HAL open science

Rationally Biased Learning

Michel de Lara

▶ To cite this version:

Michel de Lara. Rationally Biased Learning. 2017. hal-01581982v1

HAL Id: hal-01581982
https://hal.science/hal-01581982v1

Preprint submitted on 5 Sep 2017 (v1), last revised 22 Mar 2022 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Rationally Biased Learning

Michel De Lara
CERMICS, École des Ponts, UPE, Champs-sur-Marne, France.
E-mail: michel.delara@enpc.fr

September 5, 2017

Abstract

Are human perception and decision biases grounded in a form of rationality? You return to your camp after hunting or gathering. You see the grass moving. You do not know the probability that a snake is in the grass. Should you cross the grass — at the risk of being bitten by a snake — or make a long, hence costly, detour? Based on this storyline, we consider a rational decision maker maximizing expected discounted utility with learning. We show that his optimal behavior displays three biases: status quo, salience, overestimation of small probabilities. Biases can be the product of rational behavior.

Keywords: status quo bias, salience bias, overestimation of small probabilities, optimal behavior

Contents

1 Introduction **2**

2 Optimal decision-making under risk **3**
2.1 Error management theory 4
2.2 Error management theory supports adaptive biases 6
2.3 From risk to uncertainty with unknown probability 6

3 Optimal decision-making with learning **7**
3.1 A model for sequential learning 7
3.2 An expected discounted payoff maximization problem 9
3.3 Gittins index optimal strategy (GI-strategy) 11

4 Killing three biases with one stone **14**
4.1 Status quo bias 14
4.2 Salience bias 14
4.3 Overestimation of small probabilities bias 15

5 Conclusion **20**

1 Introduction

When we perceive sounds, we overestimate the change in level of rising level tones relative to equivalent falling level tones Neuhoff (1998). When we assess pros and cons in decision making, we weigh losses more than gains Kahneman and Tversky (1979).

Are some of our perception and decision biases grounded in a form of rationality? In the "heuristics and biases" literature Kahneman et al. (1982); Gilovich et al. (2002), a debate opposes two conceptions: on the one hand, some behaviors are qualified of "bias" when they depart from given "rationality benchmarks" (like expected utility theory); on the other hand, some scholars (see Gigerenzer (2004, 2008); Hutchinson and Gigerenzer (2005)) claim that those "so-called bias" were in fact advantageous in the type of environment where our ancestors lived and thrived (ecological or, rather, evolutionary, validity Barkow et al. (1992); Boutang and De Lara (2015)). In this second conception, the benchmark should be a measure of fitness reflecting survival and reproduction abilities.

We tackle a limited version of this broad issue by proposing insight from a decision model. Closing the gap between the two conceptions, we will show that a rational decision maker — maximizing expected discounted utility — can display biases.

For this purpose, we will put forward a model inspired by the one given by Haselton and Nettle in Haselton and Nettle (2006). They consider an individual who, to reach his destination, has two options: a short risky route passes through a grassy land — possibly hiding a poisonous snake inflicting serious (though non lethal) pains; whereas a safe route makes a long costly detour. They define a critical probability related to costs of encounter and avoidance. They show that, for a "bad" outcome with probability lower than the critical probability, the (optimal) individual should take the risky route. The general conclusion is nicely expressed by Martie G. Haselton (on her personal webpage) when she claims that "selection has led to adaptations that are biased by design and functioned to help ancestral humans avoid particularly costly errors". However, the so-called "error management theory" analysis in Haselton and Nettle (2006) is performed supposing known the probability that a snake is in the grass.

What happens when the individual does not know *a priori* the objective probability driving the occurence of a bad outcome? In this paper, we will consider a decision maker (DM) who has to make successive decisions. At the beginning of every period, he has to choose between two options: if he makes a "risky decision", he receives a random payoff, depending on a bad (snake) or good (no snake) outcome revealed at the end of the period (learning); if he makes a "safe decision", he receives a deterministic payoff (safe way with a long costly detour) and does not observe the outcome (good or bad). The occurrence of the bad or good outcome is drawn from a Bernoulli distribution with unknown probability. We suppose that the DM is Bayesian in that he holds a prior — an estimate of the (unknown) Bernoulli distribution. We suppose that the DM maximizes discounted expected payoffs.

We will show that a DM following an optimal strategy displays a behavior with the following three biases:

- *status quo bias*: there are at most two consecutive phases of "status quo" — one (possibly empty) of experimenting, that is, making a "risky decision", one (possibly empty) of prudence — with at most one switch; in particular, once prudent, this is forever;

- *salience bias*: if it exists, the experimenting/learning phase can only

stop when a bad outcome materializes; in other words, the switch from riskiness to prudence occurs when a salient and vivid outcome occurs;

- *probability overestimation bias (of bad and unlikely outcomes)*: the probability of a bad and unlikely outcome is overestimated when the experimenting/learning phase stops.

The paper is organized as follows. In Sect. 2, we recall a decision model from Haselton and Nettle (2006) where the probability driving the occurence of a bad outcome is supposed to be known. In Sect. 3, we extend the model to the case of unknown probability, and we exhibit the behavior of the optimal DM. In Sect. 4, we show features of the optimal DM that have interesting psychological interpretations in terms of human biases. Interpreting costs as a measure of fitness (the lower the costs, the higher the fitness), we conclude in Sect. 5 that natural selection may have favored individuals who display status quo bias, salience bias and probability overestimation bias.

2 Optimal decision-making under risk

Before examining learning in the next Sect. 3, we first focus on decision-making under risk, by means of an example from Haselton and Nettle (2006). Haselton and Nettle make use of what they call *error management theory* (EMT) to understand "how natural selection engineers psychological adaptations for judgment under uncertainty". We will follow their terminology, though EMT amounts to optimal decision-making under risk, that is, with uncertainties following a known probability distribution.

2.1 Error management theory

In Haselton and Nettle (2006), the following situation is examined. Consider two possible outcomes (states of Nature) — a "bad" one $\mathtt{B}$ and a "good" one $\mathtt{G}$ — that we illustrate by $\mathtt{B}$ = "a snake is in the grass", and by $\mathtt{G}$ the contrary. Now, suppose that the grass is moving. According to our belief in what makes the grass moving — either "a snake is believed to be in the grass", or the contrary — two decisions are possible. Supposing a snake in the grass leads to avoid (α) the place and make a costly detour. On the other hand, passing through ("trying", "learning", "experimenting" ε) is less

costly if the snake is not, but can be very costly (painful though not lethal) if the snake is present.

Then, a belief can be adopted when it is in fact true (a true positive or TP), or it cannot be adopted and not be true (a true negative or TN). There are two possible errors: a false positive (FP) error occurs when a person adopts a belief that is not in fact true (believing there is a snake, when this is not the case); a false negative (FN) occurs when a person fails to adopt a belief that is true. This is summarized in Table 1:

- assuming there is a snake induces avoidance (α) with cost $\mathcal{C}_\alpha$ – like losing time in making a detour — let there indeed be a snake (B, true positive TP) or not (G, false positive FP);
- assuming there is no snake induces a cost $\mathcal{C}^{\mathtt{B}}$ of encounter (bad B) — being bitten, with painful and incapacitating consequences — if there indeed is a snake (B, false negative FN) and a cost $\mathcal{C}^{\mathtt{G}}$ of no encounter if not (G, true negative TN).

	snake (bad B)	no snake (good G)
avoid (α)	cost of avoidance $\mathcal{C}_\alpha$	cost of avoidance $\mathcal{C}_\alpha$
experiment (ε)	cost of encounter $\mathcal{C}^{\mathtt{B}}$	cost of no encounter $\mathcal{C}^{\mathtt{G}}$

Table 1: Costs according to decisions (rows) and states of Nature (columns)

We suppose that the cost of avoidance lies between the costs of good and bad encounters:

$$\overbrace{\mathcal{C}^{\mathtt{G}}}^{\text{cost of no encounter}} < \underbrace{\mathcal{C}_\alpha}_{\text{cost of avoidance}} < \overbrace{\mathcal{C}^{\mathtt{B}}}^{\text{cost of encounter}} . \tag{1}$$

Now, assume that a snake is in the grass with probability $p^{\mathtt{B}}$, and no snake is in the grass with probability $p^{\mathtt{G}} = 1 - p^{\mathtt{B}}$. If the above situation repeats itself independently, the empirical mean costs (over repetitions) are approximated by the theoretical expected costs, due to the Law of Large Numbers:

- assuming there is a snake makes you avoid, and induces the same cost whatever the state of Nature, which is the cost of avoidance $\mathcal{C}_\alpha$;
- assuming there is no snake makes you experiment — and induces a cost of encounter $\mathcal{C}^{\mathtt{B}}$ with probability $p^{\mathtt{B}}$, and a cost $\mathcal{C}^{\mathtt{G}}$ with probability $p^{\mathtt{G}}$ — hence a mean cost $p^{\mathtt{B}}\mathcal{C}^{\mathtt{B}} + p^{\mathtt{G}}\mathcal{C}^{\mathtt{G}}$.

Therefore, in the mean, it is better to believe there is a snake rather than not if the cost of avoidance falls below the mean cost of encounter, that is,[1]

$$\text{avoid}\alpha \iff \overbrace{p^{\mathtt{B}}\mathcal{C}^{\mathtt{B}} + p^{\mathtt{G}}\mathcal{C}^{\mathtt{G}}}^{\text{expected costs of crossing}} > \overbrace{\mathcal{C}_\alpha}^{\text{cost of avoidance}} . \tag{2}$$

Rearranging the last inequality using the property that $p^{\mathtt{B}} + p^{\mathtt{G}} = 1$, we summarize the optimal rule in the following Proposition 1.

Proposition 1. *We define the* critical probability p_c *— or* avoidability index p_c *— by the following ratio:*

$$p_c = \frac{\mathcal{C}_\alpha - \mathcal{C}^{\mathtt{G}}}{\mathcal{C}^{\mathtt{B}} - \mathcal{C}^{\mathtt{G}}} = \frac{\textit{relative costs of avoidance}}{\textit{relative costs of encounter}} \in]0, 1[. \tag{3}$$

The optimal DM under risk adopts the following rule:

$$\textit{avoid}\alpha \iff p^{\mathtt{B}} > p_c . \tag{4}$$

We have coined the critical probability p_c an avoidability index. Indeed, all things being equal, the worse a bad outcome (that is, high bad costs) and the lower the cost of avoidance, the lower p_c in (3): a low index p_c means that it is cheap to avoid the bad outcome, whereas an index p_c close to 1 means that the bad outcome can only be avoided at high costs. As a consequence, the higher the cost of encounter and the lower the cost of avoidance, the better to believe there is a snake rather than not, hence to avoid.

[1] We use a strict inequality below, because we do not consider the exceptional case where the two quantities are equal.

2.2 Error management theory supports adaptive biases

The conclusion of Haselton and Nettle is that, when errors are asymmetrical in cost, there is a tendency to favor false positive error (FP), that is, adopting a belief that is not in fact true (believing there is a snake, when this is not the case).

As Haselton claims (personal webpage), "when the costs of false positive and false negative errors were asymmetrical over evolutionary history, selection will have designed psychological adaptations biased in the direction of the less costly error".

Speaking of snakes, neuroscientist Joseph LeDoux has a nice way to express such bias, in his book *The Emotional Brain*: "It is better to have treated a stick as a snake than not to have responded to a possible snake" (LeDoux, 1996, p.166).

Such asymmetry in costs is manifest in the so-called *life-dinner* principle of Richard Dawkins — "The rabbit runs faster than the fox, because the rabbit is running for his life while the fox is only running for his dinner" — and can exert a strong selection pressure Dawkins and Krebs (1979).

2.3 From risk to uncertainty with unknown probability

The optimal decision rule (4) depends on two quantities attached to the situation: one is the probability $p^{\mathtt{B}}$ that a snake is in the grass; the other is the ratio (3) of two costs, which, being less than 1, we interpret as a probability p_c. Whereas each cost — "avoid" $\mathcal{C}_\alpha$, "experiment and snake" $\mathcal{C}^{\mathtt{B}}$, "experiment and no snake" $\mathcal{C}^{\mathtt{G}}$ — can be learnt by three single experiments, the probability $p^{\mathtt{B}}$ of encounter is usually learned as a frequency resulting from multiple encounters. This is what we discuss in the next Section.

3 Optimal decision-making with learning

To "implement" the rule (4) — namely, avoid or experiment depending on beliefs — the DM needs to know the probability $p^{\mathtt{B}}$ that a snake is in the grass. This assessment is acquired by experimenting and learning. In fact, the DM learns only if he crosses the grass. Such a mix of experimenting and

of acting is examplified in the famous "multi-armed bandit problem" Gittins (1979).

In §3.1, we lay out mathematical ingredients to set up a model with sequential learning of an unknown probability. In §3.2, we formulate an expected discounted payoff maximization problem. Finally, we recall in §3.3 that this problem displays an optimal strategy based upon a so-called Gittins index.

3.1 A model for sequential learning

We lay out mathematical ingredients to set up a model with sequential learning.

Discrete time span

We suppose that, at discrete times $t \in \mathbb{N}$, the DM makes a decision — either "avoid" (α) or "experiment" (ε) — without knowing in advance the state of Nature occurring at that time — either "bad" ($\mathtt{B}$) or "good" ($\mathtt{G}$). We denote by $t = 0, 1, 2 \dots$ the discrete time corresponding to the beginning of period $[t, t+1[$.

Sample Space

Define the *sample space*

$$\mathbb{H}_\infty = \{\mathtt{B}, \mathtt{G}\}^{\mathbb{N}^*} = \{\mathtt{B}, \mathtt{G}\} \times \{\mathtt{B}, \mathtt{G}\} \times \cdots \ , \tag{5}$$

with generic element an infinite sequence $(\omega_1, \omega_2, \dots)$ of elements in $\{\mathtt{B}, \mathtt{G}\}$. Denote by

$$X_{t+1} : \mathbb{H}_\infty \to \{\mathtt{B}, \mathtt{G}\} \ , \ \ X_{t+1}(\omega_1, \omega_2, \dots) = \omega_{t+1} \ , \tag{6}$$

the *state of Nature at time* $t = 1, 2 \dots$ The index $t+1$ stresses the fact that the state of Nature X_{t+1}, though realized at the beginning of period $[t, t+1[$, cannot be revealed before time $t+1$.

Strategies

At the beginning of each period $[t, t+1[$, the DM can either "avoid" (decision α) — in which case the DM has no information about the state of

Nature — or "experiment" (decision ε)– in which case the state of Nature X_{t+1} ($\mathtt{B}$ or $\mathtt{G}$) is revealed and experimented, at the end of the period $[t, t+1[$. We note by $\{\alpha, \varepsilon\}$ the set of decisions, and by $v_t \in \{\alpha, \varepsilon\}$ the action taken by the DM at the beginning of the period $[t, t+1[$.

We assume that the DM is not visionary and learns only from the past: he cannot know the future in advance, neither can he know the state of Nature ($\mathtt{B}$ or $\mathtt{G}$) if he decides to avoid. Define the *observation sets* at time $t = 0, 1, 2, 3 \ldots$ by $\mathbb{O}_0 = \{\partial\}$, where ∂ corresponds to no information (no observation at initial time $t = 0$), and $\mathbb{O}_t = \{\mathtt{B}, \mathtt{G}, \partial\}^t$ for $t = 1, 2, 3 \ldots$ Define the *observation mapping* $\mathcal{O} : \{\alpha, \varepsilon\} \times \{\mathtt{B}, \mathtt{G}\} \to \{\mathtt{B}, \mathtt{G}, \partial\}$ by $\mathcal{O}(\varepsilon, \mathtt{B}) = \mathtt{B}$, $\mathcal{O}(\varepsilon, \mathtt{G}) = \mathtt{G}$ and $\mathcal{O}(\alpha, \mathtt{B}) = \mathcal{O}(\alpha, \mathtt{G}) = \partial$. Thus, the observation at time $t = 0, 1, 2 \ldots$ if the DM makes decision $v_t \in \{\alpha, \varepsilon\}$ is $Y_{t+1} = \mathcal{O}(v_t, X_{t+1})$.

We allow the DM to accumulate past observations; therefore the decision v_t at time t can only be a function of $Y_1, \ldots, Y_t$ (the initial decision v_0 is made without information). A *policy at time* t is a mapping $\mathcal{S}_t : \mathbb{O}_t \to \{\alpha, \varepsilon\}$ that tells the DM what will be his next action in view of past observations. A *strategy* $\mathcal{S}$ is a sequence $\mathcal{S}_0, \mathcal{S}_1 \ldots$ of policies.

Given the *scenario* $X(\cdot) = (X_1, X_2, \ldots)$ of states of Nature, and given a strategy $\mathcal{S}$, decisions and observations are inductively given by

$$v_t = \mathcal{S}_t(Y_1, \ldots, Y_t) \in \{\alpha, \varepsilon\} \text{ and } Y_{t+1} = \mathcal{O}(v_t, X_{t+1}) \in \{\mathtt{B}, \mathtt{G}, \partial\} . \tag{7}$$

True unknown probability distribution

Consider $\overline{p}^{\mathtt{B}} \geq 0$ and $\overline{p}^{\mathtt{G}} \geq 0$ such that $\overline{p}^{\mathtt{B}} + \overline{p}^{\mathtt{G}} = 1$. We equip the sample space $\mathbb{H}_\infty$ in (5) with the probability distribution $\overline{\mathbb{P}}$ such that the random process $(X_1, X_2, \ldots)$ is a sequence of independent Bernoulli trials with $\overline{\mathbb{P}}\{X_t = \mathtt{B}\} = \overline{p}^{\mathtt{B}}$ and $\overline{\mathbb{P}}\{X_t = \mathtt{G}\} = \overline{p}^{\mathtt{G}}$, that is,

$$\overline{\mathbb{P}} = \left(\overline{p}^{\mathtt{B}} \delta_{\mathtt{B}} + \overline{p}^{\mathtt{G}} \delta_{\mathtt{G}}\right)^{\otimes \mathbb{N}^*} . \tag{8}$$

Hypothesized probability

We suppose that the DM does not know the probability $\overline{p}^{\mathtt{B}}$ (nor $\overline{p}^{\mathtt{G}}$), but that he is a Bayesian assuming that

- the parameter $(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})$ is a random variable with a distribution π_0 on the one-dimensional simplex

$$S^1 = \{p^{\mathtt{B}} \geq 0 \ , \ \ p^{\mathtt{G}} \geq 0 \mid p^{\mathtt{B}} + p^{\mathtt{G}} = 1\} \ ; \tag{9}$$

- the extended sample space $S^1 \times \mathbb{H}_\infty = S^1 \times \{\mathtt{B}, \mathtt{G}\}^{\mathbb{N}^*}$ is equipped with the probability distribution

$$\pi_0(dp^{\mathtt{B}} dp^{\mathtt{G}}) \otimes \left(p^{\mathtt{B}}\delta_{\mathtt{B}} + p^{\mathtt{G}}\delta_{\mathtt{G}}\right)^{\otimes \mathbb{N}^*} . \tag{10}$$

We denote by $\mathbb{P}_{\pi_0}$ its marginal distribution on the sample space $\mathbb{H}_\infty$ in (5).

3.2 An expected discounted payoff maximization problem

Now, to compare strategies, we will make up a criterion. Beware that, in Sect. 2, we dealt with *costs* (to be minimized), whereas here, in Sect. 3, we deal with *payoffs* (to be maximized).

Instant payoffs

In an evolutionary interpretation, payoffs are measured in "fitness" unit. For instance, costs might be measured in "number of days alive" or "number of days in a reproductive state", taken as proxies for the number of offspring. The payoffs depend both on the decision and on the state of Nature as in Table 2.

	"bad" state $\mathtt{B}$	"good" state $\mathtt{G}$
avoid α	avoidance payoff $U(\alpha, \mathtt{B}) = \mathcal{U}_\alpha$	avoidance payoff $U(\alpha, \mathtt{G}) = \mathcal{U}_\alpha$
experiment ε	encounter payoff $U(\varepsilon, \mathtt{B}) = \mathcal{U}^{\mathtt{B}}$	base payoff $U(\varepsilon, \mathtt{G}) = \mathcal{U}^{\mathtt{G}}$

Table 2: Instant payoffs according to decisions (rows "avoid" (α) or "experiment" (ε)) and states of Nature (columns "bad" $\mathtt{B}$ or "good" $\mathtt{G}$)

We assume that the payoffs attached to the couple (action, state) in Table 2 are ranked as follows:[2]

$$\overbrace{U(\varepsilon, \mathtt{G}) = \mathcal{U}^{\mathtt{G}}}^{\text{base payoff}} > \underbrace{U(\alpha, \mathtt{B}) = U(\alpha, \mathtt{G}) = \mathcal{U}_\alpha}_{\text{avoidance payoff}} > \overbrace{U(\varepsilon, \mathtt{B}) = \mathcal{U}^{\mathtt{B}}}^{\text{encounter payoff}} . \tag{11}$$

[2]The relations between payoffs and the costs of Sect. 2 are $\mathcal{U}^{\mathtt{B}} = -\mathcal{C}^{\mathtt{B}}$, $\mathcal{U}^{\mathtt{G}} = -\mathcal{C}^{\mathtt{G}}$ and $\mathcal{U}_\alpha = -\mathcal{C}_\alpha$.

In other words, avoiding yields more utility than a bad encounter but less than a good one.

Intertemporal criterion

As the payoffs in Table 2 are measured in "fitness", we suppose that they are cumulative, like days in a healthy condition or number of offspring. This is why we suppose that the DM can evaluate his lifetime performance using strategy $\mathcal{S}$ by the discounted intertemporal payoff

$$J\big(\mathcal{S}, X(\cdot)\big) = \sum_{t=0}^{+\infty} \rho^t U(v_t, X_{t+1}) \ , \tag{12}$$

where v_t is given by (7).

The rationale behind using discounted intertemporal payoff is the following. Suppose that the DM's lifetime is a random variable θ — independent of the randomness in the occurence of a bad and good outcomes — that follows a Geometric distribution with values in $\{0, 1, 2, 3 \ldots\}$. Then, we have the relation $\sum_{t=0}^{+\infty} \rho^t U(v_t, X_{t+1}) = \mathbb{E}[\sum_{t=0}^{\theta} U(v_t, X_{t+1})]$. And we can interpret the discount factor $\rho \in [0, 1[$ as the mean $\overline{\theta}$ of the DM's lifetime θ, by means of the relations $\overline{\theta} = \rho/(1-\rho)$ and $\rho = \overline{\theta}/(\overline{\theta}+1)$. For instance, a discount factor $\rho = 0.95$ corresponds to a mean lifetime $\overline{\theta} = 0.95/0.05 = 19$ periods.

Expected discounted payoff maximization problem

Since the payoff (12) is contingent on the unknown scenario $X(\cdot) = (X_1, X_2, \ldots)$, it is practically impossible that a strategy $\mathcal{S}$ performs better than another for all scenarios. We look for an *optimal strategy* $\mathcal{S}^\star$, solution of

$$\mathbb{E}^{\mathbb{P}_{\pi_0}}\big[J\big(\mathcal{S}^\star, X(\cdot)\big)\big] = \max_{\mathcal{S}} \mathbb{E}^{\mathbb{P}_{\pi_0}}\big[J\big(\mathcal{S}, X(\cdot)\big)\big] \ , \tag{13}$$

where J is given by (12), and the probability $\mathbb{P}_{\pi_0}$ is the marginal distribution of (10) on the sample space $\mathbb{H}_\infty$ in (5).

> *Remark.* In Robson (2001), Robson chooses to maximize the total expected offspring in the limit as the horizon goes to infinity, uniformly in all originally unknown distribution, as inspired by the minimax approach for bandit problems (Berry and Fristedt, 1985, chap. 9).

3.3 Gittins index optimal strategy (GI-strategy)

It is well-known that the maximum in (13) is achieved by a so-called *Gittins index* strategy Gittins (1989); Berry and Fristedt (1985); Bertsekas (2000) as follows (see (Berry and Fristedt, 1985, Theorem 5.3.1)).

Bayesian update

Let $\Delta(S^1)$ denote the set of probability distributions on the simplex S^1 in (9). An optimal strategy for the optimization problem (13) can be searched for among state feedbacks strategies of the form

$$\mathcal{S}_t(Y_1, \ldots, Y_t) = \widehat{\mathcal{S}}(\widehat{\pi}_t) \text{ with } \widehat{\mathcal{S}} : \Delta(S^1) \to \{\varepsilon, \alpha\} \ , \tag{14}$$

where the *information state* is $\widehat{\pi}_t \in \Delta(S^1)$, the conditional distribution, with respect to $Y_1, \ldots, Y_t$, of the first coordinate mapping on $S^1 \times \mathbb{H}_\infty$.

The dynamics of the *posterior* $\widehat{\pi}_t \in \Delta(S^1)$ is given by:

$$\widehat{\pi}_0 = \pi_0 \text{ and } \widehat{\pi}_{t+1} = \begin{cases} \widehat{\pi}_t & \text{if } Y_{t+1} = \partial \\ \theta^{\mathtt{B}}\widehat{\pi}_t & \text{if } Y_{t+1} = \mathtt{B} \\ \theta^{\mathtt{G}}\widehat{\pi}_t & \text{if } Y_{t+1} = \mathtt{G} \ . \end{cases} \tag{15}$$

In this formula, the mappings $\theta^{\mathtt{B}}, \theta^{\mathtt{G}} : \Delta(S^1) \to \Delta(S^1)$ map a probability π on the simplex S^1 towards probabilities $\theta^{\mathtt{B}}\pi$ and $\theta^{\mathtt{G}}\pi$, absolutely continuous with respect to π, and given by:

$$(\theta^{\mathtt{B}}\pi)(dp^{\mathtt{B}}dp^{\mathtt{G}}) = \frac{p^{\mathtt{B}}}{\int_{S^1} p^{\mathtt{B}}\pi(dp^{\mathtt{B}}dp^{\mathtt{G}})}\pi(dp^{\mathtt{B}}dp^{\mathtt{G}}) \ , \tag{16a}$$

$$(\theta^{\mathtt{G}}\pi)(dp^{\mathtt{G}}dp^{\mathtt{G}}) = \frac{p^{\mathtt{B}}}{\int_{S^1} p^{\mathtt{G}}\pi(dp^{\mathtt{B}}dp^{\mathtt{G}})}\pi(dp^{\mathtt{B}}dp^{\mathtt{G}}) \ . \tag{16b}$$

Gittins index optimal strategy (GI-strategy)

Here is the *Gittins index* strategy.

Proposition 2 (Gittins (1989))**.** *There exists a function $\mathcal{I} : \Delta(S^1) \to \mathbb{R}$ (called the* Gittins index*) — which depends on the discount factor ρ and on the payoff U — such that the following strategy is optimal for the expected discounted payoff maximization problem* (13)*:*

- *if* $\mathcal{I}(\widehat{\pi}_t) < \mathcal{U}_\alpha$ *(index < sure payoff), then select decision* α *("avoid");*
- *if* $\mathcal{I}(\widehat{\pi}_t) > \mathcal{U}_\alpha$ *(index > sure payoff), then select decision* ε *("experiment");*
- *if* $\mathcal{I}(\widehat{\pi}_t) = \mathcal{U}_\alpha$ *(index = sure payoff), then select indifferently decision* α *or decision* ε*.*

To avoid the indifference case $\mathcal{I}(\widehat{\pi}_t) = \mathcal{U}_\alpha$, we single out the following prudent Gittins index optimal strategy. It is prudent because it favors avoidance over risk taking in case of indifference.

Definition 3. *We call* Gittins index strategy (GI-strategy) *the strategy*

- *if* $\mathcal{I}(\widehat{\pi}_t) \leq \mathcal{U}_\alpha$*, then select decision* α *("avoid"),*
- *if* $\mathcal{I}(\widehat{\pi}_t) > \mathcal{U}_\alpha$*, then select decision* ε *("experiment"),*

and GI-DM *a decision-maker who adopts the Gittins index strategy.*

Behavior of the GI-DM

To describe the behavior of a decision-maker who adopts the Gittins index strategy of Definition 3, we introduce the following stopping time, which plays a pivotal role.

We define the *learning time* τ as the first time t, if it exists, where the avoidance payoff $\mathcal{U}_\alpha$ exceeds the Gittins index:

$$\tau = \inf\{t = 0, 1, 2, 3 \ldots \mid \mathcal{I}(\widehat{\pi}_t) \leq \mathcal{U}_\alpha\} \ . \tag{17}$$

In case $\mathcal{I}(\widehat{\pi}_t) > \mathcal{U}_\alpha$ for all times $t = 0, 1, 2, 3 \ldots$, the convention is $\tau = \inf \emptyset = +\infty$.

We now outline the behavior of a DM that follows the GI-strategy.

Proposition 4. *The DM that follows the GI-strategy of Definition 3 switches at most one time from experimenting to avoiding. More precisely, his behavior displays one of the three following patterns, depending on the learning time* τ *in* (17)*.*

a) Infinite learning:
if $\tau = +\infty$, that is, if $\mathcal{I}(\widehat{\pi}_t) > \mathcal{U}_\alpha$ for all times $t = 0, 1, 2, 3 \ldots$, the GI-DM never avoids.

b) No learning:
if $\tau = 0$, that is, if $\mathcal{I}(\pi_0) \leq \mathcal{U}_\alpha$, the GI-DM avoids from the start and, from then on, he keeps avoiding for all times.

c) Finite learning:
if $1 \leq \tau < +\infty$, the GI-DM

- *experiments from $t = 0$ to $\tau - 1$, that is, as long as $\mathcal{I}(\widehat{\pi}_t) > \mathcal{U}_\alpha$,*
- *switches to avoiding at time $t = \tau$, that is, as soon as $\mathcal{I}(\widehat{\pi}_t) \leq \mathcal{U}_\alpha$,*
- *from then on, keeps avoiding for all times.*

Proof.

a) By Proposition 2 and Definition 3, when $\tau = +\infty$ — that is, when the Gittins index $\mathcal{I}(\widehat{\pi}_t) > \mathcal{U}_\alpha$ for all times t — it is optimal to select decision ε and experiment forever.

b) By Proposition 2 and Definition 3, when $\tau = 0$ — that is, when $\mathcal{I}(\widehat{\pi}_0) \leq \mathcal{U}_\alpha$ — it is optimal to select decision α and avoid for all times. Indeed, once the GI-DM avoids, he does not observe the random outcomes, hence he no longer updates the posterior $\widehat{\pi}_t$ in (15), so that he keeps avoiding.

c) When $1 \leq \tau < +\infty$, we have

- $\mathcal{I}(\widehat{\pi}_t) > \mathcal{U}_\alpha$ for times $t = 0$ up to $\tau - 1$; hence, by Proposition 2 and Definition 3, it is optimal to select decision ε and experiment from times $t = 0$ up to $\tau - 1$;
- $\mathcal{I}(\widehat{\pi}_t) \leq \mathcal{U}_\alpha$ for times $t = \tau$ up to $+\infty$; hence, by Proposition 2 and Definition 3, it is optimal to select decision α and avoid for times $t = \tau$ up to $+\infty$. Indeed, once the GI-DM avoids, he does not observe the random outcomes, hence he no longer updates the posterior $\widehat{\pi}_t$ in (15), so that he keeps avoiding.

This ends the proof. □

4 Killing three biases with one stone

Here, we show three features of the GI-strategy of Definition 3 that possess interesting psychological interpretations in terms of human biases: status quo bias in §4.1, salience bias in §4.2, overestimation of small probabilities bias in §4.3. Such features provide insights into which attitudes natural selection may have favored in humans having to decide under uncertainty and learning. These insights send us back to the introductory Section 1 on the debate about benchmarks for qualifying biases.

4.1 Status quo bias

Our analysis provides theoretical support to the so-called *status quo bias*, a preference for the current state of affairs, documented in Samuelson and Zeckhauser (1988).

Proposition 5. *The behavior of the GI-DM displays at most two consecutive phases of "status quo" — one (possibly empty) of experimenting, that is, making a "risky decision" , one (possibly empty) of prudence — with at most one switch; in particular, once prudent, this is forever.*

This follows from Proposition 4. In particular, once the GI-DM selects the "avoid" option, he will never more experiment. Indeed, the optimal rule of Proposition 2 states that, once the GI-DM selects the "avoid" option, he does not observe the random outcomes, hence he no longer updates the posterior $\widehat{\pi}_t$ because of the dynamics (15) so that he keeps avoiding. Thus, once stuck in a risk avoidance attitude, the status quo holds sway.

4.2 Salience bias

Our analysis provides theoretical support to the so-called *salience bias*, or *availability heuristic* Tversky and Kahneman (1982), that makes humans sensitive to vivid, salient events.

Proposition 6. *If it exists, the experimenting phase can only stop when a bad outcome materializes; in other words, the switch from riskiness to prudence occurs when a salient and vivid outcome occurs.*

We will see that a "stay-with-a-winner" characteristics of the GI-strategy makes that a (prudent) change in behavior occurs only when a bad outcome materializes.

Proof. Suppose that, at time t the GI-DM is experimenting. We will show that, if a "good" outcome $\mathtt{G}$ materializes at the end of the interval $[t, t+1[$ ($Y_{t+1} = \mathtt{G}$), then necessarily the DM goes on experimenting at time $t+1$.

By Definition 3, we have that $\mathcal{I}(\widehat{\pi}_t) > \mathcal{U}_\alpha$ since the GI-DM is experimenting at time t. Now, the Gittins index function $\mathcal{I} : \Delta(S^1) \to \mathbb{R}$ of Proposition 2 has the property that $\mathcal{I} \circ \theta^{\mathtt{G}} \geq \mathcal{I}$, that is, the index increases when the posterior changes following a "good" outcome (Berry and Fristedt, 1985, Theorem 5.3.5, page 103). As we supposed that $Y_{t+1} = \mathtt{G}$, we have that $\widehat{\pi}_{t+1} = \theta^{\mathtt{G}}\widehat{\pi}_t$ by the dynamics (15). As a consequence, we have that $\mathcal{I}(\widehat{\pi}_{t+1}) = \mathcal{I}(\theta^{\mathtt{G}}\widehat{\pi}_t) \geq \mathcal{I}(\widehat{\pi}_t) \geq \mathcal{U}_\alpha$. As $\mathcal{I}(\widehat{\pi}_{t+1}) \geq \mathcal{U}_\alpha$, the GI-DM goes on experimenting at time $t+1$ by Definition 3.

Therefore, the switch from experimenting to avoiding can only occur when a "bad" outcome materializes. □

4.3 Overestimation of small probabilities bias

Under expected utility theory, a lottery is assessed by a non-linear transformation of the outcomes into utility, followed by a sum weighted by the probabilities. However, other theories of decision-making under risk propose to perform a non-linear transformation of the probabilities attached to a lottery when weighing outcomes Yaari (1987); Quiggin (1982). Based upon experimental observations, Kahneman and Tversky's *prospect theory* Kahneman and Tversky (1979); Tversky and Kahneman (1992) and Lopes'*security/potential and aspiration* theory Lopes (1996); Lopes and Oden (1999) produce curves of S-shaped probability deformations, exhibiting overweighting of low probabilities.

Somewhat related is the observation that low probability-vivid consequences outcomes, like plane crashes, receive disproportionate coverage and attention with respect to their statistical occurrence.

Why do we overweigh small probabilities? Why do we display such a bias? We will show that the overweighing of small probabilities is an output from the model we developed in Sect. 3.

Probability estimator and critical probability

We introduce the numbers $N_t^{\mathtt{B}}$ and $N_t^{\mathtt{G}}$ of "bad" and "good" encounters up to time t

$$N_0^{\mathtt{B}} = N_0^{\mathtt{G}} = 0 \ , \ \ N_t^{\mathtt{B}} = \sum_{s=1}^{t} \mathbf{1}_{\{Y_s=\mathtt{B}\}} \ , \ \ N_t^{\mathtt{G}} = \sum_{s=1}^{t} \mathbf{1}_{\{Y_s=\mathtt{G}\}} \ , \ \ t = 1, 2, 3, \ldots \quad (18)$$

where the observations Y_s are given by (7).

Let $n_0^{\mathtt{B}} > 0$ and $n_0^{\mathtt{G}} > 0$ be two positive scalars. We suppose that the distribution π_0 is a beta distribution $\beta(n_0^{\mathtt{B}}, n_0^{\mathtt{G}})$ on the simplex S^1 in (9), that is, for any continuous function $\varphi : S^1 \to \mathbb{R}$,

$$\int_{S^1} \varphi(p^{\mathtt{B}}, p^{\mathtt{G}}) \pi_0(dp^{\mathtt{B}} dp^{\mathtt{G}}) = \frac{\int_0^1 \varphi(p, 1-p) p^{n_0^{\mathtt{B}}-1} (1-p)^{n_0^{\mathtt{G}}-1} dp}{\int_0^1 p^{n_0^{\mathtt{B}}-1} (1-p)^{n_0^{\mathtt{G}}-1} dp} \ . \quad (19)$$

By the dynamics (15), we easily establish that the posterior $\widehat{\pi}_t$ in §3.3 is the beta distribution

$$\widehat{\pi}_t = \beta(n_0^{\mathtt{B}} + N_t^{\mathtt{B}}, n_0^{\mathtt{G}} + N_t^{\mathtt{G}}) \ . \quad (20)$$

As we are interested in what estimates of the (true) probability value $\overline{p}^{\mathtt{B}}$ in (8) the GI-DM forms, we introduce the following statistics.

Definition 7. *For $t = 0, 1, 2 \ldots$, we define the* statistics $\widehat{p}_t^{\mathtt{B}}$ *by*

$$\widehat{p}_0^{\mathtt{B}} = \frac{n_0^{\mathtt{B}}}{n_0^{\mathtt{B}} + n_0^{\mathtt{G}}} \ \ \textit{and} \ \ \widehat{p}_t^{\mathtt{B}} = \frac{n_0^{\mathtt{B}} + N_t^{\mathtt{B}}}{n_0^{\mathtt{B}} + N_t^{\mathtt{B}} + n_0^{\mathtt{G}} + N_t^{\mathtt{G}}} \ , \ \ t = 1, 2 \ldots \ . \quad (21a)$$

where the numbers of "bad" and "good" encounters up to time t are given in (18)*. We also define*

$$\widehat{p}_t^{\mathtt{G}} = 1 - \widehat{p}_t^{\mathtt{B}} \ . \quad (21b)$$

The statistics $\widehat{p}_t^{\mathtt{B}}$ is built from the parameters $n_0^{\mathtt{B}}$ and $n_0^{\mathtt{G}}$ of the prior beta distribution $\pi_0 = \beta(n_0^{\mathtt{B}}, n_0^{\mathtt{G}})$ in (19) and from the observations up to time t, summarized in the number $N_t^{\mathtt{B}}$ of "bad" encounters up to time t, and the same for $N_t^{\mathtt{G}}$ with "good", as defined in (18).

As in Sect. 2, we introduce a critical probability p_c as follows.

Definition 8. *We define the* critical probability p_c *— or* avoidability index p_c *— by the following ratio:*

$$p_c = \frac{\mathcal{U}^{\mathtt{G}} - \mathcal{U}_\alpha}{\mathcal{U}^{\mathtt{G}} - \mathcal{U}^{\mathtt{B}}} = \frac{\textit{relative costs of avoidance}}{\textit{relative costs of encounter}} \in]0,1[\ . \tag{22}$$

All things being equal, the "worse" a bad outcome (that is, high bad costs), the lower the critical probability.

How does the DM obtain the basic data to set up the optimization problem?

So, how does the DM set the two positive scalar parameters $n_0^{\mathtt{B}}$ and $n_0^{\mathtt{G}}$ of the beta distribution π_0 in (19) to make up the probability $\mathbb{P}_{\pi_0}$? How does he set the instant payoffs — avoidance payoff $\mathcal{U}_\alpha$, encounter payoff $\mathcal{U}^{\mathtt{B}}$ and base payoff $\mathcal{U}^{\mathtt{G}}$ — in Table 2?

We suppose that the DM knows the avoidance payoff $\mathcal{U}_\alpha$. He starts experimenting and

- either he first enjoys n good outcomes $\mathtt{G}$ — hence discovering the base payoff $\mathcal{U}^{\mathtt{G}}$ — before suffering a bad outcome $\mathtt{B}$ — hence discovering the encounter payoff $\mathcal{U}^{\mathtt{B}}$; in that case, he sets $n_0^{\mathtt{G}} = n$ and $n_0^{\mathtt{B}} = 1$;
- or he first suffers n bad outcomes $\mathtt{B}$ — hence discovering the encounter payoff $\mathcal{U}^{\mathtt{B}}$ — before enjoying a good outcome $\mathtt{G}$ — hence discovering the base payoff $\mathcal{U}^{\mathtt{G}}$; in that case, he sets $n_0^{\mathtt{G}} = 1$ and $n_0^{\mathtt{B}} = n$.

So, at the end of those $n_0^{\mathtt{G}} + n_0^{\mathtt{B}}$ trials, the DM disposes of the two payoffs $\mathcal{U}^{\mathtt{B}}$ and $\mathcal{U}^{\mathtt{G}}$, as well as the two parameters $n_0^{\mathtt{B}} > 0$ and $n_0^{\mathtt{G}} > 0$.

Probability estimates made by the GI-DM

The following Proposition 9 details what estimates $\widehat{p}_t^{\mathtt{B}}$ of the (true) probability value $\overline{p}^{\mathtt{B}}$ the DM forms by (21a), when he follows the GI-strategy.

Proposition 9. *Consider a GI-DM, a decision-maker who follows the Gittins index strategy of Definition 3. Here are the estimates of the (true) probability value* $\overline{p}^{\mathtt{B}}$*.*

a) Infinite learning: *when $\tau = +\infty$, the GI-DM experiments forever and, asymptotically, the statistics $\widehat{p}_t^{\mathtt{B}}$ in* (21a) *reaches the (true) probability value $\overline{p}^{\mathtt{B}}$, almost surely under the (true) probability distribution $\overline{\mathbb{P}}$ in* (8)*:*

$$\lim_{t \to +\infty} \widehat{p}_t^{\mathtt{B}} = \overline{p}^{\mathtt{B}} \ , \ \ \overline{\mathbb{P}} - p.s. \tag{23}$$

b) No learning: *when $\tau = 0$, the GI-DM never experiments and his initial estimate $\widehat{p}_0^{\mathtt{B}}$ of the (true) probability value $\overline{p}^{\mathtt{B}}$ satisfies*

$$p_c \leq \widehat{p}_0^{\mathtt{B}} \ , \tag{24}$$

where the critical probability p_c is defined in (22)*.*

c) Finite learning: *when $1 \leq \tau < +\infty$, the GI-DM experiments till time τ and his estimate $\widehat{p}_\tau^{\mathtt{B}}$ of the (true) probability value $\overline{p}^{\mathtt{B}}$ satisfies*

$$p_c \leq \widehat{p}_\tau^{\mathtt{B}} \ . \tag{25}$$

Proof.

a) By Proposition 2 and Definition 3, when $\tau = +\infty$ it is optimal to select decision ε and experiment forever. Then, asymptotically, the statistics $\widehat{p}_t^{\mathtt{B}}$ reaches the (true) probability value $\overline{p}^{\mathtt{B}}$ by the Law of large numbers, almost surely under the probability $\overline{\mathbb{P}}$ in (8). Indeed, the random variables $(X_1, X_2, \ldots)$ in (6) are i.i.d. under $\overline{\mathbb{P}}$.

b) See the proof below.

c) By definition of τ in (17), we have that

$$\tau < +\infty \Rightarrow \mathcal{U}_\alpha \geq \mathcal{I}(\widehat{\pi}_\tau) \ . \tag{26}$$

Now, it is well known that (Berry and Fristedt, 1985, Corollary 5.3.2, page 101)

$$\mathcal{I}\big(\beta(N^{\mathtt{B}}, N^{\mathtt{G}})\big) \geq \frac{N^{\mathtt{B}}}{N^{\mathtt{B}} + N^{\mathtt{G}}} \mathcal{U}^{\mathtt{B}} + \frac{N^{\mathtt{G}}}{N^{\mathtt{B}} + N^{\mathtt{G}}} \mathcal{U}^{\mathtt{G}} \ . \tag{27}$$

From the two above equations and from (20), we deduce that

$$\tau < +\infty \Rightarrow \mathcal{U}_\alpha \geq \frac{n_0^{\mathtt{B}} + N_\tau^{\mathtt{B}}}{n_0^{\mathtt{B}} + N_\tau^{\mathtt{B}} + n_0^{\mathtt{G}} + N_\tau^{\mathtt{G}}} \mathcal{U}^{\mathtt{B}} + \frac{n_0^{\mathtt{G}} + N_\tau^{\mathtt{G}}}{n_0^{\mathtt{B}} + N_\tau^{\mathtt{B}} + n_0^{\mathtt{G}} + N_\tau^{\mathtt{G}}} \mathcal{U}^{\mathtt{G}} \ . \tag{28}$$

Rearranging the terms, and using (21a) and (22), we obtain:

$$\tau < +\infty \Rightarrow \widehat{p}_\tau^{\mathtt{B}} \geq p_c \ . \tag{29}$$

This ends the proof. □

The Biased Learning Theorem

An easy consequence of Proposition 9 is the following theorem.

Theorem 10 (Biased Learning Theorem)**.** *Suppose that the* "bad" *outcome* B *is* unlikely*, in the sense that*

$$\overline{p}^{\mathtt{B}} \leq p_c \,. \tag{30}$$

A GI-DM — a decision-maker who follows the Gittins index strategy of Definition 3 — will

- *either experiment forever, and he will* accurately *estimate asymptotically the (true) probability of the unlikely bad outcome* B*;*
- *or experiment during a finite number of periods (possibly zero) and, when the experiment phase ends at $\tau < +\infty$, he will* overerestimate *the (true) probability of the unlikely bad outcome* B*:*

$$\overline{p}^{\mathtt{B}} \leq \widehat{p}^{\mathtt{B}}_{\tau} \,. \tag{31}$$

The proof combines (25) and (30) when $\tau < +\infty$. Table 3 sums up the results of Proposition 9 when the inequality (30) holds true.

Case $\overline{p}^{\mathtt{B}} \leq p_c$	estimate of $\overline{p}^{\mathtt{B}}$	comment
$\tau = +\infty$	$\lim_{t\to+\infty} \widehat{p}^{\mathtt{B}}_t = \overline{p}^{\mathtt{B}}$	exact estimation of $\overline{p}^{\mathtt{B}}$
$0 \leq \tau < +\infty$	$\widehat{p}^{\mathtt{B}}_{\tau} \geq p_c \geq \overline{p}^{\mathtt{B}}$	overestimation of $\overline{p}^{\mathtt{B}}$

Table 3: Estimate $\widehat{p}^{\mathtt{B}}_t$ of the "bad" outcome in the case $\overline{p}^{\mathtt{B}} \leq p_c$

In the case (30), the (unknown) probability $\overline{p}^{\mathtt{B}}$ of the bad outcome is low enough so that the expected utility outweighs the sure utility of avoiding, as in (2). Therefore, with foresight, an optimal DM would take the risk forever. However, the Biased Learning Theorem 10 reveals that, without foresight, an optimal DM would be more prudent, and not always take the risk forever.

We conclude that — in a situation where, with foresight, an optimal DM would take the risk forever — the GI-DM will

- either *accurately estimate (asymptotically) the probability of the bad and unlikely outcome* B *when learning never stops*,

- or *overestimate the probability of the bad and unlikely outcome* `B` *when learning stops.*

The intuition for the overestimation is that the probability of a bad and unlikely outcome is approached from below, as follows from item c) of Proposition 9 and from (28).

Relations with economics and biology literature

Economists have made the point, coined the *Incomplete Learning Theorem*, that the optimal strategy (to maximize discounted expected utility) does not necessarily lead to exactly evaluate the unknown probability Rothschild (1974); Easley and Kiefer (1988); Brezzi and Lai (2000). Thus, optimality does not necessarily lead to perfect accuracy. Our results point to a *Biased Learning Theorem*, as we prove that the departure from accuracy displays a bias towards overestimation of bad and unlikely outcomes.

Our contribution resorts to economics, focusing on optimality benchmarks and resulting optimal strategies, whereas, in the evolutionary literature, the discussion bears on the precision of Bayesian estimates of the unknown probability Trimmer et al. (2011).

5 Conclusion

Our model and analysis show that biases can be the product of rational behavior, in the sense of maximizing expected discounted utility with learning.

We have provided theoretical support to the so-called *status quo bias*.

Our formal analysis also provides theoretical support to the *salience bias*, related to the *availability bias* for vivid outcomes.

Finally, we have shown a Biased Learning Theorem that provides rational ground for the human bias that consists in attributing to bad and unlikely outcomes an importance larger than their statistical occurrence. Let us dwell on this point.

In many situations, probabilities are not known but learnt. The 2011 nuclear accident in Japan has led many countries to stop nuclear energy. This sharp switch may be interpreted as the stopping of an experiment phase where the probability of nuclear accidents has been progressively learnt. In financial economics, the equity premium puzzle comes from the observation that bonds are underrepresented in portfolios, despite the empirical fact that

stocks have outperformed bonds over the last century in the USA by a large margin Mehra and Prescott (1985). However, this analysis is done *ex post* under risk, while decision-makers make their decisions day by day under uncertainty, and sequentially learn about the probability of stocks losses. *Ex ante*, the underrepresentation of bonds can be enlightened by the Biased Learning Theorem: the (small) probability of (large) bonds losses is overestimated with respect to their statistical occurrence. To end up, our results point to the fact that overestimation depends upon the relative payoffs (costs) by the formula (22). This property could possibly be tested in experiments.

Acknowledgments. The author thanks the following colleagues for their comments: Daniel Nettle (Newcastle University), Nicolas Treich (Toulouse School of Economics) and Christopher Costello[3] (University of California Santa Barbara); Jean-Marc Tallon, Alain Chateauneuf, Michelle Cohen, Jean-Marc Bonnisseau and the organizers of the Economic Theory Workshop of Paris School of Economics (Friday 4 November 2011); John Tooby, Andrew W. Delton, Max Krasnow and the organizers of the seminar of the Center for Evolutionary Psychology, University of California Santa Barbara (Friday 18 November 2011); Arthur J. Robson and the organizers of the Economics seminar at Simon Fraser University (Tuesday 21 October 2014); Pierre Courtois, Nicolas Querou, Raphaël Soubeyran[4] and the organizers of the seminar of Lameta, Montpellier (Monday 3 October 2016); Khalil Helioui, Geoffrey Barrows[5], Jean-Pierre Ponssard, Guillaume Hollard, Guy Meunier and the organizers of the Sustainable Economic and Financial Development Seminar at École Polytechnique (Tuesday 17 January 2017); Jeanne Bovet and Luke Glowacki, organizers of the Tuesday Lunch at Institute for Advanced Study in Toulouse (Tuesday 4 July 2017).

References

Jerome H. Barkow, Leda Cosmides, and John Tooby, editors. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* Oxford University Press, 1992.

[3]I thank Christopher Costello for suggesting the title of the paper.

[4]I thank Raphaël Soubeyran for the observation that the probability of a bad and unlikely outcome is approached from below.

[5]I thank Geoffrey Barrows for pointing me the salience effect.

D. A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, 1985.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, Massachusets, second edition, 2000. Volumes 1 and 2.

J. Boutang and M. De Lara. *The Biased Mind. How Evolution Shaped our Psychology, Including Anecdotes and Tips for Making Sound Decisions*. Springer-Verlag, Berlin, 2015.

Monica Brezzi and Tze Leung Lai. Incomplete learning from endogenous data in dynamic allocation. *Econometrica*, 68(6):1511–1516, 2000.

R. Dawkins and J. R. Krebs. Arms races between and within species. *Proceedings of the Royal Society of London, Series* B, 205:489–511, 1979.

David Easley and Nicholas M Kiefer. Controlling a stochastic process with unknown parameters. *Econometrica*, 56(5):1045–64, September 1988.

Gerd Gigerenzer. Fast and frugal heuristics: The tools of bounded rationality. In Derek J. Koehler and Nigel Harvey, editors, *Blackwell handbook of judgement and decision making*, pages 62–88. Blackwell Publishing, Oxford, 2004.

Gerd Gigerenzer. Why heuristics work. *Perspectives on psychological science*, 3(1):20–29, 2008.

Thomas Gilovich, Dale W. Griffin, and Daniel Kahneman, editors. *Heuristics and Biases. The Psychology of Intuitive Judgement*. Cambridge University Press, 2002.

J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B*, 41(2):148–177, 1979.

J. C. Gittins. *Multi-armed Bandit Allocation Indices*. Wiley, New York, 1989.

M. G. Haselton and D. Nettle. The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, 10(1):47–66, 2006.

John M. C. Hutchinson and Gerd Gigerenzer. Simple heuristics and rules of thumb: Where psychologists and behavioural biologists might meet. *Behavioural Processes*, 69(2):97–124, 2005.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

Daniel Kahneman, Paul Slovic, and Amos Tversky, editors. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, April 1982.

Joseph E. LeDoux. *The Emotional Brain*. Simon & Schuster, 1996.

Lola L. Lopes. When time is of the essence: Averaging, aspiration, and the short run. *Organizational Behavior and Human Decision Processes*, 65(3): 179–189, March 1996.

Lola L. Lopes and Gregg C. Oden. The role of aspiration level in risky choice: a comparison of cumulative prospect theory and SP/A theory. *J. Math. Psychol.*, 43(2):286–313, 1999.

Rajnish Mehra and Edward C. Prescott. The equity premium: A puzzle. *Journal of Monetary Economics*, 15(2):145 – 161, 1985.

J. G. Neuhoff. A perceptual bias for rising tones. *Nature*, 395(6698):123–124, 1998.

John Quiggin. A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343, December 1982.

Arthur John Robson. Why would nature give individuals utility functions? *Journal of Political Economy*, 109(4):900–929, 2001.

Michael Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185–202, October 1974.

William Samuelson and Richard Zeckhauser. Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1):7–59, March 1988.

Pete Trimmer, Alasdair Houston, James Marshall, Mike Mendl, Elizabeth Paul, and John McNamara. Decision-making under uncertainty: biases and Bayesians. *Animal Cognition*, 14:465–476, 2011.

Amos Tversky and Daniel Kahneman. Availability: a heuristic for judging frequency and probability. In Kahneman et al. (1982), pages 163–178.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5 (4):297–323, October 1992.

Menahem E. Yaari. The dual theory of choice under risk. *Econometrica*, 55 (1):95–115, 1987.