pull beliefs towards that state.[15]

1.3 Identifying Confirmation Bias

An important feature of my model is its capability to accommodate more than two states. This is in direct contrast to the previous literature, where there are typically two states, and an individual is identified to have confirmation bias towards one of the two states when she holds

[14] This is Axiom 5 (Consistency) in [10].

[15] To see this, let $\mu_0(s_1, s_2, s_3) = (\frac{2}{3}, \frac{2}{15}, \frac{1}{5})$ and $E = \{s_1, s_2\}$. Then $\mu^B(\mu_0, E) = (\frac{5}{6}, \frac{1}{6}, 0)$. Let $\rho = (\frac{3}{4}, \frac{1}{10}, \frac{3}{20})$ to represent the preference for state s_1, then $\mu_E(s_1) = \frac{85}{99} > \frac{5}{6}$. If we instead let $\mu_0 = (\frac{2}{3}, \frac{1}{15}, \frac{4}{15})$, then $\mu^B(\mu_0, E) = (\frac{10}{11}, \frac{1}{11}, 0)$ and $\mu_E(s_1) = \frac{17}{19} < \frac{10}{11}$. Mixing with ρ actually decreases belief over s_1 now.

a higher prior belief towards that state. Accordingly, a signal is confirmatory when it implies a high Bayesian posterior probability on the state the individual is biased towards. This binary situation gives rise to a natural interpretation of which state an individual is biased towards, and which signal confirms which state. When there are more than two states, it becomes difficult to determine which hypotheses (states) an individual is biased towards.

The multiplicative term F_0, or the more general ϵ, is what drives the confirmation bias in my model. In the previous discussions, the ϵ was referred to as simply an error term, and $F_0(\omega)$ an aggregation function. For the convenience of all upcoming discussions, I give F_0 a name.

Definition 5. Call $F_0(\omega) = f(\mu_0(\omega), v(\omega))$ the *total distortion* with respect to state ω.[16]

The larger the total distortion, the harder confirmation bias pushes an individual's beliefs towards that state, relative to other states. Naturally, the higher the total distortion with respect to a state is, the more the posterior belief is biased in favor of that state, which in turn implies a greater inclination to confirm that state. While it is unambiguous that the individual would want to strongly confirm the state with the highest total distortion, and disconfirm the state with the lowest total distortion, the middle states still require a threshold criterion to categorize them as "confirming" states or "disconfirming" states. I propose a criterion based on how individuals react to uninformative information.

An uninformative signal is a signal which is equally likely given every state. From the updating process in (1.2), it is apparent that uninformative signals causes shifts in beliefs under my model. This is an intuitively appealing property that is often overlooked in the literature, but is supported by experimental evidence. [5] have discovered that people became more certain of

[16] Additionally, if $F_0(\omega)$ is multiplicatively separable, that is $F_0(\omega) = g(\mu_0(\omega)) \cdot h(v(\omega))$, the total distortion on the posterior can be further dissected into a prior-induced distortion $g(\mu_0(\omega))$ and a preference-induced distortion $h(v(\omega))$, which are respectively responsible for observed UCB and MCB.

their stances on capital punishment when given a research excerpt including arguments for both sides. [21] found that subjects would shift their beliefs in favor of their affiliated political party after receiving uninformative signals. Here, reaction to noise is a desirable criterion because, on top of the supporting evidence relating such behavior to confirmation bias, it can be conveniently applied to the multi-state scenario.

Definition 6. An individual has *confirmation bias towards* state ω if $\mu^{CB}(\omega|s_0) > \mu^{B}(\omega|s_0)$, where s_0 is the uninformative signal such that $\pi(s_0|\omega_i) = \pi(s_0|\omega_j)$ for all pairs $i \neq j$. Similarly, she has *confirmation bias against* state ω if the inequality is reversed.

This definition essentially states that if an individual increases her confidence over a state after receiving uninformative information, then she has confirmation bias towards that state. Since a Bayesian individual does not update beliefs given uninformative signals, that is, the Bayesian posterior is equal to the prior $\mu^{B}(\omega|s_0) = \mu_0(\omega)$, the inequality in the definition above can be rewritten as $\mu^{CB}(\omega|s_0) > \mu_0(\omega)$. To add to this, the following result links the above definition with the total distortion of a state.

Proposition 5. *Define* $E_{\mu_0}[F_0] = \sum_{\omega'} \mu_0(\omega')F_0(\omega')$. *An individual has confirmation bias towards state* ω *if and only if* $F_0(\omega) > E_{\mu_0}[F_0]$. *Contrarily, the individual has confirmation bias against state* ω *if and only if* $F_0(\omega) < E_{\mu_0}[F_0]$.

This result establishes a threshold that allows us to identify which states the individual wants to confirm. As long as the total distortion of a state is larger than the weighted average of total distortions, weighted by the prior belief, the individual has confirmation bias towards that state.[17] It is worth mentioning that since confirmation bias is defined using the uninformative

[17]This result also makes it clear that, when applied to the two-state setting, the above definition is no different

signal s_0 here, it does not say much about how one would react to other signals. It is possible for posterior beliefs over states that the individual is biased against to be overconfident, because the threshold for overconfident posterior beliefs is $\sum_{\omega'} \mu^B(\omega'|s)F_0(\omega')$. That is,

$$\mu^{CB}(\omega|s) > \mu^B(\omega|s) \iff F_0(\omega) > \sum_{\omega'} \mu^B(\omega'|s)F_0(\omega')$$

This threshold looks similar to that in Proposition 5, but also crucially different in that it depends on the signal. However, it should be expected that the individual is more likely to be overconfident in states she is biased towards than the states she is biased against. The following observations state these insights more formally.

Remark 1. *For any state ω_i that the individual holds confirmation bias against, if $\mu^{CB}(\omega_i|s) > \mu^B(\omega_i|s)$ for some signal s, then $\mu^{CB}(\omega_j|s) > \mu^B(\omega_j|s)$ for any state ω_j that the individual is biased towards.*

Remark 2. *If $\omega_i \in \underset{\omega\in\Omega}{argmax}\{F_0(\omega)\}$, then $\mu^{CB}(\omega_i|s) \geq \mu^B(\omega_i|s)$ for all s. If $\omega_j \in \underset{\omega\in\Omega}{argmin}\{F_0(\omega)\}$, then $\mu^{CB}(\omega_j|s) \leq \mu^B(\omega_j|s)$ for all s.*

The former remark states that the set of signals that gives rise to an overconfident posterior over state ω_i, which the individual is biased against, is a subset of the set of signals that gives rise to an overconfident posterior over state ω_j, which the individual is biased towards. The latter remark states that the individual will always hold an overconfident (underconfident) posterior belief over the state with the highest (lowest) total distortion, regardless of the signal.

from the usual definition in the literature - the individual has confirmation bias towards one of two states if she has a higher prior over that state. Under my model, the state with the higher prior will correspond to a higher total distortion, and the individual would be biased towards this state.

In the same spirit, the following observation can be made when comparing two individuals with different functional forms for total distortion F_0.

Proposition 6. *Assume for two individuals i and j that $F_0^i(\omega) = k(F_0^j(\omega))$, where $k(\cdot)$ is a monotonic increasing transformation.*

- *If $k(\cdot)$ is concave, and individual j has confirmation bias towards ω, then so does individual i.*
- *If $k(\cdot)$ is convex, and individual j has confirmation bias against ω', then so does individual i.*

If individual i's total distortion is more concave than that of individual j, it means that F_0^i takes more mild values when μ_0 and v increase a lot. In other words, it is harder for individual i to show extremely high bias towards a state. This makes individual i more receptive of other middling states being realized, hence the first result. Conversely, a more convex total distortion means that individual i is more easily firmly determined to confirm a certain state, and as a result becomes less receptive of information confirming other states. A more concave F_0 denotes relatively conservative behavior where the individual exhibits higher selectivity regarding which hypotheses to disconfirm, or rule out. A more convex F_0 denotes relatively extreme behavior where the individual exhibits higher selectivity regarding which state to confirm, and would like to disconfirm many other states.

The effect of prior beliefs and valuations are simple and follows from the monotonicity of $F_0(\omega)$. If, for two individuals i and j with everything else equal, $\mu_0^j(\omega) > \mu_0^i(\omega)$ and $\mu_0^j(\omega') < \mu_0^i(\omega')$ for all $\omega' \neq \omega$ (or similarly if $v^j(\omega) > v^i(\omega)$), then individual j is more overconfident in ω than individual i: $\mu_1^j(\omega) - \mu_0^j(\omega) > \mu_1^i(\omega) - \mu_0^i(\omega)$.

1.4 Applications

1.4.1 Order of Signals Matter

There are many situations where people participate in repeated interactions with each other, and thus the optimal order at which signals should be sent is an important question.[18] Before elaborating on the details, the results can be summarized as: first impressions matter.

The effect of first impressions in belief updating experiments is well documented in the psychology literature, and is often discussed under the topic of "primacy effect". The primacy effect describes the phenomenon where earlier pieces of information are given more weight and consequently recalled more vividly than later pieces. [22], for example, performed two similar experiments where in each one, they presented the subjects sequences of information (roll of a die in the first experiment and draws from an urn in the second). They provided the sequence of information in a specific order so that the first half favors one state and the second half favors the other. A Bayesian individual would return to her initial prior of 50-50 after the sequence of draws. The experimenters found that the subjects were more inclined to believe in the state favored by the first half of the sequence of signals. [42] found that when subjects were asked to assess how intelligent a problem solver they observed were, they generally rated problems solvers who showed a descending success rate as more intelligent than both problem solvers who started poorly but showed improvement over time and those who had a relatively even success rate throughout the course. [43] showed with a simple balls and urns experiment that subjects were highly inclined to believe in the urn favored by the first draw, where they were asked to

[18] Repeated belief updating upon receiving consecutive signals would involve using the biased posterior resulting from the previous signal as the new prior belief and repeating the process in (3.5).

indicate their beliefs after two draws.

These experimental results were interpreted by their authors mainly as evidence for the primacy effect, but they can also be explained by my model of confirmation bias pretty well, as seen in the discussion that follows.[19] To provide some context to the discussion, consider the following scenario.

Example 2. Consider an individual who is applying for a job and attends a series of interviews. The individual can be "Competent" or "Incompetent", and will be hired based on the probability the HR believes she is "Competent". Each interview focuses on a different aspect of the job and the applicant knows how well they will perform in each one. The HR sees each interview performance as an independently distributed signal regarding the applicant's competence, with conditional probabilities based on past experience. The HR has confirmation bias and needs to assess whether the individual is competent or not based on the signals. The individual is allowed to organize the interview order freely. How should the individual order the interviews so as to maximize the HR's belief of the probability she is competent?

Before introducing the relevant result that may help us answer this questions, it is imperative to introduce a certain operation on a sequence of signals. Let S be a set of N elements, and $\succeq$ be a partial order on the set S.[20] Let the set of all permutations of S be $S_\sigma = \{\{s_i\}_{i=1}^N | s_i \in S\}$, where the subscript "i" denotes the position of each element in the sequence. Also, given a sequence, let $I(\{s_i\}_{i=1}^N) = \{i \in \mathbb{N} \mid 1 \leq i \leq N-1 \; and \; s_{i+1} \succeq s_i\}$ be the set of indices where the elements with these indices are "less than" their successors, under the partial order $\succeq$. Define the

[19] The concept of the primacy effect mainly concerns the process of retrieving information from memory. It is beyond the scope of this paper to dive deep into the psychological relationship between the primacy effect and confirmation bias. Nevertheless, the current model does provide theoretical support for the existence of such a connection.

[20] I abuse notation here by using S again, as it is the set of signals for the discussion of this section.

following operation:

Definition 7. Let $S_\sigma^I = \{(\{s_i\}_{i=1}^N, I(\{s_i\}_{i=1}^N))|\{s_i\}_{i=1}^N \in S_\sigma\}$. A $\succeq$-*forward swap* (at position k) is a mapping $\sigma_\succeq : S_\sigma^I \to S_\sigma$ such that:

- $\sigma_\succeq(\{s_i\}_{i=1}^N, \emptyset) = \{s_i\}_{i=1}^N$, that is, it maps the sequence onto itself if $I(\{s_i\}_{i=1}^N) = \emptyset$.
- $\sigma_\succeq(\{s_i\}_{i=1}^N, k) = \{t_i\}_{i=1}^N$ where $t_k = s_{k+1}$, $t_{k+1} = s_k$, and $t_j = s_j$ for $j \neq k, k+1$.

The formal definition may obscure the simple idea behind forward swaps. A forward swap is essentially swapping the positions of two consecutive elements in a sequence where the succeeding element is greater than the preceding one. By doing so, the larger element (as defined by the partial order) is moved forward one place. The second argument to the function determines which pair of consecutive elements to swap, if there are multiple to choose from. Subsequently, define the following partial order on a set of signals.

Definition 8. Let $\succeq_\omega$ be a partial order on S where

$$s_i \succeq_\omega s_j \iff \frac{\pi(s_i|\omega)}{\pi(s_i|\omega')} \geq \frac{\pi(s_j|\omega)}{\pi(s_j|\omega')} \; \forall \omega' \neq \omega$$

In general, it is even possible that no pairs of signals satisfy this partial order. However, when there are only two states, it becomes a total order, as there is only one $\omega' \neq \omega$. Now, I can make the following statement in regards to signal sequences:

Lemma 1. *Suppose there are two states $\{\omega, \omega'\}$ and a set of independently distributed signals S. Let $\{s_k\}_{k=1}^N$ and $\{s_j\}_{j=1}^N$ be two different permutations of S. Then,*

$$\mu^{CB}(\omega|\{s_j\}_{j=1}^N) \geq \mu^{CB}(\omega|\{s_k\}_{k=1}^N)$$

if $\{s_j\}_{j=1}^N$ *can be obtained from* $\{s_k\}_{k=1}^N$ *solely via a series of* $\succeq_\omega$*-forward swaps (at any position).*

In the previous discussion, I mentioned that my model is sensitive to the order at which signals arrive. The above result tells us exactly how the posterior reacts to changes in the sequence. Specifically, it states that arranging signals that are more indicative of a certain state earlier in the sequence will increase the recipient's belief in that specific state. The reason for this is that the effect of earlier signals last throughout the remaining updating process by influencing the total distortion with respect to each state. Lemma 1 provides a fairly simple rule to compare potentially complicated sequences of signals if the receiver has confirmation bias.[21]

Every sequence can be transformed via forward swaps into a sequence of signals in descending order under $\succeq_\omega$. Thus, by iteratively applying Lemma 1, there is a clear optimal solution.

Proposition 7. *Suppose there are only two states* $\{\omega, \omega'\}$ *and a set of independently distributed signals* S*. Receiving the signals in descending order under* $\succeq_\omega$ *results in the highest posterior belief over state* ω*.*

In *Example 2*, the interviewee should attend the interviews she is confident in doing well first, and those in which she will do poorly last. Or, stated from the point of view of the biased HR, first impressions matter the most. All the results discussed in this section arise from UCB - it is the changes to prior-induced distortions after every signal that causes the order to matter. In the appendix, I discuss the case with more than two states. The results here extend beyond binary states given certain functional forms of $f(\mu_0(\omega), v(\omega))$.

[21] Note that not all sequences can be reached from another sequence via forward swaps. Consider states A and B, with $\pi(a|A) = \pi(b|B) > 0.5$. Then a, b, b, a and b, a, a, b cannot reach each other via forward swaps under a partial order defined in Proposition 3, and thus the comparison between posteriors resulting from these two sequences is ambiguous.

1.4.2 Belief polarization

There has been strong experimental evidence that relates confirmation bias to belief polarization and persistence of wrong beliefs. Indeed, since confirmation bias causes people to be more likely to seek and accept information similar to their preexisting opinions, it is natural that it creates a feedback cycle that cause people to be increasingly certain of their initial beliefs, regardless of whether they are wrong or not. In [5], both subjects who supported and disapproved of capital punishment were found to become more certain of their existing beliefs when presented the same evidence that included both supporting claims and criticism. In [43], subjects' beliefs were also found to diverge when presented equal evidence for both sides, where the divergence was due to the difference in the order the evidence came in (this also links to the previous discussion about signal ordering). [23] documented the persistence of superstitious beliefs among villagers in the Eastern Democratic Republic of Congo, where they falsely believed in magic rituals that would make them bulletproof, and deaths, despite the ritual, were considered a result of incorrect ritual performance. In this section, I will demonstrate how belief polarization and persistence of wrong beliefs, like the cases cited above, can be explained by my model.

Assume the prior belief is μ_0, and there is a constant information source that sends an independent signal $s \in S$ each period drawn from the family of conditional probability distributions $\{\pi(\cdot|\omega)\}_{\omega\in\Omega}$. I will denote the true underlying state as ω^*. I will examine, given the true state ω^*, which other states $\omega \neq \omega^*$ the individual's belief will possibly converge towards. First, let us call all states that an individual would eventually believe to be probable as "credible":

Definition 9. The state ω is *credible* if $\mu(\omega) \overset{a.s.}{\to} c > 0$ as the number of received signals $N \to \infty$.

This definition is significant in the sense that, for a Bayesian individual, only the true state

is credible. However, for an individual with confirmation bias, this may not be the case. Now, define "credibility" as a measure of how much the biased individual wants to believe in a state.

Definition 10. The *credibility* of state ω is:

$$C(\mu_0(\omega), v(\omega)) = ln(f(\mu_0(\omega), v(\omega))) - D_{KL}(\pi(\cdot|\omega^*)||\pi(\cdot|\omega))$$

The first term corresponds to how strong the bias towards ω is, and the second term represents how similar ω is to the true state ω^*.[22] Naturally, the true state itself ω^* maximizes the second term as the divergence is 0, but it does not guarantee high credibility due to confirmation bias. The credibility of a state describes how tempting it is to believe in it, rather than simply how close to truth it is. Similarly, having confirmation bias towards state ω does not necessarily equate to the state having high credibility either, as similarity between the two distributions $\pi(\cdot|\omega^*)$ and $\pi(\cdot|\omega)$ is also important.

How much credibility makes a state credible? The credibility of a state changes as the prior belief over the state changes. More specifically, it increases in μ_0 due to the monotonicity property of $f(\mu_0(\omega), v(\omega))$. Accordingly, I can define $\underline{C}(\omega) = \lim_{\mu_0(\omega)\to 0} C(\mu_0(\omega), v(\omega))$ as the minimal credibility, and $\overline{C}(\omega) = \lim_{\mu_0(\omega)\to 1} C(\mu_0(\omega), v(\omega))$ as the maximal credibility. The following result relates credibility to a state being credible.

Proposition 8. *The state ω is credible if $\underline{C}(\omega) > \overline{C}(\omega')$ for all $\omega' \neq \omega$.*

The proposition provides a sufficient condition for when an individual can end up firmly believing a potentially "wrong" state. Note that at most one state can satisfy the conditions of

[22] States are fully characterized by their associated conditional probabilities of producing each signal. If the conditional probabilities for two states are equal for all signal realizations, the two states are indistinguishable.

Proposition 8 at the same time. Conversely, it is possible that no state is credible, as in *Example 3* below.

Proposition 8 can be strengthened if the individual exhibits no UCB. This is because the credibility would solely vary with the valuation, which is fixed for each state throughout the updating process.

Corollary 2. *If an individual only exhibits MCB, that is if $f(\mu_0(\omega), v(\omega))$ is constant over μ_0, then ω is credible if and only if*

$$\omega = argmax_{\omega'}\{C(\mu_0(\omega'), v(\omega'))\}$$

Although there is always a maximizing state for credibility, it is still possible for there to be no credible state. That is when there are multiple states that have maximal credibility. For any two states that have maximal credibility, their belief ratios generally oscillate and never converge to a point (see Appendix A).[23]

For a Bayesian individual, only the belief on the correct state $\mu(\omega^*)$ remains positive as $N \to \infty$, and the beliefs for the other states go to 0. In other words, Bayesian updating is *consistent*. For an individual with confirmation bias, as long as there is a state other than the true state that is tempting enough (high $F_0(\omega)$) and characteristically similar to the true state in terms of the conditional probabilities of signals (low $D_{KL}(\pi(\cdot|\omega^*)||\pi(\cdot|\omega))$), the individual's belief will diverge away from the correct Bayesian belief. Belief updating with confirmation bias is *inconsistent*. This also echoes how misinformation often happens in real life. Wrong beliefs are more persistent when they align with one's incentives or firmly rooted prior beliefs, or when

[23] If one were to extend the definition of "credible" to accommodate a subset of states, then one can say that the whole set of states that maximize credibility is credible, i.e. the beliefs over all other states will go to 0.

they are hardly distinguishable from the truth given the evidence available.

One may be tempted to think that, if an individual has both MCB and UCB, convergence to a wrong belief would be easier. This is not always true. While UCB certainly facilitates the process when a state with a higher prior probability is also a state with a higher valuation, it may also act against MCB and end up driving beliefs towards a state that the individual initially did not have favorable preferences over. The example below illustrates the volatility of UCB and how it potentially facilitates belief polarization.

Example 3. Expanding on the same setting as *Examples 2*, consider what happens to the principal's beliefs as she observes many signals from the agent. Let $\Omega = \{H, L\}$, $S = \{h, l\}$ with $\pi(h|H) = \pi(l|L) = 0.6$ and $f = \mu_0(\omega)^{\frac{1}{2}}$ to focus on how UCB affects the belief updating process. Allowing MCB would not change the intuition of this exercise. Notice that if prior belief $\mu_0(H) = 0.75$, then for both signals h and l,

$$\frac{\pi(s|H)}{\pi(s|L)}\frac{F_0(H)}{F_0(L)} > 1$$

Therefore, the individual will end up believing that state H is the true state with certainty after many rounds of belief updating to such signals. However, if $\mu_0(H) = 0.6$,

$$\frac{\pi(l|H)}{\pi(l|L)}\frac{F_0(H)}{F_0(L)} < 1$$

There is now a possibility that the individual ends up having $lim_{N\to\infty}\mu_N^{CB}(H) \to 0$ if the first few signals are l.

If the true state is L, there is a probability of 0.4 that the first signal is h, which results

in a posterior of $\mu_1^{CB}(H|h) \approx 0.722$, leading to $lim_{N\to\infty}\mu_N^{CB}(H) \to 1$. However, if the first 4 signals are l, the posterior belief becomes $\mu_4^{CB}(H|l,l,l,l) \approx 0.224$, and belief over state H will certainly converge to 0 from there. The probability of receiving four consecutive l signals is non-trivial (~ 0.13). In sum, due to UCB, the starting prior will influence the likelihood that $\mu^{CB}(H)$ converges to 1.[24]

Now, I extend the intuition from *Example 3* to the general case. Since credibility increases with the prior, one would expect that for a state ω with moderately high credibility, there should generally be a threshold for the belief over ω such that if the individual's prior is above the threshold, her beliefs would converge to the degenerate belief of $\mu^{CB}(\omega) = 1$. Moreover, if the individual's prior is not above the threshold, but there is a signal sequence that can push her beliefs above it, then her beliefs would converge to this degenerate belief with some positive probability. The following proposition formally states this idea.

Proposition 9. *Let ω^* be the true state. If $\overline{C}(\omega) > \underline{C}(\omega')\ \forall \omega' \neq \omega$, and $\forall \mu_0 \in M$, $\exists s$, with $\pi(s|\omega^*) > 0$, such that posterior belief $\mu^{CB}(\omega|s) > \mu_0(\omega)$, then*

$$P(\lim_{N\to\infty} \mu_N^{CB}(\omega) \to 1) > 0$$

As long as the signal space is rich enough so that there is always a signal that increases the posterior probability over state ω, then beliefs converge to the degenerate belief assigning probability one to ω with positive probability. This is because there should be a sequence of signals, occurring with positive probability, that drives beliefs close to $\mu(\omega) = 1$. Then, utilizing Theorem 1 from [44], the condition that $\overline{C}(\omega) > \underline{C}(\omega')$ for all other states ω' guarantees that

[24] To expand on this, I present some simulations in Appendix D and discuss more on how initial prior belief affects the probability that $\mu^{CB}(H)$ approaches 1 in the limit.

$\lim_{N\to\infty} \mu_N^{CB}(\omega) \to 1$ with positive probability.

The variation in beliefs over the updating process constantly changes the strength of UCB, and thus a state's credibility. Hence, it is the presence of UCB that gives rise to this result. While MCB can be considered the main driving factor of belief polarization, as seen in Proposition 8 and Corollary 2, UCB facilitates this polarization process by weakening the conditions that need to be satisfied for convergence. Proposition 9 can also be interpreted in connection with the discussion regarding signal ordering. The previous discussion concerned how signal ordering matters and sending the strongest signals that favor some state ω first would give the highest posterior belief over ω. Here, receiving a sequence of signals that favor ω early on would lead to certainty over state ω with positive probability. In a sense, both convey the notion that earlier signals play a more important role in determining the final posterior beliefs.

In summary, confirmation bias can not only result in the polarization of opinions, but also the inconsistency of beliefs, that is after many signals, beliefs may still be wrong. The model proposed in this paper provides theoretical support for claims attributing the persistence of wrong beliefs and belief polarization to confirmation bias. It is important to keep in mind, however, that the results presented here all involve the concept of credibility. As such, it is evident that confirmation bias as described by my model cannot push an individual to believe in any unrealistic state even if she is biased towards it. The wrong state is most tempting when its high similarity with the true state is paired with the individual's strong confirmation bias towards it.

1.5 Concluding Remarks

The model proposed here incorporates both unmotivated and motivated confirmation bias, characterizes confirmation bias in general state spaces, and provides a tractable model that fits empirical observations well. The model can be characterized by three simple and intuitive axioms, and makes several economically relevant predictions. Earlier signals are shown to be more salient than later ones, beliefs can be polarized by a single information source, and false beliefs can be persistent regardless of how much information the individual is exposed to. These issues have been discussed in the literature as possible results of confirmation bias, and my model provides theoretical support for their relationship.

In relation to the literature, this model is also a first step towards analyzing the closely connected concepts of confirmation bias and motivated reasoning under the same framework. This framework also allows future experiments to examine the relationship between the two and potentially resolve the current problem of having only mixed evidence for motivated confirmation bias.

Chapter 2: Dual Elicitation of Motivated and Unmotivated Confirmation Bias

2.1 Introduction

Numerous experiments have shown that people do not update their beliefs following the Bayes rule. Confirmation bias, which refers to an individual's tendency to interpret information in a way that is favorable towards a pre-existing hypothesis, is known as one of the most prominent biases that influence a person's belief updating process. [6] notes two forms of confirmation bias - motivated confirmation bias, which is driven by preference, and unmotivated confirmation bias, which is driven by prior beliefs. While the literature has provided evidence for the existence of both forms of motivated and unmotivated confirmation bias (e.g. 11, 33, 34, 38), there are still two questions left unanswered. First and foremost, as both prior beliefs and preferences are present in many economic settings, within the lab and in real life, it is hard to persuade oneself that the two forms of confirmation bias can never take effect simultaneously. Yet, the two forms of confirmation bias have either been considered separately in previous research, or as competing explanations. This is also inconsistent with how biases such as base-rate neglect are treated in experimental studies, where both base-rate neglect and unmotivated confirmation bias are allowed to be jointly estimated. Secondly, the elicitation of (mainly motivated) confirmation bias has been inconsistent, with some studies finding no obvious effect (e.g. 33, 35, 38), and some studies observing a bias in the opposite direction of what our conception of motivated confirmation bias

would predict (e.g. 34, 45, 46). In this paper, I detail a lab experiment that simultaneously elicits the two forms of confirmation bias. In doing so, I present positive evidence for the simultaneous existence of the two forms of confirmation bias, and I also explore some potential factors that may influence the elicitation of motivated confirmation bias, such as any possible interaction between motivated and unmotivated confirmation bias, heterogeneity, and reaction time.

The belief updating experiment features 8 rounds, each with a different event whose probability the subjects need to evaluate, and 6 instances of belief updating (6 signals). The details of the experimental design take into account the strengths and weaknesses of previous experiments regarding the elicitation of motivated confirmation bias. In my experiment, the specific events subjects are asked to report their beliefs on depend on real statistics regarding the U.S. economy or environment during a period lead by certain presidents. The assumption is that subjects who are more politically aligned with the Democratic (Republican) party would be more willing to believe that a Democratic (Republican) president did well during his time. Specifically, subjects are asked to answer questions with the following format: "What do you think are the chances that A is greater than B?" where A is a statistic, such as inflation rate, and B is a plausible value for A. This type of question has two main advantages. On the one hand, it has an objective answer which most people may hold subject beliefs about. The questions are not only political, but also factual and relatable. On the other hand, these questions allow for simple manipulations to silence the motivation. By changing the topic of the question to something that the subjects would never care about, I have the ability to turn off motivation without changing the task. Notably, this context has been shown to successfully elicit motivated confirmation bias (21).

Estimation of the data utilizes the model described in [47], which is the only behavioral model that can accommodate both motivated and unmotivated confirmation bias. The estimated

parameters here can be applied to general state spaces as the model is flexible enough to accommodate them. The model is also closely related to [10, 18], and the functional form of the updating rule can be conveniently converted into an estimation equation. Furthermore, the estimation model used by [21] can be viewed as a special case of the model, and has provided satisfactory results in explaining motivated confirmation bias. Specifically, since the experiment featured a 50-50 prior as well as uninformative signals, Grether's model (48) cannot be used to explain motivated confirmation bias in that situation at all. Therefore, an estimation equation based [47] would seem more reliable for the estimation of motivated confirmation bias.[1] I not only estimate the model on the aggregate level, but also on the individual level to examine heterogeneity. This is made possible by having a within-subject design, where each subject is asked both politically charged questions and non-political questions.

The main purpose of this paper is to address whether an individual can exhibit both motivated and unmotivated confirmation bias at the same time. According to [6], motivated confirmation bias refers to the case when the hypothesis a person wants to test is the preferred one. For instance, a person would naturally be more receptive of good news relative to bad news about a stock they invested in. Unmotivated confirmation bias refers to the case where there are no obvious reasons why preferences would play a role, and the hypothesis a person is testing is simply what they believe in the most. For example, someone who firmly believes the earth is flat may be reluctant to accept evidence pointing to a spherical earth, while they have no intrinsic preferences regarding the matter.

Most existing evidence of confirmation bias come from the psychology literature (e.g. 1,

[1] I also show how estimation using the Grether's model is expected to produce similar results, and provide the aggregate level estimation results with the Grether's model in the appendix.

2, 5, 28, 29, 30, 31, 49), but there is a growing experimental literature in economics as well (e.g. 33, 34, 35). While the term "confirmation bias" as used in the literature usually refers to unmotivated confirmation bias, motivated confirmation bias is also discussed, but under the topics of "motivated reasoning" or "wishful thinking" (e.g. 11, 21, 33, 34, 36, 38, 45, 46).[2] Motivated and unmotivated confirmation bias have been mostly discussed in isolation from each other or treated as competing explanations for a phenomenon (e.g. 11). Even when they are discussed together, only one of them is usually observed (e.g. 33). In relation to the literature, the experiment presented here is the one of the first to provide positive evidence that individuals can indeed exhibit both forms of confirmation bias at the same time.

Additionally, I utilize the collected data analyze some potential factors that may contribute to the elusive nature of motivated confirmation bias. While most experiments can agree on the general observation that people are conservative and exhibit some degree of base rate neglect and unmotivated confirmation bias in their updating, the findings regarding motivated confirmation bias are less consistent. [11, 21, 36] all find positive evidence for motivated confirmation bias. However, the existence of motivated confirmation bias may be heavily dependent on the context, and when the motivation comes from money, the motivated confirmation bias disappears (38). However, contexts also do not seem to explain everything. [34] compared belief updating under different contexts and found evidence for unmotivated confirmation bias, as well as the

[2]Strictly speaking, motivated reasoning refers to a reasoning process, often behind the formation or rationalization of certain beliefs, that is biased due to personal preferences. Motivated confirmation bias can be considered as a behavioral pattern that naturally follows motivated reasoning - motivated reasoning induces the specific belief that the individual's motivated confirmation bias is directed towards (32, 50). A lot of what we observe as empirical evidence for motivated reasoning can also be treated as evidence for motivated confirmation bias. I will henceforth refer to the biased treatment of information due to underlying preferences or personal interests as "motivated confirmation bias" instead of "motivated reasoning", and its counterpart where the biased treatment of information is solely due to the a biased prior belief as "unmotivated confirmation bias", to avoid confusion and to keep the discussion under the overarching theme of confirmation bias.

opposite of motivated confirmation bias - under-weighting of good news relative to bad news. Similar to [11, 36], [35, 45] asked for subjects' judgements regarding their own abilities, and the former found no evidence of motivated confirmation bias, while the latter found the opposite of motivated confirmation bias. By using a context that is shown to be effective in eliciting motivated confirmation bias (21), I attempt to explore factors other than context that may affect the aggregate estimates for motivated confirmation bias. Specifically, I explore heterogeneity among individuals, variations in behavior across rounds, and potential interactions between the two forms of confirmation bias.

In what follows, I will outline the estimation model and propose the hypotheses, describe the experimental design, and finally, present and discuss the observations made from the data.

2.2 Model and Hypotheses

In this section, I lay out the estimation model based on [47]. Let Ω denote the finite set of all possible states. In the experimental setting, this is the binary set $\{G, L\}$. The individual has prior belief $\mu_0(\omega)$ and valuation $v(\omega)$ over each state $\omega \in \Omega$. The individual receives information as signals $s \in S$ generated according to a collection of conditional probabilities $\{\pi(s|\omega)\}_{\omega\in\Omega}$. Let $F_0(\omega) = f(\mu_0(\omega), v(\omega))$ be a weight function that is non-decreasing in prior belief, $\mu_0(\cdot)$, and valuation, $v(\cdot)$, respectively. The posterior likelihood ratio is updated from the prior likelihood ratio as follows:

$$\frac{\mu_1(\omega|s)}{\mu_1(\omega'|s)} = \frac{\mu_0(\omega)\pi(s|\omega)F_0(\omega)}{\mu_0(\omega')\pi(s|\omega')F_0(\omega')} \tag{2.1}$$

Note that the weight function F_0 is essentially a multiplicative error term that represents the deviation from Bayesian updating caused by confirmation bias. Motivated and unmotivated

confirmation bias are represented by the monotonicity of F_0 in valuation and prior belief, respectively. The higher the prior or valuation over a state ω, the larger the distortion F_0 applied to the belief about ω, which in turn translates into a higher posterior belief over the state. The process coincides with Bayesian updating if F_0 is constant across states.

Take the logarithm of (2.1) to obtain the following:

$$ln\left(\frac{\mu_1(\omega|s)}{\mu_1(\omega'|s)}\right) = ln\left(\frac{\mu_0(\omega)}{\mu_0(\omega')}\right) + ln\left(\frac{\pi(s|\omega)}{\pi(s|\omega')}\right) + ln\left(\frac{f(\mu_0(\omega), v(\omega))}{f(\mu_0(\omega'), v(\omega'))}\right)$$

Assume a simple linear form for $ln(f(\mu_0(\omega), v(\omega)))$ in the analyses that follow.

$$ln(f(\mu_0(\omega), v(\omega))) = \eta\mu_0(\omega) + \phi v(\omega)$$

Denote $\delta_\mu = \mu_0(\omega) - \mu_0(\omega')$ and $\delta_v = v(\omega) - v(\omega')$. The result is a clean estimation equation:

$$ln\left(\frac{\mu_1(\omega|s)}{\mu_1(\omega'|s)}\right) = \beta_0 + \beta_1 ln\left(\frac{\mu_0(\omega)}{\mu_0(\omega')}\right) + \beta_2 ln\left(\frac{\pi(s|\omega)}{\pi(s|\omega')}\right) + \beta_3\delta_\mu + \beta_4\delta_v \tag{2.2}$$

The model gives rise to the following correspondence between estimates and model parameters: $\beta_3 = \eta$, and $\beta_4 = \phi$. In other words, β_3 and β_4 represent the degree of unmotivated and motivated confirmation bias, respective. To note some special values, a Bayesian individual would have $\beta_0 = \beta_3 = \beta_4 = 0$ and $\beta_1 = \beta_2 = 1$. Observing $\beta_0 \neq 0$ would imply a systematic shift in each belief updating instance unexplained by the model; $\beta_1 < 1$ implies some degree of base-rate neglect; $\beta_2 < 1$ implies dome degree of conservatism; β_3 and β_4 being positive would mean that there exists unmotivated and motivated confirmation bias respectively.

Hypothesis 1. $\beta_3 > 0$ and $\beta_4 > 0$, that is, both unmotivated and motivated confirmation bias should be able to co-exist.

These are the main estimates of interest. Current empirical literature provides evidence for the existence of both motivated and unmotivated confirmation bias, but cannot provide a strong argument against the coexistence of the two biases. Therefore, I hypothesize that both could be observed simultaneously, under the correct situations.

Existing empirical work generally agree on the existence of some degree of base rate neglect and conservatism, as represented by $\beta_1 < 1$ and $\beta_2 < 1$ respectively. That is, both the prior likelihood ratio and signal probability ratio are discounted relative to Bayesian updating. Even though this is not present in the original model, these two updating patterns are nevertheless expected.[3]

Hypothesis 2. $\beta_1 \leq 1$ and $\beta_2 \leq 1$, that is, the individual may show some degree of base rate neglect and conservatism.

Finally, based on both theory and previous experimental findings, I do not expect any constant shifts in beliefs.

Hypothesis 3. $\beta_0 = 0$, that is, there is no constant shift in beliefs in absence of information.

2.3 Experimental Design

The experiment is designed using O-Tree and presented to the subject on a computer screen (51). The experiment consists of two parts. The first part consists of 8 rounds of belief updating

[3]It is also worth noting that adding base rate neglect and conservatism to the model is fairly simple. One can simply exponentiate the μ_0 and π in (2.1) to achieve this.

questions, and the second part features a multiple price list to control for risk preferences. Each part contains its own set of instructions and payments.

After the subjects are seated and have signed the consent form, they are prompted to focus on their computer screen where they are presented the instructions for the first part of the experiment. The experimenter reads the instructions for the first part out loud to the subjects. Each subject is allowed to proceed at their own pace afterwards. Before the belief updating tasks, there is also a set of comprehension questions to make sure they understand the procedure. Subjects can only continue upon answering every comprehension question correctly. The instructions for the second part is not read out loud, as each subject progress at a different pace. The experiment ends with a demographic survey. Detailed description of each part is presented in the following sections.

2.3.1 Belief Updating Task

Each of the 8 rounds of the belief updating task begins with a statement indicating the data source for the question, and prompt that introduces the topic. The subjects are presented with a question in regards to the topic, asking them to assess the probability that a certain statistic f is greater than or equal to a number N. Thus, the underlying state for each round is either G (which stands for greater than), for the case of $f \geq N$, or L (which stands for less than), for the case of $f < N$. At the beginning of each round, subjects are always asked to state their beliefs before receiving any signals. This elicits the subjects' prior beliefs for the round. To prevent too many subjects from forming extreme prior beliefs about f, the prompt always presents a reference. It is undesirable to have subjects have beliefs of 0 or 1, as theoretically, beliefs no longer change

Question Type	Topic
Political	US annual carbon dioxide emission rate in 2012 US annual carbon dioxide emission rate in 2018 US annual unemployment rate in 2017 US inflation rate during June, 2018
Non-Political	Capacity of the Sydney Super Dome Student to faculty ratio of the University of Warwick in Fall, 2021 Value of Azerbaijan's total exported merchandise in 2015 Number of visitors to the British Museum over the 2018 season

Table 2.1: Topics of the 8 questions asked

when individuals hold such priors.[4] The number N is arbitrarily chosen to be somewhat close to the correct answer and the reference.

Afterwards, 6 signals are drawn independently and sequentially presented to the subjects. Each signal is framed as information coming from a different news source, where the news source can be truthful with 70% probability or untruthful with 30% probability. Information provided are messages that tell the subjects what the answer should be - whether f is greater than N (signal $s = g$) or not (signal $s = l$). Truthful news sources always provide the correct answer, while untruthful ones always lie. Essentially, signals are generated with 70% accuracy.

$$\pi(s = g|\omega = G) = \pi(s = l|\omega = L) = 0.7$$

Subjects are required to enter their updated beliefs after receiving each signal. Subjects are also reminded of their previously reported beliefs (i.e. their answers submitted for the previous round, or their prior belief if it is their first signal) upon receiving each piece of news.

The belief updating task has two within-subject variations that differ only in the type of questions asked. Each variation contains 4 questions, with one set being politically charged ques-

[4] Additionally, such a reference may serve as a comparison to emphasize on political sentiments.

tions while the other containing non-political questions. The questions with a political undertone concern the economy or the environment, and are meant to elicit motivated confirmation bias. The assumption is that subjects who have fairly clear political inclinations will have a significant bias towards the performance of presidents of their own party, and a bias against the performance of presidents of the opposite party. The non-political questions concern topics that one would expect the majority of the subjects (undergraduate students of an American university) do not care about, and are meant to only elicit unmotivated confirmation bias. The order at which each subject is presented with the 8 questions is completely random. Table 2.1 lists all 8 question topics that are asked.

An example of what is presented to the subjects at the start of each round (prior elicitation) is shown in Figure 2.1. Here, subjects are asked about the unemployment rate under Trump's term, with the unemployment rate under Obama's term as a reference. The source of the data is also stated at the top. The subjects then proceed to receive random signals, and report their updated beliefs about the topic. An example of the signal reception and belief updating page is shown in Figure 2.2. At the top, the subject is reminded of their previous beliefs. In the middle, the signal is presented, along with a reminder of the conditional probabilities of the signals.

2.3.2 Payment and Incentives

The belief reports are incentivized via the BDM method as adapted for probability elicitation (52, 53). The details of the payment scheme are not explicitly explained in detail in the instructions. Instead, there is a line stating "The payment is determined in a way such that you will maximize your chances of earning $1.50 if you simply report your beliefs truthfully. If

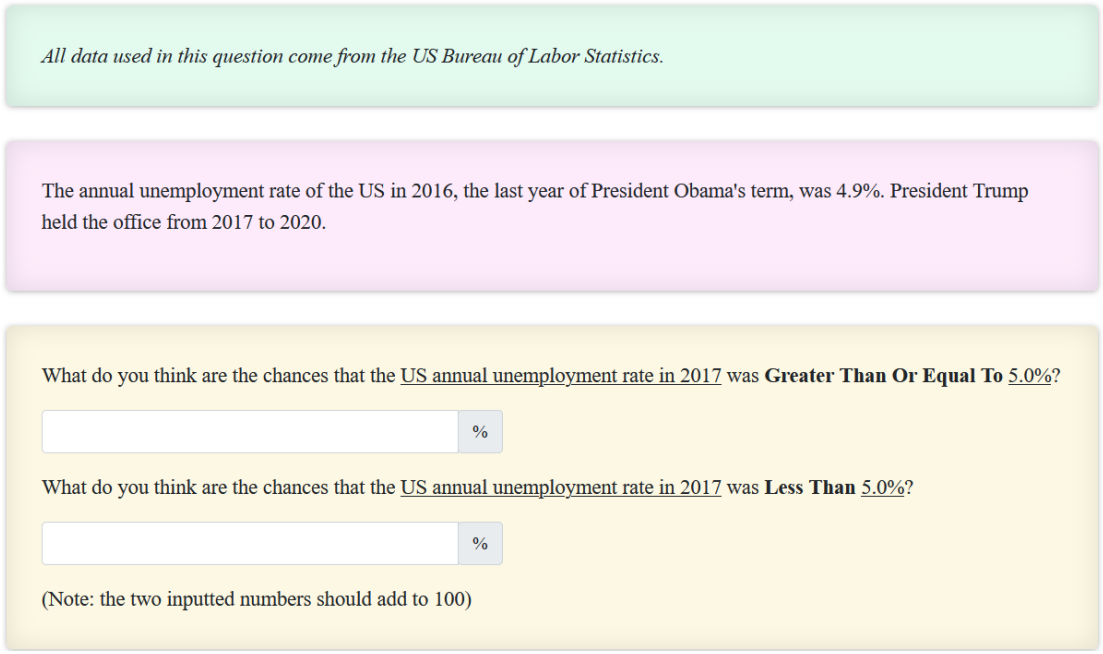


Figure 2.1: Example of prior elicitation faced by subjects.

you are interested in reading more about how the payoffs are calculated, please click here." and clicking on "here" expands a collapsible text box that contains the detailed explanations. This design choice is based on the findings of [54], where they find explanation of the payment scheme makes belief reporting less truthful. I have decided that this method strikes the best balance between avoiding confusing the subjects with the complicated payment scheme, and avoiding potential deception by not informing them at all.

When calculating the payment, one of the subject's belief reports (including the prior belief) for each of the 8 rounds is randomly selected. Their belief of state G is recorded as P. The computer generates a random number $n \in [0, 1]$. If $P \leq n$, the subject is rewarded \$1.50 with probability n, otherwise, if $P > n$, the subject is rewarded \$1.50 if the true state is G. It is easy to see how this payment scheme is incentive compatible: if the true belief is denoted P^*, then reporting higher than true belief $P > P^*$ will result in only a probability of P^* of obtaining the reward for $n \in [P^*, P]$, as opposed to a probability of n. Similarly, reporting lower than true

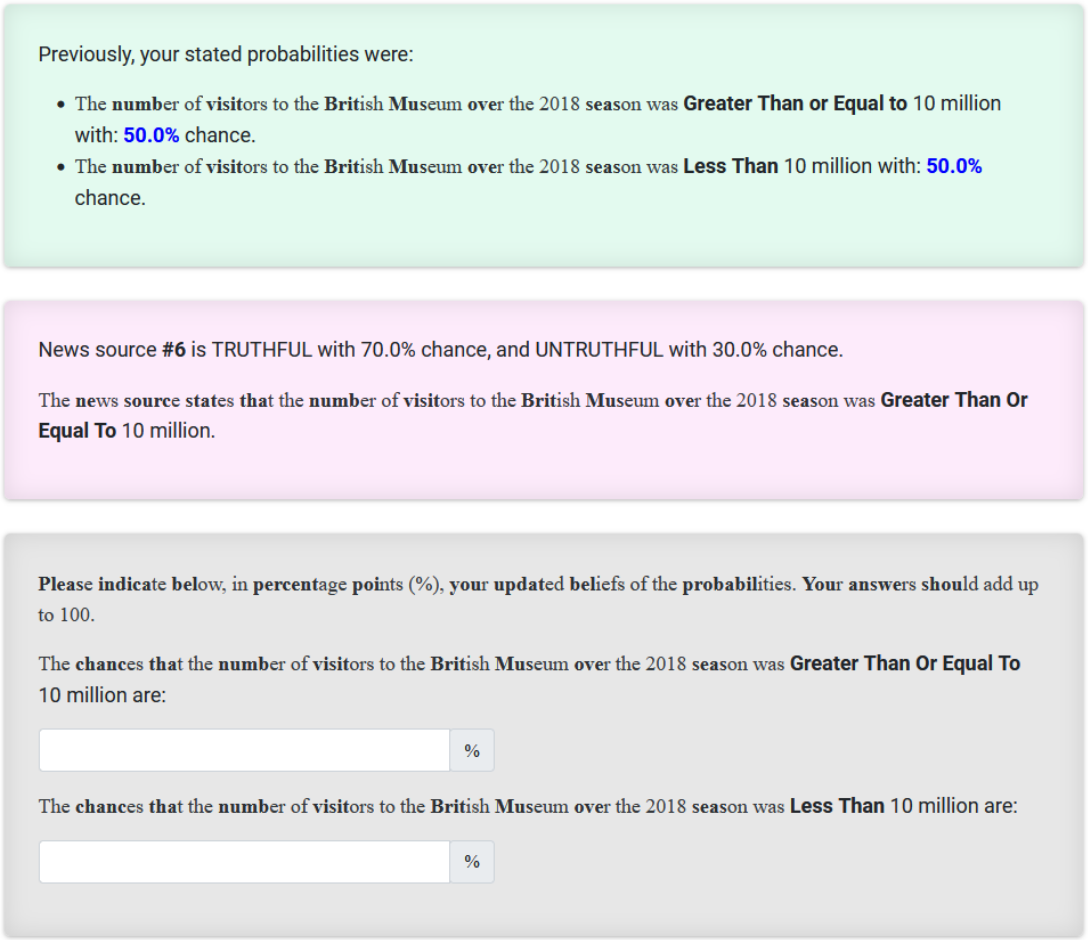


Figure 2.2: Example of signal reception and belief updating.

belief $P < P^*$ will result in only a probability of n of obtaining the reward for $n \in [P, P^*]$, as opposed to a probability of P^*. The payment scheme is not reliant on risk neutrality, unlike the quadratic scoring rule.

2.3.3 Risk Preference Elicitation

The purpose of this section is to control for the individuals' risk preferences, as it may affect their perception of probabilistic events. Subjects face a standard multiple price list presented in a table. The instructions are presented right above the list on the same page. There are 11 choices to be made between a lottery in option A and a sure payoff of option B. Lottery A for

each choice problem has ascending probability for the higher payoff. The experiment interface is programmed so that subjects only need to choose at which point they would like to switch from option A to option B, and clicking on one option in any of the binary choice questions fills the rest of the table. One of the questions is randomly chosen and the subject's choice in the question is implemented as payment.

2.3.4 Survey

The survey includes the basic demographic questions regarding race, gender, and age, as well as the subject's year in college and whether they took a course in statistics before. Importantly, subjects are asked to report their party affiliation. They can identify themselves as "Democrats", "Republicans", "Other" (a third party), or "Neutral/No preference". In addition, I ask the subjects the report their impression of the Republican and Democratic Party, on a scale of 0 (very unlikable) to 10 (very likable), as a sanity check to see if they reported their party affiliation seriously. Subjects are also asked to report how much they care about politics, the economy, and the environment on a scale of 0 (not concerned) to 10 (very concerned).

2.4 Results

2.4.1 Overview

The experiment was conducted in the Experimental Economics Lab at the University of Maryland. Subjects were recruited from the pool of available undergraduate students through ORSEE (55). A total of 183 subjects were recruited over 24 sessions, averaging 7.6 subjects per session. Each session lasted between 30 to 50 minutes. The show-up fee was $7, and the total

payment averaged $15.3 (show-up fee included).

Table B.1 presents the basic demographic composition of the subjects, where the party affiliation was self-reported during the survey. The majority of subjects were Democrats or women, and a test of independence between the two rejects the null hypothesis (p=0.0003), indicating a strong correlation between gender and party affiliation among the participants. Table B.2 presents subjects' ratings of the Democratic and Republican parties, and how concerned they are for the environment, the economy, and political issues respectively, on a scale from 0-10. The Democratic party was generally more likable among subjects, which aligns with their reported party affiliations. Most subjects seemed to care about the environment and the economy to a great extent, and were relatively less concerned about politics, which is also reflected in the proportion of subjects who claim to be neutral in their party affiliation. As a sanity check, note that self-reported Republican subjects indeed rated the Republican party higher than the Democratic party, while it was the other way around for all other subjects, as shown in Table B.3. Checking each individual's ratings also reveals the same story.

There were a total of 48 belief updating instances per subjects, which means there are potentially 48 data points per subject. However, not all subjects answered the questions carefully. First of all, it is expected that the subjects complete each round with increasing speed, both due to familiarity with the task and possibly due to boredom. Figure B.1 shows how the average time used per round changes, and as expected, subjects' time spent on each round dropped from 4 minutes to 1.5 minutes throughout the session. However, there are no outliers on the lower end of time spent, and hence it is hard to judge from the time usage whether they were careless during the task.

Alternatively, it is possible to examine the subjects' belief reports directly. Figure B.2

shows how many subjects have gone through the respective number of rounds entering the same set of beliefs for every signal draw. For example, the leftmost bar shows that 148 subjects changed their reported beliefs at least once in every round, while the rightmost bar shows that 1 subject did not updated her beliefs at all in 7 of the 8 rounds. In the ensuing analysis, I ignore a a subject's data in a round where the subject has not updated her beliefs once. Not responding to information at all most likely means the subject is not paying attention, or the subject has and maintains degenerate beliefs of 0 or 1, in which case the subject is certain about the issue at hand and the question failed to create an uncertain environment for the sake of belief elicitation.[5] However, I do preserve all the answers where subjects reported 0 or 1 as long as they incurred belief changes in that round, and analyzed those data points by changing 0 to 0.001 and 1 to 0.999. With this, I end up with 8280 belief updating instances.[6]

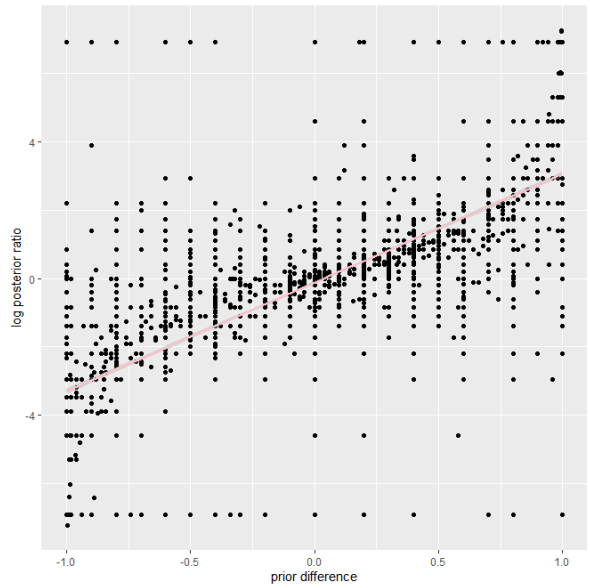
2.4.2 Estimation - Aggregate Level

I now present the main observations on the aggregate level. In the appendix, I perform this regression analysis for the Grether's model as well. Recall the estimation equation (2.3). With the states in the experiment being $\{G, L\}$, I estimate:

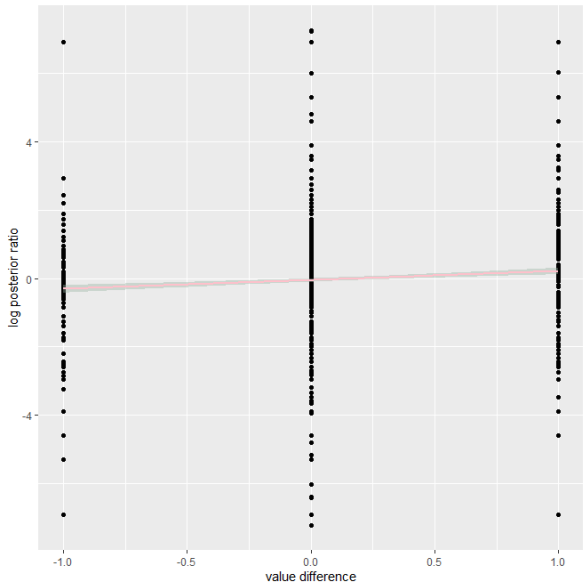
$$ln\left(\frac{\mu_1(G|s)}{\mu_1(L|s)}\right) = \beta_0 + \beta_1 ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) + \beta_2 ln\left(\frac{\pi(s|G)}{\pi(s|L)}\right) + \beta_3\delta_\mu + \beta_4\delta_v \qquad (2.3)$$

[5] Additionally, there were 16 subjects who had over one-fifth of their reported posteriors being more than 80 percentage points away from their prior after one signal. If I delete the subjects who had huge changes in their beliefs, the aggregate level regression results barely change.

[6] Alternatively, if I delete the observations where the subjects updated their beliefs to a 0 belief, and all the observations after that within that round, the parameter estimates do not change drastically. The only exception is the estimate for β_4 will become weaker.



(a) Scatter plot and fitted line for the dependent variable against the difference in prior.



(b) Scatter plot and fitted line for the dependent variable against the difference in value.

Figure 2.3: Scatter plots of log posterior ratio against variables.

Because I do not directly observe valuations, and it is near impossible to elicit "true" valuations associated with political ideologies, I proxy the valuations with normalized values 0 and 1. Specifically, I let the preferred state take value 1 and less favorable state take value 0. For example, Figure 2.1 presented a question regarding the unemployment rate during President Trump's term. A reported Democratic subject would have an assigned value of 1 for state G (greater than or equal to 5%), and a value of 0 for state L (less than 5%), and vice versa for a Republican subject. For questions involving no political context, and for subjects that are impartial towards both the Democratic or Republican parties, I set the value of both states to 0. In sum, δ_v is a ternary variable which takes value 1 if state G is preferred, -1 if state $L$is preferred, and 0 otherwise.

Before I present the OLS regression results, it is worth examining the corresponding scatter plot of the data to obtain a rough idea about the relationship between the dependent variable and our variables of interest δ_μ and δ_v. Figures 2.3a and 2.3b present the two scatter plots respectively.

	Dependent variable:		
	Posterior Log Likelihood		
	OLS - half	OLS - full	Clustered OLS
Prior (β_1)	0.567^{***}	0.544^{***}	0.544^{***}
	(0.022)	(0.016)	(0.051)
Signal (β_2)	0.832^{***}	0.816^{***}	0.816^{***}
	(0.025)	(0.018)	(0.049)
Prior difference (β_3)	1.024^{***}	1.080^{***}	1.080^{***}
	(0.089)	(0.065)	(0.135)
Value difference (β_4)		0.050^{*}	0.050^{*}
		(0.028)	(0.027)
Constant (β_0)	-0.010	-0.018	-0.018
	(0.021)	(0.016)	(0.015)
Observations	4,152	8,280	8,280
Adjusted R^2	0.657	0.628	
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 2.2: OLS regression results with and without clustered standard errors.

As there is large dispersion of the points on the graphs, both due to noise and other factors not accounted for in the graphs, A linear line of best fit is also presented, along with its 95% confidence interval. Both lines are significantly upward sloping, which is a good sign for the analysis that follows.[7]

The OLS results on the aggregate level are presented in Table 2.2. The first column presents the regression results with only the data points for the neutral questions. That is, for the data that do not involve motivated confirmation bias. The second column includes all the questions, that is all 8280 "good" data points, excluding the rounds where the subject did not update her beliefs.

[7] In the appendix, I also discuss alternative specifications of δ_v. All things considered, and noting the fact that [21] also identified motivated confirmation bias according to party affiliation, I present my analysis with the current specification of δ_v.

The third column is the OLS regression with clustered standard errors (grouped by individual).

As expected, the coefficients β_1 and β_2 are significantly positive and less than 1. This means that, similar to previous experiments, subjects exhibited some degree of conservatism and base rate neglect. β_3 is also significant at the 1 percent level, indicating that subjects, in general, exhibited unmotivated confirmation bias.[8] An important observation is that the estimates for all these variables, along with their significance, do not change across the three regressions. β_4 is positive, but only significant at the 10 percent level ($p = 0.0786$), providing some evidence for the existence of motivated confirmation bias. Notably, the standard error for β_4 actually decreased ($p = 0.0635$) for the clustered OLS. While the significance of β_4 is on the weaker side, this isn't unexpected and further examination will shed light on some potential factors that contribute to this relatively weak result.

Result 1. *On the aggregate level, motivated and unmotivated confirmation bias can be observed at the same time, although the elicited motivated confirmation bias is relatively weak.*

2.4.3 Interaction

I examine possible factors for the elusiveness of motivated confirmation bias (apart from context) in previous experiments, and the relatively weak evidence for it on the aggregate level within my experiment. Firstly, an immediate question to answer is whether the dual elicitation of both forms of confirmation bias might have caused them to interfere with each other. If there is any form of interaction, it may explain why only one of the two is observed in several previous research (e.g. 11).

[8] I also present the regression including the risk preference data in the appendix, and show that it has slight effects on the estimates for β_1 and β_3.

To account for potential interactions between the two forms of confirmation bias, I additionally assume a simple multiplicative term in $ln(f(\cdot))$:

$$ln(f(\mu_0(\omega), v(\omega))) = \eta\mu_0(\omega) + \phi v(\omega) + \rho\mu_0(\omega)v(\omega)$$

Denoting $\delta_I = \mu_0(\omega)v(\omega) - \mu_0(\omega')v(\omega')$, the equation estimated becomes:

$$ln\left(\frac{\mu_1(G|s)}{\mu_1(L|s)}\right) = \beta_0 + \beta_1 ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) + \beta_2 ln\left(\frac{\pi(s|G)}{\pi(s|L)}\right) + \beta_3\delta_\mu + \beta_4\delta_v + \beta_5\delta_I \tag{2.3}$$

Table 2.3 presents the corresponding regression results. All of β_0-β_4 are the same as before (as seen in Table 2.2), while the interaction term is not significant at all. This seems to indicate that it the issues regarding motivated confirmation bias in previous experiments is likely not due to the existence of its unmotivated counterpart.[9]

Result 2. *There is no evidence of interactions between motivated and unmotivated confirmation bias.*

2.4.4 Heterogeneity Across Individuals

I turn to estimating the regression model (2.3) on individuals to examine the heterogeneity. Since the subjects updated their beliefs 6 times each round, there is (at most) 48 data points for each individual. Hence, I ignore any subject who have more than 1 round of no belief updating (as seen in Figure B.2). This leaves the sample with 167 subjects left, and a total of 7902 data

[9]The regression only indicates that interactions of the form $\rho\mu_0(\omega)v(\omega)$ is unlikely to exist. However, as this is one of the simplest type of interactions possible, it is hard to expect that assuming a more complicated form of interaction will provide positive results.

	Dependent variable:
	Posterior Log Likelihood
Prior (β_1)	0.545^{***}
	(0.016)
Signal (β_2)	0.816^{***}
	(0.018)
Prior difference (β_3)	1.082^{***}
	(0.066)
Value difference (β_4)	0.051^{*}
	(0.030)
Interaction (β_5)	-0.0001
	(0.001)
Constant (β_0)	-0.017
	(0.018)
Observations	8,280
Adjusted R^2	0.628
F Statistic	$2{,}799.792^{***}$ (df = 5; 8274)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2.3: OLS with interaction between motivated and unmotivated confirmation bias.

points. On the aggregate level, this does not change the results greatly, as shown in Table B.4 column 1.

The cumulative distribution function of the individually estimated β_4 is presented in Figure 2.4.[10] The graph indicates that there is sizable heterogeneity in the estimates. On the one hand, a decent mass of the subjects' β_4 estimates is close to 0. On the other hand, there are still quite a few subjects with estimates greater than 0, and the distribution is slightly skewed to the positive side. It is likely that the heterogeneity affects aggregate level estimates. Conversely, one may also

[10]The distributions for all of β_1-β_4 are presented in Figure B.3 in the Appendix.

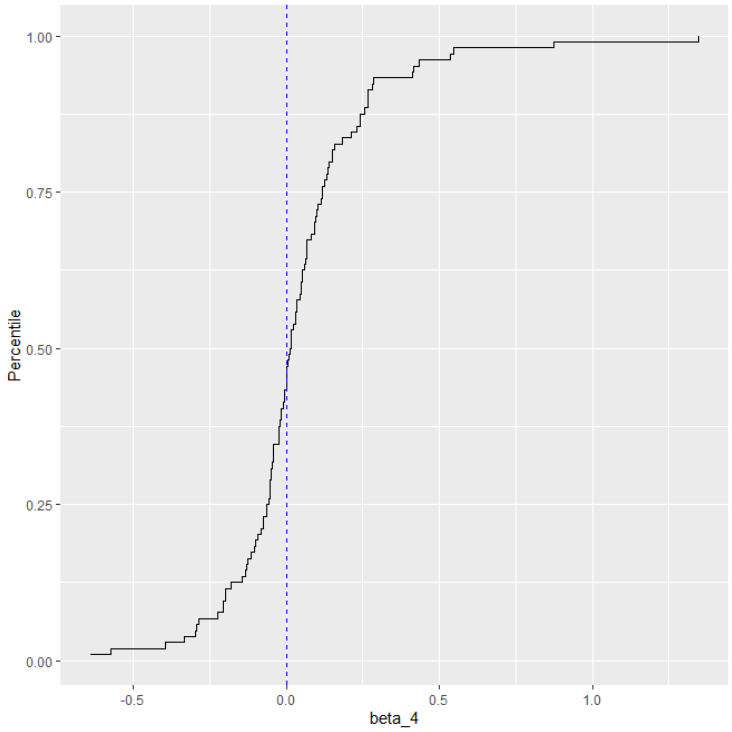


Figure 2.4: $CDF(\beta_4)$

say that aggregate level analyses mask such significant individual differences. The heterogeneity in the β_4 estimates points to the possibility that the mixed results regarding motivated confirmation bias in previous experimental works may be a result of such heterogeneity. It is important to note that [38] also observed a high level of heterogeneity in the individual estimates for motivated confirmation bias, and did not find evidence for the bias on the aggregate level. However, different from [38], my experiment presents positive evidence for the existence of both forms of confirmation bias on the aggregate level. This contrast further shows that such heterogeneity is problematic for aggregate level analyses.

Result 3. *There is sizable heterogeneity in motivated confirmation bias among individuals.*

Finally, individual level regressions also allows examination of the correlation between parameters, as shown in Table B.5. While there is a lack of statistical power to make any formal

conclusions, the lack of correlation between β_4 and any other coefficient does corroborate the observation in the section.[11] Table B.6 shows the correlation between the estimates and the individuals' demographics. From left to right, the demographic variables are gender (dummies taking value 1 for male), year in college, whether they have taken a course in statistics (dummies taking value 1 for 'Yes'), and how much environmental issues, economic issues and political issues concern them (reported on a scale from 0 to 10).[12] None of the recorded demographics seem to relate to the estimated parameters.

2.4.5 Behavioral Differences Across Rounds

As seen from Figure B.1, subjects use significantly less time in later rounds than earlier rounds. One may wonder if the subjects' behavior changes over the rounds. To examine this, I split the belief updating data into two halves, with one containing the first 4 rounds each subject participated in, and the other containing the latter four rounds. Table 2.4 compares the OLS run on each half.

The two halves do not exactly have equal observations due to the deletion of rounds without belief updating occurring. It is fascinating how the estimate for β_4 is significantly different between the two regressions, with it being hardly different from 0 in the regression on the first half of the data ($p = 0.802$), but significantly positive at the 0.05 level in the other ($p = 0.0144$). In contrast, all other estimates have very similar values between the two regressions.[13] This is

[11] The correlation between β_1 and β_3 is expected due to the high correlation between log prior ratios and prior differences at values close to 0. Individually estimated β_1 and β_3 can be especially inaccurate due to this, and this collinearity can mostly explain the extreme outliers shown in graphs B.3a and B.3c. However, this is not a problem on the aggregate level with a large amount of data. The analysis with the Grether's model also shows that the significance of unmotivated confirmation bias is not a product of this collinearity.

[12] There is only 1 non-binary subject and 1 subject who answered 'No applicable' in regards to statistics knowledge, and hence are not left out when calculating these correlations.

[13] I also show a similar result for the trimmed down data set with only subjects who update their beliefs for at least

	Dependent variable:	
	Posterior Log Likelihood	
	First Half	Second Half
Prior (β_1)	0.561^{***}	0.520^{***}
	(0.023)	(0.022)
Signal (β_2)	0.831^{***}	0.801^{***}
	(0.027)	(0.024)
Prior difference (β_3)	1.047^{***}	1.134^{***}
	(0.095)	(0.090)
Value difference (β_4)	0.011	0.090^{**}
	(0.043)	(0.037)
Constant (β_0)	-0.020	-0.015
	(0.024)	(0.021)
Observations	4,152	4,128
Adjusted R^2	0.621	0.638
F Statistic	$1{,}698.565^{***}$ (df = 4; 4147)	$1{,}816.991^{***}$ (df = 4; 4123)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 2.4: Regression on the first half and second half of the belief updating task respectively.

strong evidence for the co-existence of the two forms of confirmation bias. Additionally, this comparison also sheds light on the formation of motivated confirmation bias. Since the questions in the experiment have randomized orders, there should be no order effect, and any observed differences in behavior across rounds should be related to the round number itself. From Figure B.1, we see that the time spent on later questions are significantly shorter on average, and stabilized at around one and a half minutes per round. Answering the initial prior elicitation question and the ensuing 6 belief updating questions within such a short time frame leaves very little time for rigorous thinking. It is possible that the majority of subjects spent less time thinking and resorted

7 out of 8 rounds in Table B.4 columns 2 and 3.

to sentiments or heuristics more in later questions, and this could be connected to a higher level of motivated confirmation bias. Interestingly, this may also imply that the other deviations from Bayesian updating, as represented by β_1-β_3, are intrinsically different from motivated confirmation bias in their underlying mechanisms. Finally, Table B.7 also presents the regressions when splitting the data into the first 2 rounds and last 6 rounds, and the first 6 rounds and last 2 rounds, respectively. This serves as extra evidence for this observation, and show that it is not necessarily a product of an arbitrary division of the data.

Result 4. *Subjects exhibit significantly higher levels of motivated confirmation bias in the later rounds of the experiment, which may be tied to lower response times.*

2.5 Discussion and Concluding Remarks

This paper is an experimental investigation of motivated and unmotivated confirmation bias, and presents evidence supporting the possibility that an individual can exhibit both of them simultaneously. There does not seem to be any interactions between the two forms of confirmation bias, however. Further exploration of the data shows a high degree of individual heterogeneity, which may lead to less significant aggregate level results. Analysis on subjects' behavioral changes across rounds potentially ties motivated confirmation bias to response time, which may be indicative of the deeper mechanisms that underlie the bias. These analyses may be helpful in future research regarding the elicitation and analysis of motivated confirmation bias.

It should be noted that some factors could have affected the results but could not be analyzed. For instance, the subject pool is drawn from University of Maryland undergraduate students, who are likely to be more homogeneous and less extreme in political views compared to

the society as a whole. This may have contributed to the less significant (aggregate level) result compared to, say, [21], who employed a similar experimental task, but conducted the experiment online. Additionally, as [38] noted, there is the possibility of endogeneity via the continuous belief updating structure of the experiment. However, it is difficult to solve using the current design, as there are no exogenous beliefs to serve as instruments. The elicitation of endogenous priors also means that it is hard to perform direct comparisons of belief shifts due to the variation in prior beliefs.

All things considered, I believe a good direction for building upon the current research would be: 1) whether motivated confirmation bias is related to emotions and hasty judgments and whether it reacts to time and cognitive pressures; 2) improving the experimental design by having exogenously determined prior beliefs and one signal/belief updating per round. Further research could also focus on the potential mechanisms behind confirmation bias, and how individual differences may impact the elicitation of confirmation bias.

Chapter 3: Information Acquisition with Wishful Thinking (joint with Emel Filiz-Ozbay)

3.1 Introduction

In the age of information, discovering product quality is still a challenging task for consumers. Many platforms (such as Amazon, Yelp etc.) provide independent product reviews for third party products and online reviews are listed as the second most trusted source of information.[1] Nevertheless, anyone who has ever shopped online may know that reading those reviews is hard and they may not be fully informative even if one is willing to spend hours on them. Why does information remain so difficult to acquire and process? In this paper, we argue that the platforms may have an incentive not to provide transparent reviews in an easily accessible way and competing platforms can promote the accessibility of information but only to a certain extend.

We model an information acquisition problem of consumers (agents) who are rationally (in)attentive to information and have heterogeneous ability to process it. Moreover, a proportion of the consumers are biased when updating their initial beliefs. The platform (firm or principal) cannot choose information on behalf of the consumers but it controls the available information and may impose an upper bound on its informativeness.[2] We characterize when the platform

[1] According to a Nielsen's report, online reviews have 70% approval rate among 25,000 survey respondents. See www.nielsen.com, 2009 Report "Global Advertising Consumers Trust Real Friends and Virtual Strangers the Most"

[2] The platform may achieve this via deleting reviews or adding fake reviews to the customer review section.

may benefit from blurring the accuracy of reviews to increase the sales and show that competition among platforms sometimes (but not always) help consumers gain better information. Our findings are consistent with the recent federal register notice of the Federal Trade Commission[3] that is motivated by the lack of transparency in commercial reviews and that aims to promote honest businesses and prevent deception of consumers looking for real feedback on a product.

Our model is an application of the rational inattention approach introduced by [56] (and also 57.)[4] This model is suitable for agents who hold an initial prior on the state of the world and would like to pick an information structure to reduce uncertainty before taking an action. Since the utility depends on the action and the state of the world, it is better to pay attention to a more informative signal structure to avoid mistakes but attention costs are proportional to the reduction of uncertainty in the agent's belief. This approach has been utilized in other principal/agent problems such as dynamic bargaining (58) and sequential pricing (59, 60).[5] Rational inattention in these settings lead to non-standard results such as delayed attention in dynamic settings or uninformative prices in a quality discovery problem. Similar to these applications, the principal (the platform) benefits from sales in our setup. However, different from the other papers, it only cares about the volume of sales rather than the product price.[6] The platform can control how informative the available information is. In applications, platforms design product review pages, control the accessibility of information, customize product attribute summaries, or allow

[3] "Trade Regulation Rule on the Use of Consumer Reviews and Testimonials," 16 CFR Part 465-Matter Number: R311003, July 31, 2023.

[4] This framework has become a groundbreaking tool for information economics. It relaxed the assumption of agents being able to process all information freely. Instead, it models agents who understand the trade-off between using better information and the cost of acquiring that information.

[5] Several other papers apply rational inattention to pricing, where [61] and [62] assume the price setters to be rationally inattentive, while [63], [64], and [65] maintain the assumption that the buyers are rationally inattentive. [66], [41], [67], and [65] also build on this model.

[6] This is similar to platforms which provide information on products sold by third parties. The seller (not modeled here) determines the price and the platform receives a fixed commission per unit of sale.

customers to sample the product. These strategies or their absence might be interpreted as tools to improve or limit the informativeness of signals acquired by the consumers. To the best of our knowledge, ours is the first application of rational inattention where such constraints on consumers are analyzed.

We first show that the platform cannot benefit from imposing an upper bound on the accuracy of information if consumers are Bayesian. Such consumers will choose their attention level optimally, update their beliefs in the Bayesian sense, and follow the signal they receive when deciding whether to purchase the product. There is, however, significant evidence against the Bayesian updating assumption and even those who intend to be Bayesian may fail to employ the rule when they are actually updating their priors (see e.g., 12, 68, 69). Several non-Bayesian updating rules have been proposed and implemented in well-known decision environments in economics[7] and one of the most prevalent deviations from Bayesian updating is wishful thinking. This bias makes agents form beliefs more in line with their desire rather than rationality implied by Bayesianism. In our two-state environment, we interpret this as the possibility of some agents asymmetrically overweighting good news about the product quality relative to bad news, due to their desire to land a high quality product (see 18, 25, 47). Such bias provides an incentive for the platform to limit the informativeness of the information. The intuition is simple. Let the consumption utilities of buying high and low-quality products be 1 and -1, respectively. Let the prior belief of quality being high be slightly less than being low. If the posterior belief on high quality exceeds 0.5, the consumer will buy the product (assuming that the utility of not buying is zero). A rationally inattentive, Bayesian person would follow the signal recommendation and in expectation buy exactly at the rate implied by the marginal probability of receiving good sig-

[7]Examples include [7, 8, 10, 15, 26, 47, 48, 70]

nals. Consider a wishful thinker instead. Even if she chooses her optimal information structure, her posterior will be more optimistic than the Bayesian after a good signal and less pessimistic after a bad signal. Hence, this consumer may under-interpret a bad signal or over-interpret a good signal and end up with a posterior of high-quality being more than 0.5. The existence of such non-Bayesians benefits the platform by increasing the volume of sales in expectation. We characterize the parameters of the model that will lead to such scenarios.

Next, we introduce competition into the model where the incumbent platform might be challenged by an entrant platform. In this setup, consumers randomly choose their platform to collect information, and if they feel constrained then they switch to the other platform looking for more information. We characterize possible equilibrium outcomes of this competitive game. We show that for a certain range of parameters, the introduction of competition achieves unrestricted information structure and benefits consumers. However, there is a non-empty set of parameters where the platforms coordinate and both of them restrict the information. Given that some platforms have grown too big and competition against them by using standard tools, such as price and product variety, has become almost impossible, competition through information quality might be an effective tool against monopolies. Nevertheless, this result depends on the market power of the incumbent. For too large platforms, regulations for transparency are still justified.

Since the principal in our setup can control the information, our paper also relates to the persuasion literature (71). [72] extend the standard persuasion framework to games between a Bayesian sender and a non-Bayesian receiver. [73] analyzes optimal persuasion for rationally inattentive receivers. The game we study is related with these but ours is not a persuasion game since the platform does not know the state of the world in our setup, and it cannot choose a signal structure. It can only constrain the informativeness of the available signal structures of the

consumer.

Finally, our paper relates to the literature on exploitative contracts. In these papers, at least a proportion of consumers are either unaware of their own bias or unaware of some product attribute, and the firms exploit this unawareness. [74] discussed the implications of present bias for the optimal pricing strategy of subscription-based products such as gym memberships. [75] showed that firms can also manipulate naive consumers by actively hiding relevant information about add-on purchases. The firm of our model exploits the consumer's unawareness of their biased updating, but distinct from this literature, it manipulates the consumer with blurry information rather than manipulated price.

In the following sections, we first start with introducing the game (Subsection 3.2.1) followed by characterizing the optimal attention level of the consumer whose information might be constrained (Subsection 3.2.2). Then we introduce non-Bayesian consumers to the model and study the purchasing decisions of Bayesian and biased consumers (Subsection 3.2.3.) In Subsection 3.2.4, we study the best response of a monopolist platform and characterize when it is optimal for it to constrain the information available to the consumers. In Subsection 3.2.5, we study the competition between platforms and characterize when competition on information can help consumers. Subsection 3.3 introduces extensions of our setup to other non-Bayesian rules besides wishful thinking. Subsection 3.4 discusses alternative models such as the case where the firm knows the product quality (the state of the world) before constraining information or where consumers have other types of heterogeneity rather than information costs. The concluding remarks finalizes Section 3.4.

3.2 Rational Inattention with Wishful Thinking

3.2.1 Setup

A platform (firm) sells a third-party product through its platform. A population of consumers use the platform to gather more information about the product quality (for instance, by reading the customer reviews) and decide whether to buy the product.[8]

Specifically, the product quality, ω, is unknown and it can be "High" or "Low", $\omega \in \{H, L\}$. The consumers have a common prior belief over the quality, $f(H) \leq 0.5$, and can choose one of two actions, "buy" or "reject" denoted by $a \in \{b, r\}$.[9] We assume a normalized utility:

$$u(b|H) = 1 > 0 = u(r|H) = u(r|L) > -1 = u(b|L)$$

Before making a purchasing decision, each consumer chooses how much information about the product quality she wants to gather.[10] We employ a standard rational inattention framework for information acquisition. Formally, a consumer chooses a signal structure $\pi(s|\omega)$ for two possible signals $s \in \{h, l\}$, which costs her $\lambda I(\pi)$.[11] λ represents the consumer's marginal cost of information acquisition. We assume that λ is private information and the consumers have heterogeneous cost parameters. The firm knows the distribution of this parameter, $\lambda \sim G[0, \infty)$. $I(\pi)$ measures informativeness of the signal by calculating the degree of correlation between the

[8] The price and quality are determined by a third party that is not modeled here.

[9] The exercise can be easily extended for the case of $f(H) > 0.5$ but the solution for that would be trivial, as we will discuss in Section 3.4.

[10] We assume that consumers always acquire signals, and in the case that acquiring any meaningful information is too costly, they will simply choose a totally uninformative signal structure.

[11] We have two possible signals since having more signals than the number of actions is redundant.

signal and the state.[12]

$$I(\pi) = \sum_{s} \sum_{\omega} \pi(s|\omega) f(\omega) ln(\frac{\pi(s|\omega)}{q(s)})$$

where $q(s)$ denotes the marginal probability of signals given the prior, f, and the signal structure, π, i.e. $q(s) = \sum_{\omega} \pi(s|\omega) f(\omega)$.

Consumer's problem of choosing π: We can write the consumer's information acquisition problem as

$$\begin{aligned} max_{\pi} \sum_{\omega} \sum_{s} V(s, \pi) \pi(s|\omega) f(\omega) - \lambda I(\pi) \\ \text{subject to } I(\pi) \leq \kappa \end{aligned} \tag{3.1}$$

Here, V denotes the expected utility of the consumer who believes that she will buy or reject optimally under information structure π. Formally,

$$V(s, \pi) = max_{a \in \{b,r\}} g^B(H|s) u(a, H) + g^B(L|s) u(a, L) \tag{3.2}$$

where $g^B(\omega|s)$ is the Bayesian prosterior generated by π.

κ in the constraint of problem (3.1) is an upper bound on informativeness of the signal structure. Shortly, we will introduce the firm as the party which sets κ optimally. Before doing that, first we characterize the solution to problem (3.1).

[12] This is called the "mutual information" between the signal and the state. It is also equivalent to the "relative entropy", or the Kullback-Leibler divergence, between the joint distribution of s and ω and the product of their marginal distributions. Entropy is a measure of uncertainty implied by a distribution, and for distribution $\gamma(x)$, Entropy is defined as $-\sum_x \gamma(x) ln(\gamma(x))$.

3.2.2 Characterizing the Optimal Signal Structure, π

Note that [41] Lemma 1 applies to our setting, and hence, distinct signals should not lead to the same action if the information is selected optimally because it is inefficient to acquire information that will not be acted upon.[13] One can solve problem (3.1) using the standard Lagrangian method and the solution to it is summarized in Lemma 2 below. Note that the optimal information for this problem is a piece-wise function with a cutoff cost parameter, $\underline{\lambda}$. This cutoff is the cost parameter of the consumer who will optimally choose an information structure with exactly $I(\pi) = \kappa$.

In the proof of Lemma 2, we first observe that the upper bound on the informativeness of the information structure translates into an equivalent lower bound on the marginal cost of information, λ, pushing all those who have a lower cost act like they have the cost of $\underline{\lambda}$. The details are shown in the appendix. Intuitively, only those who have low enough λ would like to acquire a large amount of information and hence, they will be the ones who are constrained by the firm's choice of κ. The next best option for these constrained consumers is to mimic those at the threshold with $\lambda = \underline{\lambda}$. Those who do not acquire that much information to start with are unaffected by the constraint.

Lemma 2. *Given a constraint κ, the optimal information structure of a consumer with cost*

[13] This property is implied by the convexity of the entropy-based cost function that is used in the rational inattention model.

parameter $\lambda \in [0, \infty)$*, is a piece-wise function defined by* $\underline{\lambda}$ *such that* $I(\underline{\lambda}) = \kappa$ *and*

$$\pi^*(h|H) = \begin{cases} 0 & q^*(h) = 0 \\ \frac{q^*(h)e^{1/max\{\underline{\lambda},\lambda\}}}{f(H)(1+e^{1/max\{\underline{\lambda},\lambda\}})} & q^*(h) \in (0,1) \\ 1 & q^*(h) = 1 \end{cases} \quad \pi^*(l|L) = \begin{cases} 0 & q^*(l) = 0 \\ \frac{q^*(l)e^{1/max\{\underline{\lambda},\lambda\}}}{f(L)(1+e^{1/max\{\underline{\lambda},\lambda\}})} & q^*(l) \in (0,1) \\ 1 & q^*(l) = 1 \end{cases} \tag{3.3}$$

with $\pi^*(h|\omega) = 1 - \pi^*(l|\omega)$*, where*

$$q^*(h) = \begin{cases} max\left\{0, min\left\{1, \frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}}\right\}\right\} & \lambda \in [\underline{\lambda}, \infty) \\ max\left\{0, min\left\{1, \frac{f(L)-f(H)e^{1/\underline{\lambda}}}{1-e^{1/\underline{\lambda}}}\right\}\right\} & \lambda \in [0, \underline{\lambda}) \end{cases} \tag{3.4}$$

and $q^*(l) = 1 - q^*(h)$.

The marginal probability of observing signal h equals to the fractional term within the brackets in Equation (3.4) provided that it is in $(0, 1)$ interval. q^* is continuous in f and λ, and one may further show that π^* is also continuous in f and λ. One should also note that under our assumption of $f(H) \leq 0.5$, $q^*(h) \neq 1$. Nevertheless, the result of Lemma 2 applies without this assumption, hence, we state it in a more general way.

3.2.3 Bayesian and non-Bayesian consumers and their purchasing decisions

After picking the optimal information as described by Lemma 2, consumers observe their signals and update their beliefs. At this stage proportion $p \in (0, 1)$ of the consumers update their belief in a biased way, while the remaining $1 - p$ are Bayesian. Importantly, we let the biased consumers be naive, and not factor in their bias when choosing the signal structure.[14] In this

[14]These consumers are unaware of their own bias, and thus expect themselves to interpret information objectively when they choose the optimal π while solving problem (3.1). Cognitive biases are largely subconscious processes.

section, we will use an updating rule that describes wishful thinking (18). Wishful thinking is a natural explanation for why a person would hold optimistic beliefs about a product's quality. However, the results we will derive for wishful thinkers can be generalized to a much larger set of updating rules, which will be discussed in Section 3.3.

Let $g^{\delta}(\omega|s)$ denote the posterior belief of a consumer who distorts the Bayesian posterior with parameter vector $\delta = (\delta_H, \delta_L)$.

$$g^{\delta}(\omega|s) = \frac{f(\omega)\pi(s|\omega)\delta_{\omega}}{f(H)\pi(s|H)\delta_H + f(L)\pi(s|L)\delta_L} \tag{3.5}$$

Here, δ is the distortion parameter vector of a wishful thinker who uses a state-specific weight, as in [18]. g^{δ} is equal to the Bayesian posterior when $\delta_H = \delta_L$. We assume that biased consumers have $\delta_H \geq \delta_L > 0$, which represents the wishful thinking that the product is high quality.[15] Finally, we assume that being a biased consumer and the realization of λ are independent events, as it is intuitive to assume that having a high information processing cost of λ and being a wishful thinker are not correlated.

After obtaining a signal generated by the information structure, the consumers update their beliefs and decide to buy or reject by maximizing their (possibly biased) expected utility:

$$V^{\delta}(s, \pi) = max_{a \in \{b,r\}} g^{\delta}(H|s)u(a, H) + g^{\delta}(L|s)u(s, L)$$

The expected utility (V^{δ}) above depends on the posterior beliefs which might be biased as

According to [76]: "The intuitive processes that precede biased judgments and decisions, and that are at the basis of the behavior that is ultimately shown, are largely implicit and unconscious".

[15] For instance, a person who browses Amazon and sees a pair of shoes that she loves would also wish to obtain confirmation from the customer reviews that the product has good quality.

the bias only hits a portion of the consumers when they are updating their priors after collecting a signal. Hence, for a non-Bayesian consumer, $V^{\delta}(s,\pi)$ may be different from $V(s,\pi)$ calculated in Equation (3.2).

It is easy to observe that the optimal action a^* is to "buy" the product if $g^{\delta}(H|s) > g^{\delta}(L|s)$, and it is to "reject" it otherwise, due to the normalization of the utility function. Hence, we need to analyze which signals will lead to each of these for Bayesian and non-Bayesian consumers.

Before stating the overall optimal behavior of consumers in Lemma 3, we will provide the intuition behind the purchasing decision. First, one can obtain the actual posterior belief of each consumer using (3.5), and determine their expected utility of buying as a function of the signal received. The proof of Lemma 3 derives Equation (3.6) for $q(h) \in (0,1)$ and Equation (3.7) for $q(h) = 0$.

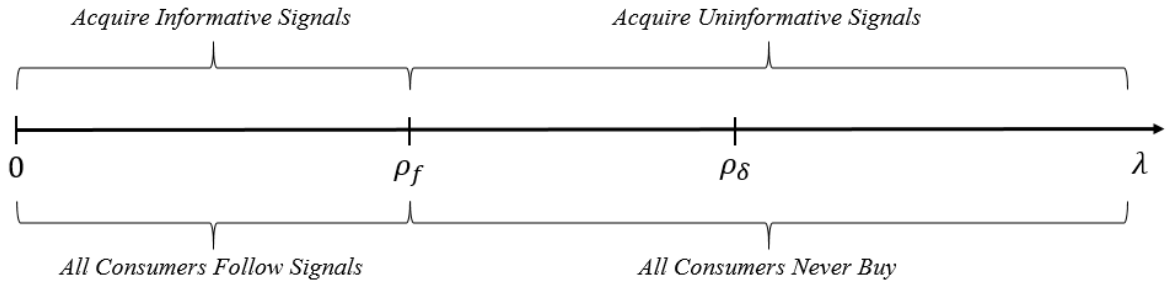
$$Eu^{\lambda}(b|s=l) = \frac{\delta_H - e^{\frac{1}{max\{\lambda,\lambda\}}}\delta_L}{\delta_H + e^{\frac{1}{max\{\lambda,\lambda\}}}\delta_L} \quad \text{and} \quad Eu^{\lambda}(b|s=h) = \frac{e^{\frac{1}{max\{\lambda,\lambda\}}}\delta_H - \delta_L}{e^{\frac{1}{max\{\lambda,\lambda\}}}\delta_H + \delta_L} \tag{3.6}$$

$$Eu^{\lambda}(b|l) = \frac{f(H)\delta_H - f(L)\delta_L}{f(H)\delta_H + f(L)\delta_L} \tag{3.7}$$

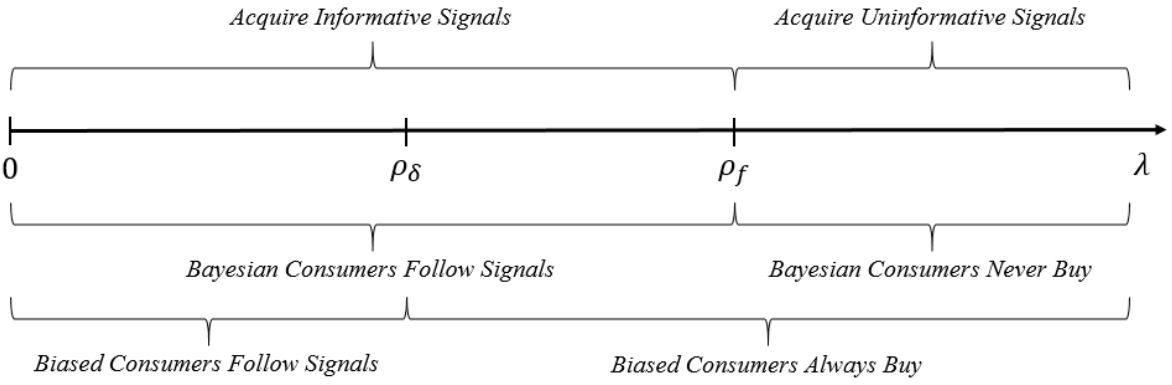
Recall that the utility of rejecting is always 0. Given the choice of π^*, the consumers' optimal action can be discussed under four cases, summarized in Table 3.1. The empty cells correspond to cases not applicable to the Bayesian consumers because they do not have a bias ($\delta_H = \delta_L$). Whether a consumer acquires informative or uninformative signal and her purchasing decision after collecting a signal varies with her cost, λ, and the relation between her prior and

	$q(h) \in (0,1)$		$q(h) \in \{0,1\}$	
	$\delta_H < e^{1/max\{\lambda,\underline{\lambda}\}}\delta_L$	$\delta_H \geq e^{1/max\{\lambda,\underline{\lambda}\}}\delta_L$	$f(H)\delta_H \geq f(L)\delta_L$	$f(H)\delta_H < f(L)\delta_L$
Bayesian ($\delta_H = \delta_L$)	Follow s	–	–	Always Reject
Biased ($\delta_H > \delta_L$)	Follow s	Always Buy	Always Buy	Always Reject

Table 3.1: Consumer's best response.



(a) $\rho_f < \rho_\delta$



(b) $\rho_f \geq \rho_\delta$

Figure 3.1: Information acquisition and purchasing decisions depending on λ.

bias. We illustrate each case in Table (3.1) also on Figures (3.1a) and (3.1b) for different cost parameters. These figures will be useful when we analyze the firm's optimal strategy later.

Below we explain the intuition behind the behavior described in each cell of Table 3.1.

Case 1: $0 < q(h) < 1$

Sub-Case 1.1: $\delta_H \geq e^{1/max\{\lambda,\underline{\lambda}\}}\delta_L$

This results in $Eu^{\lambda}(b|s = h) > Eu^{\lambda}(b|s = l) > 0$, and the biased consumer will always

buy regardless of signal. In other words, the consumer's bias is large enough for them to ignore a bad signal. This case doesn't occur for a Bayesian since $\delta_H = \delta_L$.

Sub-Case 1.2: $\delta_H < e^{1/max\{\lambda,\underline{\lambda}\}}\delta_L$

This results in $Eu^{\lambda}(b|s = h) > 0 > Eu^{\lambda}(b|s = l)$, and the biased consumer will follow the action recommended by their acquired signal, i.e. "buy" if $s = h$ and "reject" if $s = l$. For a Bayesian person, $\delta_H = \delta_L$, and this is always the case.

Next is the extreme case of $q(h) \in \{0, 1\}$. Recall that, according to Lemma 2 the optimal signal structure is the uninformative one for this case.

Case 2: $q(h) \in \{0, 1\}$

Observe that for this case Equation (3.7) applies and the utility does not depend on λ. For a Bayesian person, the expected utility is $2f(H) - 1 \leq 0$, and thus will never buy.[16] For a biased person, there are again two cases to consider.

Sub-Case 2.1: $f(H)\delta_H \geq f(L)\delta_L$

The biased consumer always buys in this case, as the expected utility of buying is greater than or equal to 0.

Sub-Case 2.2: $f(H)\delta_H < f(L)\delta_L$

The biased consumer never buys in this case, as the expected utility of buying is less than 0.

To simplify the notation for the statement of Lemma 3, define the following:

$$\rho_\delta = \frac{1}{ln(\delta_H) - ln(\delta_L)} \quad \text{and} \quad \rho_f = \frac{1}{ln(f(L)) - ln(f(H))} \tag{3.8}$$

[16] Note that $2f(H)-1 < 0$ when $f(H) < 0.5$, and it is equal to 0 when $f(H) = 0.5$. However, when $f(H) = 0.5$, $q(h) = q(l) = 0.5$, and the consumers will always acquire somewhat informative signals, and will never reach the case of $q(h) \in \{0, 1\}$.

By our assumptions on f and δ, these are positive (with the exception that the former is undefined for the Bayesian individual). We can potentially re-write the thresholds for each of the above cases with ρ_δ, ρ_f and λ. The following lemma summarizes our discussion above, while emphasizing on the important thresholds in relation to λ.

Lemma 3. *The individual acquires informative signals, i.e., $q(h) \in (0,1)$, if and only if $max\{\lambda, \underline{\lambda}\} < \rho_f$. The optimal purchasing behavior of Bayesian and non-Bayesian consumers are as follows:*

- *If the individual is Bayesian, they will follow the signal recommendation (i.e. buy the product if and only if the signal is high).*
- *If the individual is biased, they will ignore the bad signal of l, and always buy the product if and only if $max\{\underline{\lambda}, \lambda\} \geq \rho_\delta$ or $\rho_f \geq \rho_\delta$. They will follow the signal recommendation otherwise.*

The proof is in the appendix.

Since we have derived the optimal information choice and purchasing decisions of the consumers, next, we will examine the firm's problem of selecting the optimal constraint κ. The figures presented earlier are useful to visualize the information choice strategies and purchasing decisions of Bayesian and non-Bayesian consumers. We will refer to these figures in our derivation of the firm's optimal constraint, κ.

3.2.4 Firm's optimal constraint on information

The firm can limit information acquisition by setting κ. This may correspond to instances where platforms manipulate customer reviews to their advantage. This inevitably affects the

views of the potential buyers, and the probability of purchasing the product. Recall that a consumer with cost λ purchases when her posterior satisfies $g^{\delta}(H|s) > g^{\delta}(L|s))$. Then the firm's problem is:

$$max_{\kappa} \int_{\lambda} E[\mathbb{1}(g^{\delta}(H|s) > g^{\delta}(L|s))|\lambda] dG(\lambda) \tag{3.9}$$

According to the previous discussions, for any κ that the firm chooses, there exists a unique $\underline{\lambda}$ such that $I(\pi^{\underline{\lambda}}) = \kappa$. Therefore, the ensuing analyses will assume that the firm is directly setting $\underline{\lambda}$. As observed in Lemma 2, having a constraint only affects those consumers whose constraints bind, i.e., $\lambda \leq \underline{\lambda}$. Proposition 10 identifies a threshold for the prior above which the optimal $\underline{\lambda}$ is non-zero, and hence, the firm wants to constrain the consumers. Otherwise, the firm cannot benefit from blurring the consumer's information. The proof is in the appendix and it analyzes the change in expected profit when the firm imposes a $\underline{\lambda}$ versus when it doesn't constrain at all.

Proposition 10. *Let $p > 0$, $\exists \hat{f} \in [\frac{\delta_L}{\delta_H + \delta_L}, 0.5]$ such that for all prior beliefs such that $f(H) \in [\hat{f}, 0.5]$, $\underline{\lambda}^* = \rho_{\delta}$. Otherwise, for $f(H) \notin [\hat{f}, 0.5]$, the firm is better off without imposing a constraint.*

In the proof of this result, we first notice that the firm cannot benefit from constraining the Bayesian consumers. This is because the Bayesian consumers either acquire an informative signal and follow it or acquire an uninformative signal and reject the product. Under a constraint either the behavior of such consumer does not change or it switches from following the signal to not buying at all (see Table 3.1 and Figures 3.1a and 3.1b.) That cannot benefit the firm. Hence, it is crucial to have biased consumers for the firm to set non-trivial constraint optimally, i.e. we need $p > 0$.

The proof of Proposition 10 shows next that the firm can benefit from exploiting those biased consumers who would collect informative signal due to low marginal cost and follow the signal when unconstrained, and would have bought the product independent of the signal when constrained. Such consumers are the ones who have $\lambda \leq \rho_\delta$ when a constraint in the interval of (ρ_δ, ρ_f) is implemented. We note that it is sub-optimal to set $\underline{\lambda} \in (0, \rho_\delta)$, because the constrained consumers still have a low enough marginal cost to acquire informative signals and follow them, and decreasing $\underline{\lambda}$ will simply increase the probability that h signals are realized and dominate such $\underline{\lambda}$. Similarly, it is sub-optimal to set $\underline{\lambda} > \rho_\delta$, because further increasing $\underline{\lambda}$ only decreases the chances that Bayesian individuals would buy the product. In sum, the firm is deciding between $\underline{\lambda} \in \{0, \rho_\delta\}$, and it will find it more profitable to set $\underline{\lambda} = \rho_\delta$ when the prior belief of the good being high quality is not too low (so the case in Figure 3.1b rather than Figure 3.1a occurs.) Otherwise, the exploitable biased consumers may not have the intention to buy the product in the first place.

3.2.5 Entry Game

We have just discussed the conditions under which the monopolistic firm finds it profitable to constrain information acquisition, which was summarized in Proposition 10. In sum, this case would occur only for $\rho_f \geq \rho_\delta$. In what follows, we will analyze how the entry of a competitor firm affects the equilibrium strategies of the incumbent firm, which would constrain the consumers when it is a monopoly, i.e., $\rho_f \geq \rho_\delta$.

Consider a simple entry game with an entrant and an incumbent firm. Assume the incumbent firm has been facing the problem described in the previous section, with $p \in (0, 1)$, and the

parameters are such that the optimal strategy for a monopolist is setting $\underline{\lambda}^* = \rho_\delta$. The entrant can choose to enter the market or not. The profit of not entering is set to zero. If the entrant chooses to enter, it also needs to make a choice of constraining information or not. To differentiate the two firms, denote the incumbent's choice as $\underline{\lambda}^I$ and the entrant's choice as $\underline{\lambda}^E$. Assume that initially, a proportion of $\eta \in [0, 1]$ of the consumers randomly choose the entrant's platform, while the rest use the incumbent's. However, each consumer will try out the other firm's sales platform if they feel constrained by the current one. Consumers stick with their initially picked platform when they are constrained by both platforms. As a result, the firm with the lower $\underline{\lambda}$ gets all the consumers with $min\{\underline{\lambda}^I, \underline{\lambda}^E\} \leq \lambda \leq max\{\underline{\lambda}^I, \underline{\lambda}^E\}$, and for the remaining consumers ($\lambda > max\{\underline{\lambda}^I, \underline{\lambda}^E\}$ or $\lambda < min\{\underline{\lambda}^I, \underline{\lambda}^E\}$), the entrant gets a share of $\eta \in [0, 1]$. We proceed to find conditions that characterize the equilibrium when the firms move sequentially and the incumbent moves first. The detailed proof of the following discussion is in the appendix (proof of Proposition 11), and we will outline the general logic here.

Firstly, note that neither firm will choose the strictly dominated strategy of $\underline{\lambda} > \rho_\delta$, which was established in the previous section. The addition of another firm only makes this choice less desirable due to the other firm's ability to undercut and attract all the constrained consumers with a less restrictive constraint. Similarly, any $0 < \underline{\lambda} < \rho_\delta$ is dominated by $\underline{\lambda} = 0$, due to the reasons stated earlier, as well as the undercutting by the competitor firm. Hence, it suffices to consider $\{0, \rho_\delta\}$ as the set of potential strategies for equilibrium. We first find the best responses of the entrant given the incumbent's choice of either ρ_δ (case 1) and 0 (case 2.) Afterward, we will characterize the optimal strategy of the incumbent given the best responses of the entrant.

Case 1: $\underline{\lambda}^I = \rho_\delta$

The entrant has two responses: 0 or ρ_δ. The expected profits of the entrant, $\Pi^E(\underline{\lambda}^I, \underline{\lambda}^E)$ for

each strategy are:

$$\Pi^E(\rho_\delta, 0) = \int_0^{\rho_\delta} \frac{1-f-fe^{1/\lambda}}{1-e^{1/\lambda}} dG(\lambda) + \eta \left[p \int_{\rho_\delta}^{\infty} dG(\lambda) + (1-p) \int_{\rho_\delta}^{\rho_f} \frac{1-f-fe^{1/\lambda}}{1-e^{1/\lambda}} dG(\lambda) \right]$$

$$\Pi^E(\rho_\delta, \rho_\delta) = \eta p \int_0^{\infty} dG(\lambda) + \eta(1-p) \left[\int_0^{\rho_\delta} \frac{1-f-fe^{1/\rho_\delta}}{1-e^{1/\rho_\delta}} dG(\lambda) + \int_{\rho_\delta}^{\rho_f} \frac{1-f-fe^{1/\lambda}}{1-e^{1/\lambda}} dG(\lambda) \right]$$

The first equation is the expected profit of the entrant when it doesn't constrain the consumers, while the incumbent sets $\underline{\lambda}^I = \rho_\delta$. The first term is the expected probability of the entrant selling the product to the consumers who are constrained by the incumbent and, thus, come to acquire information and purchase the product from the entrant. The second term describes the consumers who are indifferent between the two firms because they are not constrained by either. A proportion of η of those consumers go to the entrant. p of them are biased (first term in the bracket) and always buy the product, and $1-p$ of them are Bayesian (second term in the bracket) and follow their signals. One may refer to Table 3.1 for these purchasing decisions.

The second equation is the expected profit of the entrant when it mimics the incumbent's strategy and sets $\underline{\lambda}^E = \rho_\delta$. The first term indicates that all the biased consumers will buy the product regardless of signal due to the firm's constraint. Again, only η share of them go to the entrant. The second term corresponds to the Bayesian consumers who go to the entrant and follow their signals. The first integral in the bracket is the profit generated by the constrained Bayesian consumers and hence, their λ ranges from zero to ρ_δ. The second integral in the bracket is the profit generated by the unconstrained Bayesian consumers and hence, their λ ranges from ρ_δ to ρ_f. Any Bayesian consumer with higher λ will pick an uninformative signal structure and reject the product. The entrant will find constraining information acquisition not profitable and

hence, provide better information to the consumers as long as Inequality (3.10) is satisfied.

$$\Pi^E(\rho_\delta, 0) > \Pi^E(\rho_\delta, \rho_\delta) \iff$$
$$\int_0^{\rho_\delta} \left[-\eta p + \frac{1 - f - fe^{1/\lambda}}{1 - e^{1/\lambda}} - \eta(1-p)\frac{1 - f - fe^{1/\rho_\delta}}{1 - e^{1/\rho_\delta}} \right] dG(\lambda) > 0 \quad (3.10)$$

Denote the set of values for η and p that violate condition (3.10) as $C = \{(\eta, p) | \Pi^E(\rho_\delta, 0) = \Pi^E(\rho_\delta, \rho_\delta)\}$. The dashed curve in Figure 3.2 is the boundary of C.[17] Since the left hand side of Inequality (3.10) is linear in η and p, for all (η, p) such that $(\eta, p) \geq (\hat{\eta}, \hat{p})$ for some $(\hat{\eta}, \hat{p}) \in C$, the entrant finds it optimal to also constrain information acquisition. For other (η, p) values, the entrant finds it optimal to not constrain information acquisition. Realistically, entrants would have a fairly low share η, and thus, it is possible that the inequality would hold for all positive p.

Case 2: $\underline{\lambda}^I = 0$

This case analyzes the entrant's strategy if the incumbent imposes no constraint to pre-empt a competition through information. First, we calculate the expected profits of the entrant, $\Pi^E(\underline{\lambda}^I, \underline{\lambda}^E)$ when it sets no constraint (i.e., $\underline{\lambda}^E = 0$) and when it sets $\underline{\lambda}^E = \rho_\delta$, given $\underline{\lambda}^I = 0$.

$$\Pi^E(0, 0) = \eta \left[\int_0^{\rho_\delta} \frac{1 - f - fe^{1/\lambda}}{1 - e^{1/\lambda}} dG(\lambda) + p \int_{\rho_\delta}^{\infty} dG(\lambda) + (1-p) \int_{\rho_\delta}^{\rho_f} \frac{1 - f - fe^{1/\lambda}}{1 - e^{1/\lambda}} dG(\lambda) \right]$$

$$\Pi^E(0, \rho_\delta) = \eta \left[p \int_{\rho_\delta}^{\infty} dG(\lambda) + (1-p) \int_{\rho_\delta}^{\rho_f} \frac{1 - f - fe^{1/\lambda}}{1 - e^{1/\lambda}} dG(\lambda) \right]$$

The first equation is the expected profit of the entrant when neither firm constrains information. The first term describes the consumers who follow their signals, whether they are biased

[17] In the proof of Proposition 11, we will show that the boundary of set C can be expressed as a decreasing function $\hat{p}(\eta)$ and has the form illustrated in the figure.

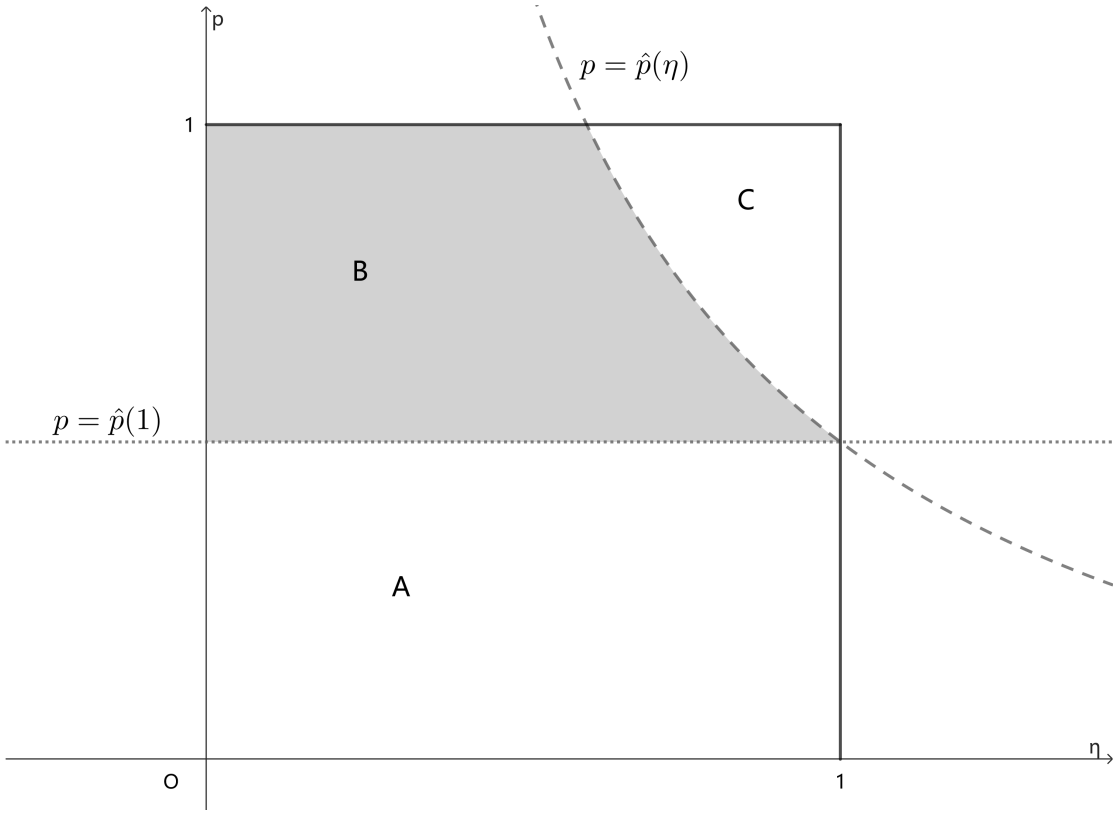


Figure 3.2: Region A: No constraint equilibrium under monopoly and competition. Region B: Equilibrium with constraint under monopoly but equilibrium without constraint under competition. Region C: Equilibrium with constraint in both monopoly and competition.

or Bayesian. The second term describes the biased consumers who always buy (i.e., biased consumers with large λ, see Table 3.1.) The third term describes Bayesian consumers who follow their signal. Since the consumers are indifferent between the two firms, we multiply the whole parenthesis by η to calculate the entrant's share in the market. The second equation is the expected profit of the entrant when it constrains information acquisition by setting $\underline{\lambda}^E = \rho_\delta$ while the incumbent doesn't constrain. The first term represents the biased consumers who always buy and aren't constrained. The second term describes the Bayesian consumers who aren't constrained and follow their signals.

It is always the case that $\Pi^E(0,0) > \Pi^E(0,\rho_\delta)$. Hence, the entrant will not constrain information when the incumbent is not constraining. In sum, not constraining information is

a dominant strategy for the entrant when neither η nor p are too high, with the thresholds as mentioned earlier.

Next, we will check the incumbent's expected profits to determine whether the incumbent will constrain information or not, anticipating the entrant's best responses above. By symmetry,

$$\Pi^I(0,0) = \frac{(1-\eta)}{\eta}\Pi^E(0,0) > \frac{(1-\eta)}{\eta}\Pi^E(0,\rho_\delta) = \Pi^I(\rho_\delta,0)$$

Since $\Pi^I(0,0) > \Pi^I(\rho_\delta,0)$, the incumbent will not constrain information acquisition in anticipation of the entrant not constraining.

If the incumbent anticipates the entrant to constrain (recall that Case 1 characterized that strategy for large η and p), then the entrant's best response to $\underline{\lambda}^I = \rho_\delta$ is also $\underline{\lambda}^E = \rho_\delta$. Hence, we have either both not constraining, or both constraining at ρ_δ. The incumbent will determine the equilibrium by moving first.

The incumbent essentially chooses which symmetric strategy to be played by checking Inequality (3.11).

$$\Pi^I(0,0) > \Pi^I(\rho_\delta,\rho_\delta) \iff \int_0^{\rho_\delta}\left[\frac{1-f-fe^{1/\lambda}}{1-e^{1/\lambda}} - (1-p)\frac{1-f-fe^{1/\rho_\delta}}{1-e^{1/\rho_\delta}} - p\right] dG(\lambda) > 0 \quad (3.11)$$

Note that this inequality is very similar to (3.10), but the left-hand side of (3.11) does not have η. If (3.11) holds, then (3.10) holds. Therefore, if $\underline{\lambda}^E = \rho_\delta$ is a best response to $\underline{\lambda}^I = \rho_\delta$ (which requires (3.10) to fail), $(\rho_\delta,\rho_\delta)$ is the equilibrium. Recall that this case occurred if there are enough exploitable biased consumers in the market. Proposition 11 summarizes this

observation.

Let $p > 0$, and $\hat{p}$ and $\hat{\eta}$ be the thresholds described to make inequality (3.10) satisfied. These give us full characterization of the equilibrium of this sequential game.

Proposition 11. *Let $p > 0$ and $\eta \in (0,1)$. There exists $\hat{p}(\eta)$, decreasing in η such that the equilibrium of the sequential game is:*

$$(\underline{\lambda}^I, \underline{\lambda}^E) = \begin{cases} (0,0) & \text{if } p < \hat{p}(1) \\ (0,0) & \text{if } \hat{p}(1) \leq p < \hat{p}(\eta) \\ (\rho_\delta, \rho_\delta) & \text{if } \hat{p}(\eta) \leq p \end{cases}$$

Moreover, when $p < \hat{p}(1)$ a monopolist would not constrain information but when $p \geq \hat{p}(1)$, a monopolist would constrain information.

Proposition 11 identifies the set of parameters for which competition would or would not help consumers collect unconstrained information. It depends on whether there are enough biased people, and whether the entrant commands a large enough market share upon entry. The former guarantees that the benefit from exploiting the biased consumers is large, and the latter ensures that the entrant has no incentive to steal consumers by providing better information. For both conditions to be satisfied, the monopolist should have the incentive to constrain in the first place, i.e. $f \geq \hat{f}$. In all other cases, neither firm will constrain information acquisition. Figure 3.2 illustrates the equilibrium characterized by Proposition 11. With the two axes representing the values of η and p, the dotted line and the dashed curve divide the graph into three regions. Region A denotes the parameter values for which there will not be information constraint either by the monopolist or under the competition. Region B denotes the parameter values where com-

petition will alleviate information exploitation. Region C denotes the region where there will be information constraints under both monopoly and competition. We can observe from the graph that competition can only be beneficial in alleviating information exploitation.

3.3 Other Non-Bayesian Updating Rules

In this section, we show that our earlier results generalize to some other non-Bayesian updating rules beyond wishful thinking described in (3.5). Denote a general updating rule:

$$g(\omega|s) = F_\omega(f, s)$$

where g is the posterior belief on state w after observing signal s given a prior distribution f. Properties 1-3 below are sufficient conditions for Proposition 10 to generalize and a monopolist to optimally constrain information acquisition for some parameters. Recall that the key idea behind Lemma 2 and Proposition 10 is that there needs to be a non-empty set of priors and λ values such that the biased posterior belief after receiving an l signal favors state H, while at the same time, the expected utility of buying after receiving an h signal is maintained positive. The properties below suffice to guarantee this idea.

Property 1. *$F_\omega(f, s)$ is continuous in both the prior probabilities $f(\omega')$ and the conditional probabilities $\pi(s|\omega')$ for all ω'.*

Property 2. *For any f and s, $F_H(f, s) \geq g^B(H|s)$. Furthermore, if the Bayesian posterior $g^B(\omega|s) \in (0, 1)$, then $F_H(f, s) > g^B(H|s)$.*

Property 3. *For any f and s, the posterior belief ratio $\frac{F_\omega(f,s)}{F_{\omega'}(f,s)}$ is non-decreasing in $\frac{\pi(s|\omega)}{\pi(s|\omega')}$.*

These are fairly loose restrictions. Property 1 is a continuity requirement for a well-behaved updating rule. The second property requires that unless the Bayesian individual is certain the product is of a certain quality, the biased individual will always be strictly more optimistic than a Bayesian person. The third property is also intuitive, as it means the biased individual understands the informational content of signals so that changing from "good news" to "better news" would not decrease the individual's optimism. For an updating function with these properties, we can show that, for $p > 0$, there exists a non-empty set of priors such that the monopolistic firm finds it profitable to constrain information acquisition. This is analogous to Proposition 10.

Proposition 12. *Let $p > 0$ and the updating rule of the biased consumers satisfy Properties 1-3, then there exists a non-empty set of model parameters and priors such that a monopolist optimally picks $\kappa > 0$, i.e. constrains consumers' information structure.*

It is straightforward to check that the wishful thinking model in Equation (3.5) satisfies these properties. Other updating rules that may satisfy the above properties under certain specifications include the following.

1. Affine Transformation: Define an updating rule that distorts the Bayesian posterior with an affine shift:

$$g^A(\omega|s) = \alpha g^B(\omega|s) + (1-\alpha)\bar{g}(\omega)$$

where $\alpha \in (0,1)$ is the weight and $\bar{g}$ is an exogenous probability distribution. If, for instance, $\bar{g}(H) = 1$, the updating rule would satisfy all the properties.[18]

2. Distorted Prior: In this updating rule, the individual employs the Bayes rule on a distorted

[18] Any anchor would work as long as the anchoring belief $\bar{g}(H)$ is greater than the maximal $g^B(H|s)$. When we limit our discussion to a certain range of $f(H)$, it may be impossible to obtain a Bayesian posterior $g^B(H|s)$ close to 1 given any signal. In such cases, there would be a wide range of eligible values for $\bar{g}(H)$, generalizing this example to more applications.

prior (as in 47 and 10).

$$g^P(\omega|s) = g^B(\omega|s, \chi(f)) = \frac{\chi(f(\omega))\pi(s|\omega)}{\sum \chi(f(\omega'))\pi(s|\omega')}$$

where χ is a distortion function. If the distortion function effectively shifts the prior towards state H, that is if $\frac{\chi(f(H))}{\chi(f(\omega))} > \frac{f(H)}{f(\omega)}$ for all $\omega \neq H$, the resulting Bayesian updating upon the distorted prior would satisfy the above properties.

3. Psychological Utility Maximization: If the individual directly gains utility from certain beliefs (as in 39) and chooses beliefs that maximize such utility, then this can distort beliefs in directions dependent on which beliefs are preferred. As discussed in [72], this can accommodate a motivated belief updater who has a target belief $\bar{g}$ and suffers a cost from deviating away from both $\bar{g}$ and the Bayesian posterior g^B.

$$g^U = argmax_g U(g, \bar{g}, g^B)$$

If $\bar{g}(H) = 1$, then similar to the first example, the updating rule would satisfy the above properties. In general, having $\bar{g}(H) > g^B(H)$ would suffice.

3.4 Discussions and Conclusion

In this section, we will revisit some of our modeling assumptions and discuss alternative ways of modeling. In our analysis, we implicitly assumed that firms do not have perfect knowledge of the products' quality. This makes sense for the main scenario we are considering, that is, where the firm is a digital platform that hosts many products potentially from many different

upstream sellers. Platforms like Amazon typically do not, and cannot, monitor the quality of every product sold on their site. However, it is still worthwhile to mention the possibility that the firm knows the product quality, as is usually the case when the seller is also the producer. This is a classic setting for signalling games and has been discussed extensively in the literature. The equilibrium typically would involve the firm's actions inevitably revealing information regarding their product quality. A similar signaling game is, for example, studied by [59]. It is an open question to study rationally inattentive biased consumers in a signaling game where the firm can directly constrain the informativeness of the signals through κ.

Another variation that is worth mentioning is the possibility that each consumer may have a different state realization. Some may be unlucky and end up with a faulty product, while others are lucky enough to obtain a perfectly manufactured one. Assuming that the firms do not control the quality of each product the consumers buy and have to offer the same set of information structures to everyone, they can then only maximize expected profits over the unconditional probabilities. Then this exercise becomes the same as our model. Alternatively, if the firms could control, or have knowledge of, the quality of each product that each consumer gets and hence, personalize the information constraint, the problem coincides with the signalling game mentioned above.

One may also consider heterogeneity in other aspects of the model. In our setting, the only sources of heterogeneity are λ (the cost parameter) and being biased or not. When individuals all have different priors, we again have two possible scenarios. If the firm knows each consumer's prior and is capable of customizing its constraint for each individual, the firm's problem is reduced to the base problem we have discussed, but with one person. Of course, the firm has a much higher profit in this case, because it can choose to only apply the information constraints

to biased consumers. However, if the firm is not capable of setting individual constraints, it will simply maximize expected profit over the distribution of priors, regardless of whether they know the consumers' idiosyncratic priors or not. Given this, the threshold for which the firm finds it profitable to constrain information acquisition will become a condition regarding the distribution of priors. It is easy to conceive that the more individuals holding higher priors, the more profitable such information exploitation is. Assuming that biased individuals have different magnitudes of bias (different δ_H and δ_L) will produce an analogous result, where the firm either faces many instances of the single consumer problem if they can apply individual constraints and knows each individual's bias, or it maximizes expected profit over the distribution of δ otherwise.

Alternatively, one may extend the discussion to continuous states (quality of the product). The optimal information structure (Lemma 2), given that the action set does not change, would retain the same form (the solution in the appendix also applies to continuous states, except with integrals instead of summations over the states). For continuous states, it may be difficult to obtain a closed form solution to both the marginal probabilities of signals q and the conditional probabilities π, unless one is willing to assume certain distributions for the prior belief. This is a standard complexity observed previously by the rational inattention literature. However, we expect that the intuition behind the results would not change, and there would still be a condition on the distribution of prior belief such that, for a prior distribution skewed enough towards higher product qualities, the firm would find it profitable to exploit information acquisition.

Finally, while we assumed $f(H) \leq 0.5$ in our discussion, it is certainly feasible to assume otherwise. However, having $f(H) > 0.5$ would ultimately lead to a trivial discussion, because the consumers lean towards buying the product by default. In this case, the firm can simply set $\kappa = 0$ (or equivalently $\underline{\lambda} \to \infty$), that is completely shutting down the customer review section

and not allowing anyone to acquire information before purchasing. Assuming this in itself does not convey malicious intentions, all consumers would buy the product. In fact, it is beneficial for the firm to do this even when all consumers are Bayesian, since information serves to warn the consumers of potential low quality products now (as opposed to signal high quality products), and stripping away the consumers' ability to acquire meaningful information is always desirable for the firm.

In sum, we provide a model where the firms may find it profitable to constrain the quality of information they provide about the product they sell. We characterize when competition among firms cannot mitigate the vagueness of information. Our results rely on the existence of biased consumers in the market and they generalize to some other non-Bayesian updating rules beyond wishful thinking. We also list applicable open questions for more complex or alternative modeling assumptions. To the best of our knowledge, this is the first model introducing competition through quality of information which has fruitful applications for modern business models.

Appendix A: Chapter 1 Appendices

Appendix I - Proofs

Proposition 1 - Proof

First, I solve for (1.3).

Note that there cannot be a corner solution, that is $\mu^*(\omega) \in (0,1)$ for all states. This is because as $\mu^*(\omega) \to 0$ for some state, or $\mu^*(\omega) \to 1$ for a state and in which case $\mu^*(\omega') \to 0$ for all other states, the KL divergence goes to $-\infty$. So we can focus on looking for an interior solution.

FOC's for each ω, with κ as the Lagrangian multiplier:

$$g(\mu_0(\omega), v(\omega)) - \lambda(ln(\frac{\mu(\omega)}{\mu^B(\omega)}) + 1) - \kappa = 0$$

Solving this gives:

$$\frac{\mu^*(\omega)}{\mu^*(\omega')} = \frac{\mu^B(\omega)}{\mu^B(\omega')} e^{g(\mu_0(\omega), v(\omega)) - g(\mu_0(\omega'), v(\omega'))}$$

for any $\omega, \omega' \in \Omega$.

Equivalence can be shown in the same way.

For equivalence 1: solving the maximization problem gives

$$\frac{\mu_0^*(\omega)}{\mu_0^*(\omega')} = \frac{\mu_0(\omega)}{\mu_0(\omega')} e^{g(\mu_0(\omega),v(\omega))-g(\mu_0(\omega'),v(\omega'))}$$

Normal Bayesian updating upon μ_0^* gives the same posterior as in Prop. 1:

$$\frac{\mu_1(\omega)}{\mu_1(\omega')} = \frac{\mu_0^*(\omega)\pi(s|\omega)}{\mu_0^*(\omega')\pi(s|\omega')}$$

$$= \frac{\mu_0(\omega)\pi(s|\omega)}{\mu_0(\omega')\pi(s|\omega')} e^{g(\mu_0(\omega),v(\omega))-g(\mu_0(\omega'),v(\omega'))}$$

$$= \frac{\mu^B(\omega|s)}{\mu^B(\omega'|s)} e^{g(\mu_0(\omega),v(\omega))-g(\mu_0(\omega'),v(\omega'))}$$

For equivalence 2: solving the maximization problem gives

$$\frac{\pi^*(s^*|\omega)}{\pi^*(s^*|\omega')} = \frac{\pi(s|\omega)}{\pi(s|\omega')} e^{g(\mu_0(\omega),v(\omega))-g(\mu_0(\omega'),v(\omega'))}$$

Bayesian updating with this distorted signal structure gives:

$$\frac{\mu_1(\omega)}{\mu_1(\omega')} = \frac{\mu_0(\omega)\pi^*(s^*|\omega)}{\mu_0(\omega')\pi^*(s^*|\omega')}$$

$$= \frac{\mu_0(\omega)\pi(s|\omega)}{\mu_0(\omega')\pi(s|\omega')} e^{g(\mu_0(\omega),v(\omega))-g(\mu_0(\omega'),v(\omega'))}$$

$$= \frac{\mu^B(\omega|s)}{\mu^B(\omega'|s)} e^{g(\mu_0(\omega),v(\omega))-g(\mu_0(\omega'),v(\omega'))}$$

□

Proposition 2 - Proof

"If":

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)}\frac{\mu(\omega_j|s')}{\mu(\omega_i|s')} = \frac{\epsilon(\omega_i,\mu_0)\mu_0(\omega_i)\pi(s|\omega_i)}{\epsilon(\omega_j,\mu_0)\mu_0(\omega_j)\pi(s|\omega_j)}\frac{\epsilon(\omega_j,\mu_0)\mu_0(\omega_j)\pi(s'|\omega_j)}{\epsilon(\omega_i,\mu_0)\mu_0(\omega_i)\pi(s'|\omega_i)}$$

$$= \frac{\pi(s|\omega_i)}{\pi(s|\omega_j)}\frac{\pi(s'|\omega_j)}{\pi(s'|\omega_i)}$$

"Only if":

By Axiom 1, for any given μ_0, $\omega_i, \omega_j \in \Omega$, and two signals s, s',

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)}\frac{\pi(s|\omega_j)}{\pi(s|\omega_i)} = \frac{\mu(\omega_i|s')}{\mu(\omega_j|s')}\frac{\pi(s'|\omega_j)}{\pi(s'|\omega_i)}$$

Define the above ratio to be

$$\rho(\omega_i,\omega_j,\mu_0) = \frac{\mu(\omega_i|s)}{\mu(\omega_j|s)}\frac{\pi(s|\omega_j)}{\pi(s|\omega_i)} = \frac{\mu(\omega_i|s')}{\mu(\omega_j|s')}\frac{\pi(s'|\omega_j)}{\pi(s'|\omega_i)}$$

Notice that the value of $\rho(\omega_i,\omega_j,\mu_0)$ is independent of the signal under Axiom 1.

Then, for any fixed state ω_1, define

$$\epsilon(\omega_i,\mu_0) = \frac{\rho(\omega_i,\omega_1,\mu_0)}{\mu_0(\omega_i)}$$

Consequently,

$$\frac{\epsilon(\omega_i,\mu_0)\mu_0(\omega_i)}{\epsilon(\omega_j,\mu_0)\mu_0(\omega_j)} = \frac{\rho(\omega_i,\omega_1,\mu_0)}{\rho(\omega_j,\omega_1,\mu_0)} = \frac{\frac{\mu(\omega_i|s)}{\mu(\omega_1|s)}\frac{\pi(s|\omega_1)}{\pi(s|\omega_i)}}{\frac{\mu(\omega_j|s)}{\mu(\omega_1|s)}\frac{\pi(s|\omega_1)}{\pi(s|\omega_j)}} = \frac{\mu(\omega_i|s)\pi(s|\omega_j)}{\mu(\omega_j|s)\pi(s|\omega_i)}$$

And rearranging,

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} = \frac{\epsilon(\omega_i,\mu_0)\mu_0(\omega_i)\pi(s|\omega_i)}{\epsilon(\omega_j,\mu_0)\mu_0(\omega_j)\pi(s|\omega_j)}$$

Furthermore, $\epsilon(\omega_i, \mu_0)$ is unique up to a linear transformation.

Assume that another set of errors ζ_i also equates

$$\mu(\omega_i|s) = \frac{\zeta_i\mu_0(\omega_i)\pi(s|\omega_i)}{\sum_j \zeta_j\mu_0(\omega_j)\pi(s|\omega_j)}$$

for all ω_i.

Then we have, for any i, j,

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} = \frac{\epsilon(\omega_i,\mu_0)\mu_0(\omega_i)\pi(s|\omega_i)}{\epsilon(\omega_j,\mu_0)\mu_0(\omega_j)\pi(s|\omega_j)} = \frac{\zeta_i\mu_0(\omega_i)\pi(s|\omega_i)}{\zeta_j\mu_0(\omega_j)\pi(s|\omega_j)}$$

Let $\zeta_i = \alpha\epsilon(\omega_i, \mu_0)$. Then we must have $\zeta_j = \alpha\epsilon(\omega_j, \mu_0)$, and this holds for all $j \neq i$. Thus, the induced errors are unique up to a linear transformation.

□

Proposition 3 - Proof

Proof:

Let $\mu_0(\omega_i) \geq \mu_0'(\omega_i)$ and $\mu_0(\omega_j) \leq \mu_0'(\omega_j)$. Let $\mu(\cdot|\cdot)$ and $\mu'(\cdot|\cdot)$ be the posteriors given μ_0 and μ_0' respectively.

"If":

Axiom 1 is trivial (same as previous proof)

Axiom 2:

By the definition of BUME-UCB,

$$\epsilon(\omega_i, \mu_0(\omega_i)) \geq \epsilon(\omega_i, \mu_0'(\omega_i))$$

and

$$\epsilon(\omega_j, \mu_0(\omega_j)) \leq \epsilon(\omega_j, \mu_0'(\omega_j))$$

Let s, s' be two signals that satisfy the condition in A2. Thus

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} = \frac{\epsilon(\omega_i, \mu_0(\omega_i))\mu_0(\omega_i)\pi(s|\omega_i)}{\epsilon(\omega_j, \mu_0(\omega_j))\mu_0(\omega_j)\pi(s|\omega_j)} \geq \frac{\epsilon(\omega_i, \mu_0'(\omega_i))\mu_0'(\omega_i)\pi(s'|\omega_i)}{\epsilon(\omega_j, \mu_0'(\omega_j))\mu_0'(\omega_j)\pi(s'|\omega_j)} = \frac{\mu'(\omega_i|s')}{\mu'(\omega_j|s')}$$

"Only if":

Without loss of generality, fix state ω_1 and assume $\mu_0(\omega_1) \leq \mu_0'(\omega_1)$. Same as Proposition 2, I define

$$\rho(\omega_i, \omega_j, \mu_0) = \frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} \frac{\pi(s|\omega_j)}{\pi(s|\omega_i)}$$

and define

$$\epsilon(\omega_i, \mu_0) = \frac{\rho(\omega_i, \omega_1, \mu_0)}{\mu_0(\omega_i)}$$

BUME can be obtained by A1, as shown earlier. The proof below focuses on the UCB part.

An implication of Axiom 2 is that if μ_0 and μ_0' are such that $\mu_0(\omega_i) = \mu_0'(\omega_i)$ and $\mu_0(\omega_j) = \mu_0'(\omega_j)$, then for s and s' satisfying the condition in Axiom 2,

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} = \frac{\mu'(\omega_i|s')}{\mu'(\omega_j|s')}$$

By Axiom 1, $\rho(\omega_i, \omega_j, \mu_0)$ is independent of signal, and so

$$\rho(\omega_i, \omega_j, \mu_0) = \rho(\omega_i, \omega_j, \mu_0')$$

Hence, for any prior, it is possible to define

$$\rho(\omega_i, \omega_j, \mu_0) = \rho(\omega_i, \omega_j, \mu_0(\omega_i), \mu_0(\omega_j))$$

and define

$$\epsilon(\omega_i, \mu_0) = \epsilon(\omega_i, \mu_0(\omega_i))$$

Since ϵ is invariant in signals, s, s' be signals that satisfy the condition in Axiom 2. For ω_i such that $\mu_0(\omega_i) \geq \mu_0'(\omega_i)$,

$$\frac{\epsilon(\omega_i, \mu_0(\omega_i))\mu_0(\omega_i)}{\epsilon(\omega_i, \mu_0'(\omega_i))\mu_0'(\omega_i)} = \frac{\rho(\omega_i, \omega_1, \mu_0(\omega_i), \mu_0(\omega_1))}{\rho(\omega_i, \omega_1, \mu_0'(\omega_i), \mu_0'(\omega_1))}$$

$$= \frac{\frac{\mu(\omega_i|s)}{\mu(\omega_1|s)}\frac{\pi(s|\omega_1)}{\pi(s|\omega_i)}}{\frac{\mu'(\omega_i|s')}{\mu'(\omega_1|s')}\frac{\pi(s'|\omega_1)}{\pi(s'|\omega_i)}} = \frac{\mu(\omega_i|s)}{\mu(\omega_1|s)}\frac{\mu'(\omega_1|s')}{\mu'(\omega_i|s')}\frac{\mu_0(\omega_i)}{\mu_0(\omega_1)}\frac{\mu_0'(\omega_1)}{\mu_0'(\omega_i)}$$

The last equality is due to s, s' satisfying the condition in Axiom 2. This gives

$$\frac{\epsilon(\omega_i, \mu_0(\omega_i))}{\epsilon(\omega_i, \mu_0'(\omega_i))} = \frac{\mu(\omega_i|s)}{\mu(\omega_1|s)}\frac{\mu'(\omega_1|s')}{\mu'(\omega_i|s')}\frac{\mu_0'(\omega_1)}{\mu_0'(\omega_i)} \geq 1$$

This is because

$$\frac{\mu(\omega_i|s)}{\mu(\omega_1|s)}\frac{\mu'(\omega_1|s')}{\mu'(\omega_i|s')} \geq 1$$

by Axiom 2, and $\mu_0(\omega_1) \leq \mu_0'(\omega_1)$.

□

Proposition 4 - Proof

I will show the "only if" part. BUME again is obtained through the same method as in Proposition 2.

Define $\succeq_P$ as: $\omega_i \succeq_P \omega_j$ if and only if ω_i is "revealed preferred" to ω_j, that is $\epsilon(\omega_i, \mu_0) \geq \epsilon(\omega_j, \mu_0)$ for some μ_0 such that $\mu_0(\omega_i) = \mu_0(\omega_j)$. Now I show that $\succeq_P$ is rational.

Completeness: $\forall i, j, \exists \mu_0 \in M = (0,1)^{|\Omega|}$ such that $\mu_0(\omega_i) = \mu_0(\omega_j)$. Then $\forall i, j$, there are 2 cases:

1) for some μ_0 such that $\mu_0(\omega_i) = \mu_0(\omega_j)$,

$$\frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} \geq \frac{\pi(s|\omega_i)}{\pi(s|\omega_j)} \iff \epsilon(\omega_i, \mu_0) \geq \epsilon(\omega_j, \mu_0)$$

And thus $\omega_i \succeq_P \omega_j$.

By A3, we have $\forall \mu_0'$ such that $\mu_0'(\omega_i) = \mu_0'(\omega_j)$,

$$\frac{\mu'(\omega_i|s)}{\mu'(\omega_j|s)} \geq \frac{\pi(s|\omega_i)}{\pi(s|\omega_j)} \iff \epsilon(\omega_i, \mu_0') \geq \epsilon(\omega_j, \mu_0')$$

So it cannot be that $\omega_i \prec_P \omega_j$

2) for some μ_0 such that $\mu_0(\omega_i) = \mu_0(\omega_j)$, $\epsilon_i \leq \epsilon_j$, which eventually yields $\omega_i \preceq_P \omega_j$ just like above.

Transitivity: Suppose $\omega_i \succeq_P \omega_j$ and $\omega_j \succeq_P \omega_k$. Then, for some μ_0 and μ_0' such that

$\mu_0(\omega_i) = \mu_0(\omega_j)$ and $\mu_0'(\omega_j) = \mu_0'(\omega_k)$,

$$\epsilon(\omega_i, \mu_0) \geq \epsilon(\omega_j, \mu_0) \wedge \epsilon(\omega_j, \mu_0') \geq \epsilon(\omega_k, \mu_0')$$

By A3, or by completeness as just shown, the above holds for any priors μ_0'' and μ_0''' such that $\mu_0''(\omega_i) = \mu_0''(\omega_j)$ and $\mu_0'''(\omega_j) = \mu_0'''(\omega_k)$.

Let μ_0 be a prior such that $\mu_0(\omega_i) = \mu_0(\omega_j) = \mu_0(\omega_k)$,

$$\epsilon(\omega_i, \mu_0) \geq \epsilon(\omega_j, \mu_0) \geq \epsilon(\omega_k, \mu_0) \iff \frac{\mu(\omega_i|s)}{\mu(\omega_j|s)} \geq \frac{\pi(s|\omega_i)}{\pi(s|\omega_j)} \wedge \frac{\mu(\omega_j|s)}{\mu(\omega_k|s)} \geq \frac{\pi(s|\omega_j)}{\pi(s|\omega_k)}$$

$$\iff \mu(\omega_i|s)\pi(s|\omega_i) \geq \mu(\omega_j|s)\pi(s|\omega_j) \geq \mu(\omega_k|s)\pi(s|\omega_k)$$

By A3, we have for all μ_0' where $\mu_0'(\omega_i) = \mu_0'(\omega_k)$,

$$\frac{\mu'(\omega_i|s)}{\mu'(\omega_j|s)} \geq \frac{\pi(s|\omega_j)}{\pi(s|\omega_k)} \iff \epsilon(\omega_i, \mu_0) \geq \epsilon(\omega_k, \mu_0)$$

And thus $\omega_i \succeq_P \omega_k$.

□

Proposition 5 - Proof

Since $\pi(s_0|\omega)$ is constant across states,

$$\mu^{CB}(\omega|s_0) = \frac{\mu^B(\omega|s_0)F_0(\omega)}{\sum_{\omega'} \mu^B(\omega'|s_0)F_0(\omega')} = \frac{\mu_0(\omega)F_0(\omega)}{\sum_{\omega'} \mu_0(\omega')F_0(\omega')/} > \mu^B(\omega|s_0) = \mu_0(\omega)$$

$$\iff F_0(\omega) > \sum_{\omega'} \mu_0(\omega')F_0(\omega')$$

□

Remark 1 & 2 - Proof

Remark 1 follows from the Proposition 5 that

$$F_0(\omega_j) > E_{\mu_0}[F_0] > F_0(\omega_i)$$

and that $\sum_{\omega'} \mu^B(\omega'|s)F_0(\omega')$ is a fixed threshold given s.

Remark 2 follows from the fact that the threshold $\sum_{\omega'} \mu^B(\omega'|s)F_0(\omega')$ is a convex combination of F_0's.

□

Proposition 6 - Proof

When k is concave, by the definition of concavity and Proposition 5,

$$F_0^i(\omega) = k(F_0^j(\omega)) > k(E_{\mu_0}[F_0^j]) \geq E_{\mu_0}[k(F_0^j)] = E_{\mu_0}[F_0^i]$$

When k is convex, by the definition of convexity and Proposition 5,

$$F_0^i(\omega) = k(F_0^j(\omega)) < k(E_{\mu_0}[F_0^j]) \leq E_{\mu_0}[k(F_0^j)] = E_{\mu_0}[F_0^i]$$

□

Lemma 1 and Proposition 7 - Proof

For simplicity, we can express: $\mu^{CB}(A) = \mu_1$ and $\mu^{CB}(B) = 1 - \mu_1$. Similarly, express $\mu_0(A) = \mu_0$ and $\mu^B(A) = \mu^B$.

First observe that the posterior odds after any single signal is monotonically increasing in

μ_0:

$$\frac{\mu^{CB}}{1-\mu^{CB}} = \frac{\pi(s|A)\mu_0}{\pi(s|B)(1-\mu_0)} \frac{f(\mu_0, v(A))}{f(1-\mu_0, v(B))}$$

Therefore, for any two consecutive signals s_k and s_{k+1} in the the sequence, if swapping their order increases the odds ratio of the posterior belief for any prior:

$$\frac{\mu^{CB}(A|s_{k+1}, s_k)}{\mu^{CB}(B|s_{k+1}, s_k)} \geq \frac{\mu^{CB}(A|s_k, s_{k+1})}{\mu^{CB}(B|s_k, s_{k+1})}$$

Then we swapping their order will also increase the odds ratio of the final posterior.

Hence we only need to find the optimal ordering of two consecutive signals. Suppose the first signal is s and the second is s', and the interim posterior after just signal s is μ_1^{CB}.

$$\frac{\mu^{CB}(A|s, s')}{\mu^{CB}(B|s, s')} = \frac{\pi(s|A)\pi(s'|A)\mu_0}{\pi(s|B)\pi(s'|B)(1-\mu_0)} \frac{f(\mu_0, v(A))}{f(1-\mu_0, v(B))} \frac{f(\mu_1^{CB}, v(A))}{f(1-\mu_1^{CB}, v(B))}$$

Notice that, since everything else does not vary along the process of belief updating, the posterior odds only varies with μ_1^{CB}, and is increasing in this interim posterior. This immediately implies that

$$\frac{\mu^{CB}(A|s, s')}{\mu^{CB}(B|s, s')} \geq \frac{\mu^{CB}(A|s', s)}{\mu^{CB}(B|s', s)} \iff \frac{\pi(s|A)}{\pi(s|B)} \geq \frac{\pi(s'|A)}{\pi(s'|B)}$$

Since the final posterior is monotone in prior, forward swaps on any position will increase the final posterior. Hence putting signals with higher odds ratios first will maximize the shift towards state A. Repeated application of the swap gives the optimal ordering.

□

Proposition 8 - Proof

Posterior likelihood ratio between states ω and ω' after N signals:

$$\frac{\mu_0(\omega)}{\mu_0(\omega')}\Pi_i\{\frac{\pi(s_i|\omega)}{\pi(s_i|\omega')}\frac{F_{i-1}(\omega)}{F_{i-1}(\omega')}\}$$

Take the logarithm:

$$ln(\frac{\mu_0(\omega)}{\mu_0(\omega')})+\sum_i\{ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega')})+ln(F_{i-1}(\omega))-ln(F_{i-1}(\omega'))\}$$

$$=ln(\frac{\mu_0(\omega)}{\mu_0(\omega')})+N[\frac{1}{N}\sum_i\{ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega')})+ln(F_{i-1}(\omega))-ln(F_{i-1}(\omega'))\}]$$

$$=ln(\frac{\mu_0(\omega)}{\mu_0(\omega')})+N[\frac{1}{N}\sum_i\{ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega^*)})-ln(\frac{\pi(s_i|\omega')}{\pi(s_i|\omega^*)})+ln(F_{i-1}(\omega))-ln(F_{i-1}(\omega'))]$$

$$\geq ln(\frac{\mu_0(\omega)}{\mu_0(\omega')})+N[\frac{1}{N}\sum_i\{ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega^*)})-ln(\frac{\pi(s_i|\omega')}{\pi(s_i|\omega^*)})+ln(f_{i-1}(0,v(\omega)))-ln(f_{i-1}(1,v(\omega')))]$$

$$\overset{a.s.}{\to} ln(\frac{\mu_0(\omega)}{\mu_0(\omega')})+N[E_{\pi(\cdot|\omega^*)}[ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega^*)})]-E_{\pi(\cdot|\omega^*)}[ln(\frac{\pi(s_i|\omega')}{\pi(s_i|\omega^*)})]+$$

$$ln(f_{i-1}(0,v(\omega)))-ln(f_{i-1}(1,v(\omega')))]$$

$$=ln(\frac{\mu_0(\omega)}{\mu_0(\omega')})+N[ln(f_{i-1}(0,v(\omega)))-D_{KL}(\pi(\cdot|\omega^*)||\pi(\cdot|\omega))-$$

$$ln(f_{i-1}(1,v(\omega')))+D_{KL}(\pi(\cdot|\omega')||\pi(\cdot|\omega'))]$$

$$=ln(\frac{\mu_0(\omega)}{\mu_0(\omega^*)})+N[\underline{C}(\omega)-\overline{C}(\omega')]$$

as $N\to\infty$, by the strong LLN.

Since

$$\underline{C}(\omega)-\overline{C}(\omega')>0$$

The log posterior ratio $ln(\frac{\mu_N^{CB}(\omega)}{\mu_N^{CB}(\omega')}) \overset{a.s.}{\to} \infty$. This is true for all ω'. Hence $\mu_N^{CB}(\omega) \overset{a.s.}{\to} 1$.

□

Corollary 2 - Proof:

"if":

If there is no UCB, $F_{i-1}(\omega)$ is invariant with i, and so log posterior ratio would becomes

$$ln(\frac{\mu_0(\omega)}{\mu_0(\omega')}) + \sum_i \{ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega')}) + ln(F_{i-1}(\omega)) - ln(F_{i-1}(\omega'))\}$$

$$= ln(\frac{\mu_0(\omega)}{\mu_0(\omega')}) + N[\frac{1}{N}\sum_i \{ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega')}) + ln(F_{i-1}(\omega)) - ln(F_{i-1}(\omega'))\}]$$

$$= ln(\frac{\mu_0(\omega)}{\mu_0(\omega')}) + N[\frac{1}{N}\sum_i \{ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega^*)}) - ln(\frac{\pi(s_i|\omega')}{\pi(s_i|\omega^*)}) + ln(F_{i-1}(\omega)) - ln(F_{i-1}(\omega'))]$$

$$\overset{a.s.}{\to} ln(\frac{\mu_0(\omega)}{\mu_0(\omega')}) + N[E_{\pi(\cdot|\omega^*)}[ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega^*)})] - E_{\pi(\cdot|\omega^*)}[ln(\frac{\pi(s_i|\omega')}{\pi(s_i|\omega^*)})] + ln(F_0(\omega)) - ln(F_0(\omega'))]$$

$$= ln(\frac{\mu_0(\omega)}{\mu_0(\omega')}) + N[C(\mu_0(\omega), v(\omega)) - C(\mu_0(\omega'), v(\omega'))]$$

Since ω has maximal credibility, we have

$$C(\mu_0(\omega), v(\omega)) - C(\mu_0(\omega'), v(\omega')) > 0 \; \forall \omega'$$

The log posterior ratio $ln(\frac{\mu_N^{CB}(\omega)}{\mu_N^{CB}(\omega')}) \overset{a.s.}{\to} \infty$, and $\mu_N^{CB}(\omega) \overset{a.s.}{\to} 1$.

"only if":

If $C(\mu_0(\omega), v(\omega)) \neq max\{C(\mu_0(\omega'), v(\omega'))\}$, there are two scenarios:

1) There exists ω' such that $C(\mu_0(\omega'), v(\omega')) > C(\mu_0(\omega), v(\omega))$, in which case

$$lim_{N\to\infty} ln(\frac{\mu_N^{CB}(\omega)}{\mu_N^{CB}(\omega')}) \to -\infty$$

And $\mu_N^{CB}(\omega) \to 0$

2) $\omega \in \Omega^* = \{\omega' \in \Omega | \omega' = argmax C(\mu_0(\omega''), v(\omega''))\}$ and there is $\omega' \neq \omega$ that also satisfies $\omega' \in \Omega^*$.

In this case, beliefs generally do not converge point-wise, but they still do in distribution. The term

$$C(\mu_0(\omega), v(\omega)) - C(\mu_0(\omega'), v(\omega'))$$

in the "if" part is 0.

$ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega^*)}) - ln(\frac{\pi(s_i|\omega')}{\pi(s_i|\omega^*)})$ are i.i.d. random variables, and

$$E_{\pi(\cdot|\omega^*)}[ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega^*)}) - ln(\frac{\pi(s_i|\omega')}{\pi(s_i|\omega^*)}) + ln(F_0(\omega)) - ln(F_0(\omega'))] = 0$$

Let their variance be σ^2. By the Central Limit Theorem, we have

$$\lim_{N\to\infty} \sqrt{N}[\frac{1}{N}\sum_i \{ln(\frac{\pi(s_i|\omega)}{\pi(s_i|\omega^*)}) - ln(\frac{\pi(s_i|\omega')}{\pi(s_i|\omega^*)}) + ln(F_{i-1}(\omega)) - ln(F_{i-1}(\omega'))] \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

And so

$$\lim_{N\to\infty} ln(\frac{\mu_N^{CB}(\omega)}{\mu_N^{CB}(\omega')}) \xrightarrow{d} \mathcal{N}(ln(\frac{\mu_0(\omega)}{\mu_0(\omega')}, \sqrt{N}\sigma^2)$$

We see that the variance is ever growing, while the expectation remains to be the prior log-likelihood ratio, which is finite.

□

Proposition 9 - Proof

We prove this with the help of Theorem 1 from Frick et al. (2022). The theorem states that: a point mass belief on state ω, δ_ω, is locally stable, if there exists a neighborhood $B \ni \delta_\omega$ and $q > 0$ such that for all $\omega' \neq \omega$,

$$\omega \succ^q \omega' \iff \int (\frac{\hat{\pi}(s|\omega')}{\hat{\pi}(s|\omega)})^q d\Pi(s|\omega^*) \leq 1$$

where Π denotes the distribution function of signal s conditional on true state ω^*, and $\hat{\pi}$ denotes the subjective conditional probabilities of signal s. Locally stable is a term in Frick et al. that refers to: for any $c < 1$, there exists neighborhood $B \ni \delta_\omega$ such that $P(\mu \to \delta_\omega) \geq c$.

Now, we show that δ_ω (belief with $\mu^{CB}(\omega) = 1$ and $\mu^{CB}(\omega') = 0$ for $\omega' \neq \omega$) where ω is such that $\overline{C}(\omega) > \underline{C}(\omega')\ \forall \omega' \neq \omega$ is locally stable under my model.

Since credibility is continuous in prior belief and

$$\overline{C}(\omega) > \underline{C}(\omega')$$

there is neighborhood $B \ni \delta_\omega$ where $\forall \mu_0 \in B$,

$$C(\mu_0(\omega), v(\omega)) > C(\mu_0(\omega'), v(\omega'))$$

As mentioned in Proposition 1, my model can alternatively be interpreted as the individual

misinterpreting the signal. This also translates to an alternative formation of credibility.

$$C(\mu_0(\omega), v(\omega)) = ln(f(\mu_0(\omega), v(\omega))) - D_{KL}(\pi(\cdot|\omega^*)||\pi(\cdot|\omega))$$

$$= ln(f(\mu_0(\omega), v(\omega))) - \sum_s \pi(s|\omega^*) ln(\frac{\pi(s|\omega^*)}{\pi(s|\omega)})$$

$$= -\sum_s \pi(s|\omega^*) ln(\frac{\pi(s|\omega^*)}{\pi(s|\omega) f(\mu_0(\omega), v(\omega))})$$

Then we have that

$$C(\mu_0(\omega), v(\omega)) > C(\mu_0(\omega'), v(\omega')) \iff$$

$$-\sum_s \pi(s|\omega^*) ln(\frac{\pi(s|\omega^*)}{\pi(s|\omega) f(\mu_0(\omega), v(\omega))}) + \sum_s \pi(s|\omega^*) ln(\frac{\pi(s|\omega^*)}{\pi(s|\omega') f(\mu_0(\omega'), v(\omega'))}) > 0$$

$$\iff \sum_s \pi(s|\omega^*) ln(\frac{\pi(s|\omega') f(\mu_0(\omega'), v(\omega'))}{\pi(s|\omega) f(\mu_0(\omega), v(\omega))}) < 0$$

$$\iff \sum_s \pi(s|\omega^*) ln(\frac{\hat{\pi}(s|\omega')}{\hat{\pi}(s|\omega)}) < 0$$

where $\hat{\pi}(\cdot|\omega)$ denotes the distorted signal probabilities as described in Proposition 1, defined by

$$\frac{\hat{\pi}(s|\omega)}{\hat{\pi}(s|\omega')} = \frac{\pi(s|\omega') f(\mu_0(\omega'), v(\omega'))}{\pi(s|\omega) f(\mu_0(\omega), v(\omega))} \; \forall s, \omega, \omega'$$

Notice that

$$\frac{d}{dq}\Big|_{q=0} \sum_s \pi(s|\omega^*) (\frac{\hat{\pi}(s|\omega')}{\hat{\pi}(s|\omega)})^q = \sum_s \pi(s|\omega^*) ln(\frac{\hat{\pi}(s|\omega')}{\hat{\pi}(s|\omega)}) < 0$$

Then for small positive q,

$$\sum_s \pi(s|\omega^*)(\frac{\hat{\pi}(s|\omega')}{\hat{\pi}(s|\omega)})^q < \sum_s \pi(s|\omega^*)(\frac{\hat{\pi}(s|\omega')}{\hat{\pi}(s|\omega)})^0 = 1$$

(This result is also stated in Lemma 2 of Frick et al. (2022))

Hence, we have both a B and a q such that $\omega \succ^q \omega'$ for all $\omega' \neq \omega$. δ_ω is locally stable.

Take some $c^* \in (0, 1)$, then there is neighborhood $B^* \ni \delta_\omega$ such that $P(\mu \to \delta_\omega) \geq c^*$ for any prior starting within B^*. Given any prior in M, there is a strong enough signal, which appears with positive probability, that increases the posterior probability over state ω. Thus, we have a signal sequence that occurs with positive probability $\tilde{p}$ that gives us a posterior belief $\mu \in B^*$, and

$$P(\mu \to \delta_\omega) \geq \tilde{p}c^* > 0$$

(Important note: The setting in Frick et al. (2022) assumes that the subjective signal probabilities are proper probability measures that add to 1. The distorted signal probabilities in my model are, strictly speaking, NOT probability measures, as they do not add to 1. However, this does not affect the applicability of their Theorem 1 in my setting, as their results do not rely on the subjective probabilities adding to 1. What is important is the property of q-dominance, i.e. $\omega \succ^q \omega'$, and how it is used to construct a supermartingale out of log-likelihood beliefs within the neighborhood B.)

□

Appendix II - Brief Discussion of Chauvin (2020) Axioms

Chauvin (2020) characterized a whole set of updating rules that are (algebraically) isomorphic to Bayesian updating, called Virtual Bayesians. First, define a learning rule $a \in \mathcal{A}$ as a mapping from the set of beliefs $\mathcal{B}$ onto itself $a : b \to b$. In essence, it is an updating rule from the prior to the posterior given the signal realization(s). The set of axioms that characterize Virtual Bayesians are:

A1. $\mathcal{A}$ and $\mathcal{B}$ are at most countably infinite.

A2. $\exists b_0 \in \mathcal{B}$ such that $\forall b \in \mathcal{B}$, there exists a unique $a_b \in \mathcal{A}$ mapping $b_0 \to b$.

A3. $\exists a_1 \neq a_2$ such that $a^k \neq a^j$ for all positive integers j, k.

A4. $\forall a \in \mathcal{A}$ and $\forall b_1 \neq b_2 \in \mathcal{B}$, $a(b_1) \neq a(b_2)$.

A5. $\forall a_1, a_2 \in \mathcal{A}$, $a_1 \circ a_2 = a_2 \circ a_1$.

A6. $a^k \neq a \; \forall a \in \mathcal{A}$ and $k > 1$.

Note that A6 should be referring to all $a \neq a_{b_0}$, which would, by A2, map $b_0 \to b_0$. I will focus on the WT model, as the others will be similar.

$$\frac{\mu^{WT}(\omega|s)}{\mu^{WT}(\omega'|s)} = \frac{\mu_0(\omega)}{\mu_0(\omega')} \frac{\pi(s|\omega)}{\pi(s|\omega')} \frac{\delta(u(g(\omega)))}{\delta(u(g(\omega')))}$$

Notably, A2 is satisfied because for any prior and posterior, one can always find a signal with certain conditional probabilities that would result in that posterior. As seen from the updating process, different π's result in different posteriors, and hence the signal is unique as well. For A3, one can find any example such that for s and s', $\frac{\pi(s|\omega)}{\pi(s|\omega')} > 1 > \frac{\pi(s'|\omega)}{\pi(s'|\omega')}$. A4 holds as different priors result in different posteriors. For A5, assume there are two consecutive signals s and s'.

Then

$$\frac{\mu^{WT}(\omega|s,s')}{\mu^{WT}(\omega'|s,s')} = \frac{\mu_0(\omega)}{\mu_0(\omega')}\frac{\pi(s|\omega)}{\pi(s|\omega')}\frac{\pi(s'|\omega)}{\pi(s'|\omega')}(\frac{\delta(u(g(\omega)))}{\delta(u(g(\omega')))})^2 = \frac{\mu^{WT}(\omega|s',s)}{\mu^{WT}(\omega'|s',s)}$$

A5 is satisfied as the ordering doesn't influence the posterior here. Except for the signal that maps a belief onto itself, all other signals change the belief in some way. My model, however, will violate A5.

Cripps (2018) axiomatized a similar model where the dividing information into multiple pieces does not change the final posterior belief. However, neither the WT model nor my model satisfies this, since each time an updating occurs, a distortion to the belief is applied. Dividing an updating process into multiple stages thus compounds the distortion.

Appendix III - Discussion on the Ordering of Signals with $|\Omega| > 2$

The situation becomes complex when $|\Omega| > 2$. The optimal ordering essentially depends on the total distortion for each and every state. For instance, even if signal s gives a greater interim posterior $\mu_1^{CB}(\omega)$ than s', and hence a higher total distortion $F_1(\omega)$ with respect to ω during the second stage belief updating, it does not guarantee signal sequence $\{s, s'\}$ gives higher final posterior $\mu_2^{CB}(\omega)$ than $\{s', s\}$ because the signal s might also have resulted in an interim posterior that is heavily skewed towards some other states. On top of that, the monotonicity of the posterior over the prior belief, which was key to Lemma 1, is lost when we expand to more than two states. This is because, fixing the signals, how the individual continuously updates her beliefs also depends on not only the prior likelihood ratio, but also the ratio between the total distortions $\frac{F_i(\omega)}{F_i(\omega')}$. However, we could state a similar result as Lemma 1 for certain functional

forms of $f(\mu_0(\omega), v(\omega))$.

Corollary 3. *Assume that given two priors, f satisfies:*

$$\frac{\mu_0(\omega)}{\mu_0(\omega')} > \frac{\mu_0'(\omega)}{\mu_0'(\omega')} \Rightarrow \frac{f(\mu_0(\omega), v(\omega))}{f(\mu_0(\omega'), v(\omega'))} > \frac{f(\mu_0'(\omega), v(\omega))}{f(\mu_0'(\omega'), v(\omega'))}$$

Let $\{s_k\}_{k=1}^N$ and $\{s_j\}_{j=1}^N$ be two different permutations of S. Then,

$$\mu^{CB}(\omega|\{s_j\}_{j=1}^N) \geq \mu^{CB}(\omega|\{s_k\}_{k=1}^N)$$

if $\{s_j\}_{j=1}^N$ can be obtained from $\{s_k\}_{k=1}^N$ solely via a series of $\succeq_\omega$-forward swaps (at any position).

Proof: Due to the additional property of f, monotonicity is restored. To see this, let μ_0 and μ_0' be such that

$$\frac{\mu_0(\omega)}{\mu_0(\omega')} > \frac{\mu_0'(\omega)}{\mu_0'(\omega')} \forall \omega' \neq \omega$$

Then

$$\frac{f(\mu_0(\omega), v(\omega))}{f(\mu_0(\omega'), v(\omega'))} > \frac{f(\mu_0'(\omega), v(\omega))}{f(\mu_0'(\omega'), v(\omega'))} \forall \omega' \neq \omega$$

And thus the posterior ratio is higher given the same signal

$$\frac{\mu_1(\omega|s)}{\mu_1(\omega'|s)} > \frac{\mu_1'(\omega|s)}{\mu_1'(\omega'|s)} \forall \omega' \neq \omega$$

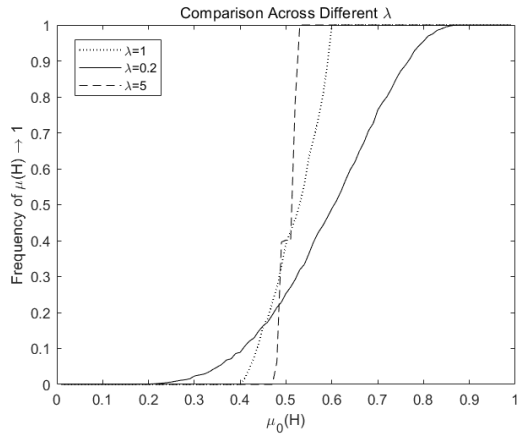
Repeating this gives a higher final posterior ratio.

What is left to show is that a forward swap increases interim posterior ratio. This also easily follows from the monotonicity above. For $s \succeq_\omega s'$, receiving s before s' would result in a higher

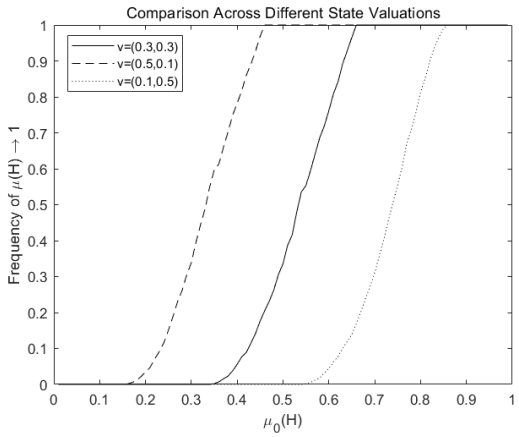
interim posterior ratio, and a higher interim total distortion ratio as a result.

□

Appendix IV - Simulations for Example 3



(a) Effect of applying a convex/concave transformation.



(b) Effect of changing v.

Figure A.1: Frequency distributions of posterior beliefs from simulations.

For each of the 99 prior beliefs between $\mu_0(H) = 0.01$ and 0.99, 10,000 instances of the posterior beliefs after 100 random signals were simulated. Signals were generated according to the information structure in *Example 3* and assuming the true state is L. Figure A.1 presents the relative frequencies of posteriors that converge to $\mu(H) = 1$, plotted against the starting prior beliefs. Figure A.1a assumes $F_0(\omega) = \mu_0(\omega)$ and specifically compares how applying a concave or convex transformation to the total distortion by raising it to a power λ, i.e. varying the strength of UCB, would affect the distribution. Observe that there are two clear thresholds for the prior belief, one where beliefs above the threshold will always converge to $\mu(H) = 1$, and another where beliefs below the threshold will never converge to $\mu(H) = 1$ (and instead converge to $\mu(H) = 0$). The greater the UCB, as represented by a smaller λ, the tighter the interval of prior

beliefs for which the limiting posterior belief is uncertain.

The probability that beliefs converge to $\mu(H) = 1$ increases relatively smoothly over the prior belief. Interestingly, there is a relatively flat segment on the curve for $\lambda = 5$. For small λ, prior beliefs close to $\mu_0(H) = 0.5$ only need one signal to cross one of the two thresholds. Signal h appears with probability 0.4, given the true state is L, and thus the probability that the priors in that range converge to $\mu(H) = 1$ is 0.4. The first signal is l with probability 0.6, in which case the belief will certainly converge to $\mu(H) = 0$.

Figure A.1b assumes $F_0(\omega) = \mu_0(\omega) + v(\omega)$ and shows that changing the valuations over states shift the curve. A higher valuation over state H, relative to state L, would systematically increase the probability that beliefs converge to $\mu(H) = 1$, hence a leftward shift. Conversely, a higher valuation over state L than H would decrease the probability that beliefs converge to $\mu(H) = 1$, hence a rightward shift.

Appendix B: Chapter 2 Appendices

Appendix I - Instructions

*The experimenter reads all of the "Experiment Guidelines", "Instructions for Part I", and its "Bonus Payment" aloud to the subjects. The subjects also see the instructions on their screens.

Experiment Guidelines:

Welcome! Thank you for participating in this experimental study. This session will take about 60 minutes. We kindly ask you to not communicate with others around you or use your mobile phones during this session. Food and drinks are not allowed during the session. If you have any questions during this session, please raise your hand. The experimenter will attend to you shortly.

This session consists of two main parts plus a simple demographic survey at the very end. There will be detailed instructions given at the beginning of each part.

Upon completion, your payments will be summarized and displayed to you. You will be paid a show-up fee of $7, along with possible bonus earnings from each of the two parts. The bonus payments for each part are also explained in their respective instructions. Once everyone has finished, the experimenter will call you one by one and hand you the payment, after which you may leave the room. Note that you will NOT be able to leave early even if you complete the study ahead of everyone else.

Instructions for Part I:

Detailed below are the instructions for Part I. The instructions contain information about the procedure and your payment. Before proceeding, there will be a short comprehension quiz to test your understanding of the instructions after you have gone through all the instructions.

Part I is about probability assessment. You will be asked to participate in 8 rounds of the same set of tasks. Every round presents you with a unique scenario and involves the exact same tasks based on the scenario. The structure of each round is as follows.

Scenario Presented
Initial Probability Assessment
Updated Probability Assessment 1
Updated Probability Assessment 2
Updated Probability Assessment 3
Updated Probability Assessment 4
Updated Probability Assessment 5
Updated Probability Assessment 6

At the start of each round, you are presented with a certain scenario, and a question related to it. Please read the scenario prompts and questions carefully, as all of them are different, despite some of them looking similar.

The question is numerical, and based on real life facts or data. It asks you to assess the probability of some real life event, and is always presented as follows: "What do you think are the chances that A is greater than or equal to (less than) B?"

An example is described below: Congo and Angola are two countries in the Central Africa. In 2010, Angola's population was approximately 23 million. What do you think are the chances

that Congo's population in 2010 was greater than or equal to 40 million? What do you think are the chances that Congo's population in 2010 was less than 40 million?

You will be prompted to enter your answers in text boxes.

Afterwards, 6 news sources are independently and randomly drawn, with replacement, from a large pool of news sources. 70% of the pool are truthful news sources, and the remaining 30% are untruthful news sources. In other words, every news source has a 70% chance to be a truthful news source, and a 30% chance to be an untruthful news source. Truthful news sources always tell the truth, and untruthful ones always lie.

Each news source will provide you one piece of information concerning the answer to the question. Each piece of information is a message that says "Greater Than or Equal to" or "Less Than". For example: If Congo's population in 2010 was indeed greater than 40 million, a truthful source will say "Greater Than or Equal to", but an untruthful news source will say "Less Than".

Immediately after receiving each piece of news, you will be asked to reassess the probabilities that you initially stated. Since there are 6 news sources that will provide you with information, you will have the opportunity to reassess the probabilities 6 times.

After all 6 probability reassessments, you will proceed to the next round.

Bonus Payment:

For each round, there is a chance you may be rewarded $1.50 based on your probability assessments. One of the 7 probability assessments of each round (including the one immediately after the question and all 6 reassessments) will be randomly selected, each with equal chance, to determine whether you receive the bonus payment. The payment is determined in a way such that you will maximize your chances of earning $1.50 if you simply report your beliefs truthfully. If you are interested in reading more about how the payoffs are calculated, please click here. You

may also consult the experimenter about this payment scheme after the session should you have any remaining questions.

Payment Scheme Explanation - accessed upon clicking "here":

The BDM payment scheme is adopted to incentive you to report your true beliefs. The details of this payment scheme are described below. It is okay if you cannot follow the details. Simply remember that it is in your best interest to truthfully report your best assessment on the probabilities that each question asks for, as truthful reports will maximize your probability of earning $1.50 in each round.

In every round, you make seven probability assessments about the same event. For each assessment, you state a probability that "A is greater than or equal to B" (denote PG) and a probability that "A is less than B" (denote PL). In each round, one of the seven reported PG values is randomly selected. Denote this reported probability as P*.

The computer randomly generates a number from 0 to 100 and call this number N. N is compared to P*.

If N is greater than or equal to P*, you will be awarded $1.50 with N% chance.

If N is less than P*, you will be awarded $1.50 if the statement "A is greater than or equal to B" is true.

To see why it is optimal to truthfully report your best assessment of probabilities, consider the following.

If you believe the probability of statement "A is greater than or equal to B" being true is 60%, and reported 20%, then for any computer generated number N between 20 and 60, you win $1.50 with probability N%, which is lower than 60%. However, if you reported 60%, then for any generated number N between 20 and 60, you win $1.50 when the statement is true, which

has a 60% chance.

On the other hand, if you reported something larger, say 80%, then for any computer generated number N between 60 and 80, you win $1.50 when the statement "A is greater than or equal to B" is true, which has a 60% chance. However, if you reported 60%, then for any generated number N between 60 and 80, you win $1.50 with N% chance, which is larger than 60%.

Instructions for Part II:

For Part II, you will face 11 decision problems. Each decision is a paired choice between "Option A" and "Option B". The options on the left (Option A) are risky and pay either $2.00 or $0.25. Note that the chance of being paid $2.00 starts at 0% and gradually increases to 100% when you move down the list. The options on the right (Option B) are fixed and pay $1.00 for sure.

For each row, think about whether you prefer Option A or Option B. Please select at which row you would like to switch from Option B to Option A. Since the chance of being paid $2.00 for option A gradually increases as you move down the rows, we will automatically fill out all the remaining rows to save you time. You can change where you want to switch until you are satisfied with your answer.

Bonus Payment:

After you completed all 11 decision problems, one of them will be randomly chosen for your payment, and you will get whatever you chose on that row. If you chose Option A for the decision selected for payment, the computer will randomly determine whether you earn $2.00 or $0.25 according to the corresponding probabilities in that decision. If you chose Option B for the decision selected for payment, you will earn $1.00.

Appendix II - Graphs and Tables

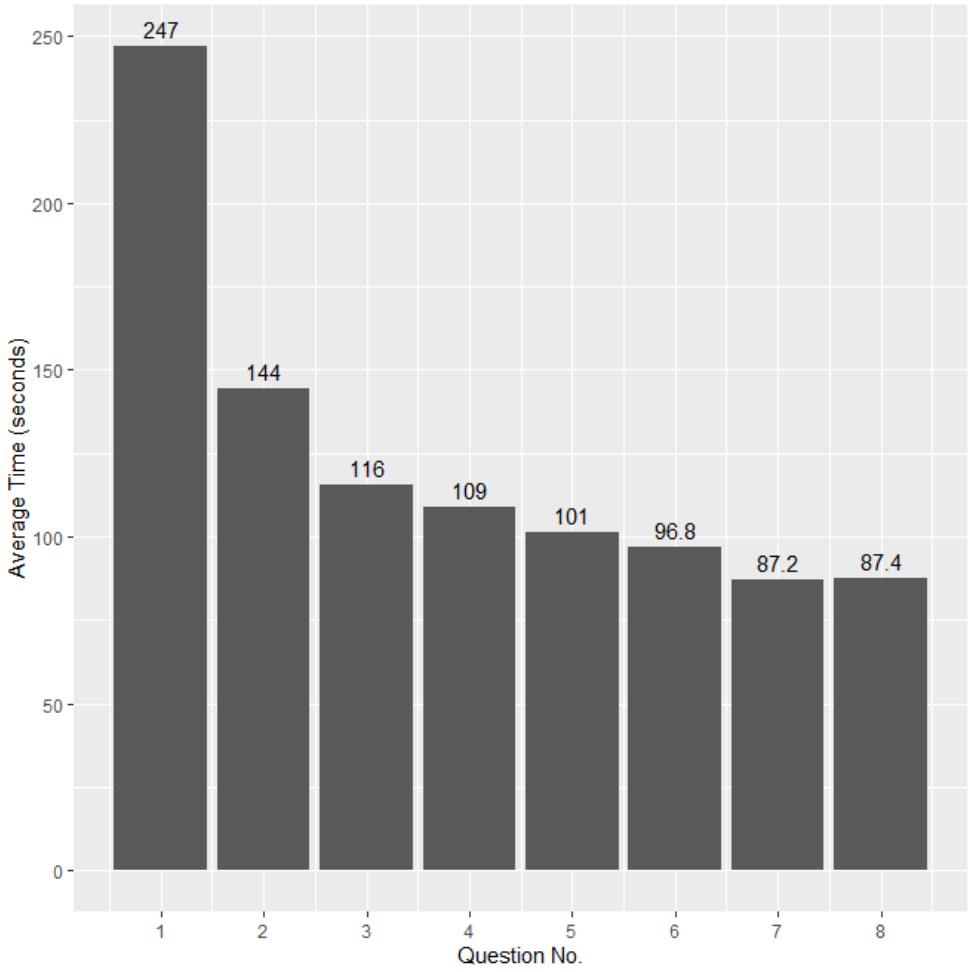


Figure B.1: Average time (seconds) spent on each round.

	Democrat	Republican	Other	Neutral	Total
Woman	72	5	13	22	112
Man	28	11	5	26	70
Non-Binary	0	0	1	0	1
Total	100	16	18	49	183

Table B.1: Cross-table of Gender by Party affiliation

Statistic	Mean	St. Dev.	Min	1st Quart.	Median	3rd Quart.	Max
Dem. Rating	5.989	2.120	0.000	5.000	6.000	7.000	10.000
Rep. Rating	3.197	2.200	0.000	2.000	3.000	5.000	10.000
Environment	7.546	2.364	0.000	6.000	8.000	10.000	10.000
Economy	7.077	2.215	0.000	6.000	7.000	9.000	10.000
Politics	5.727	2.827	0.000	4.000	6.000	8.000	10.000

Table B.2: Summary Statistics: $N = 183$

Party	Democratic	Republican	Other	Neutral/No Party
Mean Dem. Rating	6.79	3.06	5.37	5.54
Mean Rep. Rating	2.19	7.44	2.53	4.15

Table B.3: Average party rating grouped by party affiliation.

	Dependent variable:		
	Posterior Log Likelihood		
	Full Sample	First Half	Second Half
Prior (β_1)	0.582^{***}	0.579^{***}	0.581^{***}
	(0.016)	(0.028)	(0.020)
Signal (β_2)	0.787^{***}	0.807^{***}	0.775^{***}
	(0.018)	(0.031)	(0.021)
Prior difference (β_3)	0.969^{***}	1.008^{***}	0.954^{***}
	(0.064)	(0.110)	(0.078)
Value difference (β_4)	0.043	0.025	0.054^{*}
	(0.027)	(0.049)	(0.032)
Constant	-0.012	-0.040	0.003
	(0.015)	(0.027)	(0.018)
Observations	7,902	2,958	4,944
Note:			*p<0.1; **p<0.05; ***p<0.01

Table B.4: OLS while ignoring subjects who have fewer than 7 rounds of data points left.

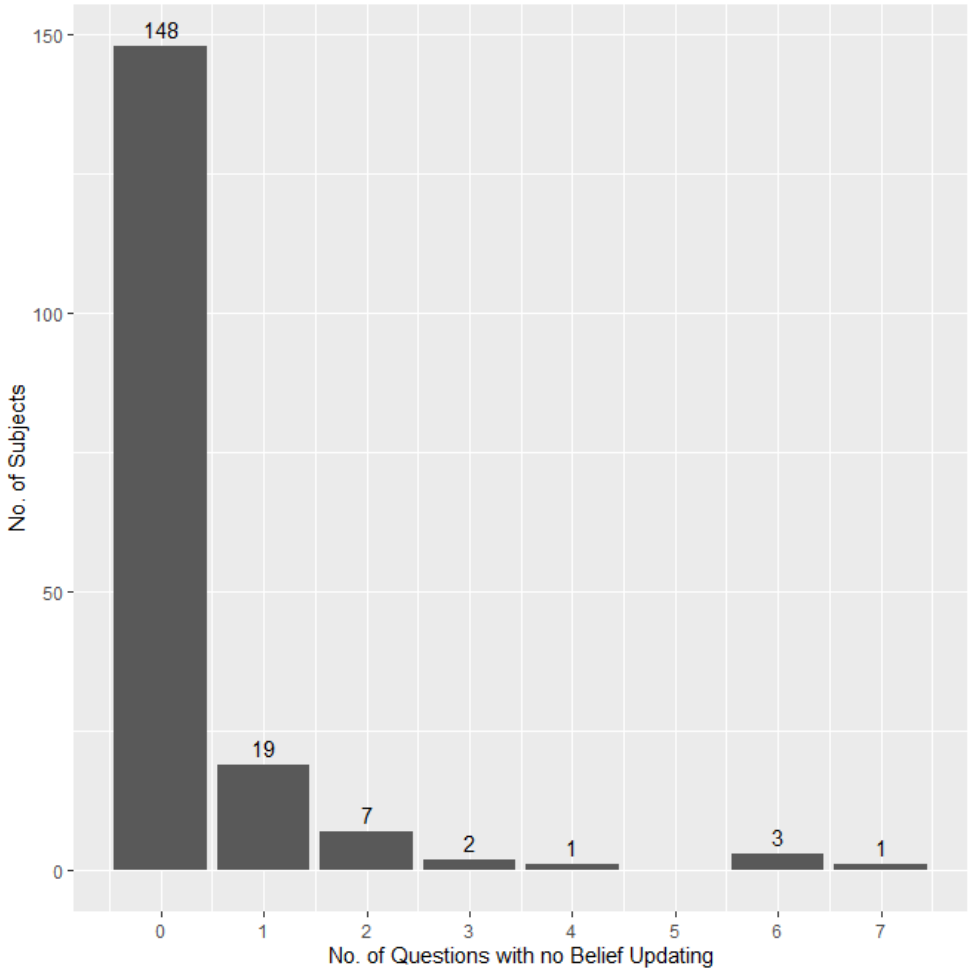
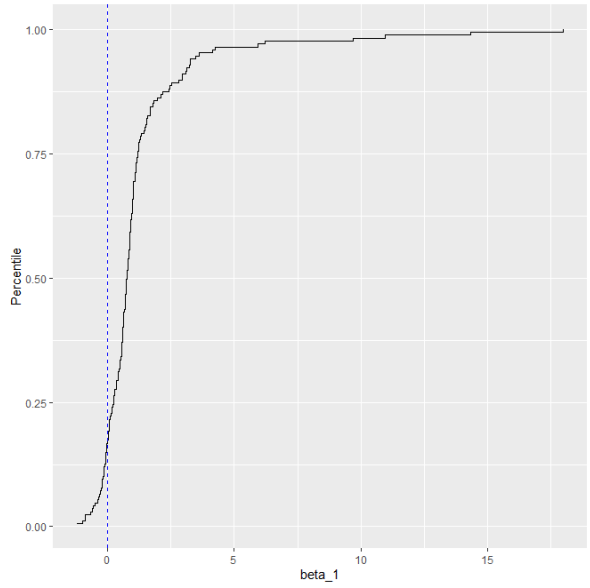


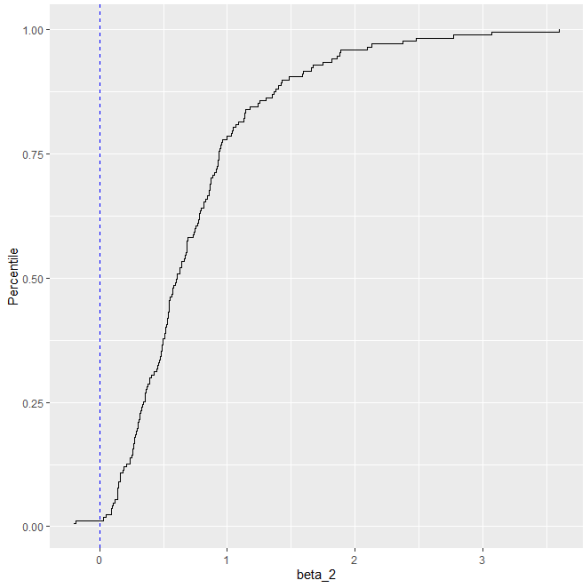
Figure B.2: Number of subjects (y-axis) whose beliefs did not change throughout the whole question, for the respective number of questions (x-axis).

	β_0	β_1	β_2	β_3	β_4
β_0	1	0.031	-0.207	-0.059	-0.256
β_1		1	-0.100	-0.981	-0.109
β_2			1	0.158	0.225
β_3				1	0.167
β_4					1

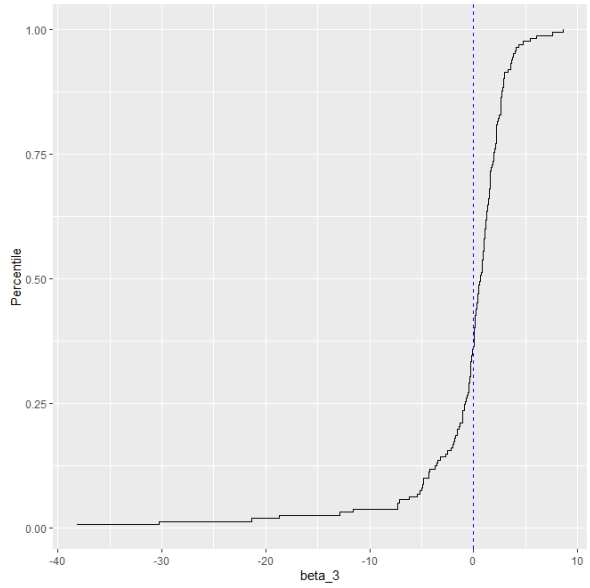
Table B.5: Correlation between estimates



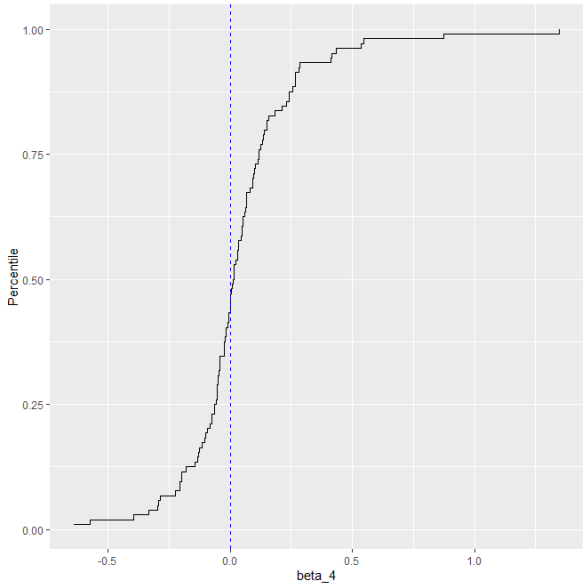
(a) $CDF(\beta_1)$



(b) $CDF(\beta_2)$



(c) $CDF(\beta_3)$



(d) $CDF(\beta_4)$

Figure B.3: CDF's of individually estimated parameters.

	gender	education	statistics knowledge	environment	economy	politics
β_0	0.106	0.148	0.090	-0.065	0.112	0.018
β_1	-0.090	0.178	0.096	0.026	0.037	0.033
β_2	0.042	-0.047	-0.036	-0.011	-0.080	-0.103
β_3	0.088	-0.198	-0.071	-0.042	-0.045	-0.034
β_4	0.060	0.044	-0.172	-0.101	-0.179	-0.166

Table B.6: Correlation between estimates and demographic variables

	Dependent variable:			
	Posterior Log Likelihood			
	Fist 2 Rounds	Last 6 Rounds	First 6 Rounds	Last 2 Rounds
Prior (β_1)	0.588^{***}	0.528^{***}	0.546^{***}	0.540^{***}
	(0.034)	(0.018)	(0.019)	(0.031)
Signal (β_2)	0.824^{***}	0.812^{***}	0.829^{***}	0.777^{***}
	(0.038)	(0.021)	(0.022)	(0.033)
Prior difference (β_3)	0.928^{***}	1.138^{***}	1.057^{***}	1.149^{***}
	(0.134)	(0.075)	(0.077)	(0.121)
Value difference (β_4)	0.028	0.060^{*}	0.032	0.101^{**}
	(0.058)	(0.032)	(0.033)	(0.051)
Constant (β_0)	-0.055^{*}	-0.004	-0.002	-0.064^{**}
	(0.034)	(0.018)	(0.019)	(0.028)
Observations	2,100	6,180	6,216	2,064
Adjusted R^2	0.610	0.635	0.619	0.662
Note:				*p<0.1; **p<0.05; ***p<0.01

Table B.7: Regressions on the split data featuring a 6-2 split and a 2-6 split.

Appendix III - Analysis with Risk Preferences

	Dependent variable:
	Posterior Log Likelihood
Prior (β_1)	0.546^{***}
	(0.016)
Risk	0.011
	(0.008)
Signal (β_2)	0.816^{***}
	(0.018)
Prior difference (β_3)	1.076^{***}
	(0.065)
Value difference (β_4)	0.049^{*}
	(0.028)
Prior*Risk	0.018^{**}
	(0.009)
Prior difference*Risk	-0.066^{*}
	(0.035)
Constant	-0.018
	(0.016)
Observations	8,280
Note:	*p<0.1; **p<0.05; ***p<0.01

Table B.8: Regression with risk preferences.

In Table B.8 above, I present the aggregate level regression with the subjects' risk preferences interacting with the variables "Prior" and "Prior difference". Other interactions were explored, but showed no significance. The variable "Risk" measures the subjects' switching point in multiple price list task. Risk neutral subjects would switch between the 6th and the 7th choice

(out of 11). The higher the value, the more subjects preferred the risky option compared to the certain option, the more risk seeking the subjects are. As indicated in the regression table, risk preference itself does not create systematic belief shifts.[1] However, it does impact the estimates for β_1 and β_3. The signs of the interaction terms indicate that a higher preference for risk results in lower levels of both base-rate neglect and unmotivated confirmation bias. It is possible that, since the majority of subjects are risk averse (see Figure B.4), the higher value corresponds to the subjects being closer to risk neutrality, which in turn drives more accurate judgements.

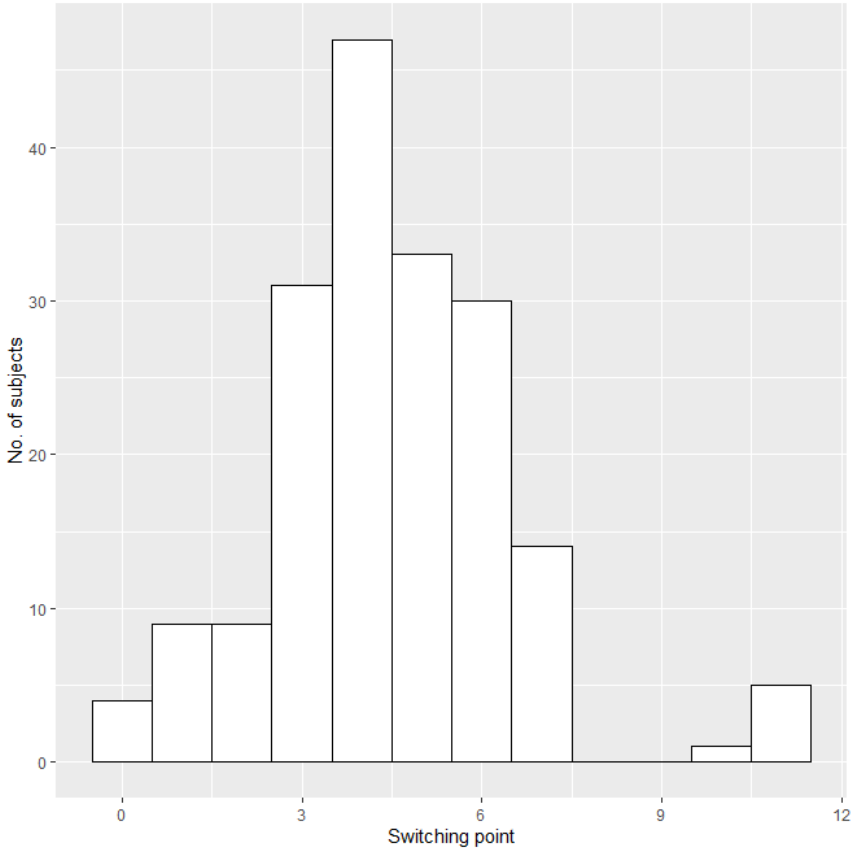


Figure B.4: Distribution of switching points among subjects. Larger value means more risk seeking.

[1]Here, the variable "Risk" is demeaned in order to provide a clear comparison with the average subject as shown in the OLS results earlier.

Appendix IV - Analysis with Grether's Model

The Grether's model (48) has been the workhorse model for estimating belief data. Here I first show that when applied to the setting of my experiment, it is not too different from that used in the main analysis, that is the model in [47]. Then, I support this point by showing the aggregate level regression results are similar.

Based to the previous literature, the regression equation utilizing the Grether's model would be as follows:

$$ln\left(\frac{\mu_1(G|s)}{\mu_1(L|s)}\right) = \beta_0 + \beta_1 ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) + (\beta_2 + \beta_3 d_{\mu_0^+} + \beta_4 d_{v^+} - \beta_5 d_{\mu_0^-} - \beta_6 d_{v^-}) ln\left(\frac{\pi(s|G)}{\pi(s|L)}\right) \quad \text{(B.1)}$$

where $d_{\mu_0^+}$ is the dummy variable that takes value 1 if the signal is confirmatory (i.e. $s = g$ when $\mu_0(G) > 0.5$ or $s = l$ when $\mu_0(G) < 0.5$), $d_{\mu_0^-}$ is the dummy variable that takes value 1 if the signal is dis-confirmatory (i.e. $s = g$ when $\mu_0(G) < 0.5$ or $s = l$ when $\mu_0(G) > 0.5$), d_{v^+} is the dummy that takes value 1 if the signal aligns with the person's motivation (i.e. $s = g$ when the person prefers state G or $s = l$ when the person prefers state L), and d_{v^-} is the dummy that takes value 1 if the signal does not align with what the person prefers (i.e. $s = g$ when the person prefers state L or $s = l$ when the person prefers state G).

By the definitions of $d_{\mu_0^+}$ and $d_{\mu_0^-}$, we have

$$d_{\mu_0^+} = \mathbb{1}[ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) \cdot ln\left(\frac{\pi(s|G)}{\pi(s|L)}\right) > 0]$$

$$d_{\mu_0^-} = \mathbb{1}[ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) \cdot ln\left(\frac{\pi(s|G)}{\pi(s|L)}\right) < 0]$$

Additionally, $|ln\left(\frac{\pi(s|G)}{\pi(s|L)}\right)| = ln(\frac{7}{3}) \approx 0.847$ due to the signal structure. Hence, when $s = g$,

$$(\beta_3 d_{\mu_0^+} - \beta_5 d_{\mu_0^-}) ln\left(\frac{\pi(s|G)}{\pi(s|L)}\right) = 0.847 \cdot [\beta_3 \mathbb{1}(ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) > 0) - \beta_5 \mathbb{1}(ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) < 0)]$$

When $s = l$,

$$(\beta_3 d_{\mu_0^+} - \beta_5 d_{\mu_0^-}) ln\left(\frac{\pi(s|G)}{\pi(s|L)}\right) = -0.847 \cdot [\beta_3 \mathbb{1}(ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) < 0) - \beta_5 \mathbb{1}(ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) > 0)]$$

$$= 0.847 \cdot [\beta_5 \mathbb{1}(ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) > 0) - \beta_3 \mathbb{1}(ln\left(\frac{\mu_0(G)}{\mu_0(L)}\right) < 0)]$$

Note that the two coincide with each other exactly if we assume $\beta_3 = \beta_5$, that is if we assume that the effect of unmotivated confirmation bias is independent of the signal label, and only dependent on the prior and signal accuracy. This assumption would also make the equation a special case of [47] within this specific setting. This method of estimation essentially partitions the range of prior beliefs into $\mu_0(G) < 0.5$, $\mu_0(G) = 0.5$, and $\mu_0(G) > 0.5$, and shifts the posterior beliefs by a constant amount depending on where the prior belief lies.

As for the estimation equation (2.3), the bias in the belief shift is a smooth function of the prior, instead of a step function. Specifically, for any two priors μ_0 and μ_0', with $\mu_0(G) > 0.5$ and $\mu_0'(G) < 0.5$, the difference in the belief shift is

$$\beta_3 \cdot 2(\mu_0(G) - \mu_0'(G))$$

Therefore, it is clear that Grether's model is estimating unmotivated confirmation bias as a step function of the prior belief, while (2.3) is estimating the bias as a smooth linear function.

Similarly, we can view d_{v^+} and d_{v^-} the same way. The former is 1 when $v(G) - v(L) = \delta_v > 0$, and the latter is 1 when $\delta_v < 0$. When $s = g$,

$$(\beta_4 d_{v^+} - \beta_6 d_{v^-}) ln\left(\frac{\pi(s|G)}{\pi(s|L)}\right) = 0.847 \cdot [\beta_4 \mathbb{1}(\delta_v^* > 0) - \beta_6 \mathbb{1}(\delta_v^* < 0)]$$

When $s = l$,

$$(\beta_4 d_{v^+} - \beta_6 d_{v^-}) ln\left(\frac{\pi(s|G)}{\pi(s|L)}\right) = 0.847 \cdot [\beta_6 \mathbb{1}(\delta_v^* > 0) - \beta_4 \mathbb{1}(\delta_v^* < 0)]$$

Here, if we have $\beta_4 = \beta_6$, the Grether's model would be estimating motivated confirmation bias the same way as what is implemented in the main analysis.

Column 1 of Table B.9 shows the OLS results using (B.1). We note that the qualitative observations do not change much compared to Table 2.2, and the signs of the estimates are all as expected. While the estimates for the effect of motivated confirmation bias seem to be less significant (with p values within the 0.1-0.2 range), I believe this may be partly attributed to the loss of power. As noted above, the idea of these dummies are no different from what is implemented in the main analysis, that is using a normalized variable δ_v with values $\{-1, 0, 1\}$. Therefore, unless there is reason to believe that the strength of motivated confirmation bias is dependent on the signals, the estimation using pairs of dummies may even suffer a loss in statistical power compared to using a single variable. Column 2 of Table B.9 is evidence for this.

	Dependent variable:	
	Posterior Log Likelihood	
	(1)	(2)
Prior (β_1)	0.720^{***}	0.719^{***}
	(0.010)	(0.010)
Signal (β_2)	0.757^{***}	0.757^{***}
	(0.052)	(0.052)
$d_{\mu_0^+}$ (β_3)	0.313^{***}	0.313^{***}
	(0.063)	(0.061)
$d_{\mu_0^-}$ ($-\beta_5$)	-0.155^{**}	-0.156^{**}
	(0.063)	(0.062)
d_{v^+} (β_4)	0.074	
	(0.053)	
d_{v^-} ($-\beta_6$)	-0.088	
	(0.058)	
Value difference		0.048^{*}
		(0.029)
Constant	-0.006	-0.010
	(0.016)	(0.016)
Observations	8,280	8,280
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table B.9: OLS with Grether's model

Appendix V - Alternative Specifications for the Valuation of States

A major concern regarding the construction of the "value difference" variable may be that what motivates the subjects is actually their concern for the economy or the environment, instead of the governing party for which the economic or environment achievements are attributed to. Therefore, I also briefly present four possible alternative specifications of δ_v where it instead depends on the subjects' reported degree of concern for economic or environment issues (out of 0, representing no concern at all, to 10, representing extreme concern). Although their reported values may not be comparable across individuals, they nevertheless are the best proxies I could obtain from the available data. What is reassuring is that the distributions of the ratings, as can be seen from Table B.2, are fairly representative of what one would expect from college students.

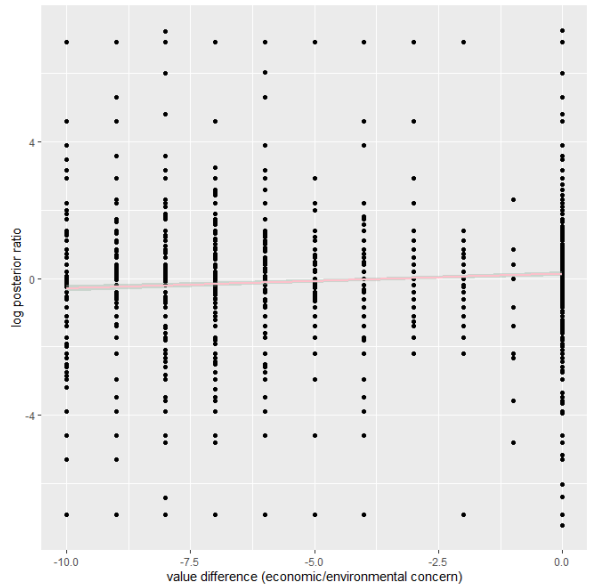
The first corresponds to the case where I assume that everyone cares about the economy and the environment, and only differ in how much they care, as indicated by their reported level of concern. As such, δ_v for all questions regarding the economy or environment are determined with $v(L)$ equal to their level of concern and $v(G) = 0$ (as they would prefer the pollution, inflation, or unemployment to be lower). The second specification is similar to the first, but with δ_v normalized. The third specification corresponds to the case where I categorize those who reported a concern level less than 5 as preferring the economy or environment to be worse, those who reported exactly 5 as being neutral, and those who reported higher than 5 as preferring the economy or environment to be better. I assign δ_v as -1, 0, or 1 accordingly. The fourth corresponds to the case where I assume the subject would be motivated by whichever they care more between politics and the environment/economy. If their reported concern for politics is higher, δ_v would depend on their party affiliation, as in the main analysis. If their reported

concern for the environment or economy is higher, δ_v would be determined in the same way as in column 2. If their reported concern for the two are equal, and both would motivate their bias in the same direction, I would use either of the aforementioned ways of determining δ_v. If their reported concern for the two are equal, but their party affiliation would motivate an opposite bias as their concern for the economy or environment would, δ_v would be 0. Notably, all of these have insignificant β_4 estimates.

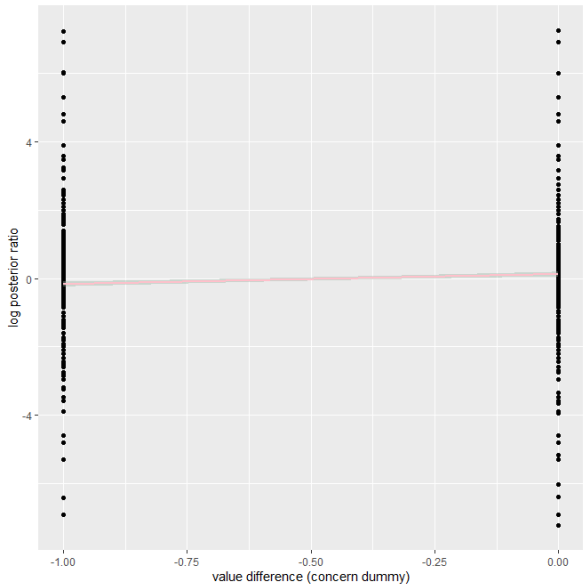
I present a series of scatter plots for all all four specifications in Figure B.5 below. In all four graphs, there is a slight positive correlation indicated by the fitted lines.[2] While all four specifications do not seem like bad alternatives, the slopes are not steep, similar to what is observed in Figure 2.3b. Therefore, I follow [21] and only present the regression results with δ_v determined by party affiliation.[3]

[2]If one instead plots the dependent variable against randomly determined δ_v (out of $\{-1, 0, 1\}$), it will be evident that what is observed cannot be obtained from randomized values.

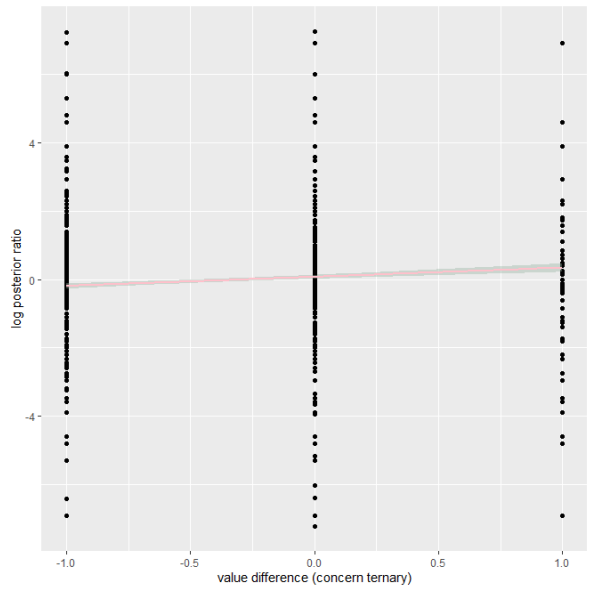
[3]Note that if one conducts aggregate level regression with these alternative specifications, the significance of β_4 would not improve, indicating that none of them are distinctly better specifications.



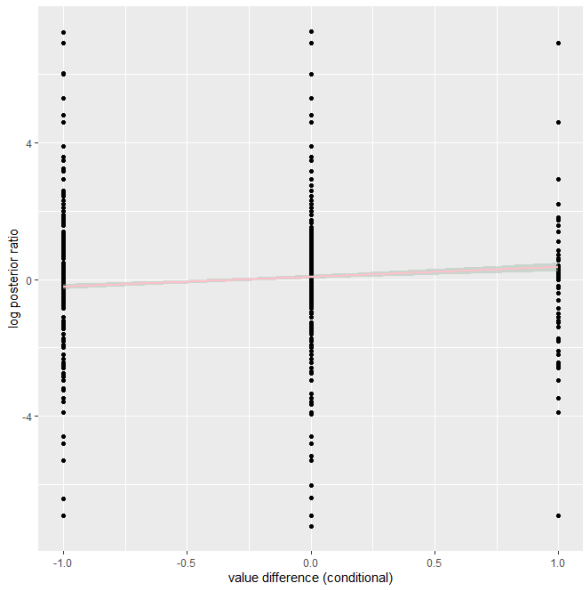
(a) Scatter plot for alternative specification 1.



(b) Scatter plot for alternative specification 2.



(c) Scatter plot for alternative specification 3.



(d) Scatter plot for alternative specification 4.

Figure B.5: Scatter plots of log posterior ratio against other specifications of δ_v.

Appendix C: Chapter 3 Appendices

Appendix I - Proofs

Proof of Lemma 2: Solution to the Constrained Problem

We first note that Lemma 1 and Corollary 1 of [41] still holds with the information constraint present. Their Lemma 1 shows that having more signals than actions while keeping the consumer's expected payoff the same will only increase information cost. Their Corollary 1 follows naturally by stating that, since the signals and actions are one-to-one, then the consumer is essentially choosing the joint probability distribution of states and actions. The optimality of the simplest signal structures, is not affected by the firm's extra constraint that we have in our model. Without loss of generality, assume that consumers would expect themselves to buy if they receive an h signals, and reject if they receive an l signal. Denote this correspondence as a_s where $a_h = b$ and $a_l = r$. The following claim provides a general solution for optimal signal structure for the constrained optimization problem. After proving the claim we will proceed with the closed form formula of π^* and q^* specific to our setup and derive Equations (C.1) and (3.4).

Claim 1: A consumer with cost parameter λ chooses optimal signal structure

$$\pi^*(s|\omega) = \begin{cases} \pi^{\lambda}(s|\omega) & \lambda \in [\underline{\lambda}, \infty) \\ \\ \pi^{\underline{\lambda}}(s|\omega) & \lambda \in [0, \underline{\lambda}) \end{cases} \tag{C.1}$$

where

$$\pi^{\lambda}(s|\omega) = \frac{q(s)e^{u(a_s|\omega)/\lambda}}{\sum_s q(s')e^{u(a_{s'}|\omega)/\lambda}}$$

Proof of Claim 1: The following is the ex-ante expected utility of choosing information structure π when the consumer expects to follow the signal either because she is a Bayesian or a naive non-Bayesian:

$$\begin{aligned} \sum_{\omega}\sum_{s} V(s,\pi)\pi(s|\omega)f(\omega) &= \sum_{s} V(s,\pi)\sum_{\omega}\pi(s|\omega)f(\omega) = \sum_{s} V(s,\pi)q(s) \\ &= \sum_{s}\sum_{\omega} g^B(\omega|s)u(a_s|\omega)q(s) = \sum_{s}\sum_{\omega} u(a_s|\omega)\pi(s|\omega)f(\omega) \end{aligned}$$

The first equality holds because $V(s,\pi)$ is based on the expectation of the state of the world. The second equation holds by the definition of $q(s)$. The third one follows from the definitions of v and a_s. Finally, the last equation is an implication of the Bayes rule. We first consider the case where $\lambda > 0$. Adding in the firm's constraint, as well as the default constraints that $\pi(s|\omega)$ needs to be non-negative and that they add to 1 across signals, we have the Lagrangian for the consumer's problem:

$$L = \sum_{\omega}\sum_{s} u(a_s|\omega)\pi(s|\omega)f(\omega) - \lambda I(\pi) + \zeta(\kappa - I(\pi)) + \sum_{s}\sum_{\omega}\gamma_{s,\omega}\pi(s|\omega)$$
$$+ \sum_{\omega}\mu_{\omega}[1 - \sum_{s}\pi(s|\omega)]$$

where ζ, $\gamma_{s,\omega}$ and μ_{ω} are the Lagrangian multipliers. The first order condition with respect to $\pi(s|\omega)$ is:

$$u(a_s|\omega)f(\omega) - (\lambda+\zeta)[ln(\pi(s|\omega) - ln(q(s))]f(\omega) + \gamma_{s,\omega} - \mu_{\omega} = 0$$

If $q(s) > 0$, and since $u(a_s|\omega) > -\infty$, we must have $\pi(s|\omega) > 0$ for all ω. Otherwise, if $\pi(s|\omega') = 0$ for some ω', $ln(\pi(s|\omega') = -\infty$, causing $\mu_{\omega'} = \infty$. This implies that, for all $s' \neq s$, $\pi(s'|\omega') = 0$. But this contradicts with $\sum_s \pi(s|\omega) = 1$ for all ω.

Therefore, $\gamma_{s,\omega} = 0$. By solving the first order condition above, we get:

$$\begin{aligned}
\frac{\mu_{\omega}}{f(\omega)} &= u(a_s|\omega) - (\lambda+\zeta)[ln(\pi(s|\omega) - ln(q(s))] \\
&\Rightarrow e^{[u(a_s|\omega) - \frac{\mu_{\omega}}{f(\omega)}]/(\lambda+\zeta)} = \frac{\pi(s|\omega)}{q(s)} \\
&\Rightarrow \sum_s e^{[u(a_s|\omega) - \frac{\mu_{\omega}}{f(\omega)}]/(\lambda+\zeta)} q(s) = 1 \\
&\Rightarrow \sum_s e^{u(a_s|\omega)/(\lambda+\zeta)} q(s) = e^{[\frac{\mu_{\omega}}{f(\omega)}]/(\lambda+\zeta)} \\
&\Rightarrow \pi(s|\omega) = \frac{q(s)e^{u(a_s|\omega)/(\lambda+\zeta)}}{\sum_s q(s')e^{u(a_{s'}|\omega)/(\lambda+\zeta)}}
\end{aligned}$$

The last equation is derived from plugging the previous observation into the second line.

If $q(s) = 0$, then we simply have $\pi(s|\omega) = 0$ for all ω. If $\lambda = 0$, the problem is trivial,

where the consumers want to gather as much information as possible. Since they are constrained by the firm, they acquire information up to the limit $I(\pi) = \kappa$.

To obtain the value of ζ, we need to note that, since it is the Lagrangian multiplier for the firm's constraint, it is only positive for consumers when it is a binding constraint. Furthermore, it acts purely as an additive amount to the marginal cost of information λ, and the consumers behave as if they have a higher marginal cost. Let π^λ denote the optimal π given λ. The constraint binds when consumers' unconstrained optimization would lead to an information structure such that $I(\pi^\lambda) > \kappa$, which happens when consumers have naturally low λ. This is because $I(\pi^\lambda)$ is decreasing in λ. To see this, assume that the optimal signal structures (without the firm's constraint) are such that, for some $\lambda > \lambda'$, $I(\pi^\lambda) > I(\pi^{\lambda'})$. Denoting the optimal expected payoffs as $EV(\pi^\lambda)$ and $EV(\pi^{\lambda'})$ respectively, we have the following due to their optimality:

$$EV(\pi^\lambda) - \lambda I(\pi^\lambda) \geq EV(\pi^{\lambda'}) - \lambda I(\pi^{\lambda'})$$

$$EV(\pi^{\lambda'}) - \lambda' I(\pi^{\lambda'}) \geq EV(\pi^\lambda) - \lambda' I(\pi^\lambda)$$

And thus

$$\lambda'[I(\pi^\lambda) - I(\pi^{\lambda'})] \geq EV(\pi^\lambda) - EV(\pi^{\lambda'}) \geq \lambda[I(\pi^\lambda) - I(\pi^{\lambda'})]$$

which is impossible as $\lambda > \lambda'$.

Define $\underline{\lambda}$ such that, without the firm's constraint, the optimal signal structure is such that $I(\pi^{\underline{\lambda}}) = \kappa$. Then, for the constrained consumers, it must be that $I(\pi^{\lambda+\zeta}) = \kappa$, which would give $\lambda + \zeta = \underline{\lambda}$ due to the monotonicity of $I(\pi)$ in λ.

Lastly, the case of $\lambda = 0$ becomes trivial, as the above solution would still hold.

In sum, the consumer who has marginal cost λ acquires the following optimal signal structure:

$$\pi^*(s|\omega) = \begin{cases} \pi^{\lambda}(s|\omega) & \lambda \in [\underline{\lambda}, \infty) \\ \pi^{\underline{\lambda}}(s|\omega) & \lambda \in [0, \underline{\lambda}) \end{cases}$$

where

$$\pi^{\lambda}(s|\omega) = \frac{q(s)e^{u(a_s|\omega)/\lambda}}{\sum_s q(s')e^{u(a_{s'}|\omega)/\lambda}}$$

This proves Claim 1.

Next, we solve for the closed form solutions of q and π within the setting of our model. Note that the following discussion will only include those with $\lambda \in [\underline{\lambda}, \infty)$ for simplicity, but the case where $\lambda \in [0, \underline{\lambda})$ carries over by replacing λ with $\underline{\lambda}$. Plugging in the signals $s \in \{h, l\}$, and the correspondence $a_h = b$, $a_l = r$ (recall that $u(r|\omega) = 0$), we obtain

$$\pi^*(s = l|\omega) = \frac{q(l)}{q(l) + q(h)e^{u(b|\omega)/\lambda}} \quad \text{and} \quad \pi^*(s = h|\omega) = \frac{q(h)e^{u(b|\omega)/\lambda}}{q(l) + q(h)e^{u(b|\omega)/\lambda}}$$

We plug this into $q(s) = \sum_\omega \pi(s|\omega)f(\omega)$ and obtain the following:

$$\frac{q(h)e^{1/\lambda}}{q(l) + q(h)e^{1/\lambda}}f(H) + \frac{q(h)e^{-1/\lambda}}{q(l) + q(h)e^{-1/\lambda}}f(L) = q(h)$$

Denoting for simplicity $q = q(h) = 1 - q(l)$, $f = f(H) = 1 - f(L)$, and assuming $1 > q > 0$, we get

$$\frac{e^{1/\lambda}f}{1 - q + qe^{1/\lambda}} + \frac{f}{1 - qe^{1/\lambda} + q} = 1$$

$$\Rightarrow e^{2/\lambda}f + e^{1/\lambda}qf - e^{2/\lambda}qf + 1 - f - q + qf + e^{1/\lambda}q - e^{1/\lambda}qf$$

$$= e^{1/\lambda} - 2e^{1/\lambda}q + e^{1/\lambda}q^2 + e^{2/\lambda}q - e^{2/\lambda}q^2 + q - q^2 + e^{1/\lambda}q^2$$

$$\Rightarrow [1 - f - e^{1/\lambda} + e^{2/\lambda}f] - q[1 - f - e^{1/\lambda} + e^{2/\lambda}f + 1 + e^{2/\lambda} - 2e^{1/\lambda}] + q^2[1 + e^{2/\lambda} - 2e^{1/\lambda}] = 0$$

$$\Rightarrow (1 - e^{1/\lambda})(1 - f - e^{1/\lambda}f)(1 - q) + (q^2 - q)(1 - e^{1/\lambda})^2 = 0$$

$$\Rightarrow q = \frac{1 - f - e^{1/\lambda}f}{1 - e^{1/\lambda}}$$

Hence,

$$q(h) = \frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}} \quad \text{and} \quad q(l) = \frac{f(H) - f(L)e^{1/\lambda}}{1 - e^{1/\lambda}}$$

The optimal $q(h)$ derived from this can actually be greater than 1 or less than 0, which can happen when $f(H)$ is too large or too small, respectively. (Under the assumption that $f(L) \geq f(H)$, having larger λ will lower $q(h)$ as well.) Since it is a probability, it must satisfy $1 > q(h) > 0$. Otherwise, we have a corner solution.

Pluging the interior solutions for q into the conditional probabilities above, the denominators of $\pi^*(s|H)$ and $\pi^*(s|L)$ are

$$q(l)+q(h)e^{1/\lambda} = 1-q(h)+q(h)e^{1/\lambda} = 1+q(h)(e^{1/\lambda}-1) = 1+f(H)e^{1/\lambda}-f(L) = f(H)(1+e^{1/\lambda})$$

$$q(l) + q(h)e^{-1/\lambda} = f(L)(1 + e^{1/\lambda}),$$

respectively. And we get:

$$\pi^*(h|H) = \frac{q(h)e^{1/\lambda}}{f(H)(1 + e^{1/\lambda})} = 1 - \pi^*(l|H) = 1 - \frac{q(l)}{f(H)(1 + e^{1/\lambda})}$$

$$\pi^*(l|L) = \frac{q(l)e^{1/\lambda}}{f(L)(1+e^{1/\lambda})} = 1 - \pi^*(h|L) = 1 - \frac{q(h)}{f(L)(1+e^{1/\lambda})}$$

for $q(h) \in (0,1)$.

Additionally, when $q(h) = 0$, $\pi^*(h|\omega) = 0 = 1 - \pi^*(l|\omega)$

And when $q(h) = 1$, $\pi^*(h|\omega) = 1 = 1 - \pi^*(l|\omega)$

In sum, the optimal marginal signal probabilities are (and including the case for constrained consumers this time)

$$q^*(h) = \begin{cases} max\left\{0, min\left\{1, \frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}}\right\}\right\} & \lambda \in [\underline{\lambda}, \infty) \\ max\left\{0, min\left\{1, \frac{f(L)-f(H)e^{1/\underline{\lambda}}}{1-e^{1/\underline{\lambda}}}\right\}\right\} & \lambda \in [0, \underline{\lambda}) \end{cases}$$

and $q^*(l) = 1 - q^*(h)$.

The optimal information structure is

$$\pi^*(h|H) = \begin{cases} 0 & q^*(h) = 0 \\ \frac{q^*(h)e^{1/\lambda}}{f(H)(1+e^{1/\lambda})} & \lambda \in [\underline{\lambda}, \infty), q^*(h) \in (0,1) \\ \frac{q^*(h)e^{1/\underline{\lambda}}}{f(H)(1+e^{1/\underline{\lambda}})} & \lambda \in [0, \underline{\lambda}), q^*(h) \in (0,1) \\ 1 & q^*(h) = 1 \end{cases}$$

$$\pi^*(l|L) = \begin{cases} 0 & q^*(l) = 0 \\ \frac{q^*(l)e^{1/\lambda}}{f(L)(1+e^{1/\lambda})} & \lambda \in [\underline{\lambda}, \infty), q^*(l) \in (0,1) \\ \frac{q^*(l)e^{1/\underline{\lambda}}}{f(L)(1+e^{1/\underline{\lambda}})} & \lambda \in [0, \underline{\lambda}), q^*(l) \in (0,1) \\ 1 & q^*(l) = 1 \end{cases}$$

with $\pi^*(h|\omega) = 1 - \pi^*(l|\omega)$.

□

Proof of Lemma 3: Optimal Purchasing Decision

Again, we mainly show the calculations for the unconstrained consumers. The calculations for the constrained consumers are analogous with $\underline{\lambda}$ instead of λ. First, we consider the case where $q(h) \in (0,1)$.

By the updating rule (3.5),

$$g^\delta(H|h) = \frac{f(H)\pi^*(h|H)\delta_H}{f(H)\pi^*(h|H)\delta_H + f(L)\pi^*(h|L)\delta_L} = \frac{e^{1/\lambda}\delta_H}{e^{1/\lambda}\delta_H + \delta_L} = 1 - g^\delta(L|h)$$

$$g^\delta(L|l) = \frac{f(L)\pi^*(l|L)\delta_L}{f(H)\pi^*(l|H)\delta_H + f(L)\pi^*(l|L)\delta_L} = \frac{e^{1/\lambda}\delta_L}{\delta_H + e^{1/\lambda}\delta_L} = 1 - g^\delta(H|l)$$

The expected utilities for buying the product given an h signal or an l signal are:

$$Eu^\lambda(s = h) = g^\delta(H|h) - g^\delta(L|h) = \frac{e^{1/\lambda}\delta_H - \delta_L}{e^{1/\lambda}\delta_H + \delta_L}$$

$$Eu^\lambda(s = l) = g^\delta(H|l) - g^\delta(L|l) = \frac{\delta_H - e^{1/\lambda}\delta_L}{e^{1/\lambda}\delta_L + \delta_H}$$

Note that since $\delta_H \geq \delta_L$, we always have $Eu^\lambda(s = h) > 0$. Hence observing an h signal

always leads to buying the product, for both biased and Bayesian consumers. When observing an l signal, it would depend on whether

$$\delta_H - e^{1/\lambda}\delta_L \geq 0$$

For Bayesian consumers ($\delta_H = \delta_L$), this will always be negative, and they always reject if they see an l signal. However, it is possible for the biased consumers to ignore the bad signal and buy regardless, as long as δ_H is large enough compared to δ_L, or when λ is large enough.

To transform this condition as a threshold for λ, let

$$\rho_\delta = \frac{1}{ln(\delta_H) - ln(\delta_L)}$$

And the condition becomes

$$\lambda \geq \rho_\delta$$

Adding in the possibility of the firm's constraint, effectively setting a lower bound for the consumers' marginal cost to $\underline{\lambda}$, we can state that the biased consumer with marginal cost λ will ignore bad signals and buy regardless if

$$max\{\lambda, \underline{\lambda}\} \geq \rho_\delta$$

Contrarily, the Bayesian consumer will always be consistent and follow their signals.

Now we discuss the case where the optimal $q(h)$ is a corner solution. It is important to note that a corner solution is not possible when $f(H) = f(L) = 0.5$, as that would put $q(h) = q(l) = 0.5$ as well. Also note that given our assumption that $f(L) \geq f(H)$, it is actually impossible for

$q(h) = 1$ and $q(l) = 0$, as this would require

$$f(L)e^{1/\lambda} < f(H)$$

And $q(h) = 1$ is only possible when $f(L) < f(H)$. However, regardless of whether $q(h) = 1$ or $q(h) = 0$, the following results don't change.

When $q(h) = 0$, that is when $f(L) - f(H)e^{1/\lambda} \leq 0$ the consumer chooses an uninformative signal structure with only l signals. First, we define

$$\rho_f = \frac{1}{ln(f(L)) - ln(f(H))}$$

And consumers will acquire uninformative signals when

$$f(L) - f(H)e^{1/\lambda} \leq 0 \iff \lambda \geq \rho_f$$

While it doesn't make a difference for Bayesian individuals if they actually acquire a signal or not, as it would be uninformative, it does matter for the biased individual, as the posteriors are different from the prior due to the bias:

$$g^{\delta}(L|l) = \frac{f(L)\pi^*(l|L)\delta_L}{f(H)\pi^*(l|H)\delta_H + f(L)\pi^*(l|L)\delta_L} = \frac{f(L)\delta_L}{f(H)\delta_H + f(L)\delta_L} = 1 - g^{\delta}(H|l)$$

The expected utility for buying is then:

$$Eu^{\lambda}(s = l) = g^{\delta}(H|l) - g^{\delta}(L|l) = \frac{f(H)\delta_H - f(L)\delta_L}{f(H)\delta_H + f(L)\delta_L}$$

This is positive if and only if

$$f(H)\delta_H - f(L)\delta_L \geq 0 \iff \rho_f \geq \rho_\delta$$

Note that for the Bayesian individual who has $\delta_H = \delta_L$, this would only be possible for the specific scenario where $f(H) = f(L) = 0.5$. However, that possibility has already been ruled out.

Hence, consumers will acquire uninformative signals when $max\{\lambda, \underline{\lambda}\} \geq \rho_f$. In that case, Bayesian individuals will not buy the product, and biased individuals will buy if and only if $\rho_f \geq \rho_\delta$. □

Proof of Proposition 10: Optimal constraint of the firm.

Choosing optimal κ is equivalent to choosing optimal $\underline{\lambda}$ since $I(\pi^{\underline{\lambda}}) = \kappa$ has a unique solution for each κ. To analyze the firm's optimal choice of $\underline{\lambda}$, we will divide the discussion into two cases depending on the relation between ρ_f and ρ_δ. The firm's profit is calculated as the expected sales, which is the unconditional probability of selling the item summed over all consumers.

Case 1: $\rho_f < \rho_\delta$

Under this assumption, we have $f(H)\delta_H < f(L)\delta_L$. Figure 3.1a illustrates this case. Furthermore, for any individual that acquires informative signals, that is with $\lambda < \rho_f$, we automatically have $\lambda < \rho_\delta$. And whenever $\lambda \geq \rho_f$, the consumer will acquire uninformative signals. Combined with $\rho_f < \rho_\delta$, the consumer will never buy. Therefore, the firm has two potential strategies:

Strategy 1.1: $0 < \underline{\lambda} < \rho_f$

By setting $0 < \underline{\lambda} < \rho_f$, the constrained consumers, with $\lambda < \underline{\lambda}$, will act as if they have a

marginal cost of $\underline{\lambda}$. By the results from Lemmas 1 and 2, and since $\underline{\lambda} < \rho_f < \rho_\delta$, these constrained consumers, whether biased or not, still acquire informative signals and buy the product only when they see an h signal.

We take the difference in expected sales between the firm setting $0 < \underline{\lambda} < \rho_f$ and imposing no constraint:

$$\begin{aligned} E[\Delta\Pi](\underline{\lambda} < \rho_f | \rho_\delta > \rho_f) &= \int_{\lambda<\underline{\lambda}} (q^{\underline{\lambda}}(h) - q^{\lambda}(h)) dG(\lambda) \\ &= \int_{\lambda<\underline{\lambda}} (\frac{f(L) - f(H)e^{1/\underline{\lambda}}}{1 - e^{1/\underline{\lambda}}} - \frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}}) dG(\lambda) < 0 \end{aligned}$$

The change only comes from people who are constrained by the $\underline{\lambda}$, hence the integral is only calculated for those $\lambda \leq \underline{\lambda}$. The first term is the marginal probability of receiving $s = h$ with the constraint, which all of the consumers will follow. The second term is the marginal probability of $s = h$ without the constraint. The difference is negative as $q(h) = \frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}}$ decreases in λ.

Strategy 1.2: $\underline{\lambda} \geq \rho_f$

As illustrated in Figure 3.1a, pushing consumers' marginal cost higher than ρ_f will cause them to obtain uninformative signals and not buy the product. This is reflected in a negative change in profit:

$$\begin{aligned} E[\Delta\Pi](\underline{\lambda} \geq \rho_f | \rho_\delta > \rho_f) &= \int_{\lambda<\rho_f} (q^{\underline{\lambda}}(h) - q^{\lambda}(h)) dG(\lambda) \\ &= \int_{\lambda<\rho_f} (0 - \frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}}) dG(\lambda) < 0 \end{aligned}$$

The upper limit for the integral is ρ_f instead, because any consumers with $\underline{\lambda} \geq \lambda \geq \rho_f$ do not buy the product with or without the firm imposing the constraint.

Case 2: $\rho_f \geq \rho_\delta$

Under this assumption, we have $f(H)\delta_H \geq f(L)\delta_L$. This case is illustrated in Figure 3.1b. Furthermore, by Lemmas 1 and 2, any biased individual with $\rho_\delta \leq \lambda < \rho_f$, will always buy the product. Any individual with $\rho_\delta \leq \rho_f \leq \lambda$, will acquire uninformative signals and buy the product regardless if she is biased, and reject the product if she is Bayesian. The firm's strategy can be discussed under three cases.

Strategy 2.1: $0 < \underline{\lambda} < \rho_\delta$

We can refer to Figure 3.1b to see that the constrained consumers are those who follow their signals without the constraint, and will continue to follow their signals with the constraint, as $\underline{\lambda} < \rho_\delta < \rho_f$. Similar to the previous case, the change in profit is negative.

$$\begin{aligned} E[\Delta\Pi](\underline{\lambda} < \rho_\delta | \rho_\delta \leq \rho_f) &= \int_{\lambda<\underline{\lambda}} (q^{\underline{\lambda}}(h) - q^{\lambda}(h)) dG(\lambda) \\ &= \int_{\lambda<\underline{\lambda}} (\frac{f(L) - f(H)e^{1/\underline{\lambda}}}{1 - e^{1/\underline{\lambda}}} - \frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}}) dG(\lambda) < 0 \end{aligned}$$

Strategy 2.2: $\rho_\delta \leq \underline{\lambda} < \rho_f$

For this case, any biased constrained consumer will ignore bad signals. If the consumer originally had $\lambda < \rho_\delta$, then she went from buying the product only when seeing an h signal to always buying the product regardless of signal. If the biased consumer originally had $\lambda \geq \rho_\delta$, their behavior doesn't change. On the other hand, constrained Bayesian consumers will continue to follow their informative signals consistently. With proportion p of the population being biased, the change in profit is as follows.

$$E[\Delta\Pi](\rho_\delta \leq \underline{\lambda} < \rho_f | \rho_\delta \leq \rho_f) =$$

$$p\int_{\lambda<\rho_\delta}(1-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}})dG(\lambda)+(1-p)\int_{\lambda<\underline{\lambda}}(\frac{f(L)-f(H)e^{1/\underline{\lambda}}}{1-e^{1/\underline{\lambda}}}-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}})dG(\lambda)$$

This can be positive as the first integral is positive.

Strategy 2.3: $\rho_\delta \leq \rho_f \leq \underline{\lambda}$

Any biased constrained consumer will acquire uninformative signals and always buy the product. However, any biased consumer with $\lambda \geq \rho_\delta$ always buys the product without the firm's constraint anyways. Meanwhile, constrained Bayesian consumers will not buy the product. This gives the following change in profit.

$$E[\Delta\Pi](\rho_f \leq \underline{\lambda}|\rho_\delta \leq \rho_f) =$$

$$p\int_{\lambda<\rho_\delta}(1-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}})dG(\lambda)+(1-p)\int_{\lambda<\underline{\lambda}}(0-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}})dG(\lambda)$$

This can be positive as the first integral is positive. However, this strategy is dominated by the best $\underline{\lambda}$ analyzed in Case 2.2. This is because the domain over which the second integral is taken is larger, and additionally,

$$0-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}} < \frac{f(L)-f(H)e^{1/\underline{\lambda}}}{1-e^{1/\underline{\lambda}}}-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}} < 0$$

Focusing on Strategy 2.2, we observe that $E[\Delta\Pi]$ is decreasing in $\underline{\lambda}$ due to the second term:

$$(1-p)\int_{\lambda<\underline{\lambda}}(\frac{f(L)-f(H)e^{1/\underline{\lambda}}}{1-e^{1/\underline{\lambda}}}-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}})$$

As λ increases, the domain of the integral is larger, and the expression inside is smaller (more

negative). Therefore, under Strategy 2.2, the highest change in expected profit is achieved by $\underline{\lambda}^* = \rho_\delta$.

For this $\underline{\lambda}^*$ to be optimal, we need $E[\Delta\Pi](\underline{\lambda}^* = \rho_\delta | \rho_\delta \leq \rho_f) \geq 0$.

$$p\int_{\lambda<\rho_\delta}(1-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}})dG(\lambda)+(1-p)\int_{\lambda<\rho_\delta}(\frac{f(L)-f(H)e^{1/\rho_\delta}}{1-e^{1/\rho_\delta}}-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}})dG(\lambda)\geq 0$$

Which gives us

$$\int_{\lambda<\rho_\delta}\left[p+(1-p)\frac{f(L)-f(H)e^{1/\rho_\delta}}{1-e^{1/\rho_\delta}}-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}}\right]dG(\lambda)\geq 0 \qquad \text{(C.2)}$$

Note that as $f(H)$ approaches 0.5, the integrand becomes:

$$\lim_{f(H)\to 0.5} p+(1-p)\frac{f(L)-f(H)e^{1/\rho_\delta}}{1-e^{1/\rho_\delta}}-\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}}=0.5p$$

This is positive whenever we have a positive p.

When $p = 0$, the term is unambiguously negative, indicating that it is only profitable to constrain information in the presence of biased consumers.

Note that the condition $\rho_\delta \leq \rho_f$ is equivalent to a lower bound on $f(H)$:

$$\delta_H f(H) \geq \delta_L f(L) \iff f(H) \geq \frac{\delta_L}{\delta_H+\delta_L}$$

When $f(H) \to \frac{\delta_L}{\delta_H+\delta_L}$, we have

$$f(L)-f(H)e^{1/\rho_\delta} \to \frac{\delta_H}{\delta_H+\delta_L}-\frac{\delta_L}{\delta_H+\delta_L}\frac{\delta_H}{\delta_L}=0$$

while for $\lambda < \rho_\delta$,

$$\frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}} \to \frac{\delta_L e^{1/\lambda} - \delta_H}{(e^{1/\lambda} - 1)(\delta_H + \delta_L)} > 0$$

The integrand becomes

$$p - \frac{\delta_L e^{1/\lambda} - \delta_H}{(e^{1/\lambda} - 1)(\delta_H + \delta_L)}$$

Therefore, the integral can be positive or negative at the limit. However, notice that the integrand is monotonically decreasing in $f(H)$, as $\frac{e^{1/\rho_\delta}}{1-e^{1/\rho_\delta}} < \frac{e^{1/\lambda}}{1-e^{1/\lambda}}$. Hence, given $p > 0$, there exists a threshold for $f(H)$ such that the expected change in profit is positive for all $f(H)$ values above the threshold.

For any $p > 0$, if $E[\Delta\Pi](\underline{\lambda}^* = \rho_\delta | \rho_\delta \leq \rho_f) \geq 0$ when $f(H) \to \frac{\delta_L}{\delta_H + \delta_L}$, then the firm finds it profitable to constrain information, setting $\underline{\lambda}^* = \rho_\delta$ whenever $p > 0$ and $\rho_\delta \leq \rho_f$. For any $p > 0$, if $E[\Delta\Pi](\underline{\lambda}^* = \rho_\delta | \rho_\delta \leq \rho_f) < 0$ when $f(H) \to \frac{\delta_L}{\delta_H + \delta_L}$, then there exists an $\hat{f} \in [\frac{\delta_L}{\delta_H + \delta_L}, 0.5)$ such that for $p > 0$ and $f(H) \geq \hat{f}$, the firm finds it profitable to constrain information by setting $\underline{\lambda}^* = \rho_\delta$. □

Proof of Proposition 11: Optimal constraints under competition.

The premise of the following derivations is the assumption that $\rho_f \geq \rho_\delta$ and $p > 0$, as these are necessary for the monopolist to have an incentive to constrain information in the first place.

Denote the incumbent's and the entrant's strategy as $\underline{\lambda}^I$ and $\underline{\lambda}^E$ respectively, and let their profits be $\Pi^I(\underline{\lambda}^I, \underline{\lambda}^E)$ and $\Pi^E(\underline{\lambda}^I, \underline{\lambda}^E)$ respectively. First, we will rule out some strategies. Note that given any $\underline{\lambda}^I$, setting $\underline{\lambda}^E = k \in (0, \rho_\delta)$ is strictly dominated by setting $\underline{\lambda}^E = 0$ for the entrant.

To see this, consider the change in expected profit from switching strategies. If $k > \underline{\lambda}^I$,

$$\Pi^E(\underline{\lambda}^I, 0) - \Pi^E(\underline{\lambda}^I, k) = \int_0^{\underline{\lambda}^I} [g^\lambda(h) - \eta g^k(h)] dG(\lambda) + \eta \int_{\underline{\lambda}^I}^k g^\lambda(h) dG(\lambda)$$

The first integral represents the entrant getting all the consumers with $\lambda \in [0, \underline{\lambda}^I]$, instead of just η of them and with lower purchase probability. The second integral represents attracting η of the consumers with $\lambda \in (\underline{\lambda}^I, k]$, instead of none of them. The change is clearly positive. Alternatively, if $k \leq \underline{\lambda}^I$, then the change becomes

$$\Pi^E(\underline{\lambda}^I, 0) - \Pi^E(\underline{\lambda}^I, k) = \int_0^k [g^\lambda(h) - \eta g^k(h)] dG(\lambda)$$

This represents the entrant attracting all the customers with $\lambda \in [0, k]$, instead of just η of them and with lower purchase probability. Again, this is positive. Furthermore, since the incumbent's profits are symmetric to the entrant's except with $1 - \eta$ instead of η, this argument holds for the incumbent as well. As such, we do not need to consider any strategy within $(0, \rho_\delta)$.

Next, we note that given any strategy by the incumbent, setting $\underline{\lambda}^E = k \in (\rho_\delta, \infty)$ is strictly dominated by setting $\underline{\lambda}^E = \rho_\delta$ for the entrant. To see this, again take the difference from switching strategies. If $k > \rho_\delta > \underline{\lambda}^I$,

$$\Pi^E(\underline{\lambda}^I, \rho_\delta) - \Pi^E(\underline{\lambda}^I, k) = \eta[p \int_{\rho_\delta}^k dG(\lambda) + (1-p) \int_{\rho_\delta}^{min\{k, \rho_f\}} g^\lambda(h) dG(\lambda)]$$

The first integral represents the entrant attracting η of the biased consumers with $\lambda \in [\rho_\delta, k]$, who will always buy. The second integral represents the entrant attracting η of the Bayesian consumers

with $\lambda \in [\rho_\delta, k]$, who will follow their signal recommendations. Alternatively, if $k > \underline{\lambda}^I > \rho_\delta$,

$$\Pi^E(\underline{\lambda}^I, \rho_\delta) - \Pi^E(\underline{\lambda}^I, k) = \eta[p \int_{\underline{\lambda}^I}^{k} dG(\lambda) + (1-p) \int_{\underline{\lambda}^I}^{min\{k,\rho_f\}} g^\lambda(h) dG(\lambda)] + \int_{\rho_\delta}^{\underline{\lambda}^I} p + (1-p) g^\lambda(h) dG(\lambda)$$

The first two integrals represent the types of customers the entrant now shares with the incumbent, while the second integral represents the types of customers the entrant has full market share over due to looser constraints. Lastly, if $\underline{\lambda}^I > k > \rho_\delta$, then

$$\Pi^E(\underline{\lambda}^I, \rho_\delta) - \Pi^E(\underline{\lambda}^I, k) = \int_{\rho_\delta}^{min\{k,\rho_f\}} p + (1-p) g^\lambda(h) dG(\lambda)$$

The entrant is fully attracting all consumers with $\lambda \in [\rho_\delta, k]$. All three of the above scenarios yield positive gains, hence we need not consider any strategy setting $\underline{\lambda}^E > \rho_\delta$. The same goes for the incumbent.

We find the equilibrium where both firms choose from one of two strategies $\{0, \rho_\delta\}$. We begin by examining the entrant's optimal response to an entrant that sets $\underline{\lambda}^I = \rho_\delta$. The profit for the entrant given $\underline{\lambda}^E \in \{0, \rho_\delta\}$ is:

$$\Pi^E(\rho_\delta, 0) = \int_0^{\rho_\delta} \frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}} dG(\lambda) + \eta \left[p \int_{\rho_\delta}^{\infty} dG(\lambda) + (1-p) \int_{\rho_\delta}^{\rho_f} \frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}} dG(\lambda) \right]$$

$$\Pi^E(\rho_\delta, \rho_\delta) = \eta p \int_0^{\infty} dG(\lambda) + \eta(1-p) \left[\int_0^{\rho_\delta} \frac{f(L) - f(H)e^{1/\rho_\delta}}{1 - e^{1/\rho_\delta}} dG(\lambda) + \int_{\rho_\delta}^{\rho_f} \frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}} dG(\lambda) \right]$$

We find the conditions for which it is optimal for the entrant to not constrain information acqui-

sition (Inequality (3.10) from the main text.)

$$\Pi^E(\rho_\delta, 0) > \Pi^E(\rho_\delta, \rho_\delta) \iff \int_0^{\rho_\delta} \left[-\eta p + \frac{1 - f - fe^{1/\lambda}}{1 - e^{1/\lambda}} - \eta(1-p)\frac{1 - f - fe^{1/\rho_\delta}}{1 - e^{1/\rho_\delta}} \right] dG(\lambda) > 0$$

Note that when either η or p is 0, the integrand is positive. However, we are not considering $p = 0$, as it would then never be profitable to impose any constraints. When both η and p are 1, the integrand is negative. The integral value is linear and decreasing in both η and p. Denote

$$\beta_0 = \int_0^{\rho_\delta} \frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}} dG(\lambda)$$

$$\beta_1 = \int_0^{\rho_\delta} dG(\lambda)$$

$$\beta_2 = \int_0^{\rho_\delta} \frac{f(L) - f(H)e^{1/\rho_\delta}}{1 - e^{1/\rho_\delta}} dG(\lambda)$$

Observe that $\beta_1 > \beta_0 > \beta_2 > 0$. Then

$$\Pi^E(\rho_\delta, 0) > \Pi^E(\rho_\delta, \rho_\delta) \iff \beta_0 - \beta_1 \eta p - \beta_2 \eta (1-p) > 0$$

The entrant's best response is not constraining information if $\eta < \frac{\beta_0}{(\beta_1 - \beta_2)p + \beta_2}$. That is, for any point (η, p) to the bottom left of the curve $\beta_0 - \beta_1 \eta p - \beta_2 \eta (1-p) = 0$.

Now consider the entrant's best response to $\underline{\lambda}^I = 0$.

$$\Pi^E(0, 0) = \eta \left[\int_0^{\rho_\delta} \frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}} dG(\lambda) + p \int_{\rho_\delta}^{\infty} dG(\lambda) + (1-p) \int_{\rho_\delta}^{\rho_f} \frac{f(L) - f(H)e^{1/\lambda}}{1 - e^{1/\lambda}} dG(\lambda) \right]$$

$$\Pi^E(0,\rho_\delta) = \eta\left[p\int_{\rho_\delta}^{\infty} dG(\lambda) + (1-p)\int_{\rho_\delta}^{\rho_f} \frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}} dG(\lambda)\right]$$

It is quite obvious that $\Pi^E(0,0) > \Pi^E(0,\rho_\delta)$. Therefore, the entrant would best respond with no constraint as well.

Lastly, we check what strategy the incumbent would employ in anticipation of the entrant's best response. When $\underline{\lambda}^E = 0$, by symmetry,

$$\Pi^I(0,0) = \frac{1-\eta}{\eta}\Pi^E(0,0)$$

$$\Pi^I(\rho_\delta,0) = \frac{1-\eta}{\eta}\Pi^E(0,\rho_\delta)$$

Immediately, we see $\Pi^I(0,0) > \Pi^I(\rho_\delta,0)$.

Therefore, if $\eta < \frac{\beta_0}{(\beta_1-\beta_2)p+\beta_2}$, the entrant has dominant strategy $\underline{\lambda}^E = 0$, and the incumbent would preemptively respond with $\underline{\lambda}^I = 0$. If $\eta \geq \frac{\beta_0}{(\beta_1-\beta_2)p+\beta_2} > 0$, the entrant's best response to $\underline{\lambda}^I = \rho_\delta$ would be to match it $\underline{\lambda}^E = \rho_\delta$. Here the incumbent gets to choose the desirable equilibrium between $(\rho_\delta,\rho_\delta)$ and $(0,0)$. By symmetry,

$$\Pi^I(\rho_\delta,\rho_\delta) = \frac{1-\eta}{\eta}\Pi^E(\rho_\delta,\rho_\delta)$$

We check the conditions for the following inequality (which was Inequality 3.11 in the main text.)

$$\Pi^I(0,0) > \Pi^I(\rho_\delta,\rho_\delta) \iff$$
$$\int_0^{\rho_\delta}\left[\frac{f(L)-f(H)e^{1/\lambda}}{1-e^{1/\lambda}} - (1-p)\frac{f(L)-f(H)e^{1/\rho_\delta}}{1-e^{1/\rho_\delta}} - p\right] dG(\lambda) > 0$$

Expressing this in terms of the β's as defined above,

$$\beta_0 - \beta_1 p - \beta_2(1-p) > 0$$

The incumbent prefers $(\rho_\delta, \rho_\delta)$ if this inequality holds, and prefers $(0,0)$ otherwise. This is the same inequality as (C.2), meaning that the incumbent will prefer $(\rho_\delta, \rho_\delta)$ over $(0,0)$ if and only if it prefers to constrain information without the entrant present.

This makes it seem like competition has no impact, but that is not the case. For $(\rho_\delta, \rho_\delta)$ (and for the information constraint under monopoly) to be preferable, we need

$$p \geq \frac{\beta_0 - \beta_2}{\beta_1 - \beta_2} > 0$$

However, (3.10) states that for the entrant to best respond to constraining information by mirroring the incumbent's move, we would need

$$p \geq \frac{\beta_0 - \beta_2\eta}{(\beta_1 - \beta_2)\eta} > \frac{\beta_0 - \beta_2}{\beta_1 - \beta_2}$$

Therefore, competition shrinks the set of eligible parameter values for which information manipulation is profitable. Specifically, for $p \in [\frac{\beta_0-\beta_2}{\beta_1-\beta_2}, \frac{\beta_0-\beta_2\eta}{(\beta_1-\beta_2)\eta})$, the only equilibrium is $(0,0)$. While the incumbent would prefer otherwise, the entrant would always choose not to constrain information.

In sum, given $p > 0$ and $\rho_f > \rho_\delta$, it is profitable to constrain information, with or without competition, if $p \geq \frac{\beta_0-\beta_2}{\beta_1-\beta_2}$. With competition from an entrant, the equilibrium is $(0,0)$ when $p < \frac{\beta_0-\beta_2\eta}{(\beta_1-\beta_2)\eta}$, and $(\rho_\delta, \rho_\delta)$ otherwise. This means that for $p \in [\frac{\beta_0-\beta_2}{\beta_1-\beta_2}, \frac{\beta_0-\beta_2\eta}{(\beta_1-\beta_2)\eta})$, the incumbent is

forced to switch from information constraint to no constraint.

□

Proof of Proposition 12: Generalized Updating Rule

Unless $f(H) = 0$ or $\lambda = 0$, the individual will not obtain perfectly accurate signals such that they have posteriors at 0 or 1. Due to the second assumption, we always have

$$Eu(b|s = h) = F_H(f, s = h) - F_L(f, s = h) > g^B(H|s = h) - g^B(L|s = h) \geq 0$$

We just need to check the behavior of biased individuals when they receive l signals.

The third assumption is a sufficient condition to guarantee that expected utility of buying after receiving an l signal is non-decreasing in λ. To see this, note that with two states

$$Eu(b|s = l) = F_H(f, s = l) - F_L(f, s = l) = 2F_H(f, s = l) - 1$$

And

$$\frac{F_H(f, s = l)}{F_L(f, s = l)} = \frac{F_H(f, s = l)}{1 - F_H(f, s = l)}$$

Both increasing in $F_H(f, s = l)$. Hence $Eu(b|s = l)$ would be non-decreasing in $\frac{\pi(l|H)}{\pi(l|L)}$. Furthermore, the optimal signal structure satisfies

$$\frac{\pi^*(l|H)}{\pi^*(l|L)} = \frac{\frac{q^*(l)}{f(H)(1+e^{1/\lambda})}}{\frac{q^*(l)e^{1/\lambda}}{f(L)(1+e^{1/\lambda})}} = \frac{f(L)}{f(H)e^{1/\lambda}}$$

which is increasing in λ. For high enough λ, the individual acquires pure noise, with $q(h) = 0$. In this region, $\frac{\pi^*(l|H)}{\pi^*(l|L)} = 1$, and the posterior is a constant function of λ. Therefore, $Eu(b|s = l)$

is non-decreasing in λ. Furthermore, since the posterior is continuous in signal probabilities, $Eu(b|s = l)$ is also continuous in λ.

Note that when we increase λ such that $\lambda \to \infty$, signals become completely uninformative at the limit. When $f(H) = 0.5$ and signals are completely uninformative, we would have $g^B(H|s) = 0.5$. Thus, $F_H(f(H) = 0.5, s) > 0.5$, and consequently, $Eu(b|s = l) > 0$ in this scenario. While the optimal signal structure is never completely uninformative when $f(H) = 0.5$ and $\lambda < \infty$, it still allows there to exist $f(H) < 0.5$ and finite λ such that $Eu(b|s = l) > 0$ by the continuity of the updating rule. However, it is crucial to point out that, with optimal information acquisition, the exogenous prior may be within a region such that the expected utility may be negative for all λ. Hence, the following discussion.

If the updating rule is such that $Eu(b|s = l) > 0$, or equivalently $F_H(f, s) > 0.5$, given any f and λ, then the problem becomes trivial, and the firm need not do anything. We focus on the scenario where the posterior over state H could be below 0.5.

Given this, we can solve the following for λ:

$$Eu(b|s = l) = 0$$

Due to continuity, and the expression being monotone in λ, we have the set of solutions $\{\lambda|Eu(b|s = l) = 0\}$ being a closed interval (a single point if the expected utility is strictly increasing in λ). Take the minimum of the set and denote it as $\rho^*(F)$.

Recall that the threshold (for λ) for the individual to seek uninformative signals is ρ_f. If $\rho_f < \rho^*(F)$, then the biased individuals with $\lambda < \rho_f$ are acquiring informative signals, while following the action recommended by the signals as well. That is, they don't buy if they see l

signals. For $\lambda \geq \rho_f$, as λ increases, the posterior does not change, as the individual is starting to acquire pure noise. Therefore, the condition $\rho_f < \rho^*(F)$ exactly indicates the region for f where it is impossible to obtain $Eu(b|s = l) \geq 0$. If $\rho_f \geq \rho^*(F)$, then for individuals with $\lambda \in [\rho^*(F), \rho_f]$, we have $Eu(b|s = l) \geq 0$ as the expected utility is non-decreasing in λ. Furthermore, since posteriors stay constant when λ increases beyond ρ_f (due to individuals acquiring pure noise from that point onward), we have $Eu(b|s = l) \geq 0$ for all $\lambda \geq \rho^*(F)$. (Note that the case where the posterior $F_H(f, s = l)$ is always higher that 0.5 can be included as the special case of $\rho^*(F) = 0$.)

Since the set $\mathcal{F} = \{f(H) \in [0, 0.5] | \rho_f \geq \rho^*(F)\}$ must include the neighborhood around $f(H) = 0.5$, the firm's strategy carries out the same as in Proposition 10. We know that the firm either chooses not to constrain information setting $\underline{\lambda} = 0$, or constrain information by setting $\underline{\lambda} = \rho^*(F)$. There exists $\underline{f} \in \mathcal{F}$ such that for all $f(H) \in \mathcal{F}$ satisfying $f(H) \geq \underline{f}$, the firm finds it profitable to set $\underline{\lambda} = \rho^*(F)$. This eligible set of priors must be non-empty, since $f(H) = 0.5$ must be in it.

□

Bibliography

[1] Deanna Kuhn. Children and adults as intuitive scientists. *Psychological review*, 96(4):674, 1989.

[2] Thomas Gilovich. Biased evaluation and persistence in gambling. *Journal of personality and social psychology*, 44(6):1110, 1983.

[3] John C Easterwood and Stacey R Nutt. Inefficiency in analysts' earnings forecasts: Systematic misreaction or systematic optimism? *The Journal of Finance*, 54(5):1777–1797, 1999.

[4] Charles S Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.

[5] Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.

[6] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.

[7] Matthew Rabin and Joel L. Schrag. First Impressions Matter: A Model of Confirmatory Bias. *The Quarterly Journal of Economics*, 114(1):37–82, 1999.

[8] Larry G Epstein. An axiomatic model of non-bayesian updating. *The Review of Economic Studies*, 73(2):413–436, 2006.

[9] Elchin Suleymanov. Robust maximum likelihood updating. *Working paper*, 2018.

[10] Adam Dominiak, Matthew Kovach, and Gerelt Tserenjigmid. Inertial updating. *working paper*, 2023.

[11] David Eil and Justin M. Rao. The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–38, May 2011.

[12] Zhenxun Liu. Dual Elicitation of Motivated and Unmotivated Confirmation Bias. *working paper, University of Maryland*, 2023.

[13] Chen Zhao. Representativeness and similarity. *Working paper*, 2018.

[14] Chen Zhao. Pseudo-bayesian updating. *Theoretical Economics*, 17(1):253–289, 2022.

[15] Martin W Cripps. Divisible updating. *Working paper*, 2018.

[16] Kyle P Chauvin. Euclidean properties of bayesian updating. *Working paper*, 2020.

[17] Alexander M Jakobsen. Coarse bayesian updating. *Working paper*, 2021.

[18] Matthew Kovach. Twisting the truth: Foundations of wishful thinking. *Theoretical Economics*, 15(3):989–1022, 2020.

[19] Francis J Anscombe, Robert J Aumann, et al. A definition of subjective probability. *Annals of mathematical statistics*, 34(1):199–205, 1963.

[20] Guy Mayraz. Wishful thinking. *Working paper*, 2011.

[21] Michael Thaler. The fake news effect: Experimentally identifying motivated reasoning using trust in news. *Forthcoming: American Economic Journal: Microeconomics*, 2024.

[22] Cameron R Peterson and Wesley M DuCharme. A primacy effect in subjective probability revision. *Journal of Experimental Psychology*, 73(1):61, 1967.

[23] Nathan Nunn and Raul Sanchez de la Sierra. Why being wrong can be right: Magical warfare technologies and the persistence of false beliefs. *American Economic Review*, 107(5):582–87, May 2017.

[24] Andrew Caplin and John V Leahy. Wishful thinking. *Working paper*, 2019.

[25] Guy Mayraz. Priors and Desires. (dp1047), March 2011.

[26] Pietro Ortoleva. Modeling the change of paradigm: Non-bayesian reactions to unexpected news. *American Economic Review*, 102(6):2410–2436, 2012.

[27] Markus K. Brunnermeier and Jonathan A. Parker. Optimal expectations. *American Economic Review*, 95(4):1092–1118, September 2005.

[28] P. C. Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140, 1960.

[29] Charles M. Wharton, Patricia W. Cheng, and Thomas D. Wickens. Hypothesis-testing strategies: Why two goals are better than one. *The Quarterly Journal of Experimental Psychology Section A*, 46(4):743–758, 1993.

[30] Clifford Mynatt, Michael Doherty, and Ryan Tweney. Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology - QUART J EXP PSYCHOL*, 29:85–95, 02 1977.

[31] John M Darley and Paget H Gross. A hypothesis-confirming bias in labeling effects. *Journal of personality and social psychology*, 44(1):20, 1983.

[32] Ziva Kunda. The case for motivated reasoning. *Psychological bulletin*, 108(3):480, 1990.

[33] Gary Charness and Chetan Dave. Confirmation bias with motivated beliefs. *Games and Economic Behavior*, 104:1–23, 2017.

[34] Alexander Coutts. Good news and bad news are still news: experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395, 2019.

[35] Zachary Grossman and David Owens. An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, 84(2):510–524, 2012.

[36] Markus M. Möbius, Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. Managing self-confidence: Theory and experimental evidence. *Management Science*, 68(11):7793–7817, 2022.

[37] Jan Engelmann, Maël Lebreton, Peter Schwardmann, Joel J van der Weele, and Li-Ang Chang. Anticipatory anxiety and wishful thinking. *Forthcoming: American Economic Review*, 2024.

[38] Kai Barron. Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics*, 24(1):31–58, 2021.

[39] Elliot Lipnowski and Laurent Mathevet. Disclosure to a psychological audience. *American Economic Journal: Microeconomics*, 10(4):67–93, November 2018.

[40] John Geanakoplos, David Pearce, and Ennio Stacchetti. Psychological games and sequential rationality. *Games and Economic Behavior*, 1(1):60–79, 1989.

[41] Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–98, January 2015.

[42] Edward E Jones, Leslie Rock, Kelly G Shaver, George R Goethals, and Lawrence M Ward. Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology*, 10(4):317, 1968.

[43] C Michael Troutman and James Shanteau. Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance*, 19(1):43–55, 1977.

[44] Mira Frick, Ryota Iijima, and Yuhta Ishii. Belief Convergence under Misspecified Learning: A Martingale Approach. *The Review of Economic Studies*, 90(2):781–814, 07 2022.

[45] Seda Ertaç. Does self-relevance affect information processing? experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior Organization*, 80(3):532–545, 2011.

[46] Camelia M. Kuhnen. Asymmetric learning from financial information. *The Journal of Finance*, 70(5):2029–2062, 2015.

[47] Zhenxun Liu. A Model of Confirmation Bias. *working paper, University of Maryland*, 2023.

[48] David M. Grether. Bayes Rule as a Descriptive Model: The Representativeness Heuristic. *The Quarterly Journal of Economics*, 95(3):537–557, 1980.

[49] Birt L Duncan. Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of blacks. *Journal of personality and social psychology*, 34(4):590, 1976.

[50] Daniel F Stone and Daniel H Wood. Cognitive dissonance, motivated reasoning, and confirmation bias: applications in industrial organization. In *Handbook of behavioral industrial organization*, pages 114–137. Edward Elgar Publishing, 2018.

[51] Daniel L. Chen, Martin Schonger, and Chris Wickens. otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016.

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

[53] Charles Holt and Angela M. Smith. Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes. *American Economic Journal: Microeconomics*, 8(1):110–39, 2016.

[54] David Danz, Lise Vesterlund, and Alistair J. Wilson. Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–83, September 2022.

[55] Ben Greiner. Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125, 2015.

[56] Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690, 2003. Swiss National Bank/Study Center Gerzensee Conference on Monetary Policy under Incomplete Information.

[57] Christopher A. Sims. Rational inattention: Beyond the linear-quadratic case. *American Economic Review*, 96(2):158–163, May 2006.

[58] Doron Ravid. Ultimatum bargaining with rational inattention. *American Economic Review*, 110(9):2948–63, September 2020.

[59] Daniel Martin. Strategic pricing with rational inattention to quality. *Games and Economic Behavior*, 104:131–145, 2017.

[60] Tamer Boyacı and Yalçın Akçay. Pricing when customers have limited attention. *Management Science*, 64(7):2995–3014, 2018.

[61] Bartosz Mackowiak and Mirko Wiederholt. Optimal sticky prices under rational inattention. *American Economic Review*, 99(3):769–803, June 2009.

[62] Michael Woodford. Information-constrained state-dependent pricing. *Journal of Monetary Economics*, 56:S100–S124, 2009. Supplement issue: December 12-13, 2008 Research Conference on 'Monetary Policy under Imperfect Information' Sponsored by the Swiss National Bank (http://www.snb.ch) and Study Center Gerzensee (www.szgerzensee.ch).

[63] Filip Matějka and Alisdair McKay. Simple market equilibria with rationally inattentive consumers. *American Economic Review*, 102(3):24–29, May 2012.

[64] Filip Matějka. Rationally Inattentive Seller: Sales and Discrete Pricing. *The Review of Economic Studies*, 83(3):1125–1155, 10 2015.

[65] Jakub Steiner, Colin Stewart, and Filip Matějka. Rational inattention dynamics: Inertia and delay in decision-making. *Econometrica*, 85(2):521–553, 2017.

[66] Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, July 2015.

[67] Henrique de Oliveira, Tommaso Denti, Maximilian Mihm, and Kemal Ozbek. Rationally inattentive preferences and hidden information costs. *Theoretical Economics*, 12(2):621–654, 2017.

[68] Colin Camerer. Bounded Rationality in Individual Decision Making. *Experimental Economics*, 1:163–183, 1998.

[69] Daniel J. Benjamin. Chapter 2 - errors in probabilistic reasoning and judgment biases. In B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, editors, *Handbook of Behavioral Economics - Foundations and Applications 2*, volume 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, pages 69–186. North-Holland, 2019.

[70] Ward Edwards. *Conservatism in human information processing*, page 359–369. Cambridge University Press, 1982.

[71] Matthew Gentzkow and Emir Kamenica. Costly persuasion. *American Economic Review*, 104(5):457–62, May 2014.

[72] Geoffroy de Clippel and Xu Zhang. Non-bayesian persuasion. *Journal of Political Economy*, 130(10):2594–2642, 2022.

[73] Dong Wei. Persuasion under costly learning. *Journal of Mathematical Economics*, 94(C), 2021.

[74] Stefano DellaVigna and Ulrike Malmendier. Contract Design and Self-Control: Theory and Evidence*. *The Quarterly Journal of Economics*, 119(2):353–402, 05 2004.

[75] Xavier Gabaix and David Laibson. Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets*. *The Quarterly Journal of Economics*, 121(2):505–540, 05 2006.

[76] Johan E Korteling and Alexander Toet. Cognitive biases. In Sergio Della Sala, editor, *Encyclopedia of behavioral neuroscience*, volume 3, pages 610–619. Elsevier Science Amsterdam-Edinburgh, 2021.