

Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
<http://economics.sas.upenn.edu/pier>

PIER Working Paper 12-027

“Belief Formation, Second Version”

by

Olivier Compte and Andrew Postlewaite

<http://ssrn.com/abstract=2101692>

Belief Formation[‡]

Olivier Compte
PSE, Paris

Andrew Postlewaite
University of Pennsylvania

March 2009; July 2012

*Compte: Paris School of Economics, 48 Bd Jourdan, 75014 Paris (e-mail: compte@enpc.fr); Postlewaite: Department of Economics, 3718 Locust Walk, University of Pennsylvania, Philadelphia, PA 19104-6207 (e-mail: apostlew@econ.upenn.edu). A large part of this work was done while Postlewaite visited the Paris School of Economics; their hospitality is gratefully acknowledged. We thank Eddie Dekel, David Dillenberger, Mehmet Ekmekci, Kfir Eliaz, Jeff Ely, Yilmaz Kocer, Yuval Salant and Alvaro Sandroni for helpful discussions, as well as the participants of numerous seminars at which the paper was presented. The authors thank the Gould Foundation for financial support, and Postlewaite thanks the National Science Foundation for financial support.

[‡]Earlier versions of this paper circulated under the titles “Mental Processes and Decision Making” and “Plausible Decision Making with Information Overload”.

Abstract

Consider an agent who is unsure of the state of the world and faces computational bounds on mental processing. The agent receives a sequence of signals imperfectly correlated with the true state that he will use to take a single decision. The agent is assumed to have a finite number of “states of mind” that quantify his beliefs about the relative likelihood of the states, and uses the signals he receives to move from one state to another. At a random stopping time, the agent will be called upon to make a decision based solely on his mental state at that time. We show that under quite general conditions it is optimal that the agent ignore signals that are not very informative, that is, signals for which the likelihood of the states is nearly equal. This model provides a possible explanation of systematic inference mistakes people may make.

It ain't so much the things we don't know that get us into trouble.
It's the things that we know that just ain't so.
Artemus Ward

1 Introduction

The past half century has seen the integration of uncertainty into a wide variety of economic models which has led to increasingly sophisticated analysis of the behavior of agents facing uncertainty. The incorporation of uncertainty into an agent's decision making typically begins with the assumption that the agent has a probability distribution over the possible outcomes that would result from decisions she might take. Savage (1954) is often given as justification for such modeling: when an agent's preferences over the random outcomes stemming from decisions the agent might take satisfy a set of seemingly plausible assumptions, the agent's choices will be as though the agent had a probabilistic belief over some underlying states and maximized expected utility. This "as if" perspective has been extraordinarily useful in shaping our models, and although Savage (1954) is not necessarily meant to be a description of the way people make decisions, nor a suggestion that agents actually form beliefs, it is standard to view expected utility maximization as a description of decision making.

This paper starts with two observations. First, the beliefs that we hold often seem to be relatively crude: we find ourselves unable to form precise probability estimates.¹ Second, beliefs sometimes seem questionable, at odds with either common sense or even scientific evidence. Consider the following example.

Robert is convinced that he has ESP (extrasensory perception) and offers the following statement to support this belief: "I was thinking of my mother last week and she called right after that." Robert is not alone in holding such beliefs. In the U.S. more people believe in ESP than in evolution, and there are twenty times as many astrologers as there are astronomers.² Readers who don't believe in ESP might dismiss Robert and other believers as under-educated anomalies, but there are sufficiently many similar examples to give pause. Nurses who work in maternity wards believe (incorrectly) that more babies are born when the moon is full³, and it is widely believed that infertile couples who adopt a child are subsequently more likely to conceive than similar couples who did not adopt (again, incorrectly).⁴

We might simply decide that people that hold such beliefs are stupid or gullible, at the risk of finding ourselves so described for some of our own beliefs.⁵

¹In the words of Savage (1954, p58):

"The postulate of personal probability imply that I can determine, to any degree of accuracy whatsoever, the probability (for me) that the next president will be a Democrat. Now it is manifest that I cannot determine that number with great accuracy, but only roughly."

²See Gilovich (1991), page 2.

³See G. O. Abell and B. Greenspan (1979).

⁴See E. J. Lamb and S. Leurgans (1979).

⁵There are numerous examples of similarly biased beliefs people hold. Research has demon-

Whether or not we are so inclined, many economic models have at their core a decision-making module, and those models must somehow take account of agents' beliefs, however unsound we may think them.

Our interest in the widespread belief in ESP goes beyond the instrumental concern for constructing accurate decision making modules for our models. The deeper question is why people hold such questionable beliefs? The simple (simplistic?) response that a large number of people are stupid is difficult to accept given the powerful intellectual tools that evolution has provided us in many domains. How is it that evolution has generated a brain that can scan the symbols on a page of paper and determine which subway connects to which bus that will systematically get an individual to work on time, and yet believe in ESP?

Our aim in this paper is to reconcile the systematic mistakes we observe in the inferences people draw from their experiences with evolutionary forces that systematically reward good decisions. We will lay out a model of how an individual processes sequences of informative signals that (a) is optimal, and (b) leads to incorrect beliefs such as Robert's. The reconciliation is possible because of computational bounds we place on mental processing. Roughly speaking, our restrictions on mental processing preclude an agent from recalling every signal he receives perfectly: he must rely on some sort of summary statistic (or belief state) that captures as well as possible the information content of all the signals that he has seen. We assume that the agent has a limited number of belief states. We also assume that he does not have a distinct mental process for each problem he might face; hence a process may do well for "typical" problems, but less well for "unusual" problems.⁶

The restrictions we impose are consistent with the view that beliefs may be relatively crude. Given the restrictions agents face in our model, they optimally ignore signals that are very uninformative. Robert's mistakes will arise naturally under these conditions.

Robert's experience of his mother calling right after he thought of her is quite strong evidence in support of his theory that he has ESP. His problem lies in his not having taken into account the number of times his mother called when he *hadn't* thought of her. Such an event may have moved Robert's posterior belief that he had ESP only slightly, but the accumulation of such small adjustments would likely have overwhelmed the small number of instances which seem important. Our primary point is that the mental processing property that we suggest leads Robert to conclude that he has ESP – ignoring signals that by

strated that people frequently estimate the connection between two events such as cloud seeding and rain mainly by the number of positive-confirming events, that is, where cloud seeding is followed by rain. Cases of cloud seeding and no rain and rain without cloud seeding tend to be ignored (Jenkins and Ward (1965) and Ward and Jenkins (1965).)

⁶Early papers that investigated a decision maker who uses a single decision protocol for a number of similar but not identical problems include Baumol and Quandt (1964). A more systematic modelling of the idea can be found in Rosenthal (1993), where it is assumed that an agent will choose among costly rules of thumb that he will employ in the set of games they face. Lipman (1995) provides a very nice review of the literature on modelling bounds on rationality resulting from limits on agents' ability to process information.

themselves have little information – will, in fact, be optimal when “designing” a mental process that must be applied to large sets of problems when there are computational bounds.

We lay out our model of mental processes in the next section. The basic model is essentially that analyzed by Wilson (2004) and by Cover and Hellman (1970), but our interest differs from that of those authors. In those papers it is assumed that an agent has bounded memory, captured by a set of memory states. Agents receive a sequence of signals that are informative about which of two possible states of Nature is the true state. The question posed in those papers is how the agent can optimally use the signals to move among the finite set of memory states, knowing that at some random time he will be called upon to make a decision, and his memory state at that time is all the information about the signals he has received that can be used.

Cover and Hellman and Wilson characterize the optimal way to transit among states of mind as additional signals arrive when the expected number of signals the agent receives before making a decision goes to infinity. Our interest differs. Our point of view is that an agent’s mental system – the set of states of mind and the transition function – has evolved as an optimal mental system for a class of problems rather than being designed for a single specific problem or a specific number of signals to be received.⁷ Our interest is then in understanding the type of bias that a given mental system exhibits across the different problems that it confronts.

Biases arise when agents tend to believe in a theory whether or not it is actually correct. Given our restriction (number of belief states, same mental process across problems), that biases arise is not surprising: the optimal process has to balance favorable and unfavorable evidence in a way that reflects the variety of signals the agent receives. If a mental system does this balancing well for some problems, there are other problems for which it will not do well. Our contribution is in explaining the nature of the biases that are likely to arise, showing first why we expect agents to ignore weak evidence, and then demonstrating that if indeed weak evidence is ignored, biases are more likely to arise in favor of theories that generate occasional strong evidence for, and (very) frequent but weak evidence against.

We lay out our basic belief formation model in Section 2, compare different mental systems in section 3 and discuss the implications of our main theorems in section 4. We discuss our analysis in section 5.

⁷When there is a limited number of belief states, optimal design requires balancing favorable and unfavorable evidence in a way that reflects the distribution over the evidence that is likely to arise, and the likely number of signals that will be received before a decision is taken. While we find it reasonable that an agent might extract relatively accurate information from a specific observation, it is implausible that they would have detailed knowledge of the distribution over other possible observations and adjust transitions to that presumed knowledge. Indeed, at the time one signal is processed, one may not even know the number of signals to be processed.

This difficulty would not arise with a continuum of states. One could define the state as the Bayesian posterior and use Bayesian updating to define transitions across states. The current state would then incorporate all relevant information about past signals, and transitions would be independent of the distribution over signals, or the number of signals received.

2 The model

Decision problem. There are two states, $\theta = 1, 2$. The true state is $\theta = 1$ with probability π^0 . An agent receives a sequence of signals imperfectly correlated with the true state, that he will use to take a single decision. The decision is a choice between two alternatives, $a \in \{1, 2\}$. To fix ideas, we assume the following payoff matrix, where $g(a, \theta)$ is the payoff to the agent when he takes action a in state θ :

$$\begin{array}{cc} g(a, \theta) & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 1 \end{array}$$

There are costs c_1 and c_2 associated with decisions 1 and 2 respectively. Let $c = (c_1, c_2)$ denote the cost profile, and $u(a, \theta, c)$ the utility associated with each decision a when the state is θ and cost profile is c . We assume that the utility function takes the form

$$u(a, \theta, c) = g(a, \theta) - c_a.$$

The cost c is assumed to be drawn from a distribution with full support on $[0, 1] \times [0, 1]$. The cost vector c is known to the agent prior to the decision. It is optimal to choose $a = 1$ when the agent's belief $\pi \geq \frac{1+c_1-c_2}{2}$. In what follows, we let $v(\pi)$ denote the payoff the agent derives from optimal decision making when π is his belief that the true state is $\theta = 1$. We have:

$$v(\pi) = E_{c_1, c_2} \max\{\pi - c_1, 1 - \pi - c_2\}.$$

It is straightforward to show that v is strictly convex. For example, if costs are uniformly distributed, $v(\pi) = v(1 - \pi)$ and a calculation shows that for $\pi \geq 1/2$, $v(\pi) = \frac{1}{6} + \frac{4}{3}(\pi - \frac{1}{2})^2(2 - \pi)$ (see figure 1).

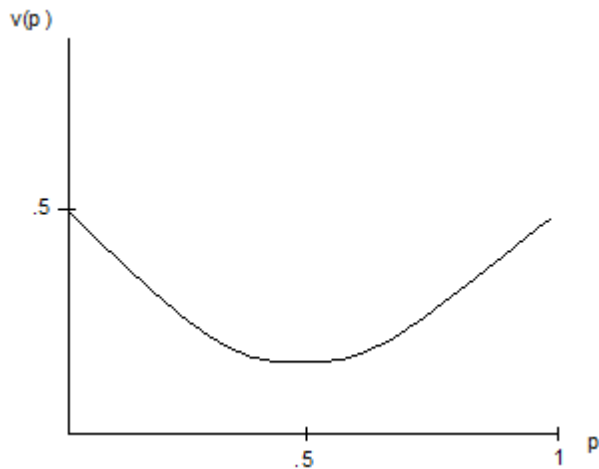


Figure 1: $v(\pi)$ when costs are distributed uniformly

The signals received. Signals are drawn independently, conditional on the true state θ , from the same distribution with density $f(\cdot | \theta)$, assumed to be positive and smooth on its support. When signal x arises, there is a state $\theta(x)$ that has highest likelihood, namely:

$$\theta(x) = \arg \max_{\theta} f(x | \theta)$$

It will be convenient to denote by $l(x)$ the likelihood ratio defined by:

$$l(x) = \frac{f(x | \theta = \theta(x))}{f(x | \theta \neq \theta(x))}.$$

The state $\theta(x)$ is the state for which signal x provides support, and the likelihood ratio $l(x)$ provides a measure of the strength of the evidence in favor of $\theta(x)$. We assume that the set of signals x that are not informative (i.e. $l(x) = 1$) has measure 0.

We assume that signals are received over time, at dates $t = 0, 1, \dots$, and that the decision must be taken at some random date $\tau \geq 1$. For simplicity, we assume that τ follows an exponential distribution with parameter $1 - \lambda$:

$$P(\tau = t | \tau \geq t) = 1 - \lambda.$$

This assumption captures the idea that the agent will have received a random number of signals prior to making his decision. The fact that this number is drawn from an exponential distribution is not important, but makes computations tractable. The parameter λ provides a measure of the number of signals the agent is likely to receive before he must make a decision: the closer λ is to 1,

the larger the expected number of signals. Note that the agent always receives at least one signal.

Perceptions. We assume that the agents correctly interpret the signals they see. That is, when they see x , their perception is that x supports theory $\tilde{\theta}(x)$ and that the strength of the evidence is $\tilde{l}(x)$, and we assume that

$$\tilde{\theta}(x) = \theta(x) \text{ and } \tilde{l}(x) = l(x).$$

Our result that it is optimal for agents to ignore weakly informative signals is robust to agents making perception errors in which they sometimes incorrectly perceive the strength of the evidence they see, and sometimes incorrectly perceive which theory the evidence supports.⁸

Limited information processing. A central element of our analysis is that agents cannot finely record and process information. Agents are assumed to have a limited number of *states of mind* or belief state, and each signal the agent receives is assumed to (possibly) trigger a change in his state of mind. We shall assume that transitions may only depend on the perception associated with the signal *just* received, so a state of mind effectively corresponds to a summary statistic. We also have in mind that those transitions apply across many decision problems that the agent may face, so the transition will not be overly problem-specific or tailored to the particular decision problem at hand.⁹

Formally a state of mind is denoted $s \in S$, where S is a finite set. For any signal x received, changes in state of mind depend on the perception $(\tilde{\theta}, \tilde{l})$ associated with x . We denote by T the transition function:

$$s' = T(s, \tilde{\theta}, \tilde{l}).$$

To fix ideas, we provide a simple example. We will later generalize the approach.

Example 1: The agent may be in one of three states of mind $\{s_0, s_1, s_2\}$. His initial state is s_0 . When he receives a signal x , he gets a perception $(\tilde{\theta}, \tilde{l})$. The event $A_0^+ = \{\tilde{\theta} = 1\}$ corresponds to evidence in favor of state $\theta = 1$, while $A_0^- = \{\tilde{\theta} \neq 1\}$ corresponds to evidence against $\theta = 1$. Transitions are as follows:

⁸See Compte and Postlewaite (2010) for a discussion.

⁹With a continuum of states, one could replicate Bayesian updating and nevertheless satisfy these two constraints (i.e., changes triggered solely by the signal just received, same transition across problems). It is the limit on the number of states that generates, along with these constraints, a particular structure on the way sequences of signals are pooled. Sequences of signals will thus be used less efficiently than a Bayesian would use them.

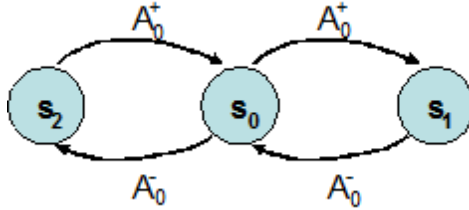


Figure 2: Transition function

As the above figure illustrates, if the agent finds himself in state s_1 when he is called upon to make his decision, there may be many histories that have led to his being in state s_1 . We assume that the agent is limited in that he is unable to distinguish more finely between histories. Consequently, S and T are simply devices that generate a particular pooling of the histories that the agent faces when making a decision.

Optimal behavior. Our aim is to understand the consequences of limits on the number of states of mind and exogenous transitions among those states. To focus on those aspects of mental processing, we assume that the agent behaves optimally contingent on the state he is in. We do not claim that there are not additional biases in the way agents process the information they receive; indeed, there is substantial work investigating whether, and how, agents may systematically manipulate the information available to them.¹⁰

Formally, the set of states of mind S , the initial state, the transition function T and Nature's choice of the true state generate a probability distribution over the states of mind the agent will be in in any period t that he might be called upon to make a decision. For each given state θ , these distributions along with the probability distribution over the periods that he must make a decision, determine a probability distribution over the state the agent will be in when he makes a decision, $\phi_\theta(\cdot) \in \Delta(S)$. This distribution along with the probability distribution over the true state θ determines a joint distribution over (s, θ) , denoted $\phi(\cdot, \cdot) \in \Delta(S \times \Theta)$, as well as a marginal distribution over states $\phi(\cdot) \in \Delta(S)$.

We assume that the agent is able to identify the optimal decision rule $a(s, c)$ ¹¹; that is, that the agent can maximize:

$$\sum_{s, \theta} \phi(s, \theta) E_c u(a(s, c), \theta, c). \quad (1)$$

¹⁰Papers in this area in psychology include Festinger (1957), Josephs et al. (1992) and Sedikides et al. (2004). Papers in economics that pursue this theme include Benabou and Tirole (2002, 2004), Brunnermeier and Parker (2005), Compte and Postlewaite (2004), and Hvide (2002).

¹¹It is indeed a strong assumption that the agent can identify the optimal decision rule. As stated above, our aim is to demonstrate that even with the heroic assumption that the agent can do this, he will systematically make mistakes in some problems.

Call $\pi(s) = \Pr\{\theta = 1 \mid s\}$ the Bayesian updated belief. The maximum expected utility can be written as:¹²

$$v(S, T) = \sum_s \phi(s)v(\pi(s)).$$

Note that while one can compute, as a modeler, the distributions ϕ_θ and posterior beliefs π , we do not assume that the agent knows them. Rather, our assumption is that the agent can identify optimal behavior and thus his behavior coincides with that of an agent who would compute posteriors and behave optimally based on these posteriors. We make the assumption that the agent's behavior coincides with that of an agent who computes posteriors optimally not because we think this is an accurate prediction of how the agent would behave, but rather to focus on the consequence of the limits on information processing that we assume. It would not be surprising that biases of various sorts might arise if no constraints are placed on how an agent makes decisions when called upon to do so. By assuming that the agent behaves optimally given his mental system, any biases that arise are necessarily a consequence of our constraints on his mental processing.

We illustrate our approach next with specific examples.

Computations. We consider the mental process of example 1, and we illustrate how one computes the distribution over states prior to decision making. Define

$$p_\theta = \Pr\{\tilde{\theta} = 1 \mid \theta\}.$$

Thus p_1 for example corresponds to the probability that the agent correctly perceives the true state as being $\theta = 1$ when the actual state is $\theta = 1$. We represent a distribution ϕ over states as a column vector:

$$\phi = \begin{pmatrix} \phi(s_1) \\ \phi(s_0) \\ \phi(s_2) \end{pmatrix},$$

and we let ϕ^0 denote the initial distribution over states of mind (i.e., that distribution puts all weight on s_0 , so that $\phi^0(s_0) = 1$). Conditional on the true state being θ , one additional signal moves the distribution over states of mind from ϕ to $M_\theta\phi$, where

$$M^\theta = \begin{pmatrix} p_\theta & p_\theta & 0 \\ 1 - p_\theta & 0 & p_\theta \\ 0 & 1 - p_\theta & 1 - p_\theta \end{pmatrix},$$

¹²Indeed, using $\phi(s, \theta) = \Pr\{\theta \mid s\}\phi(s)$, the expression (1) can be rewritten as:

$$\sum_s \phi(s) \sum_\theta \Pr\{\theta \mid s\} E_c u(a(s, c), \theta, c).$$

This expression is maximal when the agent chooses $a = 1$ when $\pi(s) - c_1 \geq 1 - \pi(s) - c_2$.

is the transition matrix associated with the mental process of example 1.

Starting from ϕ^0 , then conditional on the true state being θ , the distribution over states of mind at the time the agent takes a decision will be:

$$\phi_\theta = (1 - \lambda) \sum_{n \geq 0} \lambda^n (M^\theta)^{n+1} \phi^0$$

or equivalently,

$$\phi_\theta = (1 - \lambda)(I - \lambda M^\theta)^{-1} M^\theta \phi^0. \quad (2)$$

These expressions can then be used to derive $\phi(s, \theta)$, $\phi(s)$ and $\pi(s)$.¹³

More generally, given any mental process (S, T) , one can associate a transition matrix M^θ that summarizes how an additional signal changes the distribution over states of mind when the true state is θ , and then use (2) to derive the distributions over states ϕ_θ and further, expected welfare $v(S, T)$.

A fully symmetric case.

We illustrate our ideas under a symmetry assumption. We assume that $\pi_0 = 1/2$ and consider a symmetric signal structure:

Assumption 1: $x \in [0, 1]$, $f(x | \theta = 1) = f(1 - x | \theta = 2)$.

Figure 3 below shows an example of such density functions for each of the two states $\theta = 1, 2$, assuming that $f(x | \theta = 1) = 2x$.¹⁴

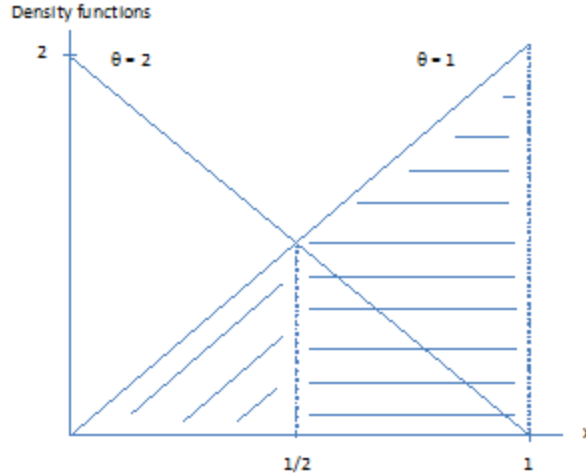


Figure 3: State contingent density functions

¹³ $\phi(s, 1) = \pi^0 \phi_1(s)$, $\phi(s, 2) = (1 - \pi^0) \phi_1(s)$, and $\phi(s) = \sum_\theta \phi(s, \theta)$

¹⁴ Given our symmetry assumption, this implies $f(x | \theta = 2) = 2(1 - x)$. So signals x above $1/2$ are evidence in favor of $\theta = 1$, and the strength of the evidence ($l(x) = \frac{x}{1-x}$) gets large when x gets close to 1.

A signal $x < 1/2$ is evidence in favor of $\theta = 2$ while a signal $x > 1/2$ is evidence in favor of $\theta = 1$. If $\theta = 1$ is the true state the horizontally shaded region below the density function given $\theta = 1$ and to the right of $1/2$ represents the probability of a signal in favor of $\theta = 1$, while the diagonally shaded region to the left of $1/2$ represents the probability that the signal is “misleading”, that is, of a signal in favor of $\theta = 2$.

The probability of a “correct” signal is thus

$$p \equiv \int_{1/2}^1 f(x | 1) dx = 3/4,$$

while the probability of a “misleading” or “wrong” signal is $1/4$. In other words, if the decision maker is in the mental state that is associated with $\theta = 1$, s_1 , there is a $1/4$ chance that the next signal will entail his leaving that mental state.

More generally, under assumption 1, one can define

$$p \equiv \Pr\{\tilde{\theta} = 1 | \theta = 1\} = \Pr\{\tilde{\theta} = 2 | \theta = 2\}$$

as the probability of a correct signal. The probability of a wrong signal is $1 - p$, and the parameter p thus fully determines the transition probabilities across states, hence contingent on the true state θ the probability distributions ϕ_θ over the states (s_2, s_0, s_1) are fully determined by p and λ . We have:

Proposition 1: For $\theta = 1$, the probability distribution ϕ_θ over the states of mind (s_2, s_0, s_1) is: $\phi_\theta = ((1 - p)\rho(1 - \lambda p), \lambda(2 - \lambda)\rho p(1 - p), p\rho(1 - \lambda(1 - p)))$ where $\rho = \frac{1}{1 - \lambda^2 p(1 - p)}$

ϕ_2 is obtained by symmetry. In the case of Figure 3 for example, $p = 3/4$. So if the expected number of signals is very large (that is, if λ is close to 1), the probability distribution over his states of mind (s_2, s_0, s_1) is close to $(1/13, 3/13, 9/13)$ if the true state is $\theta = 1$, and close to $(9/13, 3/13, 1/13)$ if the true state is $\theta = 2$.

These distributions illustrate how the true state θ affects the mental state s in which the agent is likely to be in upon taking a decision. Because signals are not perfect, the agent’s mental state is not a perfect indication to the decision maker of the true state, but it is an *unbiased* one in the following minimal sense:

$$\begin{aligned} \phi_\theta(s_1) &> \phi_\theta(s_2) \text{ when } \theta = 1 \\ \phi_\theta(s_1) &< \phi_\theta(s_2) \text{ when } \theta = 2 \end{aligned}$$

Finally, proposition 1 permits to derive the expected distribution ϕ over states,¹⁵ and thus the posterior beliefs at each mental state. It also permits welfare analysis. To get a simple expression, assume that the distribution over

¹⁵ $\phi = \frac{1}{2}(\phi_1 + \phi_2)$

costs (c_1, c_2) is symmetric, which implies that $v(\pi) = v(1 - \pi)$. Then expected welfare is:

$$2\phi(s_1)v(\pi(s_1)) + (1 - 2\phi(s_1))v\left(\frac{1}{2}\right),$$

where $\pi(s_1)$ and $\phi(s_1)$ are derived from Proposition 1.¹⁶

As one expects, welfare increases with the precision of the signal (p): as p or λ get closer to 1, Bayesian beliefs become more accurate (conditional on s_1 or s_2), and there is a greater chance that the agent will end up away from s_0 .

3 Comparing mental processes

Our objective. Our view is that a mental processing system should work well in a variety of situations, and our main interest lies in understanding which mental process (S, T) works reasonably well, or better than others. In this section, we show that there is always a welfare gain to ignoring mildly informative signals.

3.1 An improved mental process

We return to our basic mental system defined in example 1, but we now assume that a signal must be minimally informative to generate a transition, that is, to be taken as evidence for or against a particular state. Formally, we define:

$$A^+ = \{\tilde{\theta} = 1, \tilde{l} > 1 + \beta\} \text{ and } A^- = \{\tilde{\theta} = 2, \tilde{l} > 1 + \beta\}.$$

In other words, the event $A = \{\tilde{l} < 1 + \beta\}$ does not generate any transition. Call (S, T^β) the mental process associated with these transitions. Compared to the previous case ($\beta = 0$), the pooling of histories is modified. We may expect that because only more informative events are considered, the agent's welfare contingent on being in state s_1 or s_2 will be higher (posterior beliefs conditional on s_1 or s_2 are more accurate). However, since the agent is less likely to experience transitions from one state to another, the agent may have a greater chance of being in state s_0 when making a decision.

We illustrate this basic tradeoff by considering the symmetric case discussed above with $f(x | \theta = 1) = 2x$. Figure 4 below indicates the signals that are ignored for $\beta = 1$: a signal $x \in (1/3, 2/3)$ generate likelihoods in $(1/2, 2)$, and are consequently ignored by the transition T^1 .

¹⁶We have $\pi(s_1) = p \frac{1-\lambda(1-p)}{1-2\lambda p(1-p)}$ and $\phi(s_1) = \frac{1-2\lambda(1-p)p}{2-2\lambda^2(1-p)p}$. Note that $\pi(s_1) > p$ for all values of λ . Intuitively, being in state of mind $s = s_1$ means that the balance of news in favor/against state s_1 tilts in favor of s_1 by on average of more than just one signal. The reason is that if the agent is in state 1, it is because he just received a good signal, and because last period he was either in state 0 (in which case, by symmetry, the balance must be 0) or in state 1 (in which case the balance was already favorable to state s_1).

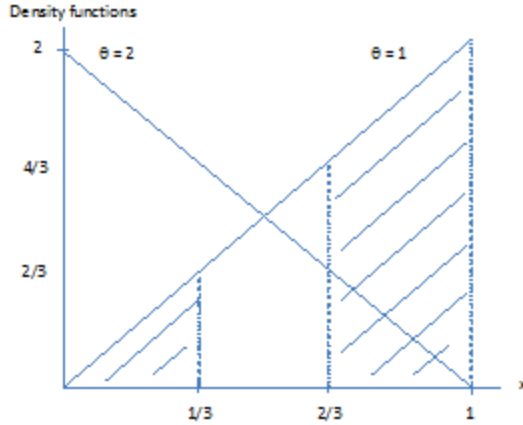


Figure 4

When $\theta = 1$, the shaded regions in Figure 5 to the right of $2/3$ to the left of $1/3$ indicate respectively the probabilities of signals in support of θ_1 and in support of θ_2 , and signals in the interval $(1/3, 2/3)$ are ignored. Now, probabilities of “correct” and “misleading” signals are $5/9$ and $1/9$ respectively, and the ergodic distribution for T^1 is $(1/31, 5/31, 25/31)$. Under T^1 , when λ is close to 1, the probability that the decision maker is in the mental state associated with the true state is nearly $5/6$, as compared with the probability under T^0 , slightly less than $3/4$. In addition, posterior beliefs are more accurate: $\pi(s_1) = 25/26$ under T^0 , and $\pi(s_1) = 9/10$ under T^0 .

Larger β would lead to even more accurate posteriors, with $\pi(s_1)$ converging to 1 as β goes to infinity. But this increase in accuracy comes at a cost. The probabilities $\phi(s)$ are those associated with the ergodic distribution, but the decision maker will get a finite (random) number signals, and the expected number of signals he will receive before making a decision will be relatively small unless λ is close to 1. When β increases, the probability that the signal will be ignored in any given period goes to 1. Consequently, there is a tradeoff in the choice of β : higher β leads to an increased probability of getting *no* signals before the decision maker must decide, but having more accurate information if he gets *some* signals.

Welfare

Assume that costs are drawn uniformly. We will plot expected welfare as a function of β for various values of λ for the example above.

Each mental process (S, T^β) and state θ generates transitions over states as a function of the signal x . Specifically, let $\alpha = \frac{\beta}{2(2+\beta)}$. When for example the current state is s_0 and x is received, the agent moves to state s_1 if $x > \frac{1}{2} + \alpha$, he moves to state s_2 if $x < \frac{1}{2} - \alpha$, and he remains in s_0 otherwise. Denote by

$y = \Pr\{\tilde{\theta} = 1, \tilde{l} < 1 + \beta \mid \theta = 1\}$ and $z = \Pr\{\tilde{\theta} = 2, \tilde{l} < 1 + \beta \mid \theta = 1\}$.¹⁷ Conditional on each state $\theta = 1, 2$, the transition matrices are given by:

$$M_{\beta}^{\theta=1} = \begin{pmatrix} p+z & p-y & 0 \\ 1-p-z & y+z & p-y \\ 0 & 1-p-z & 1-p+y \end{pmatrix}$$

and symmetrically:

$$M_{\beta}^{\theta=2} = \begin{pmatrix} 1-p+y & 1-p-z & 0 \\ p-y & y+z & 1-p-z \\ 0 & p-y & p+z \end{pmatrix}.$$

As before, these matrices can be used to compute the distributions ϕ_{θ} , the posterior belief $\pi(s_1)$ and the probability $\phi(s_1)$, hence the welfare associated with T^{β} . Increasing β typically raises $\pi(s_1)$ (which is good for welfare), but, for large values of β , it also makes it more likely to end up in s_0 (which adversely affects welfare). Figure 6 shows how welfare varies as a function of β for two values of λ , $\lambda = 0.5$ (the lower line) and $\lambda = 0.8$ (the upper line). These correspond to expected numbers of signals equal to 2 and 5 respectively.

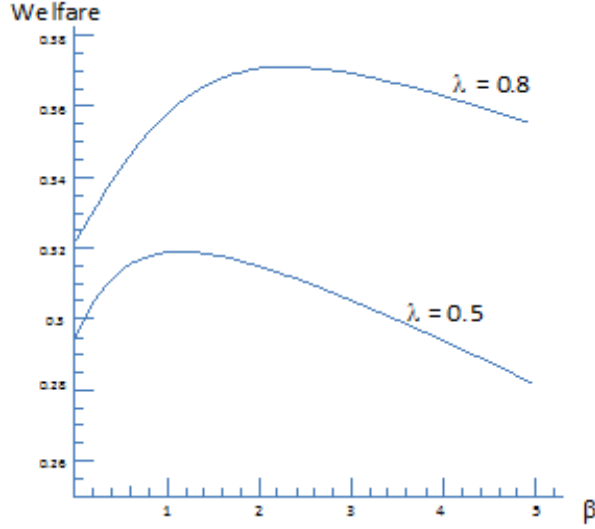


Figure 5: Welfare as a function of β

Note that for a fixed value of λ , for very high values of β there would be little chance of ever transiting to either s_1 or s_2 , hence, with high probability the decision would be taken in state s_0 . This clearly cannot be optimal, so β cannot be too large. Figure 6 also suggests that a value of β set too close to

¹⁷ $y = (1/2 + \alpha)^2 - (1/2)^2 = \alpha(1 + \alpha)$, and $z = (1/2)^2 - (1/2 - \alpha)^2 = \alpha(1 - \alpha)$.

0 would not be optimal either. The graph illustrates the basic trade-off: large β 's run the risk of never leaving the initial state while small β 's have the agent leaving the “correct” state too easily. When λ is sufficiently large, the first effect is small; consequently, the larger λ , the larger is the optimal value of β .

3.2 Simple mental processes.

The advantage of ignoring weakly informative signals in the example above does not depend on the symmetry assumption, nor on the specific mental system of the example. We first generalize the example to a class of simple mental processes, as defined below.

A simple mental process is described by a set of states of mind S and transitions T^0 which specify for each state $s \in S$ and perception $\tilde{\theta} \in \{1, 2\}$ a transition to state $T^0(s, \tilde{\theta})$. Note that the transition depends only on the perception of which state the signal is most supportive of and not on the strength of the signal. We shall restrict attention to mental processes for which T^0 has no absorbing subset. Consider any such simple mental process (S, T^0) . We define a *modified* simple mental process as a simple mental process that ignores weak evidence. Specifically, we define (S, T^β) as the mental process that coincides with (S, T^0) when the perception of the strength of the evidence is sufficiently strong, that is when $\{\tilde{l} > 1 + \beta\}$, and that does not generate a change in the agent's mental state when $\{\tilde{l} < 1 + \beta\}$.

Denote by $W(\lambda, \beta)$ the welfare associated with mental process (S, T^β) , and denote by \underline{W} the welfare that an agent with a single mental state would derive.¹⁸ The next proposition states that for any value of λ , so long as $W(\lambda, 0) > \underline{W}$ and that all states are reached with positive probability, an agent strictly benefits from having a mental process that ignores poorly informative signals.¹⁹

Proposition 2: Consider a simple mental process (S, T^0) . There exist $a > 0$ and $\beta_0 > 0$ such that for all λ and $\beta \in [0, \beta_0]$:

$$W(\lambda, \beta) - W(\lambda, 0) \geq a\beta\underline{q}(\lambda)[W(\lambda, 0) - \underline{W}],$$

where $\underline{q}(\lambda) = \min_{s \in S} \phi(s)$ denotes the minimum weight on any given state when (S, T^0) is used.

Proof: See appendix.

The left hand side of the inequality in the proposition is the welfare increase that results from modifying the transition function (S, T^0) by ignoring signals

¹⁸ $\underline{W} = v(\pi_0)$.

¹⁹In this paper we limit attention to the case that there are two possible theories. Compte and Postlewaite (2010) discuss the case when there are more than two theories. As one would expect, it continues to be optimal to ignore weakly informative signals in that case. In addition, they argue that agents may tend to see patterns in the data they get when there are no patterns.

for which the strength of the evidence is less than β . The proposition states that the gain is a positive proportion of the value of information for the initial transition function.

3.3 Generalization

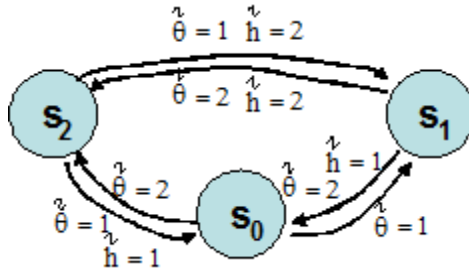
A simple mental process does not distinguish mild and strong evidence. As we shall see, a more sophisticated mental process that distinguishes between mild and strong evidence can sometimes improve welfare. We next extend the class of mental systems we consider to include a class of sophisticated mental processes and show that our insight that ignoring weak evidence improves welfare holds in this larger class.

Let us emphasize that our objective is not to characterize optimal mental processes. A mental process that is optimal for a particular problem would need to be tailored to the particular distribution f , the parameter λ for that problem as well as the particular utilities attached to each decision. Rather, our view is that a mental process will be applied to many problems, and that in general, identifying how a mental process should be amended to improve welfare, or identifying the class of problems for which one mental process would outperform another are difficult tasks. Our result illustrates that there is, however, one direction of change, ignoring weak evidence, that unambiguously improves welfare.²⁰

A class of sophisticated mental system.

We first define a class of *level-k transitions*. Consider an increasing sequence of $(\beta_1, \dots, \beta_{k-1})$, and set $\beta_0 = 0, \beta_k = +\infty$. Any perception $\tilde{l} \in (1 + \beta_{\tilde{h}-1}, 1 + \beta_{\tilde{h}})$ with $\tilde{h} \in \{1, \dots, k\}$ is labelled as a perception of strength \tilde{h} . A level-k transition is a transition function for which, given the current state, two perceptions of the same strength generate the same transitions. The agent's perception can thus be summarized by a pair $(\tilde{\theta}, \tilde{h})$.

The figure below provides an example of a level 2 mental system, where strong signals ($\tilde{h} = 2$) never induce a transition to state s_0 .



²⁰An immediate corollary of Proposition 2 is that if one considers a finite collection of problems, then Proposition 2 applies to that collection with \underline{q} defined as a minimum weight on states of mind across all problems considered.

Given any level k mental system for which $T(s, (\tilde{\theta}, \tilde{1})) \neq s$ for some s , one may consider a modified mental system T^β that would coincide with T when $\tilde{l} > 1 + \beta$, but that would ignore signals for which $\tilde{l} < 1 + \beta$.

The following proposition shows the benefit of ignoring weak information. We denote again by $W(\beta, \lambda)$ the welfare associated with the modified mental process. The original level k mental system along with an initial distribution over states of mind generates (for each value of λ) posteriors conditional on the state. Denote by $\Delta(\lambda)$ the smallest difference between these posteriors.²¹ We have:

Proposition 3: Consider any level k mental system (S, T) such that $T(s, (\tilde{\theta}, \tilde{1})) \neq s$ for some s . Then, there exists $a > 0$ and $\beta_0 > 0$ such that for all $\beta \in [0, \beta_0]$ and all λ ,

$$W(\beta, \lambda) - W(0, \lambda) \geq a\beta[\Delta(\lambda)]^2 \underline{q}(\lambda),$$

where $\underline{q}(\lambda) = \min_{s \in S} \phi(s)$ denote minimum weight on any given state when (S, T) is used.

Proof: See Appendix.

A welfare comparison.

It is easy to see that the more sophisticated mental system that permits different transitions from a state depending on the strength of the evidence may lead to higher welfare than a simple mental system that does not consider the strength. It is also easy to see that distinguishing between weaker and stronger signals, with different transitions following weaker and stronger signals may *not* be welfare optimal. Suppose that for all signals in favor of a given state, the likelihood ratios are nearly equal; it typically will not pay to treat differently the signals in the transition.

This discussion points out an important feature of an optimal mental system. The optimal transition following a particular signal does not depend only on the strength of that signal, but rather depends on the strengths of other signals that might have been received and the distribution over the possible signals. This is in stark contrast to the case of Bayesian updating, where the “transition” from a given state (the prior) to a new state (the posterior) depends only on the likelihood ratio of the signal.

4 Consequences of ignoring weak evidence

What are the consequences of the fact that $\beta > 0$ for an agent’s transition function? As we discussed in the introduction, our view is that there is a mental process that summarizes the signals an agent has received in a mental state,

²¹ $\Delta(\lambda) > 0$ except possibly for a finite number of values of λ or initial distribution over states.

and that the agent chooses the optimal action given his mental state when he is called upon to make a decision. Agents do not have a different mental process for every possible decision they might some day face. Rather, the mental process that aggregates and summarizes their information is employed for a variety of problems with different signal structures f . β is set optimally *across* a set of problems, not just for a specific distribution over signals f (that in addition would be correctly perceived). When β is set optimally across problems, it means that for some problems, where players frequently receive mild evidence and occasionally strong evidence, there is a bias (see below) towards the theory that generates occasional strong evidence, while Bayesian updating might have supported the alternative theory.²² Worse, the bias worsens when the number of signals received rises.

Specifically, each state of mind reflects a belief state (possibly more or less entrenched) about whether a particular theory is valid. In our basic three mental state example, being in state s_1 is meant to reflect the agent's belief that $\theta = 1$ is likely to hold, while s_2 is meant to reflect the belief that $\theta = 2$ is likely to hold; also, s_0 reflects some inability to form an opinion as to which is the true state. One interpretation is that these beliefs states are what the decision maker would report if asked about his inclination as to which state holds.

We defined earlier a minimal sense in which a decision maker would have unbiased beliefs, namely that he is more likely to be in mental state s_1 rather than state s_2 (hence to lean towards believing $\theta = 1$ rather than $\theta = 2$) whenever the true state is $\theta = 1$ (and similarly when the true state is $\theta = 2$).

As we saw earlier, our symmetric example resulted in unbiased beliefs. In addition, increasing β from 0 makes the mental system more accurate: if he does not receive too few messages, the probability that the decision maker is in the mental state associated with the true state increases.

As one moves away from symmetric cases, however, having a positive β may be a source of bias and receiving more signals may actually worsen the bias. Suppose that an agent faces a finite set of problems of the type analyzed above, with a probability distribution over those problems. Then there will be $\beta > 0$ for the stochastic problem, with the agent optimally ignoring signals with likelihood ratio less than $1 + \beta$. Suppose that we now add to the finite set of problems an asymmetric problem such as the ESP problem in which there is occasional strong evidence in favor of the theory, and frequent weak evidence against the theory. Specifically, suppose that the likelihood of the signals against the theory are below $1 + \beta/2$. If the probability associated with this new problem is sufficiently small, there will be a $\beta' > \beta/2$ such that the agent will optimally ignore signals with likelihood less than $1 + \beta'$, and hence, ignore evidence against the theory that the agent has ESP.

In general, a mental system that is optimal in the processing of signals for a set of problems will ignore informative signals whose strength falls below

²²Note that we do not have in mind that the decision maker would have biased posterior beliefs. We have assumed throughout the paper that the decision maker maximizes welfare in choosing an action given his mental state at the time he decides, implying that he behaves as if he had correct posterior beliefs.

some (strictly positive) threshold. Whatever is that threshold, there will be asymmetric problems that generate biased beliefs: although the theory is false, occasional evidence in favor of the theory is noted while the frequent evidence against the theory is overlooked.

The discussion above does not, of course, imply that all individuals should fall prey to the belief that they have ESP. Different individuals will receive different signals. Some individuals may read scientific articles that argue against ESP, signals that are evidence against ESP, and in addition, above the individual's threshold.

5 Discussion

In our approach, the agent receives signals (x) that he interprets as support for theories he thinks possible. We formalized this process by defining a transition function that captures how he takes a signal into account. However we do not have in mind that the agent is fully aware of this mental process. The transition function or mental system aggregates the agent's perceptions over time, and the agent's choice is only to choose the optimal behavior conditional for each belief state he may be in when a decision is called for.

The transition function is *not* optimized for a specific problem. Rather, the mental system is assumed to operate for a set of problems that typically differ in the probability distributions over the signals conditional on the true state of nature and in the expected number of signals the agent will receive prior to making a decision. Our interest then is to investigate modifications of the mental system that increase the agent's welfare across the set of problems he faces.

The limitation on the number of belief states is central to our analysis. We suggested computational bounds as the basis for this limitation, but the limitation may capture other aspects of decision making. It may, for example, be due to an inability to finely distinguish between various states of mind. Alternatively, states of mind might be instrumental in making good decisions, because as the number of states rise, the problem of finding the optimal mapping from belief state to action becomes increasingly complex.

We have explored relatively simple classes mental processes. Our insight that ignoring weak evidence improves welfare holds not only for these mental processes, but holds more generally. In particular, it is not a property that would only hold for the optimal transition: the central ingredient is whether the agent has distinct posteriors at each state of mind. This property must be satisfied for the optimal transition (as otherwise states of mind would not be used efficiently) but it holds much more generally.

Relatively simple mental processes are of interest for reasons beyond those laid out above. Simple processes can be effective when agents make errors in estimating strength of evidence (likelihood ratios).²³ Simple processes are

²³Of course the alternative causality is plausible as well: it is because we have few states

less sensitive to errors in assessing strength of evidence than more complex processes, including Bayesian updating. This in the spirit of Gilovitch’s (1991) fast and frugal trees: don’t try to look at all possible symptoms and weight them appropriately taking in consideration the detailed characteristics of a patient: just look at the most informative symptoms.

We have mentioned a number of instances in which people have beliefs that are at odds with the evidence. It has been suggested that superstitions might arise because of confirmatory bias – a tendency to interpret ambiguous evidence as confirming one’s current beliefs.²⁴ Without rejecting the possibility of confirmatory bias, it is useful to point out some differences. First, confirmatory bias is essentially neutral: whatever the initial beliefs an agent may have, they will be reinforced by subsequent evidence. This is not the case in our model. While both models might result in an agent believing he has ESP when in fact he does not, the reverse will not be the case in our model; it is precisely the asymmetry in the strength of signals that leads to biased beliefs in some problems: some signals are ignored not because they contradict first impressions but because they are weak. Second, with respect to confirmatory bias, there remains a question of how such a bias can survive evolutionary pressures. Shouldn’t individuals who were subject to this bias be selected against? The biased beliefs that arise in our model, on the other hand, are precisely a *consequence* of evolutionary pressure: given resource constraints that lead to finite belief states, biased beliefs must arise in an optimal mental system.

Extensions.

1. Consider a variant of the problem we analyze in which there is a random termination date, but the agent makes a decision in every period until then. The basic results carry over to that setup: The agent will want to ignore signals that have likelihood close to 1.

2. Consider a variant of the problem in which the agent can delay making a decision. Instead of there being a random time at which the agent must make a decision, suppose that in any period the agent delays his decision with probability λ , and the agent discounts at rate δ . λ is then endogenous and there is a tradeoff between accuracy and delay. The basic results carry over to this problem: The agent will want to ignore signals that have likelihood close to 1.

3. In early versions of this paper, we have explored an extension to cases where the set of possible theories is greater than two. This extension poses interesting challenges relative to the definition of simple transitions. A signal now potentially comes with a vector of likelihood ratios, and a plausible mental process for an agent would entail the agent extracting which theory $\tilde{\theta}$ is supported by the evidence (based on likelihoods): $\tilde{\theta}(x) = \arg \max f(x | \theta)$. It is relatively easy to see that for some distributions over signals, there will be an inherent bias against some theories, because even when they hold true, there

that we don’t care about the exact strength.

²⁴See, e.g., Rabin and Schrag (1999).

may be no signals that support them relative to all other possible theories.²⁵ Ignoring weak evidence may only reinforce that bias.

6 Appendix

Proof of Proposition 2:

Proof: We fix some initial state, say s_0 . A history h refers to a termination date $\tau \geq 1$ and a sequence of perceptions $((\tilde{\theta}_0, \tilde{l}_0), (\tilde{\theta}_1, \tilde{l}_1), \dots, (\tilde{\theta}_{\tau-1}, \tilde{l}_{\tau-1}))$. Denote by H_s^β the set of histories h that lead to state s (starting from initial state s_0) when the mental process is (S, T^β) . Also denote by $H_{s,s'}$ the set of histories that lead to state s under (S, T^0) and to state s' under (S, T^β) .

For any set of histories H , we denote by $\pi(H)$ the Bayesian posterior, conditional on the event $\{h \in H\}$:

$$\pi(H) = \frac{\Pr(H \mid \theta = 1)}{\Pr(H)} \pi_0.$$

By definition of $W(\lambda, \beta)$, we have:

$$W(\lambda, \beta) = \sum_{s \in S} \Pr(H_s^\beta) v(\pi(H_s^\beta)).$$

Now let \bar{W} denote the welfare that the agent would obtain if he could distinguish between the sets $H_{s,s'}$ for all s, s' . We have:

$$\bar{W} = \sum_{s, s' \in S} \Pr(H_{s,s'}) v(\pi(H_{s,s'})).$$

We will show that there exist constants c and c' , and β_0 such that for all $\beta < \beta_0$,

$$W(\lambda, \beta) \geq \bar{W} - c[\beta \ln \beta]^2 \quad (3)$$

and

$$W(\lambda, 0) \leq \bar{W} - c'\beta. \quad (4)$$

Intuitively, welfare is \bar{W} when the agent can distinguish between all $H_{s,s'}$. Under (S, T^β) , he cannot distinguish between all $H_{s,s'}$:

$$H_s^\beta = \cup_{s'} H_{s',s}.$$

Under (S, T^0) , he cannot distinguish between all $H_{s,s'}$ either, but the partitioning is different:

$$H_s^0 = \cup_{s'} H_{s,s'}. \quad (5)$$

²⁵For example, if one considers a signal is a sequence of two draws of 1's and 0's, and if draws can be either independant ($\theta = 1$), autocorrelated ($\theta = 2$) or anti-correlated ($\theta = 3$), then 11 and 00 are evidence of anticorrelation, while 10 and 01 are evidence of anticorrelation. So even when draws are independent, the agent will tend to see patterns.

Because each mental process corresponds to a pooling of histories coarser than $H_{s,s'}$, and because v is a convex function, both $W(\lambda, 0)$ and $W(\lambda, \beta)$ are smaller than \bar{W} . What we show below however is that the loss is negligible in latter case (of second order in β), while it is of first order in the former case.

We use three Lemmas, the proofs of which are straightforward. We let $\underline{q} = \min_s \Pr H_s^0$ and $\Delta = W(\lambda, 0) - \bar{W}$. We assume that $\underline{q} > 0$. In what follows, we choose β small so that $\gamma(\beta) \equiv \Pr[\tilde{l} < 1 + \beta]$ is small compared to \underline{q} .²⁶

Lemma 1: There exists a constant c and β_0 such that for all $\beta < \beta_0$ and λ :

$$|\pi(H_{s',s}) - \pi(H_s^\beta)| \leq c\beta |\ln \beta|. \quad (6)$$

Proof: The event $H_{s',s}$ differ from H_s^β only because there are dates t where the perception $(\tilde{\theta}^t, \tilde{l}^t)$ has $\tilde{l}^t < 1 + \beta$. For any fixed $\bar{\lambda} < 1$ and any $\lambda \leq \bar{\lambda} < 1$, there are a bounded number of such dates, in expectation, so the conclusion follows. For λ close to 1, the number of such dates may become large. However, to determine posteriors up to $O(\beta)$, there are only a number of perceptions prior to the decision comparable to $|\ln \beta|$ that matter.²⁷ So the conclusion follows as well.²⁸

Lemma 2: There exist constants c' and β_0 such that for all λ and $\beta < \beta_0$, there exist s and $s' \neq s$ with $\Pr(H_{s,s'}) \geq c'\beta$ such that :

$$|\pi(H_{s,s'}) - \pi(H_s^0)| \geq c'[W(\lambda, 0) - \bar{W}]^{1/2}.$$

Proof: We shall say that two states s and s' are consecutive when there is a perception $\tilde{\theta}$ such that $s' = T(s, \tilde{\theta})$. Let $\Delta = W(\lambda, 0) - \bar{W} > 0$. There must exist two states s, s' (not necessarily consecutive) such that $|\pi(H_{s'}^0) - \pi(H_s^0)| \geq c_0\Delta^{1/2}$. Since T has no absorbing subset, there must exist a finite sequence of consecutive states $s^{(0)}, \dots, s^{(k)}, \dots, s^{(K)}$ such that $s^{(0)} = s$ and $s^{(K)} = s'$. Hence there must exist two consecutive states s_0, s'_0 such that $|\pi(H_{s'_0}^0) - \pi(H_{s_0}^0)| \geq c_0\Delta^{1/2}/N$ (where N is the total number of states). The events $H_{s'_0}^0$ and $H_{s'_0}^\beta$ both consist of the union of $H_{s'_0, s'_0}$ and of events that have probability comparable to $\gamma(\beta) \equiv \Pr\{\tilde{l} < 1 + \beta\}$. For β small enough, $\gamma(\beta)$ can be made small compared to \underline{q} . The posteriors $\pi(H_{s'_0}^0)$ and $\pi(H_{s'_0}^\beta)$ must thus both be close to $\pi(H_{s'_0, s'_0})$

²⁶This is possible because f is smooth and because the event $\{\tilde{l} = 1\}$ has measure 0.

²⁷More precisely, consider the last date prior to the decision where the state is s_0 . Because there is a finite number of states and no absorbing subsets, the probability of not going through s_0 during T periods is at most equal to μ^T for some $\mu < 1$. So for T comparable to $|\ln \beta|$, there is a probability $o(\beta)$ to stay away from s_0 . So with probability $1 - o(\beta)$, fewer than $O(|\ln \beta|)$ perceptions matter.

²⁸Note that the bound can be improved, because the expected fraction of the time where $(\tilde{\theta}^t, \tilde{l}^t)$ has $\tilde{l}^t < 1 + \beta$ gets close to 0 with β .

hence close to each other. Applying Lemma 1, it then follows that $|\pi(H_{s_0, s'_0}) - \pi(H_{s_0}^0)| \geq c'\Delta^{1/2}$ for some c' independent of λ . Since s_0 and s'_0 are consecutive, the event H_{s_0, s'_0} must have probability at least comparable to $\gamma(\beta)$, hence at least comparable to β (since f is smooth).

Lemma 3: Let m, \bar{m} such that $\bar{m} \geq v'' \geq m$. For α small, we have, forgetting second order terms in α :

$$\alpha \frac{\bar{m}}{2} (\pi^1 - \pi^0)^2 \geq (1-\alpha)v(\pi^0) + \alpha v(\pi^1) - v(\alpha\pi^0 + (1-\alpha)\pi^1) \geq \alpha \frac{m}{2} (\pi^1 - \pi^0)^2.$$

Since

$$\sum_{s \in S} \Pr(H_{s, s'}) \pi(H_{s, s'}) = \pi\left(\bigcup_s H_{s, s'}\right) \sum_{s \in S} \Pr(H_{s, s'}) = \pi(H_{s'}^\beta) \Pr(H_{s'}^\beta),$$

we have:

$$\begin{aligned} \bar{W} - W(\lambda, \beta) &= \sum_{s' \in S} \left[\sum_{s \in S} \Pr(H_{s, s'}) [v(\pi(H_{s, s'})) - v(\pi(H_{s'}^\beta))] \right] \\ &= \sum_{s' \in S} \Pr(H_{s'}^\beta) \left[\sum_{s \in S} \frac{\Pr(H_{s, s'})}{\Pr(H_{s'}^\beta)} [v(\pi(H_{s, s'})) - v(\sum_{s \in S} \frac{\Pr(H_{s, s'})}{\Pr(H_{s'}^\beta)} \pi(H_{s, s'}))] \right]. \end{aligned}$$

Applying Lemma 3 thus yields:

$$\bar{W} - W(\lambda, \beta) \leq c \max_{s, \hat{s}, s'} |\pi(H_{s, s'}) - \pi(H_{\hat{s}, s'})|^2,$$

and Lemma 1 gives a lower bound on $W(\lambda, \beta)$.

To get the upper bound on $W(\lambda, 0)$, we use:

$$\sum_{s' \in S} \Pr(H_{s, s'}) \pi(H_{s, s'}) = \pi\left(\bigcup_{s'} H_{s, s'}\right) \sum_{s' \in S} \Pr(H_{s, s'}) = \pi(H_s^0) \Pr(H_s^0),$$

and write:

$$\bar{W} - W(\lambda, 0) = \sum_{s \in S} \left[\sum_{s' \in S} \Pr(H_{s, s'}) [v(\pi(H_{s, s'})) - v(\pi(H_s^0))] \right].$$

Lemmas 2 and 3 (using s and s' as defined in Lemma 2) then yield a lower bound:

$$\bar{W} - W(\lambda, 0) \geq c\beta[\pi(H_{s, s'}) - \pi(H_s^0)]^2.$$

Proof of Proposition 3. The proof is almost identical to that of Proposition 2. The difference is that the appropriate version of Lemma 2 is now much simpler to obtain. Assume $\Delta(\lambda) > 0$. Since $T(s, (\tilde{\theta}, \tilde{1})) \neq s$ for some s , then we immediately obtain that there are two states s, s' such that $\Pr(H_{s, s'}) = O(\beta)$ and $|\pi(H_{s'}^0) - \pi(H_s^0)| \geq \Delta(\lambda)$, which further implies, using the same argument as in Lemma 2 that:

$$|\pi(H_{s, s'}) - \pi(H_s^0)| \geq c\Delta(\lambda)$$

for some constant c .

7 Bibliography

Baumol, William J. and Richard E. Quandt (1964), "Rules of Thumb and Optimally Imperfect Decisions," *American Economic Review* 54(2), pp 23-46.

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

Benabou, R. and J. Tirole (2004), "Willpower and Personal Rules," *Journal of Political Economy*, 112 (4), 848-886.

Brunnermeier, Marcus and Jonathan Parker (2005), "Optimal Expectations," *American Economic Review*, Vol. 95, No. 4, pp. 1092-1118.

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

Compte, O. and A. Postlewaite (2010), "Mental Processes and Decision Making," mimeo, University of Pennsylvania.

Cover, T. and M. Hellman (1970), "Learning with Finite Memory," *Annals of Mathematical Statistics* 41, 765-782.

Festinger, L. (1957), *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Gilovich, T. (1991), *How We Know What Isn't So*. New York, NY: The Free Press.

Hellman, M. E. and T. M. Cover (1973), "A Review of Recent Results on Learning with Finite Memory," *Problems of Control and Information Theory*, 221-227.

Hvide, H. (2002), "Pragmatic beliefs and overconfidence," *Journal of Economic Behavior and Organization*, 2002, 48, 15-28.

Jenkins, H. M. and W. C. Ward (1965), "Judgement of contingency between responses and outcomes," *Psychological Monographs*

Josephs, R. A., Larrick, R. P., Steele, C. M., and Nisbett, R. E. (1992), "Protecting the Self from the Negative Consequences of Risky Decisions," *Journal of Personality and Social Psychology*, 62, 26-37.

Lamb, Emmet J. and Sue Leurgans (1979), "Does Adoption Affect Subsequent Fertility?", *American Journal of Obstetrics and Gynecology* 134, 138-144.

Lipman, Barton (1995), "Information Processing and Bounded Rationality: A Survey," *The Canadian Journal of Economics* 28(1), pp. 42-67.

Rabin, Matthew and Joel L. Schrag (1999), "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics* 114, 37-82.

Rosenthal, Robert W. (1993), "Rules of thumb in games," *Journal of Economic Behavior & Organization*, Elsevier, vol. 22(1), pages 1-13, September.

Savage, Leonard J. (1954), *The Foundations of Statistics*. New York: John Wiley and Sons. (Second addition in 1972, Dover. . .)

Sedikides, C., Green, J. D., and Pinter, B. T. (2004), "Self-protective memory," In: D. Beike, J. Lampinen, and D. Behrend, eds., *The self and memory*. Philadelphia: Psychology Press.

Ward, W. C. and H. M. Jenkins (1965), "The display of information and the judgment of contingency," *Canadian Journal of Psychology*

Wilson, A. (2004), "Bounded Memory and Biases in Information Processing," mimeo, University of Chicago.