On belief formation and the persistence of superstitions*

Olivier Compte†
PSE, Paris

This version: October 2021

Abstract

We propose a belief-formation model where agents attempt to discriminate between two theories, and where the asymmetry in strength between confirming and disconfirming evidence tilts beliefs in favor of theories that generate strong (and possibly rare) confirming evidence and weak (and frequent) disconfirming evidence. In our model, limitations on information processing provide incentives to censor weak evidence, with the consequence that for some discrimination problems, evidence may become mostly one-sided. Sophisticated agents who know the characteristics of the censored data-generating process are not lured by this accumulation of evidence, but less sophisticated ones end up with incorrect beliefs.

1 Introduction

Superstitions and other folk beliefs are common. These beliefs often have the structure of a particular circumstance (C) or act increasing the chance of an otherwise rare event (E); a sort of illusory correlation (Chapman and Chapman (1967)) between C and E, where one overestimates the frequency of occurence of the sequence C-E.

A common explanation for the existence of biased beliefs is that looking for patterns in the environment has fitness value – predicting the future or the imminence of danger is useful,[1] and if the cost of holding erroneous

*This paper revisits "Mental Processes and Decision Making" by Olivier Compte and Andrew Postewaite.

†Paris School of Economics, 48 Bd Jourdan, 75014 Paris (e-mail: compte@enpc.fr).

[1]See Beck and Forstmeier 2007.

beliefs is small compared to the potential benefits, taking the Pascalian bet is a good option: why not drink the miraculous water or repent if this has the slightest chance of curing illness.

In essence, the explanation is based on the idea that beliefs are inevitably incorrect to some extent and that some errors are less costly than others. The explanation is convincing. Still, one could be surprised that erroneous beliefs persist even (and sometimes even more so) among people that are repeatedly confronted with disconfirming evidence. It is not uncommon for nurses working in maternity wards to believe in lunar effects (Abell and Greenspan (1979)), for example the fact that a full moon would increase the number of (unprogrammed) baby deliveries. Or at the very least, these erroneous beliefs seem inconsistent with Bayesian modelling, where eventually, after being exposed to data for long enough, correct beliefs should prevail.

Outside the Bayesian sphere, one plausible explanation for some biased beliefs is that they have *instrumental value*: some biases may have a direct positive effect on well-being or performance either because they reduce anxiety, improve focus or give a sense of control. This includes many (personal) superstitions such as the protection from Bad Luck conferred by charms or amulets,[2] or the powers conferred by magical thoughts and other ritualized or routine behaviors.[3] Holding such beliefs generates direct (first-order) gains and, if not excessively biased – magic thoughts giving a sense of invincibility are potentially harmful, only second-order losses.[4]

Another plausible explanation for biased beliefs is the *confirmation bias*: once the seed of a belief is planted in people's mind, this belief tends to persist even when erroneous because evidence is then processed with a bias; people are more likely to see/look for/process evidence that confirms the belief, rather than disconfirm it.[5]

[2] See for example Hildburgh (1951), who suggests that amulets act an anxiety reducer, which fosters good lactation.

[3] This also includes placebo effects: an inactive treatment may have positive health effect, so long as you believe it does.

[4] This trade-off is for example examined in Compte and Postlewaite 2004, where biased beliefs about chances of success positively affect performance. See Köszegi (2006) for the case where beliefs directly affect preferences. See also Brunnermeier and Parker (2005).

[5] The negative consequences of the confirmation bias is clear (Rabin and Schrag (1999)). The possible fitness value of the confirmation bias is discussed in models where agents lack will-power (Benabou Tirole, 2002, 2004), modelled as a discrepancy between the welfare criterion and the decision rule. Plausibly however, in the same way that some biases in beliefs contribute to reduce anxiety, there could be some reassuring value to seeing one's beliefs confirmed, a reassurance that has first-order effect on welfare in the same way that confidence does. The fitness value of the confirmation bias has also been discussed within

Still, some beliefs seem more easily confirmed than others: if one starts with the belief that the full moon has *no effect* on baby deliveries, how strongly will that belief be reinforced by the observation of hospital tension on a non-full moon day? Or at least, for lunar effects, providing evidence in favor of a lunar effect seems easier than providing evidence against it. A single coincidence of a full moon and a high number of deliveries seem to be strong evidence in favor of the theory, which cannot be matched in strength by a single instance of a high number of deliveries without full moon: these kinds of bad days just happens.

This asymmetry between the strengths of confirming and disconfirming evidence is at the heart of our argument: we shall argue that beliefs are easily tilted in favor theories that generate strong (and possibly rare) confirming evidence and weak (and frequent) disconfirming evidence, even when these theories are untrue.

At a broad level, the main logic of our argument is that (i) limitations on information processing provide incentives to censor weak evidence; (ii) the consequence is that for some problems, evidence may become mostly one-sided (in the direction of strong confirming evidence), independently of the underlying state of the word; (iii) a sophisticated agent who would know the characteristics of the censored data-generating process would not be lured by the accumulation of these confirming evidence,[6] but a less sophisticated one ends up with incorrect beliefs.

This paper provides a simple model that articulates these arguments. We review below the main modelling assumptions and intuitions, and next discuss some of these assumptions.

1.1 Main modelling assumptions.

We consider a family of decision problems over two alternatives 1 and 2 where many signals are processed prior to decision making. There are two underlying states $\theta = 1,2$, defining which alternative is the better one, and signals potentially permit the agent to discriminate between the two underlying states. Our model has four main ingredients:

(a) *A coarse mental system*: the individual has a limited number of

the perspective of social interactions (see Peters (2020).

[6] For example, even if, because of censoring, a nurse mostly processes signals suggesting that full moon affects birth numbers, the Bayesian nurse should not fall prey to that mental suggestion. She should understand that her mental system is biased, ignore the mental suggestion and fear no extra work pressure upon full moon shifts.

mental states,[7] with one-step changes in mental state triggered by confirming or disconfirming evidence. Specifically, we assume $2K+1$ mental states, labelled s and ordered from $-K$ to $+K$. The initial state is $s=0$ and a signal perceived as confirming $\theta=1$ (denoted $\widetilde{\theta}=1$) triggers a move to the right (if the move is possible), while a signal perceived as confirming $\theta=2$ triggers a move to the left (if the move is possible). With $K=2$, we have:

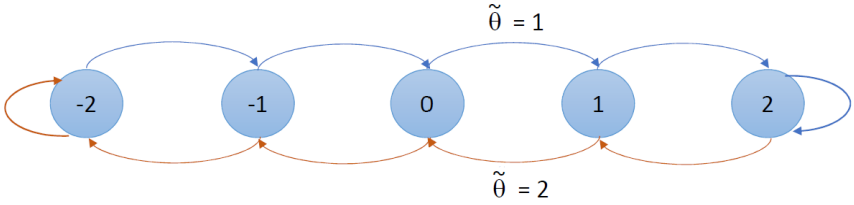


Figure 1: Mental system

(b) *an option to censor weak evidence* (if this enhances welfare) and thus focus on the more informative signals: only strong enough signals trigger changes in the mental state, with a threshold strength parameterized by a scalar β.[8]

(c) *limitations on how beliefs are formed:* each agent is equipped with a possibly noisy prior, denoted $\widetilde{\rho}$, and we postulate an "all-purpose" family of belief-formation strategies mapping priors and mental state to a posterior belief. Specifically, expressing beliefs using likelihood of 1 vs. 2, we assume that the agent's posterior belief in mental state s is

$$\widetilde{\rho}d^{s} \tag{P}$$

where d is a parameter that characterizes the degree to which the agent's mental state affects posterior beliefs, or the *discriminatory power* of the belief-formation rule. Given this (subjective) posterior belief, the agent takes a decision that (subjectively) maximizes welfare.

Note that the mental system and (P) are restrictions that jointly shape how posterior beliefs are formed: given a censoring level β, posteriors beliefs

[7] This is in the spirit of Cover and Hellman (1970) and Wilson (2004), which we shall laer discuss.

[8] Formally, we shall say that a signal x confirms (or is evidence for) θ, if signal x is more likely under θ than under $\theta' \neq \theta$). The strength of the evidence is then defined as the ratio of probabilities of receiving x under θ and under θ'.

and behavior in our model are entirely driven by d.

(d) *The signal-generating process is problem-specific and unobservable.* We have in mind that both β and d are adjusted to the set of problems that the agent might face, so as to maximize welfare *on average across problems*: belief formation (d) and censoring (β) cannot be tuned to the detailed characteristics of *each* signal-generating process.[9]

1.2 Main intuitions.

Regular and irregular problems. Given the coarse processing (a) assumed, the relevant characteristics of the signal-generating process reduce to the probabilities $p_{\theta\theta}$ of processing evidence for state θ when the state is θ (conditional on processing evidence)[10] and we shall say that a problem is *regular* iff

$$p_{\theta\theta} > 1/2 \text{ for each } \theta$$

Intuitively, this means that for a regular problem, the agent's mental state leans to the right when the underlying state is 1, and to the left when the underlying state is 2. Discrimination between the two underlying states is thus relatively easy. If, across all the problems faced, regular problems are preponderant, then the individual has incentives to set the discriminatory power parameter d above 1 (because indeed the mental system is truly informative).

However, if there are problems for which, given censoring, the probabilities $p_{\theta\theta}$ satisfy

$$p_{11} > 1/2 \text{ and } p_{22} < 1/2,$$

the individual's mental state will lean to the right *independently of the underlying state*, and he will thus erroneously end up with beliefs favoring theory 1 even in events where $\theta = 2$ holds: in these cases, processing information moves posteriors *away from the truth* and possibly deteriorates welfare. For such problems, the agent would have been better off setting $d = 1$. The agent's inability to adjust d to the characteristics $p_{\theta\theta}$ of the problem considered will be key.

Censoring. In essence, at the margin, weak evidence adds noise to the mental system, generating moves to the right and left with almost equal

[9] This is the sense in which the family (P) is "all-purpose". Technically, this means that β and d are adjusted at an ex ante stage, before the signal-generating process is selected.

[10] That is, conditional on processing evidence, we define $p_{\widetilde{\theta}\theta}$ as the probability of processing evidence in favor of $\widetilde{\theta}$ when the underlying state is θ, so $p_{1\theta} + p_{2\theta} = 1$.

probability. Censoring weak evidence eliminates these noisy moves. Does this enhance welfare?

For a Bayesian who knows the signal generating process, the answer is positive in most cases (though not all cases), and the reason is that mental states being a scarce resource, one is generally better off limiting the use of mental-state changes to sufficiently informative signals.[11]

For our less sophisticated agent, the answer depends on whether the problem is regular or not. For regular problems, censoring weak evidence induces an increase in both p_{11} and p_{22}:[12] the correlation between the underlying state and the mental state is improved and the discrimination between underlying states is improved – and more mental states help.

For irregular problems, with beliefs pointing towards, say $\widehat{\theta}$, irrespective of the underlying state, the effect is opposite, reinforcing the trend towards $\widehat{\theta}$: the balance between evidence confirming and disconfirming $\widehat{\theta}$ becomes more favorable to $\widehat{\theta}$ independently of the underlying state, and when $\widehat{\theta} \neq \theta$, this is potentially harmful for welfare (and even more so when there are more mental states).

Optimal censoring of evidence trades off the two effects above, and to the extent that regular problems are preponderant on average, the decision maker has incentives to censor weak evidence.

Supersition-prone problems. The last piece of our argument consists in observing that censoring weak evidence affects the type of problems that are irregular as well as the underlying state that gets most likely confirmed: problems for which a state mostly generate weak confirming evidence become highly irregular when this weak evidence is censored. These types of problems are thus prone to superstitious beliefs. For example, in evaluating whether a rare circumstance C has a positive influence on the probability that a rare event E occurs ($\theta = 1$), or no influence ($\theta = 2$), the only event delivering strong evidence is $C - E$, and it favors $\theta = 1$.[13]

Framing and pooling. Finally, we use our framework to discuss the importance of framing (i.e., which alternative theories θ are compared) and

[11] One intuition is that given the limited number of mental states, posterior Bayesian beliefs differ substantially from one another (across mental states). Changing state after a poorly informative signal triggers a change in posterior that seems unjustified. In most cases, avoiding these unjustified changes is welfare increasing. In some rare cases however, for example when an extreme mental state is very likely under both θ, adding noise may increase the informativeness of the mental system, as we further explain in Section 4.1.

[12] This is because $\frac{p-\Delta}{1-2\Delta} > p$ when $p > 1/2$.

[13] Since E is rare, $C - \overline{E}$ cannot be very informative, and when C is rare too, θ cannot affect much the occurence of $\overline{C} - E$.

pooling (i.e., how information or signals are structured) in fostering biased beliefs. For a Bayesian, neither framing nor pooling alter the direction of learning: beliefs on average lean towards the truth. For our less sophisticated agent, both framing and pooling may affect which signals remain strong enough evidence and get processed, possibly pushing beliefs away from the truth.[14,15]

In a similar vein, we discuss the effect of processing signals in batches rather than sequentially as they arrive. With large enough batches, problems become regular, so infrequent processing likely reduce biases.

In summary, while the incentives to ignore weak evidence seem unavoidable, we conclude that some problems are more prone to superstitions than others because of asymmetries in strength of evidence for and against them. This being said, people with a better understanding of the inherent biases of the data generating process will be less prey to these biased beliefs, exerting some form of skepticism, either by reducing the number of mental states, or, when stakes are high, attributing less power to their mental system.

1.3 Discussion of modeling assumptions.

Our assumptions regarding belief formation depart from typical decision models in several ways. First, we model agents who form beliefs in a world where the data generating process is not fully observable. In many models, the source of uncertainty is limited (for example to few underlying states) and learning the true state is often just a by-product of the law of large numbers. The law of large number however will not be particularly helpful if the set of possible data-generating processes is rich, as one may not be able to simultaneously identify the underlying state and the distribution that generates the signals.

Second, we think of belief-formation as a strategy to emphasize that how people form beliefs is a challenging task. Classic approaches take Bayes rule as a benchmark. But this rule is tuned to a well-defined data generating process. When the data generating process is unknown, how does an agent

[14] This includes censoring, that is, pooling some a priori informative signals with the many instances where signals are absent. For a Bayesian, this "missing data" event would become informative, while for our agent, it would typically be too weak to be processed.

[15] For example, assume that C is not rare but $\overline{C} - E$ and $\overline{C} - \overline{E}$ are pooled. Then the only potentially discriminating events are $C - E$ and $C - \overline{E}$. Based on these events only, a Bayesian's belief would lean towards the truth, while if E is rare enough,beliefs of our agent censoring weak evidence would lean towards $\theta = 1$ independently of the underlying state.

come up with an adapted belief-formation rule? While the classic Bayesian route remains a technically feasible modelling option,[16] we choose to define an a priori plausible family of belief-formation rules/strategies, parameterized by a one dimensional parameter d, and to assume that, given the distribution over problems faced, agents manage to adjust d in a direction that improves expected welfare.[17]

The particular restriction (P) chosen is made for pedagogical reasons: first it coincides with Bayesian updating in some special cases where the data generating process is known, second it offers a simple way to characterize the influence of mental processing on posterior beliefs. In addition, many of the insights presented in this paper, including the incentives to censor weak evidence, do not depend on the particular family chosen (nor on the fact that d is set optimally), but just on the fact that beliefs are monotone in s.

1.4 Related work.

We mentioned the instrumental value of beliefs and the confirmation bias as two plausible explanations for the persistence of superstitions. Our explanation is not inconsistent with these. We argue that some discrimination problems are more prone to superstitions than others, which also implies that for these problems, some instrumental or confirmation value will be more easily derived.

Another explanation for superstitious beliefs is due to Chapman and Chapman (1967), who coined the term "illusory correlation". Chapman and Chapman run an experiment in which subjects are presented associations of two words (sequentially), and then later asked about the most frequent pairs. Subjects tend to overweight the presence of per-existing natural associations such as "lion-tiger". In other words, the presence of "natural associations" biases the judgment about the existence of correlations in the

[16] This route involves simultaneous learning about the underlying state and the data-generating process, hence it is rather cognitively demanding for the agent and the analyst.

[17] We do not model how this adjustment is made, though reinforcement learning or evolution is a natural candidate. In that respect, this follows one of the classic route in game theory which keeps unmodelled how players come up best responses. In any event, learning within a simple family of strategies is certainly easier than if no restrictions were put on the set of feasible belief-formation strategies. This feature of the model is inspired from CP (2019), which is more generally concerned with modelling agents dealing with complex environment, making sure that behavior is not too finely tuned to details of the model that agents cannot plausibly know. The family (P) can be viewed as an attempt to address this concern in a context where agents process and aggregate a large number of signals prior to decision making.

data. Kahnman and Tversky (1973) see this an example of the availability heuristics. The "lion-tiger" pair being more natural, it is readily available in the brain and becomes over-weighted when one tries to estimate ex post its occurence in the data (consisting in a list of paired words).

In a similar vein, one could argue that "moon affecting deliveries" is a natural association (the moon affects tides, why not a woman's womb), and that as a result these events get over-represented in people's mind. We provide an informativeness-based story for this over-representation: the conjunction "full-moon and many deliveries" is more easily recorded or recalled than other events because of an *informativeness asymmetry*.

From a theory perspective, our paper is related to Robert Wilson's critique, who argues that economic theories or mechanisms build on potentially fragile ground, with optimal mechanisms tuned to details of the economic environment that the mechanism designer cannot plausibly know. A similar critique holds for agents finely adjusting strategies to details of a model they cannot plausibly know. We address this critique by considering a richer-than-usual economic environment (the signal-generating process is unknown) and by keeping the number of strategic instruments limited (β and d are the only two instruments). As a consequence, our agent is unable to adjust its belief formation strategy to each particular data-generating process.

Our model itself is closest to Andrea Wilson (2014)'s work (as well as Compte and Postlewaite (2009)), with an agent choosing an action after receiving a (random) number of signals. The issue in Wilson is the optimal use of a limited number of states, which includes the optimal design of transition probabilities between states, conditional on the signal received. When a long sequence of signals is available (as in Cover and Hellman (1970)), the optimal use of signals consists in organizing moves as in Figure 1, focusing only on the most informative signal confirming $\theta = 1$ (for moves to the right) or $\theta = 2$ (for moves to the left), and dealing with the asymmetry in strength of evidence by adjusting the probability of moving away from an extreme state. In that model, weak evidence is thus ignored (though what is considered weak depends on the direction of evidence), and the asymmetry between the frequencies of moves to the right and left are corrected by an appropriate choice of transition probabilities at extreme states. Both of these features (i.e. contingent censoring and contingent moves at extreme states) rely on precise knowledge of the distribution over signals, which we do not assume.

The role played by signals of asymmetric informational strength echoes

some insights of the *mental accounting literature*. In comparing two alternatives A and B, agents may need to process many signals related to the benefits or drawback of taking A over B, and decision anomalies may arise when few, say, positive gains are compared with numerous yet small losses that each seem negligible and eventually ignored, tilting the decision in favor of the one yielding the large positive gains.

The classification of signals into those that are informative enough to qualify as evidence for a particular state echoes some notion representativeness, which we compare to Kahneman Tversky's (1973).

Finally, we contrast our work with the literature that explain biases through agents forming beliefs based on a misspecified (or incomplete) model of the environment (Esponda and Pouzo 2016, Spiegler 2016, for example).

2 The model

2.1 Preferences and uncertainty

We consider a family of decision problems, each having a similar structure: there are two possible *states of the world* $\theta = 1, 2$ and after processing a sequence of signals, the agent eventually chooses between two *alternatives*, $a \in A = \{1, 2\}$. We assume the following *payoff matrix*, where $g(a, \theta)$ is the payoff to the agent when she takes action a in state θ:

$g(a,\theta)$	1	2
1	$1-\gamma$	0
2	0	γ

where $\gamma \in [0, 1]$. When $\gamma > 1/2$, taking the right decision is more important when the state is 2 than in state 1, and the ratio $\Gamma = \gamma/(1-\gamma)$ characterizes that relative importance.

We assume that $\theta = 1$ for a fraction π of the problems. π thus characterizes some *objective uncertainty* about the true state, and we let $\rho = \pi/(1-\pi)$ denote the odds ratio. This objective uncertainty does not necessarily coincide with the agent's *initial/prior belief*: we allow for some discrepancy between the objective uncertainty π and the agent's initial perception of it. Formally, we denote by $\widetilde{\pi}$ the agent's initial belief that the state is 1 (or, expressed in odds ratio, $\widetilde{\rho}$) and assume a stochastic relationship between $\widetilde{\rho}$ and ρ:

$$\widetilde{\rho} = \eta\rho$$

where η is a positive random variable.[18] When η is concentrated on 1, the agent has correct priors.

For any belief $\widehat{\pi}$ about state 1 that the agent might hold upon taking a decision, we assume that the agent chooses the welfare maximizing action given this belief, i.e., choose action 1 when $\widehat{\pi}(1-\gamma) > (1-\widehat{\pi})\gamma$, or equivalently, denoting $\widehat{\rho} \equiv \widehat{\pi}/(1-\widehat{\pi})$ the odds ratio, when

$$\widehat{\rho} > \Gamma \tag{1}$$

Thus, in the absence of any signals to be processed, the agent chooses action 1 when $\widetilde{\rho} > \Gamma$, hence on average across realizations of $\widetilde{\rho}$, he obtains:[19]

$$\underline{W} \equiv (1-\gamma)\pi \Pr(\widetilde{\rho} > \Gamma) + \gamma(1-\pi)\Pr(\widetilde{\rho} < \Gamma)$$

In case the agent has correct priors, he achieves an expected welfare equal to

$$\underline{W}_0 \equiv \max(\pi(1-\gamma), (1-\pi)\gamma) \geq \underline{W}$$

2.2 Signals

Prior to decision making, the agent receives a sequence of signals imperfectly correlated with θ that she may use to form a posterior belief. The sequence is denoted X, assumed to be arbitrarily long, and conditional on the true state θ, each signal $x \in X$ is drawn independently from the same distribution with density $f(\cdot \mid \theta)$, assumed to be strictly positive and smooth on its support $[0,1]$. In addition, the odd ratio

$$L(x) \equiv f(x \mid 1)/f(x \mid 2)$$

is assumed to be strictly increasing in x. The distributions $\{f(\cdot \mid \theta)\}_\theta$ are problem specific and we think of them as objective characteristics of the problem faced.

One aspect of our analysis will be the possibility that a signal does not get to the agent's attention, or that it is simply not processed, for example because its *strength* is too weak, i.e., not informative enough. Another aspect will be that even when a signal is processed, its informative value is not perfectly assessed.

Signal characteristics. When signal x arises, there is a state $\overline{\theta}(x) \in$

[18] In simulations to come, we assume that $\log \eta$ is normally distributed.

[19] Note that whether we assume that the agent has noisy peceptions of Γ or noisy perceptions of ρ, the consequence regarding expected welfare has a similar expression.

$\{1,2\}$ that has highest likelihood, namely:

$$\overline{\theta}(x) = \arg\max_{\theta\in\{1,2\}} f(x \mid \theta).$$

We say that signal x is evidence for state $\overline{\theta}(x)$.[20] To measure the *strength of the evidence*, we define

$$l(x) = \frac{f(x \mid \theta = \overline{\theta}(x))}{f(x \mid \theta \neq \overline{\theta}(x))} = \max(L(x), 1/L(x)).$$

Any signal x thus has an "objective" (informational) characteristics

$$h \equiv (\overline{\theta}, l).$$

We shall denote by H the sequence of characteristics associated with the sequence X.

Censoring. We shall later endogenize the incentives to censor weak evidence. For now, we define an exogenous threshold strength $1+\beta$ above which the signal gets to the agent's attention, and we denote by $\widetilde{X}_\beta$ the subsequence of signal actually processed:

$$\widetilde{X}_\beta = \{x \in X, l(x) \geq 1+\beta\}$$

where $\beta \geq 0$ characterizes the degree to which weak signals are censored or go unnoticed.

We assume the agent takes a decision after N signals have been processed, and unless otherwise mentioned (i.e., in Section 6.1), we consider the limit case where N is arbitrarily large.

Noisy perceptions. We allow for the possibility that the signal characteristic h is not accurately or fully perceived by the agent. We assume that the agent perceives

$$\widetilde{h} = (\widetilde{\theta}, \widetilde{l})$$

imperfectly correlated with $h = (\overline{\theta}, l)$. We let $\widetilde{H}$ denote the sequence of perceived characteristics associated with sequence $\widetilde{X}_\beta$. Although we briefly discuss this general case, we shall mostly focus on a *coarse* version of the model where the agent only processes the perceived direction of evidence $\widetilde{\theta}$.

[20] Given our assumption on L, there is a unique uninformative signal x_0 (i.e., $L(x_0) = 1$): all signals above x_0 provide evidence for $\theta = 1$, and all signals below x_0 provide evidence for $\theta = 2$.

In summary, to an outsider, a decision problem is characterized by γ, θ, f and the realized sequence of signals X (induced by θ, f). The agent faces a family of such problems, characterized by a distribution ω.[21] Upon deciding, the agent observes γ and has noisy perception $(\widetilde{\rho}, \widetilde{H})$ which he can use to form a belief about θ and then decide which alternative he takes. We now turn to belief formation.

2.3 Belief formation and welfare.

We are interested in how beliefs are formed and in the performance of belief-formation strategies σ that map the perception $(\widetilde{\rho}, \widetilde{H})$ to a posterior belief $\widehat{\rho} = \sigma(\widetilde{\rho}, \widetilde{H})$, assuming that the agent eventually chooses actions according to (1). For any fixed f and β, expected welfare is given by

$$W(\sigma, f, \beta) = (1-\gamma)\pi \Pr_{f,\beta,\sigma}(\widehat{\rho} > \Gamma) + \gamma(1-\pi)\Pr_{f,\beta,\sigma}(\widehat{\rho} < \Gamma) \qquad (2)$$

We shall be interested in the performance of σ and β on average over the possible realizations of f, that is:

$$W(\sigma, \beta) = E_f W(\sigma, f, \beta)$$

Before putting structure on the belief-formation strategies σ that we shall consider, let us clarify how we depart from the classic approach to information aggregation and motivate why we put structure on them.

2.4 The classic approach to information aggregation.

In the classic approach to information aggregation, signals are not censored ($\beta = 0$), the agent has correct priors ($\widetilde{\rho} = \rho$) and knows f. For any sequence X, she correctly perceives the sequence $H = (\overline{\theta}(x), l(x))_{x \in X}$ and computes

$$L(H) \equiv \prod_{x \in X} L(x) = \prod_{(1,l) \in H} l \quad / \quad \prod_{(2,l) \in H} l \qquad (3)$$

and form a (Bayesian) posterior

$$\sigma^*(\rho, H) = \rho L(H) \qquad (4)$$

Next, following (1), the agent undertakes action 1 when $\sigma^*(\rho, H) > \Gamma$.

[21] The marginal over θ is π, and θ and f are independently distributed.

One may view σ^* as an algorithm for aggregating signals – the *Bayesian algorithm*. As is well-known, this Bayesian algorithm achieves the best possible expected welfare for the agent. As is also well-known, signals on average improves welfare, and with a large enough number of signals, the agent eventually takes the correct decision.[22]

While it is standard to assume that agents would form beliefs in the way described above, the nice welfare properties of this belief-formation algorithm rely on agents perfectly assessing $h \equiv (\overline{\theta}, l)$. When perceptions of strength are biased, for example, applying the Bayesian algorithm to $(\widetilde{\rho}, \widetilde{H})$ may result in poor decision making.[23]

2.5 Putting structure on belief-formation rules.

As analysts, we often motivate the use of strategies by their welfare performance – a classic reinforcement learning argument. Our view is that the Bayesian algorithm defines a strategy for aggregating signals, and one motivation for this rule is that it generates maximum welfare. When the agent just has noisy perceptions $\widetilde{\rho}$ and $\widetilde{H}$, applying naively the Bayesian algorithm to these perceptions may hurt welfare and an "appropriately" adjusted strategy $\sigma(\widetilde{\rho}, \widetilde{H})$ may be called for. But given the size of the strategy space, it is not clear how one can invoke reinforcement learning arguments to motivate the use of a welfare-maximizing strategy $\sigma(\widetilde{\rho}, \widetilde{H})$.

The route we propose below is to put a priori structure on the set of belief-formation rules, hopefully making the reinforcement learning argument more compelling to justify the use of particular rules over others. This is what we do next, first defining a coarse mental system that only exploits the perceptions $\widetilde{\theta}$.

2.6 Coarse mental system and belief formation

We consider agents attempting to form beliefs based on the direction of the evidence $\widetilde{\theta}$ only. This can be either because the agent has no perception of strength, or because the agent thinks that $\widetilde{l}$ is not reliable enough. $\widetilde{H}$ thus consists of a realizations of $\widetilde{\theta}$ favoring either 1 or 2, and the agent faces two

[22] This is because $E(\log L(x)|\theta = 1) > 0 > E(\log L(x)|\theta) = 2$. See Appendix.

[23] If the strength of evidence is perceived with a bias (say $\widetilde{l} = \eta l$ with $\eta > 1$) for example, if the agent forms beliefs by applying the Bayesian algorithm to $(\widetilde{\rho}, \widetilde{H})$, and if the expected tally of evidence for 1 and against 1 is positive independently of the underlying state, the bias η may become the preponderant force, sending beliefs towards $\theta = 1$ even when the state is $\theta = 2$, hence away from the true state. In these cases, the agent would be better off just using her priors rather than attempting to exploit signals using Bayes rule.

issues: (i) how to aggregate these coarse signals? (ii) Given this aggregation, what posterior belief should he hold?

To answer (i) we posit the *simple mental system* described in Introduction. The agent starts at $s = 0$, moving one step up (if possible and) if $\widetilde{\theta} = 1$, moving one step down (if possible and) if $\widetilde{\theta} = 2$, where $s \in S \equiv \{-K, .., 0, ...K\}$.

To answer (ii), we assume that when in state s prior to decision making, the agent uses a *simple belief-formation strategy*:

$$\sigma^d(\widetilde{\rho}, s) = \widetilde{\rho} d^s$$

for some $d \geq 1$. When $d > 1$, a positive (negative) mental state thus moves the agent away from her prior, towards believing that $\theta = 1$ ($\theta = 2$). The parameter d captures the degree to which the agent's mental system influences beliefs. When $d = 1$, the agent keeps her prior and effectively ignores all signals received.

For the sake of exposition, we also introduce a more *sophisticated updating rule* in which the agent updates his initial probabilistic belief $\widetilde{\rho}$ according to

$$\sigma^{d,\Lambda}(\widetilde{\rho}, s) = \widetilde{\rho} \Lambda d^s.$$

With rules of this kind, the agent has two instruments: the degree d to which her mental system influences beliefs, and the degree to which she biases her decision: if her mental system tends to generate higher mental states on average, she will have an incentive to choose Λ below 1.

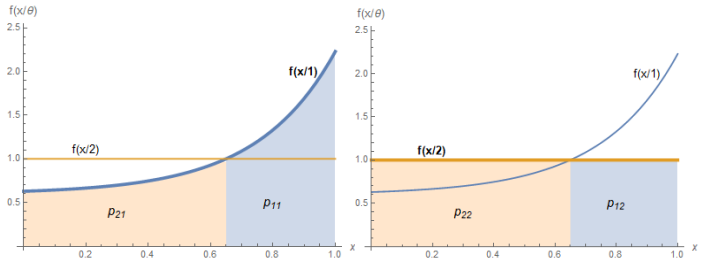
3 Optimal belief formation

For a given (f, β), the mental-state dynamic is entirely driven by the transition probabilities

$$p_{\widetilde{\theta}\theta} = \Pr(\widetilde{\theta} \mid \theta, f, x \in \widetilde{X}_\beta), \text{ for } \theta \in \{1, 2\} \text{ and } \widetilde{\theta} \in \{1, 2\},$$

and since $p_{1\theta} + p_{2\theta} = 1$, the dynamic is actually driven by only two parameters, say p_{11} and p_{22}. Welfare is thus jointly determined by σ and $p \equiv (p_{11}, p_{22})$, and for notational convenience, we shall now write $W(\sigma, p)$ for welfare. We provide below a graphic representation of the probabilities $p_{\widetilde{\theta}\theta}$ for a given f, assuming $\beta = 0$ (no censoring) and correct attributions

$\widetilde{\theta} = \overline{\theta}$:[24]



$\theta = 1$ $\theta = 2$

Throughout the paper, we consider problems where observing evidence for state θ is more likely under state θ than under state $\theta' \neq \theta$, that is, $p_{\theta\theta} > p_{\theta\theta'}$. For a given $p = (p_{11}, p_{22})$, this condition can be written

$$d_\theta^p \equiv p_{\theta\theta}/p_{\theta\theta'} > 1 \tag{C}$$

and it holds by construction when $\beta = 0$ and when the agent has correct perceptions of evidence ($\widetilde{\theta} \equiv \overline{\theta}$).[25] Note that when (C) holds either $p_{11} > 1/2$ or $p_{22} > 1/2$.

3.1 p-optimal strategies

We start studying optimal belief-formation strategies for the case where priors are correct ($\widetilde{\rho} = \rho$) and strategies can be tuned to p. This corresponds to a classic Bayesian case, under the constraints imposed by censoring and mental processing. For any p, define

$$d_p = d_1^p d_2^p \text{ and } \Lambda_p = \sum_{s\in S} (d_2^p)^s / \sum_{s\in S} (d_1^p)^s$$

We have:

Proposition 1: *Assume the agent has correct priors. For any fixed pair p satisfying (C), the strategy $\sigma_p^* \equiv \sigma^{d_p,\Lambda_p}$ achieves maximum welfare across all possible belief-formation rules σ.*

[24] In each figure ($\theta = 1$ or 2), the blue area corresponds to the probability of moving upward ($\overline{\theta} = 1$).

[25] The assumption also holds mechanically when there is noise in the perception of $\overline{\theta}$ and the noise is independent of θ.

Proof. Given p, define $\phi_\theta^p(s)$ as the long-run distribution over belief states under θ. Let $\Lambda^p(s) = \phi_1^p(s)/\phi_2^p(s)$. The agent's welfare is at most equal to

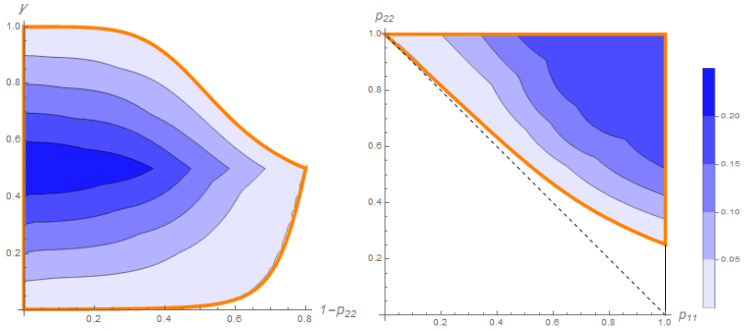
$$\overline{W}(p) = \sum_{s\in S} \max((1-\gamma)\pi\phi_1^p(s), \gamma(1-\pi)\phi_2^p(s))$$

and the rule $\sigma^*(\rho, s) = \rho\Lambda^p(s)$ permits to achieve this maximum.[26] It is easy to check $\Lambda^p(s) = \Lambda_p d_p^s$, so when $\widetilde{\rho} = \rho$, σ_p^* coincides with σ^* and thus achieves maximum possible welfare.

To assess the magnitude of these gains, we assume $K = 2$ (five belief states) and compute numerically the welfare gains

$$\Delta(p) \equiv W(\sigma_p^*, p) - \underline{W}$$

associated with σ_p^* compared to only relying the prior. With correct priors, the agent cannot be worse off using σ_p^* so $\Delta(p)$ is non negative. Since beliefs states are correlated with the underlying state, the agent may obtain a strictly higher welfare when this correlation is strong enough. Fixing $\pi = 1/2$, the following figures report the domain of strict welfare gains when $p_{11} = 0.8$ (on left) and when $\gamma = 0.6$ (on right), as well as the magnitude of these gains.

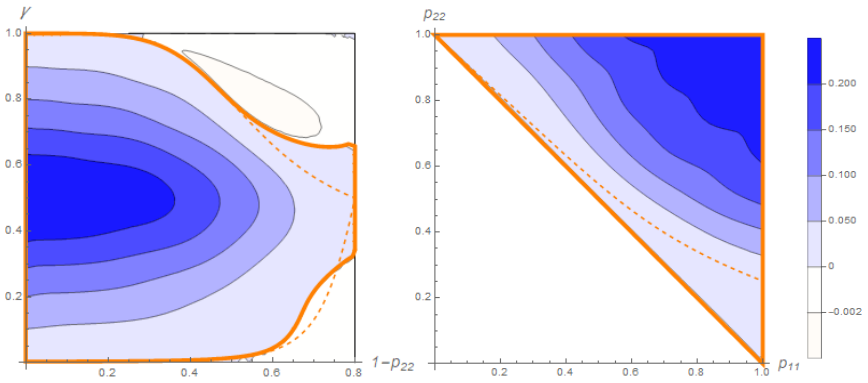


The orange line defines the boundary of the domain for which $\Delta(p)$ is strictly positive. Outside this domain, the decision maker plays the same action irrespective of her mental state: the mental system is not informative enough to tilt the decision away from what the prior suggests. This happens when $1 - p_{22}$ is too close to p_{11}, as the informativeness of the mental system is

[26] This rule corresponds to Bayesian updating given the constraint imposed by simple mental processing

then small (i.e., d^* is close to 1) or when γ lies away from 1/2, as it then requires substantial evidence to override the prior.

Proposition 1 corroborates the standard view that processing signals correlated with the underlying state cannot hurt decision making. With noisy priors, σ_p^* is not the welfare optimizing rule. Nevertheless, the following figures show that comparable gains obtain in that case as well:



Intuitively, when priors are noisy, there are two effects at work: (i) relying on priors is a worse option than before ($\underline{W} < \underline{W}_0$); (ii) for a fixed (π, γ), the change in belief required to switch decision is modified.

(i) implies an expansion (in most directions) of the set of parameters for which σ_p^* helps compared to relying on priors only, as well as comparable welfare gains for most parameters (ii) implies that (for a small range of parameters), the agent may be (slightly) worse off using the mental system than his (noisy) prior. For these parameters, the agent has the illusion that the mental system is powerful enough to override the prior, while he would be better off ignoring the mental system.[27]

3.2 When p is unknown

We now explore the case where priors are noisy and the updating strategy cannot be tuned to p. We shall make two observations. First, the simple updating strategy $\widetilde{\rho} d^s$ works well for many problems, *even when d cannot be tuned to* (p_{11}, p_{22}). For some problems however, the agent would be better off ignoring signals and only trusting her prior.

We examine the welfare gain (or loss) $\Delta^d(p) = W(\sigma^d, p) - \underline{W}$ associ-

[27] When priors are noisy, σ_p^* is not the welfare optimizing rule. This is why $\Delta(p)$ can be negative.

ated with using a given strategy σ^d. We first observe that for a range of parameters p, Δ^d is positive irrespective of d. Formally, let

$$B = \{p, d_p > \max(\frac{\Gamma}{\rho\Lambda_p}, \frac{\rho\Lambda_p}{\Gamma}) \tag{5}$$

Proposition: For any $p \in B$, $\Delta^d(p) > 0$ for all $d > 0$.

The set B corresponds to cases where the flow of evidence remains somewhat balanced (Λ_p not too far from 1) and signals are sufficiently informative (d large enough).[28]

Formally, Δ^d can be expressed as

$$\Delta^d(p) = \textstyle\sum_s \psi^p(s) J^d(s)$$

where

$$\begin{aligned} \psi^p(s) &= (1-\gamma)\pi\phi_1^p(s) - \gamma(1-\pi)\phi_2^p(s) \text{ and} \\ J^d(s) &= \Pr(\mu \geq \Gamma/(\rho d^s)) - \Pr(\mu \geq \Gamma/\rho). \end{aligned}$$

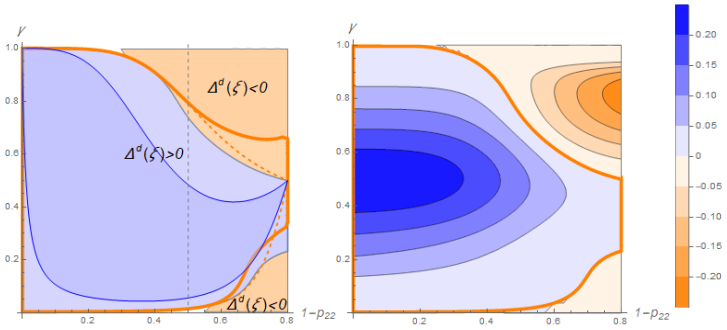
For $d > 1$, $J^d(s)$ has the same sign as s, and by construction, for $p \in B$, $\psi^p(s)$ also has the same as s, which proves the proposition.[29]

For a fixed d, Δ^d can be positive even if welfare gains are not positive for each belief state: mostly matters states that are more likely to be reached, so welfare may increase over a range much larger than B. In contrast to the Bayesian case however, there is now a significant range of parameters for which σ^d hurts welfare. The reason is that for some p, the mental system may generate evidence towards $\widetilde{\theta} = 1$ irrespective of the underlying state, and σ^d does not correct for that. The figures below summarizes these observations assuming $d = 3$ and $p_{11} = 0.8$.[30]

[28] The condition is more easily satisfied when stakes and priors are favoring too much a given alternative.

[29] Note that when the number of mental states rises, Condition (5) becomes a more stringent one. The reason is that when the number of mental states rises, being in state 0 can become quite informative if p_{11} and p_{22} differ (i.e., $\max(\Lambda_p, 1/\Lambda_p)$ becomes large).

[30] As before, we fix $\pi = 1/2$, $p_{11} = 0.8$, $K = 5$, so problems are parameterized by (p_{22}, γ). Priors are noisy with $Log\mu \sim \mathcal{N}(0, 0.5)$.



The orange boundary defines the domain for which σ^{d_*,Λ_*} improves decision making. In the left figure, the blue domain indicates the set of problems for which σ^d improves decision making, and the darker blue region defines domain B. In the rest of the domain, using σ^d decreases welfare, and the right figure provides the magnitude of these changes: the magnitude of the losses can be as large as that of the gains.

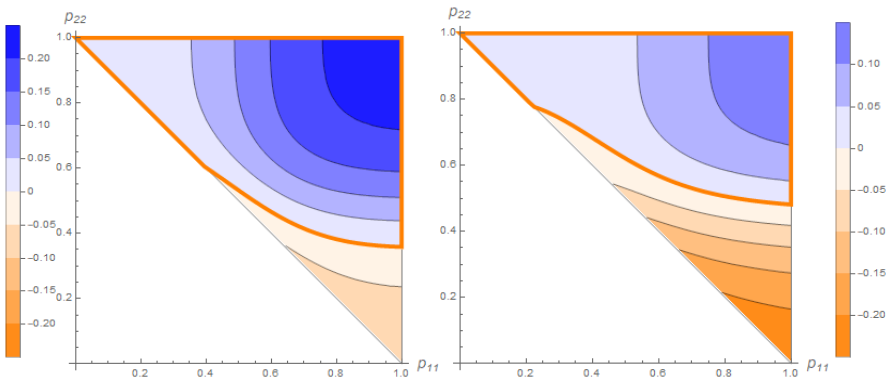
In the absence of optimization tuned to (p_{11}, p_{22}), one expects that agents end up with mistaken beliefs for some (irregular) problems: they will form a posterior belief that bents towards one state of the world, mistakenly thinking that their mental system permits to discriminate well between states of the world, while their belief state primarily results from the fact that evidence on average points towards the same direction irrespective of the underlying state.

3.3 A motive for stake-contingent skepticism.

To conclude this Section, we observe that in contrast to the classic Bayesian formulation that separates belief formation and preferences, the issues become intertwined with our less sophisticated agent: the agent has incentives to decrease d when the stakes γ are larger.

To see why, we fix again $d = 3$ and compare the magnitudes of gains and

losses for $\gamma = 0.6$ and $\gamma = 0.75$ across all possible p:



Incentives to set d depends on the distribution over problems p faced, but it should be clear for the figure that when γ is high, losses become preponderant, thus providing the agent to decrease d and give a more prominent role to priors.

Intuitively, the agent faces two kinds of problems: some for which evidence is somewhat balanced (i.e., Λ_p is not too far from 1) and some for which evidence on average points in a given direction $\widehat{\theta}$ independently of the state θ. When the agent ends up in a high mental state, this is evidence in favor of $\theta = 1$ for the first set of problems, but this is evidence for $\widehat{\theta} = 1$ for the second set of problems.

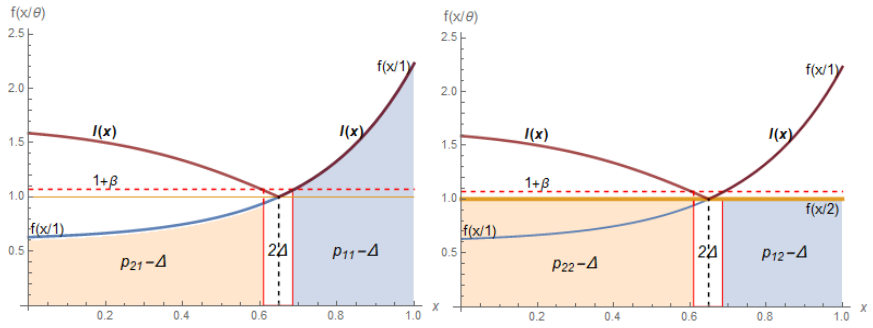
When $\Gamma/\rho = 1$, relying on priors gives the lowest possible welfare, so, for the second set of problems, being erroneously influenced by the mental state is not costly. When Γ/ρ is large however, this influence is costly: for the second set of problems, and if $\widehat{\theta} = 1$, the agent is mislead into thinking that $\theta = 1$ while he would have been better off following priors.

In other words, for asymmetric-stake cases, the agent may benefit from being more cautious and exert some stake-contingent skepticism, which can be done by reducing d when stakes are higher.

4 Incentives to ignore weak evidence

Ignoring weak evidence modifies the distribution of signals processed, hence also the transition probabilities (p_{11}, p_{22}) between mental states. Starting from a situation where signals are not censored ($\beta = 0$), we study the welfare consequence of marginally censoring weak evidence. Graphically, we illustrate below why and how transition probabilities are affected by

censoring, for a small value of β:



$\theta = 1$ $\theta = 2$

Because of censoring, only signals for which $l(x) > 1 + \beta$ are processed. In the figure, the chance a signal is not processed is approximately 2Δ. One key observation is that when β is small, the weak evidence censored is equally likely to favor $\theta = 1$ or $\theta = 2$, implying that both p_{11} and p_{21} are reduced by Δ.[31] Said differently, processing weak evidence is equivalent to adding state-independent noise to the mental system.

This observation implies (see Appendix):

Proposition: *At* $\beta = 0$, (i) $\frac{\partial d_p}{\partial \beta} > 0$;(ii) $\frac{\partial p_{kk}}{\partial \beta}$ *has the same sign as* $p_{kk} - 1/2$, *and (iii)* $\frac{\partial \Lambda_p}{\partial \beta} > 0$ *iff* $\Lambda_p > 1$

So when weak information is censored, d_p rises, which implies that from a Bayesian perspective, the spread in posterior beliefs is larger. But it also implies that for most problems (i.e., unless $p_{11} = p_{22}$), Λ_* lies further away from 1, that is, the mental system is less balanced.

As we shall show, the consequence of the larger spread is that, in the Bayesian case (where the agent can tune his strategy to p), censoring weak evidence improves welfare for most values of p.[32] The consequence of the reduced balancedness of the mental system is that in the non-generate case where the agent follows a given strategy σ^d (which does not correct for this imbalance), censoring weak evidence hurt welfare for many problems.

Nevertheless, we will show that σ^d improves welfare for *all "regular prob-*

[31] These observations rely on our assumption that L is smooth and strictly increasing.
[32] For most pairs, but not for all, as we shall explain.

lems", that is, problems for which

$$p_{11} > 1/2 \text{ and } p_{22} > 1/2 \tag{R}$$

One may thus conclude that to the extent that regular problems are preponderant, incentives to censor weak evidence are present even when the agent cannot finely tune his belief-formation strategy to p.

4.1 The Bayesian case

Since weak evidence is just adding noise to the transitions between mental state, it would seem that from a Bayesian perspective, censoring weak evidence always improves welfare. In particular, since d_p rises, the spread in posterior beliefs must increase, so the set of problem for which the mental system helps would seem to increase as well.

Also, the intuition that one might derive from Wilson is that there is a value to keep as much as possible past accumulated evidence, and this ought to provide agents with incentives to prevent moves from the edges once they are reached, rather than throwing in noise into the mental system.

It turns out however that, for a small set of problems, censoring weak evidence actually hurts welfare. Technically, letting $\overline{\Lambda}_p = \Lambda_p(d_p)^{\overline{s}}$, the reason is that the set

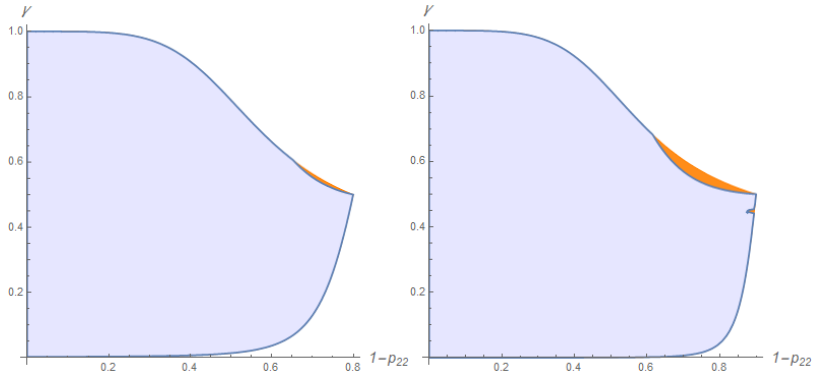
$$D = \{p, \frac{\partial \overline{\Lambda}_p}{\partial \beta} < 0\}$$

is not empty. That is, although censoring increases d_p, the largest possible shift in posterior beliefs decreases, so for these marginal problems where Γ/ρ is below but close to $\overline{\Lambda}_p$, censoring makes the mental system useless.

Intuitively, this happens for problems for which evidence points on average in the same direction (here $\widehat{\theta} = 1$), independently of the actual state of the world. Then censoring evidence reinforces that trend, and this makes being in the high mental state K potentially less informative: at the limit where p_{11} is close to 1 and p_{22} is small, the player likely ends up in the highest mental state K, irrespective of the underlying state. Being in the high state K is thus not very informative, and even less so when weak information is censored. In contrast, adding noise increases the informativeness of the most likely signal $s = K$, and for some γ's close to the posterior belief at $s = K$, the agent is better off not censoring.[33]

[33] Said differently, the ex ante long-run distribution over mental states can be viewed as a prior, and adding noise (or censoring) endogenously modifies the prior, making (for some p) learning easier (or more difficult).

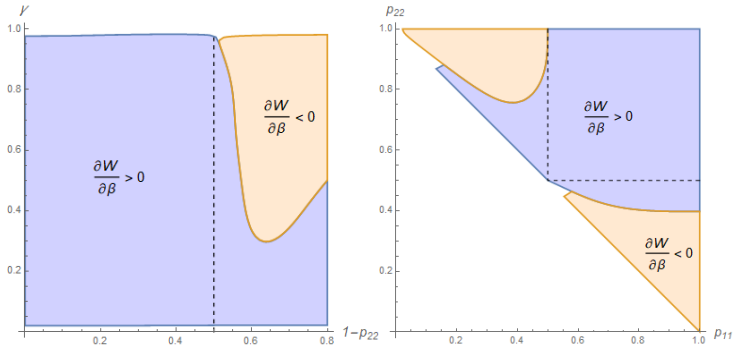
We illustrate this by plotting the locus of problems for which censoring help (blue) and hurts (orange) for two values of p_{11}.



$p_{11} = 0.8$ $p_{11} = 0.9$

4.2 When p is unkown

Ignoring weak evidence may further increase the imbalance of the mental system. This implies that for simple belief-formation strategies, which do not correct for this imbalance, welfare may decrease as the following figures confirm:[34]



The Figures show a negative effect of censoring on welfare for a signif-

[34] As in previous figures, on the left, we set $p_{22} = 0.8$ and $\pi = 1/2$, and examine whether welfare increases (blue) or decreases (orange) depending on parameters p_{22} and γ. On the right, we fix $\gamma = 0.8$, and examine variations in the (p_{11}, p_{22}) space.

icant range of problems. Nevertheless, we show below that for *all regular problems* (both $p_{\theta\theta}$ above $1/2$)[35], and in spite of the increased imbalance that censoring generates, welfare increases. Intuitively, the reason is that for these problems, ignoring weak evidence always increases both p_{11} and p_{22}, so it increases the correlation between the underlying state $\theta = 1$ (respectively $\theta = 2$) and being in a positive mental state (respectively negative mental state).

Formally, consider any monotone strategy σ and any realization $\widetilde{\rho}$. Under $(\sigma, \widetilde{\rho})$, the decision maker chooses action 1 if and only if the belief state is high enough, say $s \geq k_{\sigma,\widetilde{\rho}}$, and the welfare is given by $W(k_{\sigma,\widetilde{\rho}}, p)$ where

$$W(k,p) \equiv \pi(1-\gamma)\Phi_1^p(k) + (1-\pi)\gamma(1-\Phi_2^p(k)) \tag{6}$$

where $\Phi_\theta^p(k) \equiv \sum_{s\geq k} \phi_\theta^p(s)$ is the probability to end up in a mental state $s \geq k$ when the underlying state is θ.[36] From the definition of $\phi_\theta^p(s)$, and recalling that $d_\theta^p = \frac{p_{\theta\theta}}{1-p_{\theta\theta}}$, we have

$$\Phi_1^p(k) = \Phi(k, d_1^p) \text{ and } \Phi_2^p(k) = \Phi(k, 1/d_2^p)$$

where

$$\Phi(k,d) = \frac{\sum_{s\geq k} d^s}{\sum_{s\in S} d^s}$$

We have:

Lemma: *For any $k > -K$, with $k \leq K$, $\Phi(k,d)$ strictly increases with d.*

Since for regular problems, censoring weak evidence strictly increases $p_{\theta\theta}$ for both $\theta = 1, 2$, hence also d_θ^p, and we immediately conclude that $\Phi_1^p(k)$ increases and $\Phi_2^p(k)$ decreases, so welfare increases for any realization of $\widetilde{\rho}$, hence it also increases on average over realizations of $\widetilde{\rho}$.

Proposition: *For any monotone belief formation strategy σ and any regular problem, censoring weak evidence marginally increases welfare.*

[35] The frontiers of regular problems are indicated by the dashed line in the Figures. These problems lie within the blue (welfare improving) region.

[36] This means that, over realizations of $\widetilde{\rho}$, the agent obtains an expected welfare equal to $E_{\widetilde{\rho}} W(k_{\sigma,\widetilde{\rho}}, p)$.

5 The persistence of superstitions.

5.1 The role of asymmetries.

Our hypothesis is that it is difficult for agents to adjust censoring and the belief-formation rule to each problem that one faces, that is, to unobservable characteristics of the data generating process. We think of these kinds of adjustment as being more plausibly made on average across problems.

Given this hypothesis, the general message conveyed by previous Sections is that, to the extent that agents face a substantial fraction of regular problems, agents have incentives to both censor weak evidence ($\beta > 0$) and raise the power of their mental system ($d > 1$).[37]

The adverse consequence however is that for some problems, agents would be better off not trusting their mental process and ignoring the updating that it suggests. We illustrate below the type of problems for which this occurs.

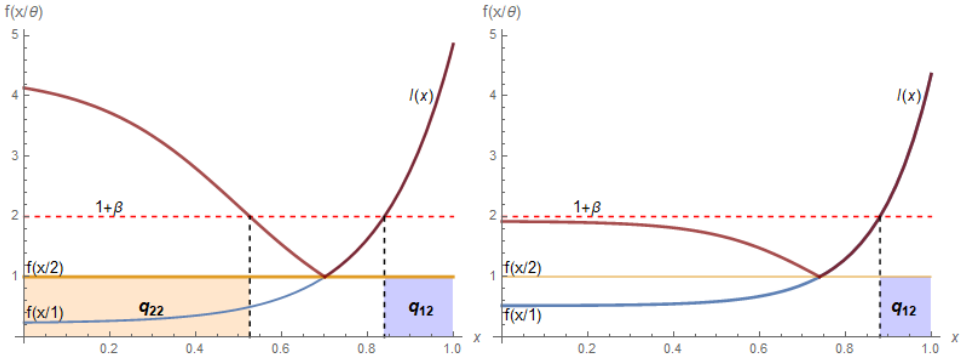
Let us first observe the consequence of raising β on (p_{11}, p_{22}) for two different distributions. The two figures below depict how one obtains the transition probabilities $q_{\overline{\theta},2} = \Pr(\overline{\theta} \text{ and } l > 1 + \beta \mid \theta = 2)$, which in turn permits to compute

$$p_{22} \equiv \frac{q_{22}}{q_{22} + q_{21}}.$$

In the left figure, whether signals provide evidence in favor of $\theta = 1$ or 2, the strength of evidence $l(x)$ is somewhat comparable, and p_{22} remains above $1/2$ (and both p_{11} and p_{22} actually rise). For the more asymmetric distribution $f(. \mid 1)$ considered on the right figure, the strength of evidence is less evenly distributed, and a similar rise in β sends p_{22} to 0: all signals processed are evidence in favor of $\theta = 1$.

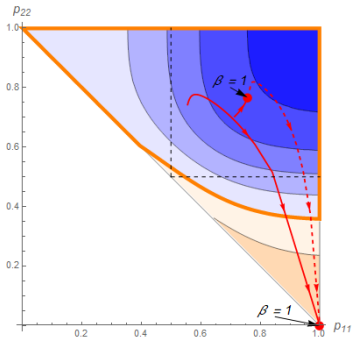
[37] Note that censoring weak evidence actually increases the benefits of raising d for regular problems.

Note also that in addition, more sophisticated agents may have incentives to exert skepticism with respect to their mental system whenever stakes are perceived as unbalanced ($L_0 < 1$ when $\Gamma/\widetilde{\rho} > 1$), but this skeptcism may be counter productive when $\widetilde{\rho}$ is too noisy.



"Symmetric" distribution Asymmetric distribution

The next figure plots, for each distribution, the effect of raising β above 0 on the pair (p_{11}, p_{22}) (see the red curves). The figure also recalls welfare levels as a function of (p_{11}, p_{22}) for γ set to 0.6.



That is, in the absence of censoring, each problem considered in regular (i.e., at $\beta = 0$, $q_{\theta\theta} > 1/2$ for each θ), and marginally censoring weak evidence generates a welfare gain. The consequence of more significant censoring differs across problems: for the "symmetric" distribution, at $\beta = 1$ (indicated by the red dot), a further increase in censoring still increases welfare; for the more asymmetric distribution, $\beta = 1$ already sends W to the worst possible welfare level (given $\gamma > 1/2$).

Ideally, the agent would wish to adjust censoring to each problem that he faces. When characteristics of the whole data generating process are not observable, performing this adjustment is difficult and this is a potential

source of biases, in particular for discrimination problems that occasionally generate strong evidence in favor of one alternative (here $\theta = 1$), and frequently generate evidence in favor of the other alternative ($\theta = 2$) but only weakly so.

5.2 Examples.

How does this related to superstition, superstitious beliefs, or more generally, folk beliefs? Our claim is that such beliefs typically arise for problems of the asymmetric kind described above. Leaving aside the question of the role of superstitions and why they arise, we wish to suggest that some theories are likely to give rise to persistent superstitions, when the asymmetry in strength between evidence that supports it and evidence that goes against it, combined with a general incentive to ignore weak evidence, may be responsible for beliefs in favor of these theories, independently of their actual veracity.

We first illustrate this with a common folk belief, a theory that the (full) moon has a positive influence over on the number of deliveries.[38]

Lunar effects. Consider an individual trying to discriminate between two states of the world. Under state 1, full moon generates on average a 20% increase in the number of deliveries, while under state 2, there is no effect. Assume that over the year, the average number of babies is 10, with each day a realized number n following a Poisson distribution. The hospital/staff is calibrated to handle 12 babies, and any number $n > 12$ creates tension. We call $X = \max(n - 12, 0)$ the level of tension, Y the event as to whether there is full moon ($Y = 1$) or no full moon ($Y = 0$).[39] A signal is a pair $x = (X, Y)$ and for each x we can compute the pair $(\overline{\theta}, l)$. The signals that are evidence in favor of $\theta = 2$ have the following strength:

X,Y	0,1	8,0	7,0,	6,0	5,0	4,0	3,0	2,0	1,0
l	1.31	1.22	1.20	1.17	1.15	1.13	1.10	1.08	1.06

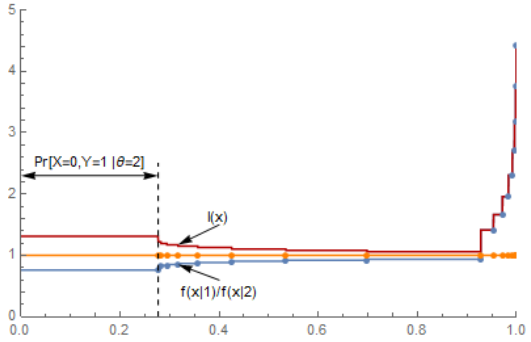
[38] The absence of lunar effects and beliefs in lunar effects have both been documented. The issue is not just anectodical, as if such an influence exists, hospitals would want to take it into account to modify work load as a function of moon phases.

[39] We assume that a full moon lasts 3 days out of 30, so under state 1, so if we denote by q_1 (respectively q_0) the expected number of deliveries on a full moon day (respectively on other days), we have $q_1 = 1.2\ q_0$ and $q_1 + 9q_0 = 10q$.

and the strongest of these is $(0,1)$ (no tension on a full moon day). Signal that are evidence in favor of $\theta = 1$ have the following strength:

X,Y	0,0	1,1	2,1	3,1	4,1	5,1	6,1	7,1	8,1
l	1.02	1.41	1.67	1.96	2.31	2.71	3.19	3.76	4.42

so, apart from signal (0,0), which is almost uninformative, they are strong compared to those in favor of $\theta = 2$. This can be also seen by plotting both distributions:[40]



Under the simple mental processing considered earlier, and if $\beta > 0.31$, evidence in favor of $\theta = 2$ is never processed, leading the individual to believe in a lunar effect independently of the true state. Note that a smaller threshold would not make the problem regular: for smaller values of the threshold β, evidence is preponderantly in favor of $\theta = 2$, *independently of the underlying state.*[41]

Illusory correlation and pattern identification. Superstitions and other folk beliefs can also be interpreted as an instance of illusory correlation (Chapman and Chapman) between two events, or more generally, the illusory identification of a pattern in the environment.

Formally, one can think of a pattern as a sequence of two events P and C where P is a premise and C a consequence. The issue is whether the premise makes the consequence more likely. Sometimes PC is observed, but at other

[40] The length of an horizontal segment indicates the probability of occurence of signal x under state $\theta = 2$) – and for lisibility, we omit the most likely signal $(0,0)$ and report distributions conditional on $x \neq (0,0)$.

[41] One can check that at $\beta = 1.31$, $q_{22} = 0.27$ and $q_{12} = 0.07$, so $p_{22} = q_{22}/(q_{22} + q_{21}) = 0.65$, and $q_{11} = 0.14$ and $q_{21} = 0.21$, so $p_{11} = 0.39$.

times, $\overline{P}C$ or $P\overline{C}$ or $\overline{PC}$ can also be observed.[42] As we explain below, when P and C are both rare events, the only event which has significant strength is the observation of PC and it favors the theory that an influence exists. To see this, assume that under state $\theta = 1$, an influence exits, while under state 2 it does not:

$$\begin{aligned} \Pr(C \mid P, \theta = 1) &= \alpha \Pr(C \mid \overline{P}, \theta = 1) \text{ and} \\ \Pr(C \mid P, \theta = 2) &= \Pr(C \mid \overline{P}, \theta = 2) \end{aligned}$$

with $\alpha > 1$. Denote by $r = \Pr(P)$ and $q = \Pr(C)$. Letting $q_1 = \Pr(C \mid P, \theta = 1)$ and $\overline{q}_1 = \Pr(C \mid \overline{P}, \theta = 1)$, we have $rq_1 + (1-r)\overline{q}_1 = q$, implying

$$\overline{q}_1 = \frac{q}{1 + (\alpha - 1)r} < q < q_1 = \frac{\alpha q}{1 + (\alpha - 1)r}.$$

This gives us the direction and strength of evidence $(\overline{\theta}, l)$ for each signal $x \in \{PC, \overline{P}C, P\overline{C}, \overline{PC}\}$:

	$\overline{P}C$	$P\overline{C}$	$\overline{PC}$	PC
$\overline{\theta}$	2	2	1	1
l	$\frac{q}{\overline{q}_1}$	$\frac{1-q}{1-q_1}$	$\frac{1-\overline{q}_1}{1-q}$	$\frac{q_1}{q}$

When P and C are both rare events (i.e., q and r small), $\overline{q}_1$ and q_1 are small as well, so $\frac{1-q}{1-q_1}$ and $\frac{1-\overline{q}_1}{1-q}$ are close to 1. In addition, $\frac{q}{\overline{q}_1} = 1 + (\alpha - 1)r$ remains close to 1 while $q_1/q = \frac{\alpha}{1+(\alpha-1)r}$ is comparable to α. It follows that the strength of PC is significantly higher than that of all other signals, and it favors theory $\theta = 1$. Of course, a proper weighting of all evidence along with the Bayesian aggregation rule should eventually lead individuals to avoid erroneous beliefs. However, under simple processing and if weak evidence is ignored, the mental system inevitably points towards high belief states (for which individuals are inclined to think that influence exists ($\theta = 1$).

5.3 Framing and pooling

For a Bayesian, the frequency with which updating occurs is irrelevant. Nor does it matter whether signals are pooled or not: to the extent that the distributions $f(\mid \theta)$ over signals are statistically distinguishable, a Bayesian learns the correct state. Which alternative θ' is the true state θ pitted against does not matter either. So long as θ is the true state, a Bayesian

[42] We denote by $\overline{P}$ the absence of a premise and $\overline{C}$ the absence of the consequence.

will learn it. Under our simple belief-formation assumption, the frequency of updating, how signals are pooled and how the problem is framed may all affect long-run beliefs.

Batch processing. Assume that instead of processing signals $x_1, .., x_n, ...$ sequentially (and updating the mental state after each one), signals are processed by batches of J signals, say $X_1 = (x_1, .., x_J)$, $X_2 = (x_{J+1}, .., x_{2J})$ etc.... For any given problem, if J is sufficiently large, then by the law of large number, most batches generated under θ are strong evidence in favor of θ, hence the problem becomes regular even if it was not regular under frequent processing. Dealing with batches of signals may of course be cognitively more demanding, but to the extent that the agent categorizes batches correctly (i.e., $\widetilde{\theta} = \overline{\theta}$) or with some errors but without introducing systematic biases, biased beliefs can be avoided. Conversely, this illustrates that frequent updating may contribute to biased beliefs.

Pooling signals. Another source of bias may come from the way signals are pooled. In our lunar effect example, the no-tension event $X = 0$ pools all events where the number of deliveries n is below or equal to 12. If these events were not pooled, and if a low n were processed on a full moon, then events $(n, 1)$ with low n could be processed, and this would be reasonably strong evidence if favor of the no-lunar effect hypothesis. So the way signals are pooled affects their strength, and, under our simple belief-formation rules, this affects long-run beliefs.

In the case of deliveries, we chose to pool all realizations $n \leq 12$ into $X = 0$. One justification would be that observing low n is difficult, as there are always programmed deliveries that makes the number of unprogrammed ones difficult to observe. But there may be other reasons. People tend to be looking for explanations for unlikely events that they observe, and the act of looking for explanations may be event dependent may affect which signals are actually recorded and/or processed.

For example, imagine that we do not even wonder whether there is a full moon (or an absence of full moon) when there is no tension (as we do not see a priori see the full moon as a plausible cause for lack of tension). This means that signals $(X, Y) = (0, 1)$ and $(0, 0)$ are pooled into $X = 0$. For a Bayesian that understands this selection process, this is not an issue, as X remains (weakly) informative, and in the long run, she would correctly assess that a lunar effect does not exist if there is none. For our less sophisticated agent, one can argue that signal $X = 0$ has only weak informative value, hence likely falls under the radar.[43]

[43] An alternative interpretation is that all events that seem irrelevant are pooled with

More generally, prior views may structure the signals that agents actually process, and differences in prior views thus shape the relative strength of evidence in favor of each alternative, hence eventually the persistence of erroneous beliefs, as well as the persistence of disagreement among people holding different prior views despite the presence of common signals.

Framing. To illustrate as simply as possible the effect of framing, assume that we draw a biased coin with a probability $p_0 = 0.7$ of showing a Tail (rather than a Head). Imagine that each signal is a draw and that we test this theory ($\theta = 0$, i.e., $p_0 = 0.7$) against the alternative theory $\theta = 1$ with $p_1 = 0.3$. Then the event T is evidence for $\theta = 0$, while H is evidence for $\theta = 1$, and belief states therefore point towards the correct state. In contrast, if the alternative theory is $p_1 = 0.8$, the agent will more frequently see evidence for $\theta = 1$ than against it, and could thus erroneously conclude that $\theta = 1$ is the more likely state.

If a signal is a sequence of draw rather than a signal draw, the issue persists so long as the sequence remains small enough, with a key role played by the censoring threshold β in shaping the long-run distribution over belief states.

6 Discussion and Extensions

6.1 Fewer signals

Our analysis has so far assumed an arbitrarily large number of signals. We discuss below the consequence of individuals only processing a limited number of signals.

Formally, call $\phi_\theta^{p,N}$ the distribution over mental states when θ is the underlying state and N the number of signals processed, and let $\Phi_\theta^{p,N}(s) = \sum_{k \geq s} \phi_\theta^{p,N}(k)$. Under $(\sigma, \widetilde{\rho})$, the decision maker chooses action 1 if and only if the belief state is above or equal to $k_{\sigma,\widetilde{\rho}}$ (as before – $k_{\sigma,\widetilde{\rho}}$ is unaffected by N), but the welfare obtained is now given by $W^N(k_{\sigma,\widetilde{\rho}}, p)$ where

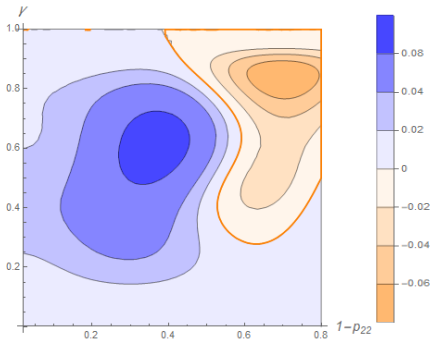
$$W^N(k,p) \equiv \pi(1-\gamma)\Phi_1^{N,p}(k) + (1-\pi)\gamma(1-\Phi_2^{N,p}(k))$$

Welfare thus changes, but since $\Phi_\theta^{N,p}$ converges "quickly" to Φ_θ^p, the change is limited (with enough signals). With 5 states and 10 signals for example, the maximum difference between cumulatives $\Phi_\theta^{N,p}(s)$ and $\Phi_\theta^p(s)$ is at most

truly irrelevant ones. Again, for a Bayesian, this makes the "irrelevant pool" not so irrelevant, but it affects the long run belief state of agents that ignore these events.

equal to 4% *uniformly over the transition probabilities* p. So this gives an upperbound of 4% on the possible change in welfare when 10 signals are processed, compared to the limit case where infinitely many are processed.[44]

For a fixed strategy σ^d, the direction of change however depends on p. For regular problems, more signals help because they tend to increase the probability to end up in an extreme state (hence opposite extreme states when the problem is regular). For non-regular problems however, getting more signals may increase the loss, as we now illustrate by reporting the ratio $(W - W^N)/\underline{W}$



For these irregular problems ($p_{11} = 0.8$ and $p_{22} < 1/2$), processing more signals tends to generate positive mental states independently of the underlying state: the agent's decision is more subject to the mental system's bias, hence the higher losses when $\gamma > 1/2$.

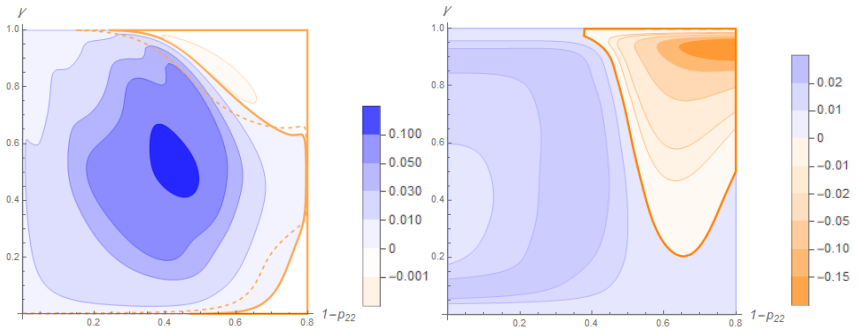
Regarding the incentives to censor weak evidence, the locus of problems for which the marginal effect of ignoring weak evidence is positive remains similar to the large N case. We report the figure in Appendix.

6.2 More belief/mental states

With more mental states, the mental system is potentially more efficient in aggregating information. For example, in the Bayesian benchmark where (d_p, L_p) can be adjusted to both p **and** the number of mental states, the set of problems for which the mental system helps mostly expends, with a significant percentage gain for many problems. We illustrate this below with a change from 5 to 7 states. We report the welfare gain ratio $r = \Delta W/\underline{W}$ as a function of p_{22} and γ, assuming noisy priors. The left figure is the

[44] When the number of mental states increases, convergence takes proportionally more signals.

Bayesian benchmark (with mostly gains). We keep σ^d fixed to 3 in the right figure:



The right figure illustrates that increasing the number of states has mixed consequences. The trade-off is similar to the one discussed in previous sections: a higher number of states increases welfare for (most) regular problems, but it diminishes welfare for some irregular problems.

This suggests that *even in the absence of costs associated with maintaining a larger number of states*, there may be a cost associated with the more complex mental system. It performs slightly better on many problems, but significantly worse for some.

Of course, the individual could adjust d downward when he has 7 mental states rather than 5. By reducing d down to $d^{2/3}$, the spread in beliefs remains the same whether he has 5 or 7 states: this would limit the gains in the region of regular problems, but avoid the adverse consequence of keeping a large d in case the problem is irregular.

Nevertheless, to the extent that d is an instrument that one finds difficult to adjust, *limiting the number of states* can be viewed as an alternative instrument for reducing the risk of falling prey to mental processing biases.

Thus, beyond the classic motive that mental states are scarce cognitive resources, we suggest here an alternative motive for reducing the number of mental states: with uncertainty about the data-generating process, too many states may actually hurt welfare.

6.3 More complex mental systems

The previous discuss echoes the classic observation that complexity comes with lower fitness. Following up along this line of thought, a natural extension of the simple mental process would be to allow for mental state changes that are functions of the perceived strength of signals, for example:

- If $\widetilde{l} \in (d^{1/2}, d^{3/2})$ move one step,
- if $\widetilde{l} > d^{3/2}$, move two steps.

With sufficiently acute perception of strength, this type of mental processing is likely to be helpful for some problems, as the mental state transitions are better tuned to the real informativeness of the signals being processed. There are two caveats however:

(i) Estimating the strength of evidence seems much more demanding than estimating the direction of evidence

(ii) Even if correct, the issue we raised remains: if the agent is unable to perceive correctly the resulting ex ante balance between confirming and disconfirming evidence, the distinction between regular and irregular problems will remain relevant.

(ii) With sufficiently noisy perception of strength, the process gives rise to random moves of $0, 1$ or 2 steps, and this more complex mental processing may actually deteriorate welfare compared to the simple mental processing we discussed (See Compte and Postlewaite (2009) for an example along those lines).

6.4 More states of world

We have considered an agent attempting to discriminate between only two states. What if the agent attempts to discriminate between more than two states, say three? This is a challenging task as in principle a sophisticated agent would want to keep track of all relative likelihoods, and any signal could in principle modify all these relative likelihoods.[45] It is also a challenging modelling task to come up with an intuitive and simple belief formation process. We provide a suggestion below with two purposes in mind: to show that a one-dimensional belief formation rule remains feasible and would perform well under some conditions; to highlight that even if weak information is not ignored, a bias towards theories that generate strong evidence to likely to arise.

Regarding the processing of signals, to each signal x we associate a direction and strength of evidence, with $\overline{\theta}$ defined as before and $l = f(x \mid \overline{\theta}) / \max_{\theta \neq \overline{\theta}} f(x \mid \theta)$. Regarding the mental system, we assume $3K+1$ mental states, with states labelled as 0 or (i, k) with $i \in \{1, 2, 3\}$ and $k \in \{1, .., K\}$.

[45] A Bayesian needs to keep track of the likelihood of each θ against each θ', which means keeping track of L_{12} and L_{13} for example. For a less sophisticated agent, keeping track of the relative likehihoods against all alternatives theories might be difficult, hence our suggestion.

We interpret a mental state $s = (i, k)$ as indicating overall evidence pointing towards state $\theta = i$, to a degree k. Accordingly, starting from $s = 0$, we assume that when the agent processes a signal in favor of $\theta = i$, his state moves up one step on the i-ladder if $s = 0$ or (i, k) with $k < K$, and otherwise (i.e., if on a j-ladder with $j \neq i$) moves down one step (possibly reverting to $s = 0$).

Regarding how beliefs are formed, let $\rho_{ij} = \Pr(i)/\Pr(j)$ denote the prior likelihood. We assume that when in state (i, k), the posterior likelihood of i against j is:

$$\rho_{ij} d^k$$

In other words, the mental state can reinforce a belief in one particular state, but it cannot modify the relative probabilities of low probability states.

Let again $p_{\overline{\theta}\theta} = \Pr(\overline{\theta} \mid \theta)$. It should be clear that if $p_{kk} > 1/2$ for all k, then the mental system, however limited, improves welfare as it creates a positive correlation between the underlying state θ and the set of mental states $\{(\theta, k)\}_{k \geq 1}$. But it is also easy to come up with problems for which evidence for some theory is always inexistent, independently of the underlying state.

For example, consider an agent receiving a sequence $x = (x_1, ..., x_5)$ of 6 draws of 1's and 0's possibly autocorrelated. Let $\rho = \Pr(x_{m+1} = x_m)$ and assuming that possible values of ρ are $2/3$ $(\theta = 1), 1/3$ $(\theta = 2)$ and 0.5 $(\theta = 3)$. The following table gathers the pairs $(\overline{\theta}, l)$ for each sequence received as a function of the number n of reversals (i.e., $x_{m+1} \neq x_m$).

n	0	1	2	3	4	5
$\overline{\theta}$	1	1	1	2	2	2
l	4.2	2.1	1.05	1.05	2.1	4.2
$\Pr(x \mid \theta = 3)$	0.03	0.16	0.31	0.31	0.16	0.03

With sequences of 6 draws, there is no sequence that provides clear evidence in favor of independence, and the agent is lead to believe either in positive or negative autocorrelation even when there is no autocorrelation. With sequences of limited length, there is no sequence that is an obvious representative of an independent sequence of draws.

As the number of elements in a sequence increases, evidence in favor of independence surfaces for some draws and become frequent, but even for sequence of 10 draws, the evidence tends to be weak compared to evidence

in favor of other states. When x consists of a sequence of 10 draws, we have:

n	0	1	2	3	4	5	6	7	8	9
$\overline{\theta}$	1	1	1	1	3	3	2	2	2	2
l	13	6.7	3.3	1.7	1.2	1.2	1.7	3.3	6.7	13
$\Pr(x \mid \theta = 3)$	0.002	0.02	0.07	0.16	0.25	0.25	0.16	0.07	0.02	0.002

When $\theta = 3$, evidence in favor of 3 is frequent, but never quite striking, unlike evidence for other states of the world. Evidence in favor of more extreme states is more striking.

6.5 Further discussion

6.5.1 Classifocation and representativeness.

We defined $\widetilde{\theta} \in \{1, 2\}$ as indicating whether a given signal x was perceived as evidence for $\theta = 1$ or $\theta = 2$. One interpretation is that the agent categorizes processed signals as being representative of either state 1 or state 2 and that signals that are not sufficiently representative of either state (because they are too poorly informative) are not processed.

So one can think of our agent repeatedly using a *representativeness heuristic* to classify signals, and raising β means that the agent is categorizing signals only when they are sufficiently representative of a particular state. By endogenizing β, we are thus endogenizing the categorization made, and also what is meant by being representative: a signal is representative of θ if it is sufficiently informative of theory θ (as opposed to θ'). This classification of signals (and thus this notion of representativeness) is not attached to θ, but to the discrimination problem (θ versus θ') considered, which is why framing matters.[46]

With batch processing in mind (with x being a sequence of signals), another interpretation is that the agent is using a *law of small numbers* to categorize signals (with a smaller number sufficient when β is smaller). When many signals are processed, the agent repeatedly this law of small number, and this improves welfare so long as problems are regular (that is, so long as evidence is balanced), but it generally deteriorates welfare when problems are not regular.

[46] Note that this notion of representativeness differs from Kahneman and Tsversky's (1973), who argue that a signal x may be considered as being more representative of θ than another signal x' (hence also more "likely") even though $f(x|\theta) < f(x'|\theta)$.

6.5.2 Misspecified models.

Let us contrast our work with the literature that explain biases through agents forming beliefs based on a misspecified (or incomplete) model of the environment (See Spiegler (2020) for a review). In our work, agents do not have a mispecified model in the sense that the true state of the world θ may lie within the set of alternatives compared. Or from a broader perspective, if one think of all (θ, f) as "extended" states of the world, all these extended states are considered possible by the agent.

Rather, it is the agent's assumed inability to adjust belief formation (and in particular the bias Λ) to each realization of f that creates biased beliefs for some problems. This issue is thus related to the classic heuristics and bias debate:

- from a close look, the heuristic that the agents use is not adapted to the particular data-generating process that they are contronted to: they behave as if their mental system was not biased, that is, as if they had a special (and thus mispecified view) of f, one where the ex ante tally between confirming and disconfirming evidence is balanced (i.e., as if $\Lambda_p = 1$), and this misspecification is a source of biased beliefs.

- from a broader perspective, and since agents are not observing directly f nor the consequence of censoring on the data-generating processing, they are doing the best they can with what they see, within a reasonable family of belief formation rules, and on average, there is no reason to introduce a bias $\Lambda \neq 1$ in decision rules.

6.5.3 Conclusion.

We have modelled agents whose behavior is governed by two heuristics, one that governs the classification of signals (through censoring β), and one that governs caution in decision making (through the discrimination power d that the agent assigns to the mental system), having in mind that these two heuristics adjust on average over the various discrimination problems that the agent faces, without being able to adjust these two instruments to the specific data-generating process associated with each one of them. The optimal heuristic can only be good on average, and our analysis highlights the type of discrimination problems for which biases are generated, as well as how pooling and framing can be used to distort one's belief. A more systematic study of this last phenomenon deserves further research.

7 Bibliography

Abell, G. O., & Greenspan, B. S. (1979). Human Births and the Phase of the Moon. *New England Journal of Medicine*, 300(2).

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.

Beck, Jan & Forstmeier, Wolfgang. (2007) "Superstition and Belief as Inevitable By-Products of an Adaptive Learning Strategy," *Human Nature.* 18. 35-46.

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

Chapman, L. J., & Chapman, J. P. (1967). "Genesis of popular but erroneous psychodiagnostic observations," *Journal of Abnormal Psychology*, 72(3), 193–204.

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.

Compte, O. and A. Postlewaite (2018), *Ignorance and Uncertainty.* Econometric Society Monograph. Cambridge University Press.

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

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

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

Gigerenzer, Gert (2007), *Gut feelings: The intelligence of the unconscious.* New York, NY: Viking.

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.

Hildburgh, W. L. (1951). "Some Spanish Amulets Connected with Lactation," *Folklore*, 62(4), 430–448.

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.

Kahneman, D. and A. Tversky (1972), "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology*, 3, 430-454.

Köszegi, B. (2006). Ego Utility, Overconfidence, and Task Choice. *Journal of the European Economic Association*, 4(4), 673–707.

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.

Peters, U. (2020) "What Is the Function of Confirmation Bias?" *Erkenntnis*

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.

Spiegler, R. (2016) "Bayesian Networks and Boundedly Rational Expectations," *The Quarterly Journal of Economics*, 131, 3, 1243–1290

Spiegler, R. (2020), "Behavioral Implications of Causal Misperceptions," *Annual Review of Economics* 12:1, 81-106

Tversky, A. and D. Kahneman (1971), "Belief in the Law of Small Numbers," *Psychological Bulletin*, 76, 105-110.

Tversky, A. and D. Kahneman (1973), "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology*, 5, 207-232.

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

Wilson, A. (2014) "Bounded memory and biases in information processing," *Econometrica*, 82(6), 2257–2294.

Wilson, R. (1987) "Game-Theoretic Approaches to Trading Processes," Advances in Economic Theory: Fifth World Congress, Ch. 2, Cambridge, Cambridge University Press, 33-77.

8 Appendix

Proof of Lemma: Relabelling mental states from $n = 0$ to $N = 2K$, we have $n = k + K$ and let $f(n, d) \equiv \frac{\partial \Psi(n-K,d)}{\partial d}$. We have $f(n, d) = 1 - \frac{d^n - 1}{d^{N+1}-1}$ and so $\frac{\partial f}{\partial d}$ has the same sign as $g(d) = (N+1-n)d^{N+1} - (N+1)d^{N+1-n} + n$. Since $g(1) = 0$ and $g'(d) = (N+1)(N+1-n)d^{N-n})(d^n - 1) > 0$ for $d > 1$, so $g(d) > 0$ for all $d > 1$, so $f'(d)$ is positive for all $d > 11$ and $n \in \{1, N\}$ which concludes the proof.

Proof of $E[\ln L|\theta = 1] > 0 > E[\ln L|\theta = 2]$: We have $E[\ln L|\theta = 1] = E[L \ln L|\theta = 2]$, which is strictly positive because $L- > L \ln L$ is a convex function. Similarly, $E[\ln L|\theta = 2] = -E[\frac{1}{L} \ln 1/L|\theta = 2] < 0$.