entropy

MDPI

Article

Biases and Variability from Costly Bayesian Inference

Arthur Prat-Carrabin [1,2], Florent Meyniel [3], Misha Tsodyks [4,5] and Rava Azeredo da Silveira [2,4,6,7,*]

1 Department of Economics, Columbia University, New York, NY 10027, USA; arthur.p@columbia.edu
2 Laboratoire de Physique de l'École Normale Supérieure, Université Paris Sciences & Lettres, Centre National de la Recherche Scientifique, 75005 Paris, France
3 Cognitive Neuroimaging Unit, Institut National de la Santé et de la Recherche Médicale, Commissariat à l'Energie Atomique et aux Energies Alternatives, Université Paris-Saclay, NeuroSpin Center, 91191 Gif-sur-Yvette, France; florent.meyniel@cea.fr
4 Department of Neurobiology, Weizmann Institute of Science, Rehovot 76000, Israel; misha@weizmann.ac.il
5 The Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540, USA
6 Institute of Molecular and Clinical Ophthalmology Basel, 4056 Basel, Switzerland
7 Faculty of Science, University of Basel, 4001 Basel, Switzerland
* Correspondence: rava@ens.fr

Abstract: When humans infer underlying probabilities from stochastic observations, they exhibit biases and variability that cannot be explained on the basis of sound, Bayesian manipulations of probability. This is especially salient when beliefs are updated as a function of sequential observations. We introduce a theoretical framework in which biases and variability emerge from a trade-off between Bayesian inference and the cognitive cost of carrying out probabilistic computations. We consider two forms of the cost: a precision cost and an unpredictability cost; these penalize beliefs that are less entropic and less deterministic, respectively. We apply our framework to the case of a Bernoulli variable: the bias of a coin is inferred from a sequence of coin flips. Theoretical predictions are qualitatively different depending on the form of the cost. A precision cost induces overestimation of small probabilities, on average, and a limited memory of past observations, and, consequently, a fluctuating bias. An unpredictability cost induces underestimation of small probabilities and a fixed bias that remains appreciable even for nearly unbiased observations. The case of a fair (equiprobable) coin, however, is singular, with non-trivial and slow fluctuations in the inferred bias. The proposed framework of costly Bayesian inference illustrates the richness of a 'resource-rational' (or 'bounded-rational') picture of seemingly irrational human cognition.

Keywords: approximate inference; Bayesian inference; online inference; resource rationality; cognitive cost

check for updates

Citation: Prat-Carrabin, A.; Meyniel, F.; Tsodyks, M.; Azeredo da Silveira, R. Biases and Variability from Costly Bayesian Inference. *Entropy* **2021**, *23*, 603. https://doi.org/10.3390/e23050603

Academic Editors: Francesco Donnarumma, Domenico Maisto and Ivilin Stoianov

Received: 21 March 2021
Accepted: 26 April 2021
Published: 13 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

While the faculty of rational thinking defines, at least to an extent, our human nature, it suffers from a remarkably long list of so-called 'cognitive biases'—systematic deviations from rational information processing and behavior [1]. Much as optical illusions are informative of the neurobiological mechanisms of vision [2], one expects cognitive biases to reveal aspects of the algorithmic basis of cognition and behavior [3]. A notable category of biases comprises those that govern the way we manipulate probabilistic quantities: these biases affect our inference of the probability of events, our decision-making process, and, more generally, our behavior in situations where stimuli obey (seeming or unknown) stochastic rules [4–12].

In such situations, human subjects violate a normative predicament viewed by the experimenter as the rational one (e.g., "maximize the number of correct responses per unit time in a given task") [13]. In a task involving random stimuli, a subject's brain may fail to manipulate the probabilities it infers from the stimuli in a sound (Bayesian) and precise mathematical manner, and, as a result, predictions or decisions based upon the

inferred probabilities may suffer from biases. Several origins of this phenomenon have been proposed. One common suggestion is that the brain manipulates probabilities in the correct, Bayesian way, but that the outcome of the inference process is altered by the choice of an 'erroneous' prior probability [13–15]. In other words, the subject may reason correctly, but base the reasoning upon incorrect prior beliefs. It is also possible that the brain carries out Bayesian calculations, but with noisy processing units (neurons) that yield stochastic, approximate responses [16] or posteriors [17–19]. Alternatively, the brain may simply not be concerned with a mathematically sound manipulation of probabilities, but rather may operate with the use of a set of heuristic algorithms [5,20].

Here, we propose another, but possibly complementary, cognitive mechanism for the emergence of biases in the inference of probabilities. We assume that in order to perform Bayesian inference, the brain needs to represent probability distributions, but expends a computational cost to do so. When inferring a probability, p, based on a sequence of observations, the Bayesian observer updates a probability distribution, $P(p)$, after each new observation. For the brain, however, producing a representation of a probability distribution is presumably an operation that comes with a metabolic cost, or that is subject to cognitive constraints. The brain, thus, might prefer to choose an alternative distribution, $\hat{P}(p)$, that is less costly to represent, but that still captures much of the behaviorally relevant structure in the Bayesian posterior, $P(p)$. We formalize this compromise as an optimization problem, in which the objective of choosing a posterior that is close to the optimum competes with the cost of the posterior. We examine two cost functions, that capture two natural assumptions about the cognitive constraints at play. The *precision cost*, first, follows from the assumption that a more precise encoding by the brain requires greater metabolic resources: in general, in information-theoretic terms, a larger number of bits is needed to resolve finer uncertainty. Mechanistically, a more precise encoding of information in a brain area may require the use of more neurons or more spikes, involving a greater metabolic cost. The *unpredictability cost*, second, penalizes beliefs that imply more uncertain (or less deterministic) environments, which are more difficult for the brain to cope with and represent an ecological challenge.

We examine the ways in which these two costs impact inference. For the sake of simplicity, we focus on the scenario of an observer facing a series of binary signals (i.e., p is the parameter of a Bernoulli distribution). In contrast to other studies, in which the statistics of the inferred variable undergo changes [16,19,21], here they are constant. Yet, because of the presence of a cost, the subject's estimate of p will generally be biased, and the posterior, in some cases, does not even converge. In paying special attention to an equiprobable environment ($p = 1/2$)—the case of a fair coin—we show the sometimes surprising dynamics of the inference in this singular case. Our study proposes a fresh theoretical understanding of the biases that the humans exhibit when carrying out inferences about probabilities.

2. Results

2.1. Costly Bayesian Inference

We propose to regularize Bayesian inference by a cost on probability distributions. We reformulate the inference procedure as a trade-off between Bayesian updating of the probability distribution and a cost term that penalizes some distributions more than others (we consider two natural examples of possible costs below). While this modification of the Bayesian procedure may appear harmless, and does not assume any explicit bias, we show below that it may profoundly alter the outcome of the inference of probabilities. We emphasize that our aim, here, is not to claim that humans carry out the altered Bayesian inference that precisely matches our mathematical formulation, presented below, nor that our prescription fits behavioral data better than another model; rather, we would like to present the idea of regularized Bayesian inference in which the trade-off emerges from a cognitive constraint as a possible, additional ingredient among a number of rationalizations of cognitive biases.

In this spirit, we consider the simplest possible scenario of online inference: successive flips of a (fair or biased) coin, from which a model subject is asked to infer the probability of observing 'head', p, (or 'tail', $1-p$) in the upcoming flip. (Equivalently, the model subject infers the 'bias' of the coin, i.e., the extent to which the probability differs from 0.5.) We assume that the model subject carries out online inference by updating a probability distribution, $\hat{P}_N(p)$, at successive coin flips indexed by N. A natural choice of distribution, after each coin flip, is the posterior prescribed by Bayes' rule, P_N; our model subject would select this distribution if it were not for the cognitive cost, $C[P]$, associated with any distribution P. Instead, the model subject chooses a compromise between bearing the cost, $C[\hat{P}_N]$, of the inferred posterior, $\hat{P}_N$, and minimizing the discrepancy between the inferred posterior and the Bayesian posterior, P_N, as measured by a function that we denote by $D(\hat{P}_N, P_N)$. Specifically, the model subject minimizes the loss function

$$L = D(\hat{P}_N, P_N) + \lambda C[\hat{P}_N], \tag{1}$$

where $\lambda \geq 0$, the only parameter in the theory, determines the strength of the regularization from a cognitive cost. We define the distance as the Kullback–Leibler divergence between the inferred distribution and the Bayesian posterior, $D_{KL}(\hat{P}_N||P_N)$. A similar loss function has been proposed in the literature, using the Kullback–Leibler divergence as a distance metric, in modeling the sub-optimality of cognition [22]. We will show that even a cost that does not assume any explicit bias may alter the outcome of the inference qualitatively. Depending on the nature of the cognitive toll borne by the brain to represent a probability distribution, and depending on the coin's degree of bias (or lack thereof), the inference process of the model subject may converge to an under- or an over-estimation of the bias, or may not converge at all. We obtain a diversity of behavioral patterns—some, counter-intuitive—by examining just two costs that derive from natural hypotheses on the constraints at work when the brain manipulates probability distributions. We now present these costs and the behaviors they yield.

2.2. Inference of Probability under a Precision Cost

A natural hypothesis is that costly distributions are the ones that reduce the uncertainty on the inferred latent parameter (here, p). Specifically, for a distribution, $P(p)$, we propose a 'precision cost' equal to the negative of its Shannon entropy, i.e.,

$$C[P] \equiv -H[P] = \int P(p) \ln P(p) \mathrm{d}p. \tag{2}$$

Under this hypothesis, the uniform distribution is the least costly distribution (with cost zero), and we note that this cost function is equal to the Kullback–Leibler divergence between the uniform distribution and the inferred distribution, P. More 'precise' distributions, concentrated around some value, come with a higher cost. The distribution that minimizes the loss function (Equation (1)), with this cost, is proportional to the Bayesian posterior raised to an exponent:

$$\hat{P}_N(p) \propto P_N(p)^{\frac{1}{\lambda+1}}. \tag{3}$$

In the absence of cost ($\lambda = 0$), the inferred distribution coincides with the Bayesian posterior; with an infinitely large cost ($\lambda \to \infty$), the solution is the uniform distribution; finite, positive values of λ yield intermediate distributions. After a new coin flip is observed, the updated Bayesian posterior is proportional to the product of the prior and the likelihood, and thus the updated inferred posterior is proportional to these two quantities, raised to the exponent $1/(\lambda+1)$. A similar form of the posterior has been posited to capture human biases in probability estimations [23]; here, it is obtained as the solution of an optimization argument.

Before the first coin flip, we assume that our model subject holds a uniform prior, i.e., $\hat{P}_0(p) = 1$. In this case, the probability distribution after the Nth flip is a Beta distribution,

$$\hat{P}_N(p) \propto p^{\hat{n}_1}(1-p)^{\hat{n}_0}, \tag{4}$$

where $\hat{n}_1$ and $\hat{n}_0$ are exponentially filtered counts of 'heads' and 'tails',

$$\hat{n}_1 = \sum_{k=0}^{N-1} \left(\frac{1}{\lambda+1}\right)^{k+1} x_{N-k}, \tag{5}$$

$$\text{and } \hat{n}_0 = \sum_{k=0}^{N-1} \left(\frac{1}{\lambda+1}\right)^{k+1} (1-x_{N-k}), \tag{6}$$

where x_i is 1 if 'head' occurs at the ith flip, and 0 if it is 'tail'.

In these counts, the outcomes of past coin flips are gradually 'forgotten'; as a result, the posterior does not converge as more evidence is accumulated, but fluctuates, instead, as a function of the recent history of coin flips. Hence, although the posterior of a perfectly Bayesian observer would be entirely determined by the total number of flips, N, and the number of 'heads' among them, n, for our model subject, different orders of 'heads' and 'tails' in the sequence result in different posteriors (Figure 1A). In addition, at long times, the variance of the posterior is controlled by the strength of the cost, λ (and by the probability, p). Even after a large number of coin flips, the width of the posterior does not vanish; it is an increasing function of λ (Figure 1A; see Methods for an approximate expression of the variance of the posterior as a function of λ). Furthermore, the mean of the posterior, $\hat{p}$, varies as a function of the recent history of flips; on average, however, it is a biased estimate of the probability, except in the case of a fair coin (a case we examine more closely below). With a biased coin, large probabilities are underestimated, i.e., $p > \mathbb{E}[\hat{p}] > 1/2$, and small probabilities are overestimated, i.e., $p < \mathbb{E}[\hat{p}] < 1/2$ (the expectations are taken over the space of all sequences, or, equivalently, over a very long sequence). In other words, the model subject underestimates the bias of the coin, and more so for a stronger cost (Figure 1B). We emphasize that we did not start by positing the exponentially filtered counts of the observations (Equation (5)), or by making a related assumption on the limited memory of the model subject. The decaying counts naturally result from the assumption of a precision cost, from which the underestimation of the coin's bias and the fluctuations in the posterior also follow.

2.3. Inference of Probability under an Unpredictability Cost

We considered, above, the hypothesis that more concentrated posteriors entail a higher computational toll than broader ones. We now turn to a cost that varies not with the entropy of the posterior, but with that of the inferred probability. More precisely, as decisions are more easily made when the (inferred) environment is more predictable, we examine the hypothesis that the brain favors the probabilities that have more predictive power, and penalizes those that imply, conversely, unpredictability in the environment. We quantify the degree of unpredictability of a coin flip by the entropy of the Bernoulli random variable corresponding to the coin flips,

$$H(p) = -p\ln(p) - (1-p)\ln(1-p), \tag{7}$$

and we define the cost of a probability distribution, $P(p)$, as the average of this entropy over the posterior distribution, i.e.,

$$C[P] = \mathbb{E}_P[H(p)] = \int H(p)P(p)\mathrm{d}p. \tag{8}$$

(While the precision cost is a function of the entropy of the *posterior*, $H[P]$, the unpredictability cost is a function of the entropy implied by the *probability*, $H(p)$, averaged over the posterior).

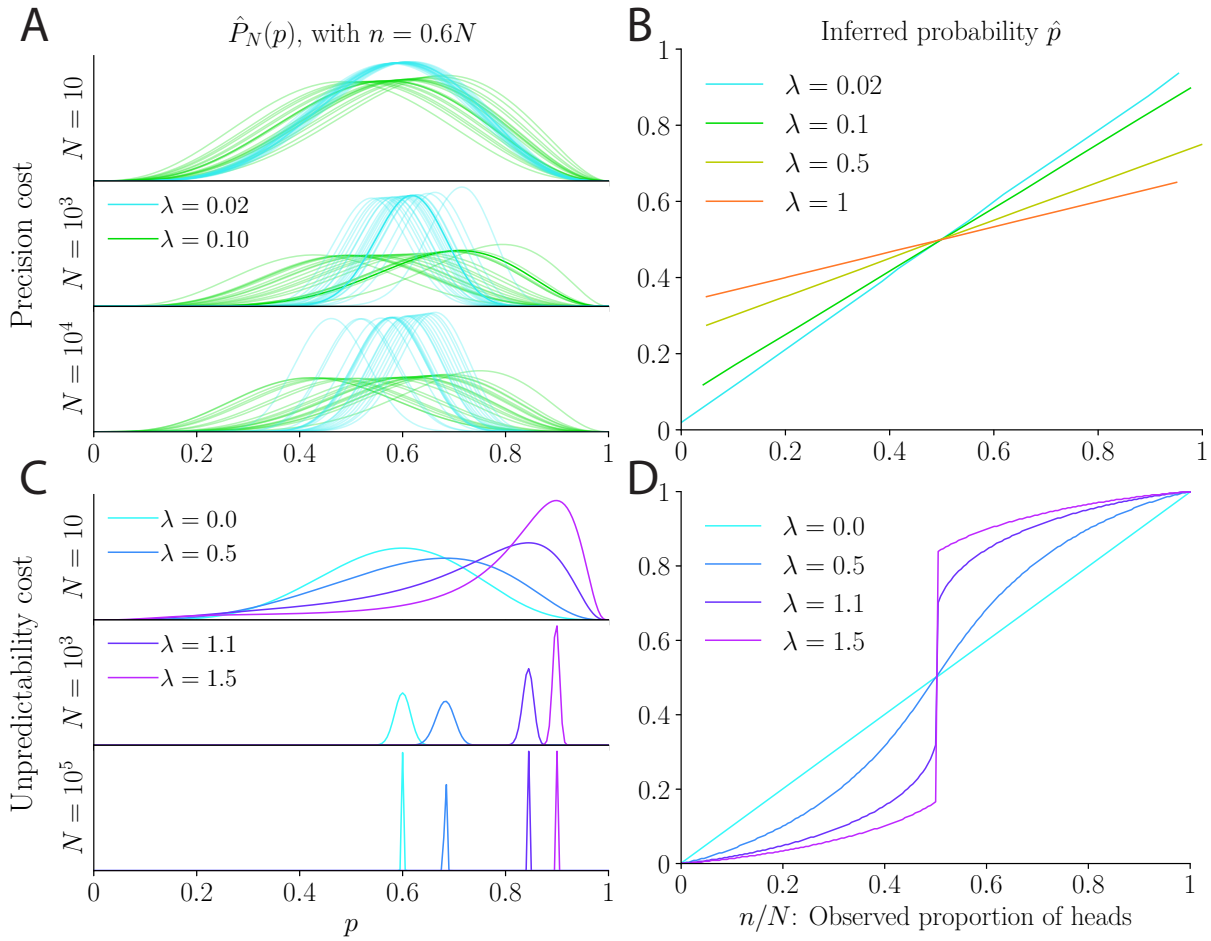


Figure 1. Biased coin: probability inference under a precision cost (**A**,**B**) and an unpredictability cost (**C**,**D**). (**A**) Inferred distribution, $\hat{P}_N(p)$, for the precision cost with $\lambda = 0.02$ and 0.10, after $N = 10, 1000$, and 10,000 coin flips, and 60% of 'heads', in 20 sequences of coin flips. Different sequences result in different posteriors, and a larger cost yields wider posteriors. As more evidence is accumulated, the width of the posterior decreases at first, but plateaus at large N. (**B**) Average estimate of the model with the precision cost, $\hat{p}$, as a function of the proportion n/N of 'heads' observed. (**C**) Inferred distribution $\hat{P}_N(p)$ for the unpredictability cost, after $N = 10, 1000$, and 100,000 coin flips and 60% of 'heads'. With this cost, given the number of coin flips, N, the ratio n/N fully determines the posterior. As more evidence is accumulated, the posterior is more peaked around $\hat{p}$. (**D**) Inferred estimate of the model with the unpredictability cost, $\hat{p}$, as a function of the proportion n/N of 'heads' observed. For $\lambda > 0$, the cost exacerbates the true bias of the coin. For $\lambda > 1$, at exactly $n = N/2$ there are two local maxima. When n fluctuates around $N/2$, the inferred probability will switch from one maximum to the other (see Figures 2 and 3).

Online Bayesian inference, when regularized by this cost (i.e., when the quantity in Equation (1), using this cost, is minimized), yields a probability distribution of the form $\hat{P}_N(p) \propto [\varphi_{n/N}(p)]^N$ (up to a normalization factor) at the Nth flip of the coin, where

$$\varphi_{n/N}(p) \propto p^{n/N}(1-p)^{1-n/N} e^{-\lambda H(p)} \tag{9}$$

and n is the number of times 'head' is observed among a total of N flips. Because $\hat{P}_N(p)$ is obtained by raising $\varphi_{n/N}(p)$ to its Nth power, after a large number of coin flips, the inference is dominated by the maxima of $\varphi_{n/N}(p)$. This function can have one or two local maxima, depending upon the value of the *empirical bias*, n/N, and the weigth of the cost, λ. In the limit of large N, $\hat{P}_N(p)$ converges to a delta function centered at the inferred probability, $\hat{p} = \text{argmax}[\varphi_{n/N}(p)]$, which is subjected to the fluctuations in the empirical bias, n/N. This inference calculation leads to qualitatively different scenarios for a biased coin and a fair coin (we discuss the latter case in the next section).

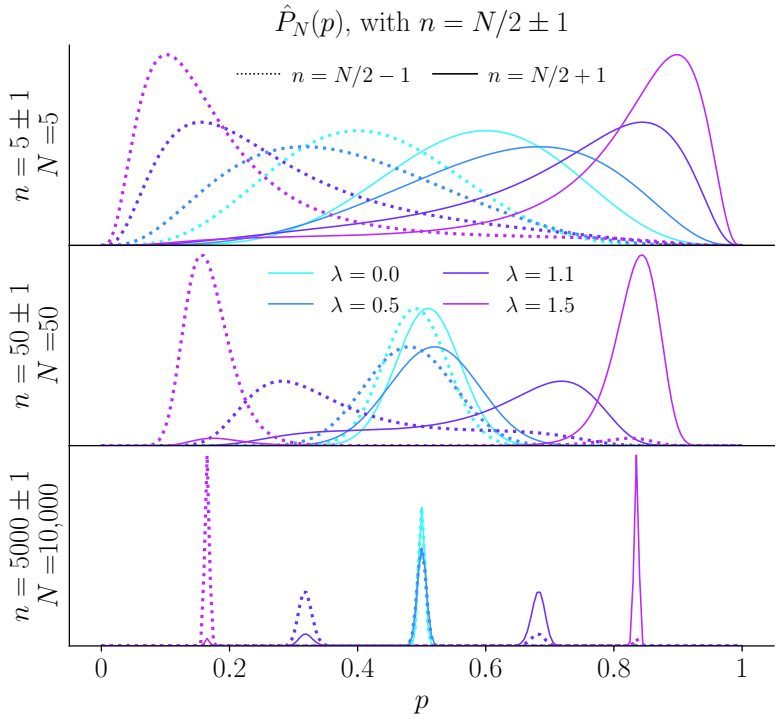


Figure 2. Inferred bias from a fair coin under the unpredictability cost. Inferred distribution $\hat{P}_N(p)$ for $N = 10$ (top), 100 (middle), and 10,000 (bottom) coin flips, when the number n of 'heads' is $N/2$ plus one (solid lines) or minus one (dashed line), with the unpredictability cost. Depending on the sign of $n - N/2$, the maximum of $\hat{P}_N(p)$ is greater or lower than $1/2$. With more coin flips, for $\lambda < 1$ the two maxima (above and below $1/2$) converge to $1/2$. For $\lambda > 1$, the two maxima stay separated, and thus the inferred probability can switch from one maximum to the other, depending on the fluctuations of n around $N/2$.

In the case of a *biased coin*, the inference always converges to a probability, $\hat{p}$, controlled by (but in general not equal to) the expectation of n/N (the true bias of the coin). The value of $\hat{p}$ is not equal to the true bias of the coin because it also depends upon the strength of the unpredictability cost, λ: the larger λ, the more exacerbated the inferred bias. If 'head' is more probable than 'tail', then the inference will enhance this bias ($\hat{p} > p > 1/2$), and similarly if 'tail' is more probable than 'head' ($\hat{p} < p < 1/2$). For sufficiently strong regularization ($\lambda > 1$), the inferred bias is always appreciable: even for an infinitesimally small (empirical) bias, the inferred bias remains non-vanishing. Thus, the unpredictability cost boosts even the slightest bias present in the observations into an appreciable bias in the inferred probabilistic origin of these observations (Figure 1D).

The extreme sensitivity of the inferred probability to the empirical bias, noted above in the context of the long-time solution (Figure 1C,D), also has implications for transients in the inference. Consider a small true bias of the coin: the probability of 'head' is $1/2 + \varepsilon$, with $\varepsilon \ll 1$. Typically, in half of the experiments, a fluctuation in the empirical bias overcompensates the true bias for as long as $N \approx \varepsilon^{-2}$. During this transient, the model subject will believe that the more likely bias is opposed to the true one, i.e., favors 'tail'. Furthermore, a model subject with large unpredictability cost will assign a non-vanishing inferred bias for 'tail', because of its extreme sensitivity to the empirical bias. This sensitivity yields counter-intuitive behavioral patterns when the true bias of the coin vanishes, i.e., with a fair coin: we now turn to this singular case.

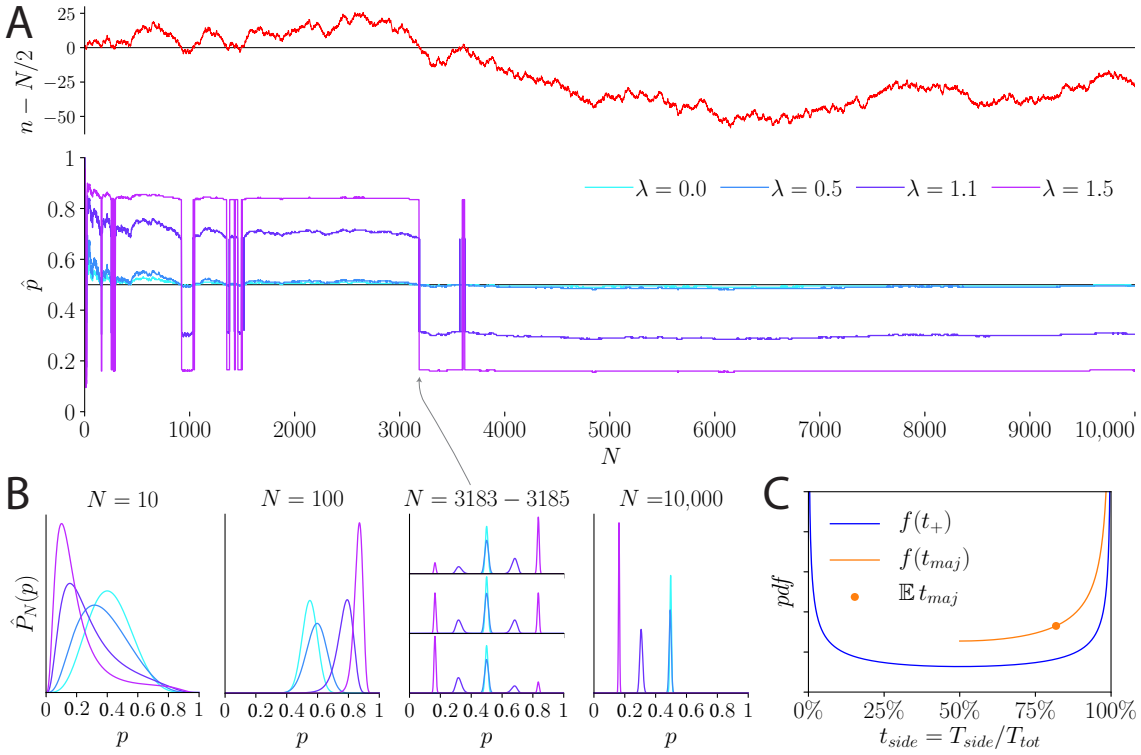


Figure 3. Behavior of the unpredictability-cost model over 10,000 flips of a fair coin. (**A**) *Top:* Random walk described by the fluctuations of the number n of 'heads' around $N/2$. The quantity $n - N/2$ has excursions of varying lengths above and below zero. *Bottom:* Inferred probability, $\hat{p}$, over the course of the 10,000 coin flips. Depending on the sign of $n - N/2$, the inferred probability is above or below $1/2$. For $\lambda < 1$, $\hat{p}$ converges to $1/2$, while for $\lambda > 1$ it never converges but switches between two maxima, $p_>$ and $p_<$, above and below $1/2$. (**B**) Inferred distribution, $\hat{P}_N(p)$, at various times during the course of the coin flips. From $N = 10$ to $N = 100$, the density becomes narrower; it is subject to the fluctuations in the coin flips, but for $\lambda < 1$ the density approaches $1/2$. For $N = 3183$ to 3185, the number of observed 'heads' goes from just above $N/2$ to just below. As a result, for $\lambda > 1$ the maximum at $p_>$, above $1/2$, decreases and the other maximum at $p_<$ increases, becoming the global maximum. For exactly $n = N/2$, the density is symmetric and both maxima have the same height. For $\lambda < 1$, there is only one maximum. Finally, for $N = 10{,}000$, in this instance the random walk is relatively far from 0; thus, for all $\lambda > 1$, the distribution is concentrated around a maximum below $1/2$. (**C**) Probability density function of the proportion of time spent on a given side—above or below 0—of the random walk, and on the most visited side. The density of the time T_{side} spent on a given side, relative to the total duration T_{tot} of the random walk, follows an arcsine distribution (blue line). The extremes are more likely than the center, i.e., a random walk is not likely to spend around half of its time on a given side: it is much more likely to spend either a very long time or a very short time on this side. One side, the 'majority' side, will thus dominate. The proportion t_{maj} of time spent on this majority side has an arcsine density 'folded' on values above 50% (orange line). Its expected mean is 81.8% (orange dot), i.e., on average, unbiased random walks spend 81.8% of their time on one side.

2.4. The Case of a Fair Coin: Fluctuating and Biased Beliefs

We examine the behavior of a model subject whose inference is regularized by the unpredictability cost or by the precision cost, when the coin is fair, i.e., when $p = 1/2$. With the unpredictability cost, the case of a fair coin presents a non-trivial inference scenario when the cost is strong ($\lambda > 1$). (If $\lambda < 1$, inference converges to the true solution, with $\hat{p} = 1/2$.) The function $\varphi_{n/N}(p)$ has two maxima, but now their relative amplitudes are controlled by the *fluctuations* in the empirical bias (n/N), rather than by its mean (Figure 2). It is straightforward to show that the locations of the two maxima converge to two values, $p_<(\lambda) < 1/2$ and $p_>(\lambda) > 1/2$, respectively, and that the difference in their heights follows an unbiased random walk proportional to $n - N/2$. Whenever this random walk takes negative values, the inferred, *biased* probability of 'head' is $p_<(\lambda)$; conversely,

when the random walk takes positive values, the inferred probability of 'head' is $p_>(\lambda)$ (Figure 3A,B).

In a long experiment, the inferred bias never converges (if $\lambda > 1$); it switches between two values, and the switching times follow the non-trivial—and rather counter-intuitive—statistics of return times of a random walk [24]. The distribution of the duration between successive switches in the inferred bias is heavy-tailed, with a diverging mean. As a result, the belief that the fair coin has a given bias typically lasts for long durations on the order of the duration of the entire experiment. In particular, whenever the experiment is stopped, it is likely that the model subject has maintained a belief in an inferred bias, either $p_<(\lambda)$ or $p_>(\lambda)$, throughout most of the experiment. On average, a model subject believes in one of the two biased probabilities for about 82% of the duration of the experiment, no matter how long the experiment (Figure 3C).

We emphasize that the regularization of the inference process by the unpredictability cost does not enforce this fluctuating behavior explicitly. The latter emerges naturally from the dynamics of the inference: the inferred bias is governed by the fluctuations in the stimulus, and switches between two values without ever converging. However, because of the peculiar properties of the return times of random walks, it is much more likely to observe a small number of such switches, in any given experiment, rather than a large number. Typically, any given experiment is dominated by one of two inferred biases, either $p_<(\lambda)$ or $p_>(\lambda)$ (Figure 3).

The behavior of a model subject under the precision cost, in the presence of a fair coin, is qualitatively different from that under the unpredictability cost. For large N, the posterior fluctuates with the recent history of coin flips; its expectation, $\hat{p}$, follows an autoregressive process of order 1, whose two parameters are its mean, which is the correct probability, $1/2$, and its coefficient, equal to $1/(1+\lambda)$ (see Methods). The variations in the mean result from 'local' variations in the empirical bias that reflect the recent history of coin flips over a timescale determined by the strength of the cost, λ. When λ is small, this timescale is long, and short-term variations are small in comparison to the overall variability in $\hat{p}$. Successive estimates are highly correlated, resulting in long 'flights' away from the mean, $1/2$, and infrequent changes in the sign of the inferred bias. With larger costs, the estimate depends less on past estimates and more on recent outcomes of the coin flips, resulting in frequent changes in the belief regarding the sign of the coin's bias (Figure 4A,B).

More precisely, as the model subject gradually 'forgets' past flips, its counts of 'heads' and 'tails' are exponentially filtered (Equation (5)), and thus they are bounded: even with an infinite sequence of 'heads', the effective count of 'heads' is finite (it is the sum of a geometric series with a ratio smaller than unity). As a result, the estimate of the probability, $\hat{p}$, is confined between two values placed symmetrically above and below $1/2$, and the distance between these two values is smaller for a strong cost, λ. If the cost is large ($\lambda \geq 1$), the estimate is always close enough to $1/2$ so that just one coin flip can make it switch to the opposite inferred bias. For instance, after a long streak of 'heads', the model subject infers that the coin is biased towards the 'heads' ($\hat{p} \simeq 2/3$, for $\lambda = 1$), but one 'tail' observation suffices to reverse this inference, in favor of a bias towards the 'tails' (Figure 4A). Hence, for all $\lambda \geq 1$, the model subject is always one coin flip away from changing its belief, and this happens with probability $1/2$. Therefore, the probability that the inferred bias stays in either direction for a flight duration of exactly T successive coin flips is $(1/2)^T$. As a result, the model subject holds a belief in a given sign of the bias for an average flight duration of two flips (Figure 4C,D). For weaker costs ($\lambda < 1$), longer flights are more likely. The tail of the distribution of flight durations is heavier for small values of λ, and long flights become more probable (Figure 4C; note the logarithmic scale). This results, as λ nears zero, in large expectations of flight durations, i.e., the model subject maintains its belief in a given sign of the coin's bias for relatively long average durations (Figure 4D; the expected flight duration remains finite as long as λ is non-vanishing [25]—if λ vanishes, i.e., in the

optimal, Bayesian case, the inferred probability does not follow an autoregressive process, and the expected flight duration diverges).

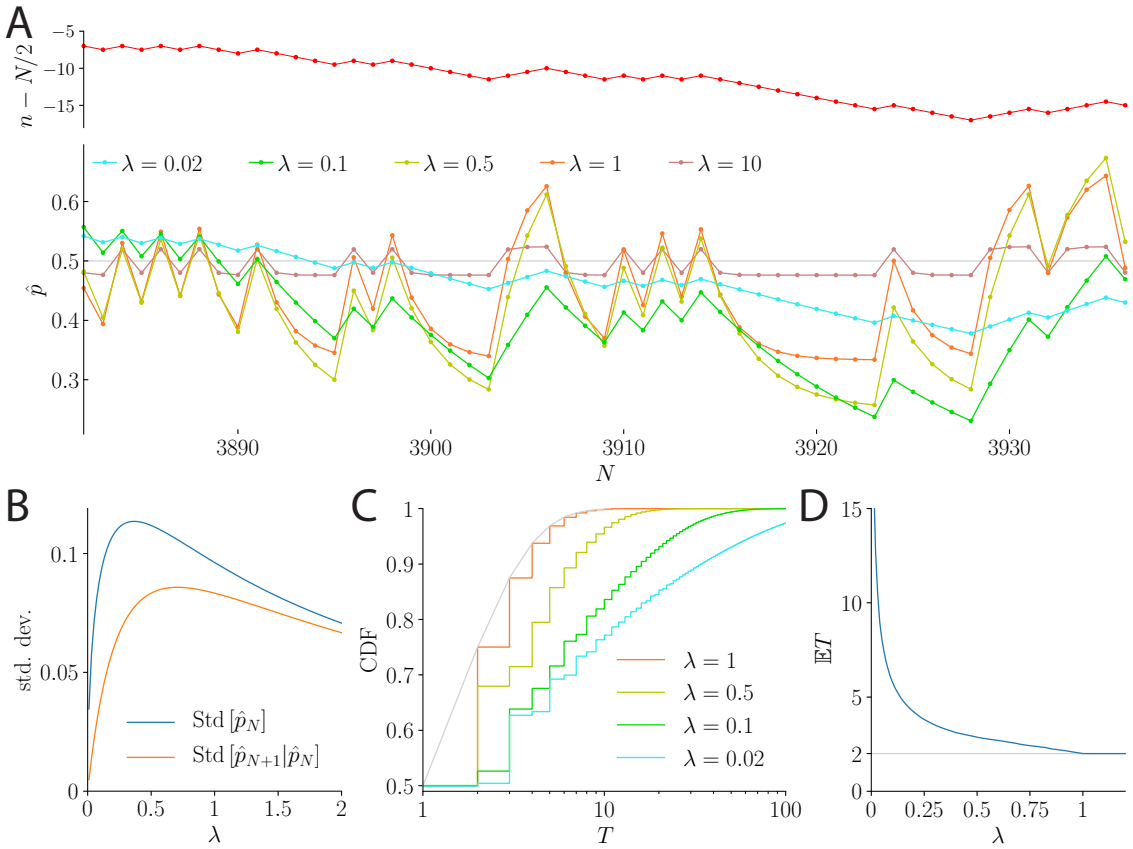


Figure 4. Behavior of the precision-cost model over 55 flips of a fair coin. (**A**) *Top:* Random walk described by the fluctuations in the number n of 'heads', relative to $N/2$, over 55 successive coin flips. *Bottom:* Inferred probability, $\hat{p}$, over the course of the sequence, for various values of λ. Under a weak cost (λ small), short-term variations in the inferred probability are small, and flights away from $1/2$ are long. Under stronger costs, the inferred probability is susceptible to larger changes, as it varies with the recent history of coin flips, and flights are shorter, but the overall magnitude of the fluctuations becomes smaller for larger λ. (**B**) Standard deviation of the inferred probability, $\hat{p}_N$, unconditional (blue), and conditional on the preceding inferred probability (orange). The latter is also the standard deviation of the change in the inferred probability following a coin flip, conditional on the preceding inferred probability, $\mathrm{Std}[\hat{p}_{N+1} - \hat{p}_N|\hat{p}_N]$. (**C**) Cumulative distribution function of the flight duration, T (i.e., the number of flips during which the model subject maintains a belief in a given sign of the coin's bias), for several values of λ. (**D**) Expected flight duration, $\mathbb{E}[T]$, as a function of the strength of the cost, λ.

3. Discussion

3.1. Summary

We have proposed and investigated a theoretical account of cognitive biases that emerge when humans make inferences about probabilities. The main postulate of this theory is that the brain aims to carry out sound, Bayesian inference, but cognitive cost—a functional of the inferred posterior—hinders its ability to do so. Instead of choosing the Bayesian posterior, the brain chooses another, less costly, distribution; we formalize this trade-off as ensuing from the minimization of a loss function (Equation (1)). We consider two different costs, and examine the resulting behaviors in the case of coin flips, i.e., when a Bernoulli probability is inferred.

The *precision cost* penalizes less entropic distributions (e.g., distributions concentrated around some value). Under this cost, the posterior does not converge, but fluctuates with the recent history of coin flips, and the average inferred probability is less extreme than the true probability (i.e., it is closer to $1/2$). In the case of a fair coin ($p = 1/2$), the model subject believes that the coin has a bias of a given sign for durations that decrease with

the strength of the cost. With a strong cost ($\lambda > 1$, i.e., the weight of the cost in the loss function is greater than the weight of the distance measure), just one coin flip is enough to switch the sign of the inferred bias. The *unpredictability cost* penalizes posteriors that imply unpredictable environments. Under this cost, if the coin is biased ($p \neq 1/2$), the inferred posterior converges to a probability that is more extreme than the true probability (i.e., it is further from $1/2$, and closer to 0 or to 1), and if the cost is strong, an infinitesimal true bias of the coin results in an appreciable inferred bias. In the case of a fair coin ($p = 1/2$), and with a strong cost ($\lambda > 1$), the posterior does not converge, and the inferred probability abruptly switches between two values, one above and one below $1/2$, as the true empirical bias fluctuates between small positive and negative values; however, the belief in a given sign of the bias typically lasts for a long duration, comparable to that of the experiment.

3.2. Costly Inference vs. Erroneous Beliefs

The *precision cost* yields an inference process in which past observations are gradually forgotten (Equation (5)). This results from the structure of the loss function (Equation (1)) in which the objective of learning the true value of the probability competes with the cost of being precise in doing so. The cost incites the model subject to discard, after each coin flip, some amount of information. (Another model subject bearing the same precision cost, but endowed with a different objective, may choose the information to be discarded differently.) This forgetfulness results from a cost that is large for narrow probability densities and, hence, can be interpreted as penalizing a form of overfitting, in line with the maximum-entropy principle [26,27].

A similar exponential forgetfulness is an essential assumption of some of the models proposed in the literature on biases in human inference; there, forgetfulness is interpreted as a rational process in the face of environments governed by statistics that vary over time. Some experimental tasks are indeed set in such a context, as when human subjects are asked to make inferences in volatile environments [16,19,21,28,29] (see also models of temporal discounting [30–34]). In such situations, older events provide less useful information than more recent ones, and, as such, forgetfulness is a way to avoid using 'outdated' information. However, a number of theoretical studies go beyond this view and assume that (exponential) forgetfulness occurs even in situations in which the environmental statistics are static [9,35]. They argue that humans are so strongly adapted to changing environments that they still hold this belief even after having been exposed to a static environment over a long period of time. What such an approach is tacit about, however, is how the specific temporal structure characterizing the belief is chosen or how it may change as a function of experience. This view is to be contrasted with the framework presented here, in which there is no explicit assumption of forgetfulness; instead, a decaying memory arises as a consequence of a cognitive cost on the precision of internal representations. While we have applied our framework to, arguably, the simplest case of a binary, i.i.d. random variable, it may be applied to more general problems.

It is worth noting that the topic of the correctness or incorrectness of beliefs in the context of inference extends well beyond the question of whether these should include a temporal dependence. Many observed biases may be explained as resulting from a particular choice of incorrect belief or prior. Indeed, the complete class theorem [36,37] shows that any set of choices is Bayes optimal under some form of prior belief. One can, therefore, derive from behavioral data the prior belief that makes the behavior optimal (at least in principle). This observation underlines the relevance of carrying out behavioral experiments that can cover a range of conditions and parameters, with the aim of examining how putative beliefs vary and what explanation of the behavior is most parsimonious.

In the spirit of the complete class theorem, the *unpredictability cost* can be interpreted as embodying the belief that predictable random processes are more likely than unpredictable ones, in the context of hierarchical inference over sources of uncertainty. A subject behaving according to our model under unpredictability cost can then, alternatively, be viewed as behaving rationally provided she holds incorrect beliefs. But if cognitive constraints *do*

exist, then one may not need to have recourse to assuming incorrect beliefs in order to explain biases. Furthermore, procedurally the two approaches are not equivalent: a costly inference model can predict trends in behavior as a function of variations of the experimental conditions, whereas models invoking erroneous beliefs must, in addition, specify how these may change as a function of experimental conditions.

3.3. Costly Inference vs. Variational Inference

Our approach, in this study, is thus to postulate that humans approximate a Bayesian observer because of cognitive limitations. This idea is the object of an active debate in cognitive neuroscience [38–41], set within a larger picture according to which structural constraints curb brain functions in perception, inference, learning, and decision-making [42–46]. In cognitive science, this picture can be traced back to the concept of 'bounded rationality', introduced by Simon [47,48], but the proposal of a 'resource-rational' approach [42–44] is perhaps closest to our approach, as it explicitly formalizes the mechanisms used by the brain as resulting from the optimization of a loss function equal to a negative objective plus a cost (in our case, Equation (1), see also [22]).

Which objective and cost best characterize the loss function optimized by the brain in a given situation remains a matter of debate [49]. Here, we have focused on an objective function chosen as the Kullback–Leibler divergence between the inferred distribution and the Bayesian posterior, $D_{KL}(\hat{P}||P)$. This choice suggests an interesting connection with so-called 'variational approaches'. In our case, the objective function is traded off with a cost. In the case of variational approaches, the often intractable optimization of the objective function is replaced by a tractable problem in which the optimization is carried out over a restricted class of (parameterized) functions [38,41,50–53]. Our formulation (Equation (1)) can thus be viewed as a variant of a variational approach: in most instances of the latter, functions belonging to a restricted class bear no cost, while all other functions come at an infinite cost; in our case, all posteriors are allowed, but incur an information-theoretic cost. Although it is possible that the brain can only carry out inferential computations using a restricted family of distributions, we have chosen to study two 'smoother' costs, that do not preclude *a priori* the use of any kind of distributions. These two costs are grounded in information theory: the precision cost varies with the entropy of the posterior, while the unpredictability cost depends on the entropy of the inferred probabilities.

The connection of our framework with variational inference is particularly relevant in the context of studies that rely on the minimization of a quantity that has become known as 'evidence bound' in machine learning [54] and 'free energy' in cognitive science [55]. This quantity is equal to the sum of the Kullback–Leibler divergence between the inferred posterior and the Bayesian posterior, $D_{KL}(\hat{P}||P)$, i.e., the objective function in Equation (1), and a term that does not depend upon the inferred posterior. Thus, the minimization of the free energy, in these studies, amounts to the minimization of our objective function, but it is carried out within a restricted class of posteriors in the case of variational inference, while in our case, the objective function is regularized by a cost. In an alternative formulation, the free energy can be written as the sum of two terms, the entropy of the inferred posterior, i.e., our precision cost, and an 'energy' term. Optimization then becomes a version of the maximum-entropy principle [26], in which the energy term (in our case, the first term on the right-hand-side of Equation (1)) acts as a constraint or regularizer, while the entropy is maximized. This interpretation of Equation (1) is dual to the one we have entertained throughout the paper, namely that entropy is used as a regularizer.

3.4. Costly Inference and Human Behavior

Our goal, in this study, was not to establish the extent to which each model captures the behavior of human subjects quantitatively. We note, however, that recent models of human inference are based on exponentially filtered counts of the outcomes [9,35], precisely of the kind that result from our precision cost (Equation (5)). In addition, this cost predicts that inferred probabilities are less extreme than the true probabilities (Figure 1B),

a behavior that was reported in experiments of decision-making under risk with human subjects [10,11]. Conversely, in decisions from experience, subjects were found to underestimate the probability of rare events [12]; the unpredictability cost predicts a bias in this direction (Figure 1D).

The different predictions of our models call for a detailed and quantitative examination of the behavior of human subjects in inference tasks. Most existing empirical studies of human inference of probabilities focus on the singular case of an equiprobable environment ($p = 1/2$) [35,56,57]. The models, however, make predictions about how the biases in inference should vary with the Bernoulli probability of outcomes, p. This begs for the design of a more flexible experimental paradigm, in which human behavior is investigated in environments that reflect a range of Bernoulli probabilities or, more broadly, an array of different generative statistics. This kind of experimental investigation, which addresses behavioral biases quantitatively and over a broad range of conditions, is not only warranted in the context of fundamental research, but may prove useful also in the nascent field of computational psychiatry. It would provide a basis to study and compare the structure of cognitive processes in patients and healthy subjects [58–61].

4. Methods

4.1. Precision Cost: Variance of the Posterior

We derive an approximation of the variance of the posterior inferred by the model subject under a precision cost with strength λ. The posterior at trial N is a Beta distribution with parameters $\hat{n}_1$ and $\hat{n}_0$ (Equations (4) and (5)). Its variance is

$$\mathrm{Var}[\hat{P}_N] = \frac{(\hat{n}_1+1)(\hat{n}_0+1)}{(\hat{n}_1+\hat{n}_0+2)^2(\hat{n}_1+\hat{n}_0+3)}. \tag{10}$$

For N large, we can calculate the expected value of this variance, as

$$\mathbb{E}\,\mathrm{Var}[\hat{P}_N] = \lambda\frac{4p(1-p)+\lambda(1+\lambda)(2+\lambda)}{(1+2\lambda)^2(1+3\lambda)(2+\lambda)}, \tag{11}$$

where p is the true value of the probability. For small λ, we obtain the approximation

$$\mathbb{E}\,\mathrm{Var}[\hat{P}_N] \simeq \lambda 4p(1-p), \tag{12}$$

while for large λ, we obtain

$$\mathbb{E}\,\mathrm{Var}[\hat{P}_N] \simeq \frac{1}{12}\Big(1-\frac{1}{3\lambda}\Big). \tag{13}$$

Note that $1/12$ is the variance of the uniform distribution on the interval $[0,1]$.

4.2. Precision Cost: Autoregressive Process

We show, here, that for large N the inference under the precision cost results in the inferred probability following an autoregressive process. At trial N, the mean, $\hat{p}_N$, of the posterior, $\hat{P}_N(p)$, defined in Equation (4), is a function of the exponentially filtered counts, $\hat{n}_0$ and $\hat{n}_1$ (Equation (5)), as

$$\hat{p}_N = \frac{\hat{n}_1+1}{\hat{n}_0+\hat{n}_1+2} \tag{14}$$

$$= \frac{\sum_{k=0}^{N-1}(\frac{1}{1+\lambda})^{k+1}x_{N-k}+1}{\sum_{k=0}^{N-1}(\frac{1}{1+\lambda})^{k+1}+2}. \tag{15}$$

Taking the expectation of this expression, we obtain, for large N, the expected inferred probability, as

$$\mathbb{E}[\hat{p}] = \frac{1}{1+2\lambda}p + \frac{2\lambda}{1+2\lambda}\frac{1}{2}, \tag{16}$$

an intermediate value between $1/2$ and the true value of the probability, p. Thus, in this model, the inferred probabilities are less 'extreme' than the true probabilities (Figure 1B). Furthermore, we can express the inferred probability at trial $N+1$ as a function of the inferred probability at trial N. We obtain, for large N,

$$\hat{p}_{N+1} = \frac{\lambda}{1+\lambda}\mathbb{E}[\hat{p}] + \frac{1}{1+\lambda}\hat{p}_N + \frac{\lambda}{(1+\lambda)(1+2\lambda)}(x_{N+1} - p), \tag{17}$$

which defines an autoregressive process of order 1 with coefficient $1/(1+\lambda)$.

Author Contributions: Conceptualization, R.A.d.S.; methodology, A.P.-C., R.A.d.S.; formal analysis, A.P.-C., M.T., R.A.d.S.; writing—original draft preparation, A.P.-C., R.A.d.S.; writing—review and editing, A.P.-C., F.M., M.T., R.A.d.S.; visualization, A.P.-C.; supervision, R.A.d.S.; project administration, R.A.d.S.; funding acquisition, R.A.d.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Alfred P. Sloan Foundation through grant G-2020-12680, the CNRS through UMR8023, the Global Scholar Program at Princeton University and the Visiting Faculty Program at the Weizmann Institute of Science. A.P.C. was supported by a Ph.D. fellowship of the Fondation Pierre-Gilles de Gennes pour la Recherche.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wikipedia. List of Cognitive Biases. Wikipedia, The Free Encyclopedia. 2016. Available online: https://en.wikipedia.org/wiki/List_of_cognitive_biases (accessed on 19 September 2020).
2. Hubel, D.H. *Eye, Brain, and Vision*; Scientific American Library/Scientific American Books: New York, NY, USA, 1995.
3. Baron, J. *Thinking and Deciding*; Cambridge University Press: Cambridge, UK, 2000.
4. Wendt, D.; Vlek, C. *Utility, Probability, and Human Decision Making: Selected Proceedings of an Interdisciplinary Research Conference, Rome, 3–6 September, 1973*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 11.
5. Gilovich, T.; Griffin, D.; Kahneman, D. *Heuristics and Biases: The Psychology of Intuitive Judgment*; Cambridge University Press: Cambridge, UK, 2002.
6. Hilbert, M. Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychol. Bull.* **2012**, *138*, 211. [CrossRef]
7. Group, T.M.A.D.; Fawcett, T.W.; Fallenstein, B.; Higginson, A.D.; Houston, A.I.; Mallpress, D.E.; Trimmer, P.C.; McNamara, J.M. The evolution of decision rules in complex environments. *Trends Cogn. Sci.* **2014**, *18*, 153–161.
8. Summerfield, C.; Tsetsos, K. Do humans make good decisions? *Trends Cogn. Sci.* **2015**, *19*, 27–34. [CrossRef]
9. Meyniel, F.; Maheu, M.; Dehaene, S. Human Inferences about Sequences: A Minimal Transition Probability Model. *PLoS Comput. Biol.* **2016**, *12*, 1–26. [CrossRef]
10. Gonzalez, R.; Wu, G. On the Shape of the Probability Weighting Function. *Cogn. Psychol.* **1999**, *38*, 129–166. [CrossRef] [PubMed]
11. Zhang, H.; Ren, X.; Maloney, L.T. The bounded rationality of probability distortion. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 22024–22034. [CrossRef] [PubMed]
12. Hertwig, R.; Barron, G.; Weber, E.U.; Erev, I. Decisions from experience and the effect of rare events in risky choice. *Psychol. Sci.* **2004**, *15*, 534–539. [CrossRef] [PubMed]
13. Ma, W.J. Organizing probabilistic models of perception. *Trends Cogn. Sci.* **2012**, *16*, 511–518. [CrossRef] [PubMed]
14. Weiss, Y.; Simoncelli, E.P.; Adelson, E.H. Motion illusions as optimal percepts. *Nat. Neurosci.* **2002**, *5*, 598–604. [CrossRef] [PubMed]
15. Stocker, A.A.; Simoncelli, E.P. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* **2006**, *9*, 578–585. [CrossRef] [PubMed]
16. Khaw, M.W.; Stevens, L.; Woodford, M. Discrete adjustment to a changing environment: Experimental evidence. *J. Monet. Econ.* **2017**, *91*, 88–103. [CrossRef]
17. Acerbi, L.; Vijayakumar, S.; Wolpert, D.M. On the origins of suboptimality in human probabilistic inference. *PLoS Comput. Biol.* **2014**, *10*, e1003661. [CrossRef]

18. Drugowitsch, J.; Wyart, V.; Devauchelle, A.D.; Koechlin, E. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron* **2016**, *92*, 1398–1411. [CrossRef]
19. Prat-Carrabin, A.; Wilson, R.C.; Cohen, J.D.; da Silveira, R.A. Human Inference in Changing Environments With Temporal Structure. *Psychol. Rev.* **2021**. [CrossRef]
20. Gigerenzer, G.; Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **2011**, *62*, 451–482. [CrossRef] [PubMed]
21. Gallistel, C.R.; Krishan, M.; Liu, Y.; Miller, R.; Latham, P.E. The perception of probability. *Psychol. Rev.* **2014**, *121*, 96–123. [CrossRef] [PubMed]
22. Icard, T.F.; Goodman, N.D. A Resource-Rational Approach to the Causal Frame Problem. In Proceedings of the 37th Annual Meeting of the Cognitive Science Society, Pasadena, CA, USA, 22–25 July 2015; pp. 962–967.
23. Benjamin, D.J. Errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics*; Elsevier B.V.: Amsterdam, The Netherlands, 2019; Volume 2, pp. 69–186. [CrossRef]
24. Feller, W. *An Introduction to Probability Theory and Its Application*, 3rd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1967; Volume 1, pp. 1–525.
25. Novikov, A.; Kordzakhia, N. Martingales and first passage times of AR(1) sequences. *Stochastics* **2007**, *80*, 197–210. [CrossRef]
26. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
27. Banavar, J.R.; Maritan, A.; Volkov, I. Applications of the principle of maximum entropy: From physics to ecology. *J. Phys. Condens. Matter* **2010**, *22*, 063101. [CrossRef] [PubMed]
28. Rushworth, M.F.; Behrens, T.E. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat. Neurosci.* **2008**, *11*, 389–397. [CrossRef]
29. Mathys, C.; Daunizeau, J.; Friston, K.J.; Stephan, K.E. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **2011**, *5*, 1–20. [CrossRef] [PubMed]
30. Commons, M.L.; Woodford, M.; Ducheny, J.R. How Reinforcers are Aggregated in Reinforcement-density discrimination and Preference Experiments. In *Quantitative Analyses of Behavior: Volume 2, Matching and Maximizing Accounts*; Ballinger: Cambridge, MA, USA, 1982; Volume 2, pp. 25–78.
31. Commons, M.L.; Woodford, M.; Trudeau, E.J. How Each Reinforcer Contributes to Value: "Noise" Must Reduce Reinforcer Value Hyperbolically. In *Signal Detection: Mechanisms, Models, and Applications*; Commons, M.L., Nevin, J.A., Davison, M.C., Eds.; Lawrence Erlbaum: Hillsdale, NJ, USA, 1991; pp. 139–168.
32. Sozou, P.D. On hyperbolic discounting and uncertain hazard rates. *Proc. R. Soc. B Biol. Sci.* **1998**, *265*, 2015–2020. [CrossRef]
33. Green, L.; Myerson, J. A discounting framework for choice with delayed and probabilistic rewards. *Psychol. Bull.* **2004**, *130*, 769–792. [CrossRef] [PubMed]
34. Gabaix, X.; Laibson, D. *Myopia and Discounting*; NBER Working Paper No. 23254; NBER: Cambridge, MA, USA, 2017. [CrossRef]
35. Yu, A.J.; Cohen, J.D. Sequential effects: Superstition or rational behavior? *Adv. Neural Inf. Process. Syst.* **2008**, *21*, 1873–1880.
36. Brown, L.D. A Complete Class Theorem for Statistical Problems with Finite Sample Spaces. *Ann. Stat.* **2007**, *9*, 1289–1300. [CrossRef]
37. Wald, A. An Essentially Complete Class of Admissible Decision Functions. *Ann. Math. Stat.* **1947**, *18*, 549–555. [CrossRef]
38. Penny, W. Bayesian Models of Brain and Behaviour. *ISRN Biomath.* **2012**, *2012*, 785791. [CrossRef]
39. Pouget, A.; Beck, J.M.; Ma, W.J.; Latham, P.E. Probabilistic brains: Knowns and unknowns. *Nat. Neurosci.* **2013**, *16*, 1170–1178. [CrossRef]
40. Sanborn, A.N. Types of approximation for probabilistic cognition: Sampling and variational. *Brain Cogn.* **2015**, 8–11. [CrossRef]
41. Gershman, S.J.; Beck, J.M. Complex Probabilistic Inference: From Cognition to Neural Computation. In *Computational Models of Brain and Behavior*; John Wiley & Sons: Hoboken, NJ, USA, 2016; pp. 1–17. [CrossRef]
42. Griffiths, T.L.; Lieder, F.; Goodman, D. Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Top. Cogn. Sci.* **2015**, *7*, 217–229. [CrossRef]
43. Lieder, F.; Griffiths, T.L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **2019**. [CrossRef]
44. Griffiths, T.L. Understanding Human Intelligence through Human Limitations. *Trends Cogn. Sci.* **2020**, *24*, 873–883. [CrossRef] [PubMed]
45. Bhui, R.; Lai, L.; Gershman, S.J. Resource-rational decision making. *Curr. Opin. Behav. Sci.* **2021**, *41*, 15–21. [CrossRef]
46. Summerfield, C.; Parpart, P. Normative principles for decision-making in natural environments. *PsyArXiv* **2021**. [CrossRef]
47. Simon, H.A. A Behavioral Model of Rational Choice. *Q. J. Econ.* **1955**, *69*, 99–118. [CrossRef]
48. Simon, H.A. *Models of Bounded Rationality: Empirically Grounded Economic Reason*; MIT Press: Cambridge, MA, USA, 1997.
49. Ma, W.J.; Woodford, M. Multiple conceptions of resource rationality. *Behav. Brain Sci.* **2020**, *43*, e15. [CrossRef] [PubMed]
50. Ghahramani, Z.; Jordan, M.I. Factorial Hidden Markov Models. *Mach. Learn.* **1997**, *29*, 245–273. [CrossRef]
51. Dauwels, J. On variational message passing on factor graphs. *IEEE Int. Symp. Inf. Theory Proc.* **2007**, 2546–2550. [CrossRef]
52. Beal, M.J. Variational Algorithms for Approximate Bayesian Inference. Ph.D. Thesis, University College London, London, UK, 2003.
53. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
54. Winn, J.; Bishop, C. Variational message passing. *J. Mach. Learn. Res.* **2005**, *6*, 661–694.
55. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev.* **2010**, *11*, 127–138. [CrossRef] [PubMed]

56. Cho, R.Y.; Nystrom, L.E.; Brown, E.T.; Jones, A.D.; Braver, T.S.; Cohen, J.D. Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task. *Cogn. Affect. Behav. Neurosci.* **2002**, *2*, 283–299. [CrossRef] [PubMed]
57. Gökaydin, D.; Ejova, A. Sequential effects in prediction. In Proceedings of the Annual Conference Cognitive Science Society, London, UK, 26–29 July 2017; pp. 397–402.
58. Stephan, K.E.; Mathys, C. Computational approaches to psychiatry. *Curr. Opin. Neurobiol.* **2014**, *25*, 85–92. [CrossRef]
59. Adams, R.A.; Huys, Q.J.; Roiser, J.P. Computational Psychiatry: Towards a mathematically informed understanding of mental illness. *J. Neurol. Neurosurg. Psychiatry* **2016**, *87*, 53–63. [CrossRef] [PubMed]
60. Schwartenbeck, P.; Friston, K. Computational phenotyping in psychiatry: A worked example. *eNeuro* **2016**, *3*, 47. [CrossRef]
61. Ashinoff, B.K.; Singletary, N.M.; Baker, S.C.; Horga, G. Rethinking delusions: A selective review of delusion research through a computational lens. *Schizophr. Res.* **2021**. [CrossRef]