Rationally Biased Learning

Michel de Lara

Rationally Biased Learning

Michel De Lara
CERMICS, Ecole des Ponts, Marne-la-Vallée, France
E-mail: michel.delara@enpc.fr

March 22, 2022

Abstract

Humans display a tendency to pay more attention to bad outcomes, often in a disproportionate way relative to their statistical occurrence. They also display euphorism, as well as a preference for the current state of affairs (status quo bias). Based on the analysis of optimal solutions of infinite horizon stationary optimization problems under imperfect state observation, we show that such human perception and decision biases can be grounded in a form of rationality (optimality). We also provide conditions (boundaries) for their possible occurence and an analysis of their robustness. Thus, biases can be the product of rational behavior.

Keywords: pessimism bias, status quo bias, euphorism bias, probability overestimation, optimal behavior, imperfect state information.

1 Introduction

When we perceive sounds, we overestimate the change in level of rising level tones relative to equivalent falling level tones [18]. When we assess pros and cons in decision making, we weigh losses more than gains [15]. We are more frightened by a snake or a spider than by a passing car or an electrical shuffle. Such human assessments are qualified of biases, because they depart from physical measurements or objective statistical estimates. Thus, there is "bias" when a behavior is not aligned with a given "rationality benchmark"

(like expected utility theory), as documented in the "heuristics and biases" literature [14, 10].

However, if such biases are found consistently in human behavior, they must certainly have a reason. Some scholars (see [8, 9, 13]) 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 [1, 3, 4]). In this conception, the benchmark should be a measure of fitness reflecting survival and reproduction abilities, and the "bias" can be explained in two ways.

- *Bias by mismatch.* The bias can result from a timelag: because the modern environment has departed so much from the environment in which natural selection had time to shape our minds, human behavior displays a mismatch (cars are objectively more dangerous than spiders in our modern environment).

- *Bias by design.* But the bias can be the feature of an optimal strategy where optimality is measured in fitness (the genes of those who accurately estimated the change in level of rising level tones have, more often than the "overestimaters", finished in the stomach of a predator).

This last conception of "bias by design" is reflected, for example, in [12]. In their attempt to understand "how natural selection engineers psychological adaptations for judgment under uncertainty", Haselton and Nettle consider an individual who has to decide between a safe option (one known payoff) and a risky option (known bad and good payoffs). They define a critical probability and observe that, if the bad outcome has higher probability, the (optimal) individual should avoid taking risks and select the safe option. The interesting point is that the critical probability is the ratio of the difference between good and safe payoffs over the difference between good and bad payoffs. As a consequence, the higher the latter difference, the more the individual should avoid taking risks. 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" and that "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".[1] However, the above

[1]Such asymmetry in costs is manifest in the so-called *life-dinner* principle of Richard

analysis is performed under the so-called "error management theory", that is, supposing known the probability of the bad outcome. What happens when the individual does not know *a priori* the objective probability driving the occurence of a bad outcome?

In this paper, we consider the classical problem of a decision maker (DM) faced with a repeated choice between a certain option (one known safe payoff) and a risky option (two known risky payoff values, but unknown probability of each). Regarding uncertainty, we are thus in the so-called ambiguity setting (and not in the risk setting). Regarding payoffs, we suppose that the three payoffs are ranked in such a way that the safe one stands between the two risky ones; thus, the lowest (risky) payoff reflects a bad outcome. We will show that a rational decision maker — in the sense of maximizing expected discounted utility (where the mathematical expectation involves a prior on the unknown probabilities) — can exhibit a behavior displaying "euphorism" and status quo biases, as well as, under suitable conditions, the pessimistic erroneous assessment of the best objective option and an overestimation bias for the probability of the bad outcome. Thus, in some settings (detailed in the paper), it is quite rational to pay more attention to bad outcomes than to good ones, and to exaggerate their importance, even if one aggregates uncertainties by means of a (balanced, risk-neutral) mathematical expectation, and aggregates payoffs by summation.

It is well-known that the problem we address can be framed as a two-armed bandit problem, as there are only two decisions, as information is triggered by decision, as the criterion is intertemporal under unknown probabilities. In the mathematics and the psychology literature, there is a huge body of work on armed bandit problems, and on its celebrated solution (when suitable hypothesis are met, see [11]) by means of a dynamic allocation index (Gittins Index Theorem). However, this not the route we follow. Indeed, we revisit this type of problem as an instance of optimization problem under imperfect state observation, and we devote a whole part to discuss which of our results are robust w.r.t. (with respect to) to assumptions like stationarity, discounting, finite or infinite horizon. By doing so, we want to reveal features of optimal strategies that are more general than those obtained by

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 [6]. Neuroscientist Joseph LeDoux has a nice way to express "bias by design" 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" ([16, p.166]).

means of the Gittins index strategy. Of course, some features — for instance that the information needed for optimal decisions can be summarized in a posterior that is updated following Bayes rule — are shared with this latter, but they do not depend on the mathematical expression of the index.

The paper is organized as follows. In Sect. 2, we consider the problem of a decision maker faced with a repeated choice between a certain option (one known safe payoff) and a risky option (yielding either a bad or a good outcome, but with unknown probabilities). We set up a formal mathematical model of stochastic sequential decision-making — under (Bayesian) ambiguity regarding random sequences of bad and good outcomes (Bernoulli trials) — and we describe an optimal strategy and the behavior of the optimal DM. This section contains known results, but with new proofs that make it possible to assess the robustness of the findings. In Sect. 3, we prove and display features of the optimal strategy — optimally designed for a Bayesian criterion, that is, for a certain (subjective) probability distribution on sequences of bad and good outcomes — when it is implemented with a Bernoulli process under objective probabilities (objective environment). We distinguish two outputs of the optimal strategy — estimation of the unknown objective probability of the bad outcome, assessement of whether the objective environment is prone to risk-taking or not. When one output is not what it would be were the objective probability distribution known, we deal with a biais. We summarize in Tables 2 and 3 our findings regarding consistency or discrepancy w.r.t. what would be optimal in the objective environment: "euphorism" and status quo biases, as well as boundaries and amplitudes of two effects, the pessimistic erroneous assessment of the best objective option and the overestimation of the probability of the bad outcome. In Sect. 4, we discuss the cognitive burden of implementing the optimal strategy (hence the possibility to be an outcome of natural selection), the robustness of our findings and possible psychological interpretations, and we conclude. Appendix A gathers technical results and proofs.

2 A mathematical model of repeated decision-making under ambiguity

In §2.1, we lay out mathematical ingredients to set up a model of sequential decision-making under unknown probability, and formulate an expected dis-

counted payoff maximization problem. In §2.2, we analyze the structure of an optimal strategy, and then we describe the behavior of a decision-maker who adopts such optimal strategy.

This Sect. 2 fixes vocabulary, notation and provides the main properties that will be used to show our main results in Sect. 3. The results exposed in this section are not new: the structure of an optimal strategy and the induced behavior are well-known, although they are generally presented as a consequence of the Gittins Index Theorem, which is not the way we prove them in Appendix A. By taking another route for the proofs, we are able to obtain (what we think are new) results on i) how the probability of different regimes in the optimal behavior depends monotonically upon some of the data (ii) which of our results in Sect. 3 are robust w.r.t. to assumptions like stationarity, discounting, finite or infinite horizon (§4.2).

2.1 An expected discounted payoff maximization problem

In [12], the following situation is examined. To reach her/his destination, an individual 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. Two decisions are possible, with different costs. If one avoids the grass, one makes a detour that is costly in time, but one suffers no pain from the (possible) snake. If one passes through the grass ("trying", "learning", "experimenting"), the time spent is shorter but one can suffer pain (though not lethal) if the snake is present. We will illustrate our mathematical setting with this story.

Sequential decision-making

We 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. We suppose that, at discrete stages $t \in \mathbb{N}$, the DM makes a decision — either "avoid" and be prudent (α) or "experiment" and take risks (ε) — without knowing in advance the state of Nature occurring at that time — either bad ($\mathtt{B}$) or good ($\mathtt{G}$). We denote by $t = 0, 1, 2 \ldots$ the stage corresponding to the beginning of the time interval $[t, t+1[$. We denote 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 time interval $[t, t+1[$. We define the *sample space*

$$\mathbb{H}_\infty = \{\mathtt{B}, \mathtt{G}\}^{\mathbb{N}^*} = \{\mathtt{B}, \mathtt{G}\} \times \{\mathtt{B}, \mathtt{G}\} \times \cdots \ , \tag{1}$$

with generic element an infinite sequence $(w_1, w_2, \dots)$ of elements in $\{\mathtt{B}, \mathtt{G}\}$. For $t = 1, 2 \dots$, we denote by[2]

$$\mathbf{W}_t : \mathbb{H}_\infty \to \{\mathtt{B}, \mathtt{G}\} \ , \ \ \mathbf{W}_t(w_1, w_2, \dots) = w_t \ , \tag{2}$$

the *state of Nature* realized at the beginning of the time interval $[t, t+1[$, but that cannot be revealed before the end of $[t, t+1[$.

Strategies

At the beginning of each time interval $[t, t+1[$, 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 $\mathbf{W}_{t+1}$ ($\mathtt{B}$ or $\mathtt{G}$) is revealed and experimented, at the end of the time interval $[t, t+1[$.

We assume that the DM is not visionary and learns only from the past: she/he cannot know the future in advance, neither can the DM know the state of Nature ($\mathtt{B}$ or $\mathtt{G}$) if the DM decides to avoid. We define the *observation sets* at stage $t = 0, 1, 2, 3 \dots$ by $\mathbb{Y}_0 = \{\partial\}$, where ∂ corresponds to no information (no observation at initial stage $t = 0$), and $\mathbb{Y}_t = \{\mathtt{B}, \mathtt{G}, \partial\}^t$ for $t = 1, 2, 3 \dots$ We 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 stage $t = 0, 1, 2 \dots$ if the DM makes decision $v_t \in \{\alpha, \varepsilon\}$ is $\mathbf{Y}_{t+1} = \mathcal{O}(v_t, \mathbf{W}_{t+1})$. This case is also known as the *partial feedback* case, where foregone payoffs are not revealed.

We allow the DM to accumulate past observations; therefore the decision v_t at stage t can only be a function of $\mathbf{Y}_1, \dots, \mathbf{Y}_t$ (the initial decision v_0 is made without information). A *policy at stage t* is a mapping $\mathcal{S}_t : \mathbb{Y}_t \to \{\alpha, \varepsilon\}$ that tells the DM what will be the next action in view of past observations. A *strategy* $\mathcal{S}$ is a sequence $\mathcal{S} = (\mathcal{S}_0, \mathcal{S}_1, \dots)$ of policies. Given a strategy $\mathcal{S}$, decisions and observations are inductively given by

$$\mathbf{V}_0 = \mathcal{S}_0 \in \{\alpha, \varepsilon\} \ , \tag{3a}$$

$$\mathbf{Y}_{t+1} = \mathcal{O}(\mathbf{V}_t, \mathbf{W}_{t+1}) \in \{\mathtt{B}, \mathtt{G}, \partial\} \ , \ \ \forall t = 0, 1, 2 \dots \ , \tag{3b}$$

$$\mathbf{V}_t = \mathcal{S}_t(\mathbf{Y}_1, \dots, \mathbf{Y}_t) \in \{\alpha, \varepsilon\} \ , \ \ \forall t = 0, 1, 2 \dots \ . \tag{3c}$$

[2]We denote random variables by uppercase bold letters.

In the full feedback case, where foregone payoffs are revealed no matter what the decision made, we have $\mathbf{Y}_t = \mathbf{W}_t$, for $t = 0, 1, 2 \ldots$.

Hypothesized probability

We introduce the one-dimensional simplex

$$\Sigma^1 = \left\{ (p^{\mathtt{B}}, p^{\mathtt{G}}) \in \mathbb{R}^2 \,\middle|\, p^{\mathtt{B}} \geq 0 \,,\ \ p^{\mathtt{G}} \geq 0 \,,\ \ p^{\mathtt{B}} + p^{\mathtt{G}} = 1 \right\} . \tag{4}$$

The simplex Σ^1 is identified with the unit segment $[0, 1]$ by the mapping (measurable bijection with measurable inverse) $\Sigma^1 \ni (p^{\mathtt{B}}, p^{\mathtt{G}}) \mapsto p^{\mathtt{B}} \in [0, 1]$. For any $(p^{\mathtt{B}}, p^{\mathtt{G}}) \in \Sigma^1$, we denote by

$$\mathcal{B}(p^{\mathtt{B}}, p^{\mathtt{G}}) = \bigotimes_{t=0}^{\infty} \left(p^{\mathtt{B}} \delta_{\mathtt{B}} + p^{\mathtt{G}} \delta_{\mathtt{G}} \right) \tag{5}$$

the probability $\mathbb{P}$ on the sample space $\mathbb{H}_\infty$ in (1) which makes the stochastic process $(\mathbf{W}_1, \mathbf{W}_2, \ldots)$ of states of Nature, as in (2), a sequence of independent Bernoulli trials with marginals given by $\mathbb{P}\{\mathbf{W}_t = \mathtt{B}\} = p^{\mathtt{B}}$ and $\mathbb{P}\{\mathbf{W}_t = \mathtt{G}\} = p^{\mathtt{G}}$.

We suppose that the DM makes the assumption that the stochastic process $(\mathbf{W}_1, \mathbf{W}_2, \ldots)$ is governed by $\mathcal{B}(p^{\mathtt{B}}, p^{\mathtt{G}})$, but that the DM does not know the probabilities $(p^{\mathtt{B}}, p^{\mathtt{G}})$. Moreover, we suppose that the DM is a Bayesian who makes the assumption that the unknown couple $(p^{\mathtt{B}}, p^{\mathtt{G}})$ is a random variable with a distribution π_0 on the one-dimensional simplex Σ^1 in (4). This is why we consider the extended sample space $\Sigma^1 \times \mathbb{H}_\infty = \Sigma^1 \times \{\mathtt{B}, \mathtt{G}\}^{\mathbb{N}^*}$ equipped with the probability distribution $\pi_0\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \otimes \mathcal{B}(p^{\mathtt{B}}, p^{\mathtt{G}})$, whose marginal distribution on the sample space $\mathbb{H}_\infty$ in (1) we denote by $\mathbb{P}^{\pi_0}$. Thus, for any measurable bounded function $g : \mathbb{H}_\infty \to \mathbb{R}$, we have that

$$\mathbb{E}^{\mathbb{P}^{\pi_0}}[g] = \int_{\Sigma^1} \pi_0\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \mathbb{E}^{\mathcal{B}(p^{\mathtt{B}}, p^{\mathtt{G}})}[g] \,. \tag{6}$$

Instantaneous payoffs

Now, to compare strategies, we will make up a criterion, or an objective function for the DM. In an evolutionary interpretation, payoffs are measured in "fitness" unit, for instance "number of days alive" or "number of days in a reproductive state", taken as proxies for the number of offspring. The

	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	low payoff $U(\varepsilon, \mathtt{B}) = \mathcal{U}^{\mathtt{B}}$	high payoff $U(\varepsilon, \mathtt{G}) = \mathcal{U}^{\mathtt{G}}$

Table 1: Instant payoffs according to decisions (rows avoid (α) or experiment (ε)) and states of Nature (columns bad $\mathtt{B}$ or good $\mathtt{G}$)

payoffs depend both on the decision and on the state of Nature as in Table 1.

We assume that the payoffs attached to the couple (action, state) in Table 1 are ranked as follows:

$$\overbrace{U(\varepsilon, \mathtt{G}) = \mathcal{U}^{\mathtt{G}}}^{\text{high payoff}} > \underbrace{U(\alpha, \mathtt{B}) = U(\alpha, \mathtt{G}) = \mathcal{U}_\alpha}_{\text{avoidance (middle) payoff}} > \overbrace{U(\varepsilon, \mathtt{B}) = \mathcal{U}^{\mathtt{B}}}^{\text{low payoff}} . \tag{7}$$

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

Intertemporal criterion

As the payoffs in Table 1 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 her/his lifetime performance using strategy $\mathcal{S}$ by the discounted intertemporal payoff

$$j\big(\mathcal{S}, \mathbf{W}\big) = \sum_{t=0}^{+\infty} \rho^t U(\mathbf{V}_t, \mathbf{W}_{t+1}) \, , \tag{8}$$

where $\mathbf{V}_t$ is given by (3). Beyond "fitness", our analysis extends to the maximization of any objective function which can be expressed as an infinite sum over time of discounted payoffs.

The rationale behind using *discounted* intertemporal payoff is the following. Suppose that the DM makes decisions up to a random ultimate stage $\mathbf{T}$ like, for instance, the DM's lifetime (measured in number of decision stages).

If we suppose that the random variable $\mathbf{T}$ is independent of the randomness in the occurence of a bad and good outcomes, and follows a (memoryless) Geometric distribution with values in $\{0, 1, 2, 3 \ldots\}$, then it is easy to establish the equality $\sum_{t=0}^{+\infty} \rho^t U(\mathbf{V}_t, \mathbf{W}_{t+1}) = \mathbb{E}_{\mathbf{T}}[\sum_{t=0}^{\mathbf{T}} U(\mathbf{V}_t, \mathbf{W}_{t+1})]$, where the mathematical expectation $\mathbb{E}_{\mathbf{T}}$ is only w.r.t. the random variable $\mathbf{T}$. Then, we can interpret the discount factor $\rho \in [0, 1[$ in term of the expected value $\overline{\mathbf{T}}$ of the random number $\mathbf{T}$ of stages during which the DM has to make decisions, by means of the equations $\overline{\mathbf{T}} = \rho/(1-\rho)$ and $\rho = \overline{\mathbf{T}}/(\overline{\mathbf{T}}+1)$. For instance, for an individual making daily decisions during a mean time of one year (resp. fifty years), we have $\overline{\mathbf{T}} = 365$ (resp. $\overline{\mathbf{T}} = 365 \times 50$), hence $\rho \approx 0.9972$ (resp. $\rho \approx 0.9999$).

Expected discounted payoff maximization problem

As the payoff (8) is contingent on the unknown scenario $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2, \ldots)$, it is practically impossible that a strategy $\mathcal{S}$ performs better than another for all scenarios. We look for an *optimal strategy* $\mathcal{S}^*$, solution of

$$\mathbb{E}^{\mathbb{P}^{\pi_0}}\left[j(\mathcal{S}^*, \mathbf{W})\right] = \max_{\mathcal{S}} \mathbb{E}^{\mathbb{P}^{\pi_0}}\left[j(\mathcal{S}, \mathbf{W})\right] , \tag{9}$$

where $j(\mathcal{S}, \mathbf{W})$ is given by (8), and the probability $\mathbb{P}^{\pi_0}$, on the sample space $\mathbb{H}_\infty$ in (1), is defined by (6).

2.2 Structure of an optimal strategy and behavior of an optimal decision-maker

Here, we analyze the structure of an optimal strategy, and then we describe the behavior of a decision-maker who adopts such optimal strategy.

Structure of an optimal strategy. Let $\Delta(\Sigma^1)$ denote the set of probability distributions on the simplex Σ^1 in (4). For any $\pi \in \Delta(\Sigma^1)$, we define

$$[\![\pi]\!] = \int_{\Sigma^1} (p^{\mathtt{B}}, p^{\mathtt{G}}) \pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) = \big([\![\pi]\!]^{\mathtt{B}}, [\![\pi]\!]^{\mathtt{G}}\big) \in \Delta(\Sigma^1) , \tag{10a}$$

$$[\![\pi]\!]^{\mathtt{B}} = \int_{\Sigma^1} p^{\mathtt{B}} \pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \in [0, 1] , \tag{10b}$$

$$[\![\pi]\!]^{\mathtt{G}} = \int_{\Sigma^1} p^{\mathtt{G}} \pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \in [0, 1] , \tag{10c}$$

that is, the mean of the random variable $(p^{\mathtt{B}}, p^{\mathtt{G}})$ under probability π, and the means of its two components (with $[\![\pi_\tau]\!]^{\mathtt{B}} + [\![\pi_\tau]\!]^{\mathtt{G}} = 1$). We also define the two *shift mappings* $\theta^{\mathtt{B}}, \theta^{\mathtt{G}} : \Delta(\Sigma^1) \to \Delta(\Sigma^1)$ by

$$(\theta^{\mathtt{B}}\pi)\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) = \frac{p^{\mathtt{B}}}{[\![\pi]\!]^{\mathtt{B}}}\pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \;, \tag{11a}$$

$$(\theta^{\mathtt{G}}\pi)\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) = \frac{p^{\mathtt{G}}}{[\![\pi]\!]^{\mathtt{G}}}\pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \;. \tag{11b}$$

Thus, $\theta^{\mathtt{B}}\pi$ and $\theta^{\mathtt{G}}\pi$, are absolutely continuous with respect to π. When $[\![\pi]\!]^{\mathtt{B}} = 0$, that is, when $\pi = \delta_{(0,1)}$, we set $\theta^{\mathtt{B}}\delta_{(0,1)} = \delta_{(0,1)}$ and, when $[\![\pi]\!]^{\mathtt{G}} = 0$, that is, when $\pi = \delta_{(1,0)}$, we set $\theta^{\mathtt{G}}\delta_{(1,0)} = \delta_{(1,0)}$.

As we said at the beginning of this section, the following result is not new, but we give a proof (in §A.1) that does not rely on the Gittins Index Theorem.

Proposition 1. *There exists an optimal strategy* $\mathcal{S}^* = (\mathcal{S}_0^*, \mathcal{S}_1^*, \ldots)$ *solution of the optimization problem* (9) *made of stationary feedback policies of the form*

$$\mathcal{S}_t^*(\mathbf{Y}_1, \ldots, \mathbf{Y}_t) = \widehat{\mathcal{S}}(\pi_t) \;, \;\; \forall t = 0, 1, 2 \ldots \;, \tag{12}$$

where $\pi_t \in \Delta(\Sigma^1)$ *is given by the dynamical equation*

$$\pi_0 = \pi_0 \;\; \textit{and} \;\; \pi_{t+1} = f(\pi_t, \mathbf{Y}_{t+1}) = \begin{cases} \pi_t & \textit{if } \mathbf{Y}_{t+1} = \partial \;, \\ \theta^{\mathtt{B}}\pi_t & \textit{if } \mathbf{Y}_{t+1} = \mathtt{B} \;, \\ \theta^{\mathtt{G}}\pi_t & \textit{if } \mathbf{Y}_{t+1} = \mathtt{G} \;. \end{cases} \tag{13}$$

Regarding the stationary feedback $\widehat{\mathcal{S}} : \Delta(\Sigma^1) \to \{\varepsilon, \alpha\}$*, there exists a subset* $\Pi_\alpha \subset \Delta(\Sigma^1)$*, and its complementary subset* $\Pi_\varepsilon = \Delta(\Sigma^1) \setminus \Pi_\alpha$*, such that*

- $\widehat{\mathcal{S}}(\pi) = \alpha$ *(that is, select decision "avoid") if* $\pi \in \Pi_\alpha$*,*
- $\widehat{\mathcal{S}}(\pi) = \varepsilon$ *(that is, select decision "experiment") if* $\pi \in \Pi_\varepsilon$*.*

Regarding the complementary subsets Π_α *and* Π_ε*, there exists a function* $V : \Delta(\Sigma^1) \to \mathbb{R}$ *such that*

$$\pi \in \Pi_\alpha \iff V(\pi) = \frac{\mathcal{U}_\alpha}{1-\rho} \;, \;\; \pi \in \Pi_\varepsilon \iff V(\pi) > \frac{\mathcal{U}_\alpha}{1-\rho} \;. \tag{14}$$

The so-called *information state* $\pi_t \in \Delta(\Sigma^1)$ is the conditional distribution, with respect to $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$, of the first coordinate mapping on $\Sigma^1 \times \mathbb{H}_\infty$, that is, the *posterior* of $(p^{\mathtt{B}}, p^{\mathtt{G}})$ at stage t.

Behavior of an optimal decision-maker. We call *optimal DM* a decision-maker who adopts the optimal strategy of Proposition 1. To describe the behavior of an optimal DM, we introduce the *first avoidance stage*, or *first prudent stage*, as the random variable defined by

$$\tau = \inf \left\{ t = 0, 1, 2 \dots \,\middle|\, \pi_t \in \Pi_\alpha \right\} . \tag{15}$$

In case $\pi_t \notin \Pi_\alpha$ for all stages $t = 0, 1, 2 \dots$, the convention is $\tau = \inf \emptyset = +\infty$.

As we said at the beginning of this section, the following result is not new, but we give a proof (in §A.2) that does not rely on the Gittins Index Theorem.

Proposition 2. *The DM that follows the optimal strategy of Proposition 1 switches at most once from experimenting to avoiding. More precisely, her/his behavior displays one of the three following patterns, depending on the first avoidance stage τ in* (15)*.*

a) Infinite risky behavior:
if $\tau = +\infty$, that is, if $\pi_t \in \Pi_\varepsilon$ for all stages $t = 0, 1, 2 \dots$, the optimal DM always experiments (taking risks), hence never avoids.

b) No risky behavior:
if $\tau = 0$, that is, if $\pi_0 \notin \Pi_\alpha$ (that is, $\pi_0 \in \Pi_\varepsilon$), the optimal DM avoids from the start and, from then on, the DM keeps avoiding (prudence) for all times.

c) Finite risky behavior:
if $1 \leq \tau < +\infty$, the optimal DM

- *experiments (taking risks) from $t = 0$ to $\tau - 1$, that is, as long as $\pi_t \in \Pi_\varepsilon$,*
- *switches to avoiding at stage $t = \tau$, that is, as soon as $\pi_t \in \Pi_\alpha$,*
- *from then on, keeps avoiding (prudence) for all times.*

3 Conditions for biased or accurate assessments

Now, we show features of the optimal DM behavior that possess interesting psychological interpretations in terms of human biases: "euphorism" and

status quo biases in §3.1; possible erroneous assessments of the objective best option and of the objective probabilities in §3.2. Contrarily to Sect. 2, the results exposed in this Sect. 3 are new.

3.1 "Euphorism" and status quo biases

Our analysis provides theoretical support to a mix of the so-called *status quo bias* — a preference for the current state of affairs documented in [20] — and to an inclination that we coin "euphorism" bias, related to the "stay-with-a-winner" rule — if an individual experiments a good outcome, it is rational to go on taking risks.

The proof of the following Proposition 3 can be found in A.3.

Proposition 3. *If the optimal DM experiments a good outcome, the DM will go on experimenting ("euphorism"). As a consequence, the experimenting phase (in case it exists) of the optimal DM can only stop when a bad outcome materializes: the switch from riskiness to prudence can only be triggered by the occurrence of a bad outcome.*

Therefore, the behavior of the optimal DM displays at most two consecutive phases of "status quo" — one (possibly empty) of experimenting, that is, taking risks, one (possibly empty) of prudence — with at most one switch; in particular, once prudent, this is forever.

3.2 Possible erroneous assessments of the objective best option and of the objective probabilities

In §3.2.1, we formalize what is an objective environment, with objective best option and probabilities. By refering to an objective environment, we are able to provide a formal definition of a bias in §3.2.2. Finally, we study possible erroneous assessments of the objective best option and of the objective probabilities, for environments prone to prudence in §3.2.3 and for environments prone to risk-taking in §3.2.4.

3.2.1 Objective environment

We suppose that Nature produces bad and good outcomes that are sequences of independent Bernoulli trials governed by a given $(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}}) \in \Sigma^1$. Thus, we equip the sample space $\mathbb{H}_\infty$ in (1) with the probability distribution $\mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}} = \mathcal{B}(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})$ as in (5).

Definition 4. *We call the couple* $\overline{p} = (\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}}) \in \Sigma^1$ *the* objective *or* true *probabilities. We call* environment *the triplet* $(\rho, U, \overline{p})$ *consisting of the discount factor* $\rho \in [0,1[$*, the payoff function* U *in Table 1 (that is, avoidance payoff* $\mathcal{U}_\alpha$*, low payoff* $\mathcal{U}^{\mathtt{B}}$ *and high payoff* $\mathcal{U}^{\mathtt{G}}$*), and the objective probabilities* $\overline{p} = (\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})$*.*

We define the critical probability p_c *by the ratio*

$$p_c = \frac{\mathcal{U}^{\mathtt{G}} - \mathcal{U}_\alpha}{\mathcal{U}^{\mathtt{G}} - \mathcal{U}^{\mathtt{B}}} = \frac{\textit{relative payoff of avoidance}}{\textit{relative payoff of bad outcome}} \in]0,1[\ , \tag{16a}$$

so that we have the equivalence

$$\overline{p}^{\mathtt{B}} < p_c \iff \overline{p}^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + \overline{p}^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} > \mathcal{U}_\alpha\ . \tag{16b}$$

When $\overline{p}^{\mathtt{B}} < p_c$ *(resp.* $\geq$*) or, equivalently, when* $\overline{p}^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + \overline{p}^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} > \mathcal{U}_\alpha$ *(resp.* $\leq$*), we say that the risky (resp. certain) option is the* objectively best option *and that* the environment is prone to risk-taking *(resp.* prudence*).*

All things being equal, the worse a bad outcome (that is, low payoff of bad outcome), the lower the critical probability (16a). When $p_c \approx 0$, the bad outcome is so bad that the positive difference between the payoff of the good outcome and the avoidance payoff is negligible w.r.t. the positive difference between the payoff of the good and the bad outcomes; hence, prudence (avoidance) is the best objective option for most of the values $\overline{p}^{\mathtt{B}}$, since $\overline{p}^{\mathtt{B}} \geq p_c \approx 0$. When $p_c \approx 1$, the good outcome is so good that avoiding the bad outcome costs almost as well as suffering it; taking risks is the best objective option for most of the values $\overline{p}^{\mathtt{B}}$, since $\overline{p}^{\mathtt{B}} < p_c \approx 1$.

3.2.2 Formal definition of bias

We consider the situation where the optimal DM adopts the strategy of Proposition 1, optimal for a given prior beta distribution π_0. More precisely, let $n_0^{\mathtt{B}} > 0$ and $n_0^{\mathtt{G}} > 0$ be two positive scalars. We suppose that the distribution π_0 is the beta distribution $\beta(n_0^{\mathtt{B}}, n_0^{\mathtt{G}})$ on the simplex Σ^1 in (4), that is, for any measurable and integrable function $\varphi : \Sigma^1 \to \mathbb{R}$, we have that

$$\int_{\Sigma^1} \varphi(p^{\mathtt{B}}, p^{\mathtt{G}}) d\pi_0(p^{\mathtt{B}}, p^{\mathtt{G}}) = \frac{\int_0^1 \varphi(p, 1-p) p^{n_0^{\mathtt{B}}-1}(1-p)^{n_0^{\mathtt{G}}-1} dp}{\int_0^1 p^{n_0^{\mathtt{B}}-1}(1-p)^{n_0^{\mathtt{G}}-1} dp}\ . \tag{17}$$

Now, we are equipped to formally define what we call a bias. On the one hand, the optimal strategy of Proposition 1, depends on the discount

factor $\rho \in [0,1[$, on the payoff function U in Table 1, and on the prior beta distribution $\pi_0 = \beta(n_0^{\mathtt{B}}, n_0^{\mathtt{G}})$, but not on the objective probabilities $(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})$. In other words, optimality is w.r.t. the criterion (9), where the mathematical expectation is taken w.r.t. the (subjective) probability $\mathbb{P}^{\pi_0}$, on the sample space $\mathbb{H}_\infty$ in (1), as defined by (6). On the other hand, Nature produces bad and good outcomes that are sequences of independent Bernoulli trials governed by the objective probabilties $(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}}) \in \Sigma^1$. Would the DM know $(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})$, she/he would design a strategy maximizing the criterion (9), but where the mathematical expectation would be taken w.r.t. the (objective) probability $\mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}} = \mathcal{B}(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})$ as in (5). We say that the optimal strategy of Proposition 1 displays a bias when one of its outputs is discrepant with what it would be if the objective probability distribution were known.

Thus, the probability distribution $\mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}} = \mathcal{B}(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})$ and the stochastic process $(\mathbf{W}_1, \mathbf{W}_2, \ldots)$ governed by $\mathcal{B}(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})$ play the role of a background reference objective environment against which one can assess the outputs of the optimal strategy (optimal for π_0), and possibly qualify them of biased or not. We distinguish two outputs of the optimal strategy. One such output is $[\![\pi_t]\!]^{\mathtt{B}}$, the mathematical expectation (10) of the random variable $p^{\mathtt{B}}$. By optimally updating the posterior distribution π_t as in (13), the optimal DM also updates her/his estimate $[\![\pi_t]\!]^{\mathtt{B}}$ of the unknown probability $\overline{p}^{\mathtt{B}}$ of the bad outcome $\mathtt{B}$. Another output is whether $[\![\pi_t]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_t]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} > \mathcal{U}_\alpha$ or $[\![\pi_t]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_t]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} \leq \mathcal{U}_\alpha$, that is, how the DM assesses whether the environment is prone to risk-taking or to prudence.

3.2.3 The case of environments prone to prudence

In Table 2, we sum up the results of Proposition 7 in the case of a prudence-prone environment. Apart from the "euphorism" and status quo biases (row 2, column 2) already discussed in §3.1, the optimal DM displays no bias in the following sense: she/he makes the accurate assessment that the environment is prone to prudence (row 5, column 2); nothing can be said of how $[\![\pi_\tau]\!]^{\mathtt{B}}$ — the posterior of the bad event when learning stops — is related to the objective probability $\overline{p}^{\mathtt{B}}$ (hence the empty box in row 4, column 3).

3.2.4 The case of environments prone to risk-taking

In Table 3, we sum up the results of Proposition 7 in the case of an environment prone to risk-taking, and we point out two possible biases. On

Environment prone to prudence	Endless risk-taking $\tau = +\infty$	Risk-taking then endless prudence $1 \leq \tau < +\infty$
$\overline{p}^{\mathtt{B}} \geq p_c$ $\Longleftrightarrow$ $\overline{p}^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + \overline{p}^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}$ $\leq \mathcal{U}_\alpha$	Behavior *discrepant* with the feature of the environment	Behavior *consistent* with the feature of the environment
	The more likely the bad outcome, the *lower the probability* of *discrepant* endless risk-taking: $\overline{p}^{\mathtt{B}} \nearrow \Longrightarrow$ $\mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}}\{\tau = +\infty\} \searrow$	The more likely the bad outcome, the *higher the probability* of *consistent* endless prudence: $\overline{p}^{\mathtt{B}} \nearrow \Longrightarrow$ $\mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}}\{\tau < +\infty\} \nearrow$
	Accurate asymptotic estimation of $\overline{p}^{\mathtt{B}}$: $\lim_{t\to+\infty} [\![\pi_t]\!]^{\mathtt{B}} = \overline{p}^{\mathtt{B}}$	
	Asymptotic *accurate* assessment that the environment is prone to prudence: $\lim_{t\to+\infty} \left([\![\pi_t]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_t]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}\right)$ $\leq \mathcal{U}_\alpha$	*Accurate* assessment that the environment is prone to prudence: $[\![\pi_\tau]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_\tau]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} \leq \mathcal{U}_\alpha$

Table 2: Optimal behavior in an environment objectively prone to prudence

Environment prone to risk	Endless risk-taking $\tau = +\infty$	Risk-taking then endless prudence $1 \leq \tau < +\infty$
	Behavior *consistent* with the feature of the environment	Behavior *discrepant* with the feature of the environment
$\overline{p}^{\mathtt{B}} < p_c$ $\Longleftrightarrow$ $\overline{p}^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + \overline{p}^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} > \mathcal{U}_\alpha$	The more unlikely the bad outcome the *higher the probability* of *consistent* endless risk-taking: $\overline{p}^{\mathtt{B}} \searrow \Longrightarrow \mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}}\{\tau = +\infty\} \nearrow$	The more unlikely the bad outcome the *lower the probability* of *discrepant* endless prudence: $\overline{p}^{\mathtt{B}} \searrow \Longrightarrow \mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}}\{\tau < +\infty\} \searrow$
	$\overline{p}^{\mathtt{B}} \downarrow 0 \implies \mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}}\{\tau = +\infty\} \uparrow 1$	*Vanishing discrepancy:* $\overline{p}^{\mathtt{B}} \downarrow 0 \implies \mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}}\{\tau < +\infty\} \downarrow 0$
Possible bias	Asymptotic *accurate* estimation of the probability $\overline{p}^{\mathtt{B}}$ of the bad outcome: $\lim_{t\to+\infty}[\![\pi_t]\!]^{\mathtt{B}} = \overline{p}^{\mathtt{B}}$	*Overestimation* of the probability $\overline{p}^{\mathtt{B}}$ of the bad outcome: $\overline{p}^{\mathtt{B}} < p_c \leq [\![\pi_\tau]\!]^{\mathtt{B}}$
Possible bias	Asymptotic *accurate* assessment that the environment is prone to risk-taking: $\lim_{t\to+\infty}\left([\![\pi_t]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_t]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}\right) > \mathcal{U}_\alpha$	*Erroneous* assessment that the environment is prone to prudence: $[\![\pi_\tau]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_\tau]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} \leq \mathcal{U}_\alpha$

Table 3: Optimal behavior in an environment objectively prone to risk-taking

top of the "euphorism" and status quo biases (row 2, column 3) already discussed in §3.1, the optimal DM displays an additional form of pessimism bias. Indeed, the first (and last) time the optimal DM stops choosing the risky option, she/he will erroneously assess that the environment is prone to prudence (row 6, column 3), and will overestimate the probability of the bad outcome (row 5, column 3). More precisely, we establish the following *Biased Learning Theorem*. Its proof is a consequence of Proposition 7 to be found in §A.5. To our knowledge, these results are new.

Theorem 5 (Biased Learning Theorem)**.** *Suppose that the assumptions of Proposition 7 are satisfied and that $\pi_0 \in \Pi_\varepsilon$, so that learning happens, either infinite or finite. Suppose that the environment is prone to risk-taking, that is, (see Definition 4)*

$$\overline{p}^{\mathtt{B}} < p_c \ . \tag{18}$$

Then, the optimal DM (of Proposition 1) can only display two behaviors.

1. *Either the optimal DM will experiment forever, and will* accurately *estimate asymptotically the objective probability $\overline{p}^{\mathtt{B}}$ of the bad outcome* B*; this experiment phase happens with a probability $\mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}}\big(\pi_t \in \Pi_\varepsilon \ , \ \ \forall t = 0,1,2\ldots\big)$ which goes up to 1 when the objective probability $\overline{p}^{\mathtt{B}}$ of the bad outcome* B *goes down to 0. In that case, we conclude that the more likely a bad outcome, the more likely the optimal DM makes an accurate estimation of its objective probability.*

2. *Or the DM will experiment during a finite number of stages and then stop experimenting forever; this stopping phase happens with a probability $\mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}}\big(\exists t = 1,2,\ldots \ , \ \pi_t \in \Pi_\alpha\big)$ which goes down to 0 when the objective probability $\overline{p}^{\mathtt{B}}$ of the bad outcome* B *goes down to 0. In that case, we conclude that, if the objective probability $\overline{p}^{\mathtt{B}}$ of the bad outcome* B *is so low that $\overline{p}^{\mathtt{B}} \leq p_c$, then, when the experiment phase ends at $\tau < +\infty$, the optimal DM will* overerestimate *the objective probability $\overline{p}^{\mathtt{B}}$ of the bad outcome* B *because of the inequalities*

$$\overline{p}^{\mathtt{B}} \leq p_c \leq [\![\pi_\tau]\!]^{\mathtt{B}} \ . \tag{19}$$

However, such an overerestimation happens with a vanishing probability as $\overline{p}^{\mathtt{B}} \downarrow 0$.

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 [19, 7, 5]. 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 the bad outcome when learning stops. However, learning stops (hence overerestimation happens) with nonincreasing and vanishing probability as the objective probability of the bad outcome goes down to zero.

4 Discussion

In §4.1, we discuss the possible implementation of an optimal strategy by humans, and, in §4.2, the robustness of our findings, before concluding in §4.3.

4.1 Possible implementation of an optimal strategy by humans

We discuss the data necessary to design an optimal strategy, as in Proposition 1, and the cognitive burden of implementing it, to see if they are insuperable impediments to its progressive selection during the course of human evolution.

How does the DM obtain the basic data needed to implement an optimal strategy?

The optimal strategy of Proposition 1 depends on the discount factor ρ in (8) and on the payoff function U in Table 1 (that is, avoidance payoff $\mathcal{U}_\alpha$, low payoff $\mathcal{U}^{\mathtt{B}}$ and high payoff $\mathcal{U}^{\mathtt{G}}$). Also, under the assumptions of Proposition 7, the DM holds the prior beta distribution $\pi_0 = \beta(n_0^{\mathtt{B}}, n_0^{\mathtt{G}})$ in (17), where $n_0^{\mathtt{B}} > 0$ and $n_0^{\mathtt{G}} > 0$ are two positive scalars.

We have already discussed, right after Equation (8), how the discount factor ρ can be related to the mean number of stages during which the DM has to make decisions. We suppose that the DM knows the avoidance (middle) payoff $\mathcal{U}_\alpha$. Then, we suggest a way for the DM to jointly determine two integers $n_0^{\mathtt{B}}$ and $n_0^{\mathtt{G}}$ for the beta distribution $\beta(n_0^{\mathtt{B}}, n_0^{\mathtt{G}})$, and both the low

payoff $\mathcal{U}^{\mathtt{B}}$ and the base payoff $\mathcal{U}^{\mathtt{G}}$. The DM starts by making a risky decision and

- either the DM first enjoys n good outcomes $\mathtt{G}$ — hence discovering the high payoff $\mathcal{U}^{\mathtt{G}}$ — before suffering a bad outcome $\mathtt{B}$ — hence discovering the low payoff $\mathcal{U}^{\mathtt{B}}$; in that case, the DM sets $n_0^{\mathtt{G}} = n$ and $n_0^{\mathtt{B}} = 1$;
- or the DM first suffers n bad outcomes $\mathtt{B}$ — hence discovering the low payoff $\mathcal{U}^{\mathtt{B}}$ — before enjoying a good outcome $\mathtt{G}$ — hence discovering the high payoff $\mathcal{U}^{\mathtt{G}}$; in that case, the DM 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 integer parameters $n_0^{\mathtt{B}} > 0$ and $n_0^{\mathtt{G}} > 0$.

What is the stage by stage cognitive load of the optimal DM?

We suppose that the DM holds the prior beta distribution $\pi_0 = \beta(n_0^{\mathtt{B}}, n_0^{\mathtt{G}})$, where $n_0^{\mathtt{B}}, n_0^{\mathtt{G}}$ are integers. Then, we know from Proposition 7 that the posterior π_t is the beta distribution $\beta(n_0^{\mathtt{B}} + N_t^{\mathtt{B}}, n_0^{\mathtt{G}} + N_t^{\mathtt{G}})$ as in (37). Thus, at each decision stage t, the cognitive load of the optimal DM is to keep track of the two integers $n_0^{\mathtt{B}} + N_t^{\mathtt{B}}$ and $n_0^{\mathtt{G}} + N_t^{\mathtt{G}}$ since, by Proposition 1, the optimal decision at stage t is function of the posterior π_t.

How can the DM make an optimal decision at stage t?

By Proposition 1, the DM has to determine if the current posterior $\pi_t \in \Delta(\Sigma^1)$ either belongs to the subset $\Pi_\alpha \subset \Delta(\Sigma^1)$ or to the complementary subset $\Pi_\varepsilon = \Delta(\Sigma^1) \setminus \Pi_\alpha$, to make an optimal decision at stage t. As $\pi_t = \beta(n_0^{\mathtt{B}} + N_t^{\mathtt{B}}, n_0^{\mathtt{G}} + N_t^{\mathtt{G}})$, the DM needs to identify in which of two subsets of $\mathbb{N}^2$ — that is, the couples of integers corresponding to the subsets Π_α and Π_ε of $\Delta(\Sigma^1)$ — does the couple $(n_0^{\mathtt{B}} + N_t^{\mathtt{B}}, n_0^{\mathtt{G}} + N_t^{\mathtt{G}})$ belong.

It is hard to say if our mind can design — using the discount factor ρ, the avoidance (middle) payoff $\mathcal{U}_\alpha$, the low payoff $\mathcal{U}^{\mathtt{B}}$ and the high payoff $\mathcal{U}^{\mathtt{G}}$ — and if our brain can hold such a "mental planar map" made of couples of integers. For instance, for an individual making daily decisions during a mean time of one year (resp. fifty years), this planar map would consist of $365^2 = 133,225$ (resp. $(5 \times 365)^2 3 \approx 333\ 10^6$ couples of integers labelled with a binary label. Even if they are not astronomical, these numbers are huge.[3]

[3]We can easily arrive at astronomical figures with general policies. Indeed, recall that

However, it is possible that a close suboptimal strategy be much more simply encoded by the following rule: if

$$[\![\pi_t]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_t]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} = \frac{n_0^{\mathtt{B}} + N_t^{\mathtt{B}}}{n_0^{\mathtt{B}} + n_0^{\mathtt{G}} + t}\mathcal{U}^{\mathtt{B}} + \frac{n_0^{\mathtt{G}} + N_t^{\mathtt{G}}}{n_0^{\mathtt{B}} + n_0^{\mathtt{G}} + t} > \mathcal{U}_\alpha$$

then make the risky decision $\mathbf{V}_t = \varepsilon$, else avoid. A DM adopting this strategy would be more prudent than the optimal DM because, by (27b), we have that

$$\frac{[\![\pi_t]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_t]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}}{1-\rho} > \frac{\mathcal{U}_\alpha}{1-\rho} \implies V(\pi_t) > \frac{\mathcal{U}_\alpha}{1-\rho} \implies \pi_t \in \Pi_\varepsilon \ .$$

4.2 Robusteness of the results obtained

We discuss which of our results are robust w.r.t. to assumptions like stationarity (of the primitive random variables, of the payoffs), discounting, finite or infinite horizon.

"Euphorism" and status quo biases

The property that, if the optimal DM experiments a good outcome, she/he will go on taking risks is a consequence of the "stay-with-a-winner" property (29): the value function cannot decrease when the posterior changes following a good outcome. Screening the proof of (29) shows that this property only depends on the ranking (7) of payoffs, hence that our finding ("euphorism" bias) is robust w.r.t. nonstationarity (as long as it does not change the ranking), absence of discounting, and finite or infinite horizon.

The property that, if one selects the prudent decision once, one will no longer make risky decisions afterwards is a consequence of both the stopping of posterior updating and of the stationarity of the avoidance domain Π_α.

a policy at stage t is a mapping $\mathcal{S}_t : \{\mathtt{B}, \mathtt{G}, \partial\}^t \to \{\alpha, \varepsilon\}$ that tells the DM what will be the next action in view of past observations. Disregarding the irrelevant "observation" ∂, a policy at stage t is a mapping from a set of cardinal 2^t towards a binary set. If an interval $[t, t+1[$ represents one day, the storage of policies for one year would be astronomically prohibitive. This is why the existence of a stationary feedback optimal policy seems good news. However, the argument of such policy is now an element of $\Delta(\Sigma^1)$, that is, a probability distribution on the one-dimensional simplex. Equivalently, being able to implement the optimal strategy of Proposition 1 amounts to being able to characterize the two complementary subsets Π_α and Π_ε, which is out of question except with simple rules.

We discuss both of them. The property that the posterior is a sufficient information state for optimization is quite robust, as it holds true under nonstationarity, absence of discounting, and finite or infinite horizon ([2, Chap. 10]). The stopping of posterior updating follows from the information structure, as avoidance freezes observation, hence is robust. However, these are stationarity and infinite horizon that lead to status quo. Indeed, were the avoidance domain Π_α dependent on the stage t that we could no longer conclude to status quo. Thus, the status quo bias is less robust than the "euphorism" bias.

Pessimistic erroneous assessment of the environment and overestimation bias for the probability of the bad outcome

The property that the probability of a bad outcome is overestimated when the risky phase stops (hence the erroneous assessment that the environment is prone to prudence) comes from the inequality (27b), itself a consequence of stationarity, discounting and infinite horizon. In this sense, it is less robust than the "euphorism" bias.

4.3 Conclusion

Our model and analysis show that certain biases can be the product of rational behavior, here in the sense of maximizing expected discounted utility (that is, being risk neutral) with learning. Indeed, our theoretical results provide support to "euphorism" and status quo biases, as well as, under narrow boundary conditions, the pessimistic erroneous assessment of the best objective option and an overestimation bias for the probability of bad outcomes. In particular, we have shown a Biased Learning Theorem that provides rational ground for the human bias that consists in attributing to bad 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 [17]. 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 relative payoffs by the formula (16a). 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[4] (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 and the organizers of the seminar of Lameta, Montpellier (Monday 3 October 2016); Khalil Helioui, Geoffrey Barrows, 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).

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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

A Technical results and proofs

A.1 Proof of Proposition 1

Proof. We follow the approach of [2, Chap. 10] for the analysis of imperfect state information models.

The infinite horizon imperfect state information stochastic optimization problem (9)-(8)-(3) can be written as[5]

$$\sup \mathbb{E}^{\mathbb{P}^{\pi_0}}\Big[\sum_{t=0}^{+\infty} \rho^t U(\mathbf{V}_t, \mathbf{W}_{t+1})\Big] = \sup \int_{\Sigma^1} \Big[\sum_{t=0}^{+\infty} \rho^t L(x_t, v_t)\Big] \pi_0(dx_0) \tag{20a}$$

where we have introduced the one-dimensional simplex Σ^1 in (4) as state space, and the state dynamics

$$x_0 = (p^{\mathtt{B}}, p^{\mathtt{G}}) \in \Sigma^1 \ , \ \ x_{t+1} = x_t \ , \ \ \forall t = 0, 1, 2 \dots \ , \tag{20b}$$

or, equivalently, the state space Σ^1 and the state transition kernels

$$k_X(dx \mid x, v) = k_X(d(p^{\mathtt{B}}, p^{\mathtt{G}}) \mid (p^{\mathtt{B}}, p^{\mathtt{G}}), v) = \delta_x = \delta_{(p^{\mathtt{B}}, p^{\mathtt{G}})} \ , \tag{20c}$$

the control space $\{\alpha, \varepsilon\}$ and the controls

$$v_t \in \{\alpha, \varepsilon\} \ , \ \ \forall t = 0, 1, 2 \dots \ , \tag{20d}$$

the one-stage payoff

$$\begin{aligned} L(x, v) = L\big((p^{\mathtt{B}}, p^{\mathtt{G}}), v\big) &= p^{\mathtt{B}} U(v, \mathtt{B}) + p^{\mathtt{G}} U(v, \mathtt{G}) \\ &= \begin{cases} \mathcal{U}_\alpha & \text{if } v = \alpha \ , \\ p^{\mathtt{B}} \mathcal{U}^{\mathtt{B}} + p^{\mathtt{G}} \mathcal{U}^{\mathtt{G}} & \text{if } v = \varepsilon \ , \end{cases} \end{aligned} \tag{20e}$$

and the observation space $\{\mathtt{B}, \mathtt{G}, \partial\}$ and the observation stochastic kernel

$$\begin{aligned} & k_Y(dy \mid x, v) = k_Y(dy \mid (p^{\mathtt{B}}, p^{\mathtt{G}}), v) \\ = & \begin{cases} \delta_\partial(dy) & \text{if } v = \alpha \ , \\ p^{\mathtt{B}} \delta_{\mathtt{B}}(dy) + p^{\mathtt{G}} \delta_{\mathtt{G}}(dy) & \text{if } v = \varepsilon \ . \end{cases} \end{aligned} \tag{20f}$$

[5] We do not detail over which possible solutions the two suprema are taken. The left hand side supremum is w.r.t. (3), whereas the right hand side supremum is w.r.t. suitable stochastic kernels deduced from the information structure below.

The stochastic kernel

$$k_X(dx \mid \pi, v, y) = \begin{cases} \pi(dx) & \text{if } v = \alpha \text{ and } y = \partial\,, \\ (\theta^{\mathtt{B}}\pi)(dx) & \text{if } v = \varepsilon \text{ and } y = \mathtt{B}\,, \\ (\theta^{\mathtt{G}}\pi)(dx) & \text{if } v = \varepsilon \text{ and } y = \mathtt{G}\,. \end{cases} \tag{20g}$$

satisfies [2, Lemma 10.3], because it can be checked that, for any measurable subset $\Sigma \subset \Sigma^1$ and subset $C \subset \{\mathtt{B}, \mathtt{G}, \partial\}$, and any $v \in \{\alpha, \varepsilon\}$, one has that

$$\int_\Sigma k_Y(C \mid x, v)\pi(dx) = \int_{\Sigma^1} \Big[\int_C k_X(\Sigma \mid \pi, v, y) k_Y(dy \mid x, v)\Big]\pi(dx)\,. \tag{21}$$

Indeed, for $v = \alpha$, Equation (21) is satisfied because

$$\begin{aligned} &\int_{\Sigma^1} \Big[\int_C k_X(\Sigma \mid \pi, \alpha, y) k_Y(dy \mid x, \alpha)\Big]\pi(dx) \\ &= \int_{\Sigma^1} \Big[\int_C \pi(\Sigma)\delta_\partial(dy)\Big]\pi(dx) \end{aligned}$$

by the expressions (20g) for $k_X(dx \mid \pi, \alpha, y)$ and (20f) for $k_Y(dy \mid x, \alpha)$

$$\begin{aligned} &= \delta_\partial(C)\pi(\Sigma) = \int_\Sigma \delta_\partial(C)\pi(dx) \\ &= \int_\Sigma k_Y(C \mid x, \alpha)\pi(dx)\,. \qquad \text{(by the expression (20f) for } k_Y(dy \mid x, \alpha)) \end{aligned}$$

For $v = \varepsilon$, we show that Equation (21) is satisfied for $C = \{\mathtt{B}\}, \{\mathtt{G}\}, \{\partial\}$. For $C = \{\partial\}$, both sides of the Equation (21) are zero as $k_Y(\{\partial\} \mid x, \varepsilon) = 0$ by the expression (20f) for $k_Y(dy \mid x, \varepsilon)$. For $C = \{\mathtt{B}\}$, we calculate

$$\begin{aligned} &\int_{\Sigma^1} \Big[\int_{\{\mathtt{B}\}} k_X(\Sigma \mid \pi, \varepsilon, y) k_Y(dy \mid x, \varepsilon)\Big]\pi(dx) \\ &= \int_{\Sigma^1} k_X(\Sigma \mid \pi, \varepsilon, \mathtt{B}) k_Y(\{\mathtt{B}\} \mid x, \varepsilon)\pi(dx) \\ &= \int_{\Sigma^1} (\theta^{\mathtt{B}}\pi)(\Sigma)\, p^{\mathtt{B}}\pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \end{aligned}$$

by the expressions (20g) for $k_X(dx \mid \pi, \varepsilon, y)$ and (20f) for $k_Y(dy \mid x, \varepsilon)$

$$
\begin{aligned}
&= \frac{1}{\int_{\Sigma^1} p^{\mathtt{B}} \pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big)} \int_{\Sigma^1} \Big[\int_{\Sigma} q^{\mathtt{B}} \pi\big(d(q^{\mathtt{B}}, q^{\mathtt{G}})\big)\Big] p^{\mathtt{B}} \pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \\
&\qquad\qquad \text{(by the expression (11) of } \theta^{\mathtt{B}}\pi) \\
&= \int_{\Sigma} q^{\mathtt{B}} \pi\big(d(q^{\mathtt{B}}, q^{\mathtt{G}})\big) \\
&= \int_{\Sigma} k_Y(\{\mathtt{B}\} \mid (p^{\mathtt{B}}, p^{\mathtt{G}}), \varepsilon) \pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \\
&\qquad\qquad \text{(by the expression (20f) for } k_Y(dy \mid (p^{\mathtt{B}}, p^{\mathtt{G}}), \varepsilon)) \\
&= \int_{\Sigma} k_Y(\{\mathtt{B}\} \mid x, \varepsilon) \pi(dx) \; .
\end{aligned}
$$

For $C = \{\mathtt{G}\}$, we obtain Equation (21) in the same way.

By [2, Propositions 10.5 and 10.6], we conclude that the imperfect state information model can be reduced to a perfect state one, with new information state $\pi \in \Delta(\Sigma^1)$, new information state transition kernels

$$
k_{\Pi}(d\pi \mid \pi, v) = \begin{cases} \pi & \text{if } v = \alpha \; , \\ [\![\pi]\!]^{\mathtt{B}} \delta_{\theta^{\mathtt{B}}\pi} + [\![\pi]\!]^{\mathtt{G}} \delta_{\theta^{\mathtt{G}}\pi} & \text{if } v = \varepsilon \; , \end{cases} \tag{22}
$$

where $[\![\pi]\!]^{\mathtt{B}}$ and $[\![\pi]\!]^{\mathtt{G}}$ have been defined in (10), and new one-stage payoff

$$
\begin{aligned}
\tilde{L}(\pi, v) &= \int_{\Sigma^1} L(x, v) \pi(dx) && (23) \\
&= \int_{\Sigma^1} \big[p^{\mathtt{B}} U(v, \mathtt{B}) + p^{\mathtt{G}} U(v, \mathtt{G})\big] \pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) && \text{(by (20e))} \\
&= \begin{cases} \mathcal{U}_{\alpha} & \text{if } v = \alpha \; , \\ [\![\pi]\!]^{\mathtt{B}} \mathcal{U}^{\mathtt{B}} + [\![\pi]\!]^{\mathtt{G}} \mathcal{U}^{\mathtt{G}} & \text{if } v = \varepsilon \; . \end{cases}
\end{aligned}
$$

The value function $V : \Delta(\Sigma^1) \to \mathbb{R}$ given by

$$
V(\pi) = \sup \mathbb{E}^{\mathbb{P}^{\pi}} \Big[\sum_{t=0}^{+\infty} \rho^t U(\mathbf{V}_t, \mathbf{W}_{t+1})\Big] = \sup \int_{\Sigma^1} \Big[\sum_{t=0}^{+\infty} \rho^t L(x_t, v_t)\Big] \pi(dx_0) \tag{24}
$$

satisfies, by [2, Proposition 9.8], the dynamic programming equation

$$
V(\pi) = \max_{v \in \{\alpha, \varepsilon\}} \left(\tilde{L}(\pi, v) + \int_{\Sigma^1} k_{\Pi}(d\pi' \mid \pi, v) V(\pi')\right) , \tag{25}
$$

that is, by (23) and (22),

$$V(\pi) = \max\left\{\mathcal{U}_\alpha + \rho V(\pi), [\![\pi]\!]^{\mathtt{B}}\big(\mathcal{U}^{\mathtt{B}} + \rho V(\theta^{\mathtt{B}}\pi)\big) + [\![\pi]\!]^{\mathtt{G}}\big(\mathcal{U}^{\mathtt{G}} + \rho V(\theta^{\mathtt{G}}\pi)\big)\right\}. \tag{26}$$

By definition (24) of the value function $V : \Delta(\Sigma^1) \to \mathbb{R}$, we have that

$$V(\pi) \geq \sum_{t=0}^{+\infty} \rho^t \int_{\Sigma^1} L(x_t, \alpha)\pi(dx_0) = \sum_{t=0}^{+\infty} \rho^t \mathcal{U}_\alpha = \frac{\mathcal{U}_\alpha}{1-\rho} \tag{27a}$$

by applying the open-loop strategy $\mathcal{S}_t = \alpha$ for all t, and we also have that

$$V(\pi) \geq \frac{[\![\pi]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}}{1-\rho}\,. \tag{27b}$$

Indeed, we have that

$$\begin{aligned} V(\pi) &\geq \sum_{t=0}^{+\infty} \rho^t \int_{\Sigma^1} L(x_t, \varepsilon)\pi(dx_0) \\ &\qquad \text{(by applying the open-loop strategy } \mathcal{S}_t = \varepsilon \text{ for all } t) \\ &= \sum_{t=0}^{+\infty} \rho^t \int_{\Sigma^1} \big[p^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + p^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}\big]\pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \qquad \text{(by (20e))} \\ &= \frac{[\![\pi]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}}{1-\rho} \qquad \text{(by (10).)} \end{aligned}$$

The existence of the proposed stationary optimal policy is given by [2, Proposition 9.12, Corollary 9.12.1, Corollary 9.17.1]: depending whether the maximum in the dynamic programming equation (25) is achieved for $v = \alpha$ or for $v = \varepsilon$, we select an optimal strategy accordingly. This is why we define the subset $\Pi_\alpha \subset \Delta(\Sigma^1)$ by

$$\pi \in \Pi_\alpha \iff V(\pi) = \mathcal{U}_\alpha + \rho V(\pi) \iff V(\pi) = \frac{\mathcal{U}_\alpha}{1-\rho}\,,$$

which gives the first part of (14). From the inequality (27a), we deduce the second part of (14):

$$\pi \notin \Pi_\alpha \iff \pi \in \Pi_\varepsilon \iff V(\pi) > \frac{\mathcal{U}_\alpha}{1-\rho}\,.$$

This ends the proof. □

A.2 Proof of Proposition 2

Proof.

a) By Proposition 1, when $\tau = +\infty$ — that is, when $\pi_t \in \Pi_\varepsilon$ for all stages t — it is optimal to select decision ε and to experiment forever.

b) By Proposition 1, when $\tau = 0$ — that is, when $\pi_0 \in \Pi_\alpha$ — it is optimal to select decision α and to avoid for all times. Indeed, once the optimal DM avoids, the DM does not observe the random outcomes, hence the DM no longer updates the posterior π_t in (13), so that the DM keeps avoiding.

c) When $1 \leq \tau < +\infty$, we have, by definition (15) of the first avoidance stage τ,

   - $\pi_t \in \Pi_\varepsilon$ for stages $t = 0$ up to $\tau - 1$; hence, by Proposition 1, it is optimal to select decision ε and experiment from stages $t = 0$ up to $\tau - 1$;
   - $\pi_t \in \Pi_\alpha$ for stages $t = \tau$ up to $+\infty$; hence, by Proposition 1, it is optimal to select decision α and avoid for stages $t = \tau$ up to $+\infty$. Indeed, once the optimal DM avoids, the DM does not observe the random outcomes, hence the DM no longer updates the posterior π_t in (13), so that the DM keeps avoiding.

This ends the proof. □

A.3 Proof of Proposition 3

Proof. First, we prove that, if the optimal DM experiments a good outcome, the DM will go on experimenting.

Suppose that, at stage t the optimal DM is experimenting. We will show that, if a good outcome $\mathtt{G}$ materializes at the end of the interval $[t, t+1[$ (that is, if $\mathbf{Y}_{t+1} = \mathtt{G}$), then necessarily the optimal DM goes on experimenting at stage $t + 1$. In what follows, the value function $V : \Delta(\Sigma^1) \to \mathbb{R}$ has been introduced in Proposition 1, and is defined in (24). We have that

$$V(\pi_{t+1}) = V(\theta^{\mathtt{G}}\pi_t)$$

because, as we supposed that $\mathbf{Y}_{t+1} = \mathtt{G}$, we have that $\pi_{t+1} = \theta^{\mathtt{G}}\pi_t$ by the dynamics (13)

$$\begin{aligned} &\geq V(\pi_t) \qquad \text{(by the property } V \circ \theta^{\mathtt{G}} \geq V\text{, shown afterward)} \\ &> \frac{\mathcal{U}_\alpha}{1-\rho} \end{aligned}$$

by (14), because, as we supposed that the optimal DM is experimenting at stage t, we have that $\pi_t \in \Pi_\varepsilon$. Thus, we have obtained that $V(\pi_{t+1}) > \frac{\mathcal{U}_\alpha}{1-\rho}$. By (14), we conclude that the optimal DM goes on experimenting at stage $t+1$ by Proposition 1.

We now prove that the value function (24) has the property

$$V \circ \theta^{\mathtt{G}} \geq V \,, \tag{29}$$

that is, the value function cannot decrease when the posterior changes following a good outcome. We will use this "stay-with-a-winner" property when we discuss the robustness of our findings in §4.2.

Before that, we recall that two random variables $\mathbf{C}$ and $\mathbf{D}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, are said to be *comonotonic* when we have that $\big(\mathbf{C}(\omega) - \mathbf{C}(\omega')\big)\big(\mathbf{D}(\omega) - \mathbf{D}(\omega')\big) \geq 0$, for any $(\omega, \omega') \in \Omega^2$. In that case, it is easily shown that $\mathbb{E}\big[\mathbf{CD}\big] \geq \mathbb{E}\big[\mathbf{C}\big]\mathbb{E}\big[\mathbf{D}\big]$, when $\mathbf{C}$ and $\mathbf{D}$ are square integrable.

By definition (24) of the value function V, to prove (29) it suffices to show that

$$\int_{\Sigma^1} L(x,v)(\theta^{\mathtt{G}}\pi)(dx) \geq \int_{\Sigma^1} L(x,v)\pi(dx) \,, \;\; \forall v \in \{\alpha, \varepsilon\} \,.$$

This is obvious for $v = \alpha$ since $L(x,\alpha) = \mathcal{U}_\alpha$ by (20e). For $v = \varepsilon$, we have that

$$\begin{aligned} \int_{\Sigma^1} L(x,\varepsilon)(\theta^{\mathtt{G}}\pi)(dx) &= \int_{\Sigma^1} \big[p^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + p^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}\big] \frac{p^{\mathtt{G}}}{[\![\pi]\!]^{\mathtt{G}}} \pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \\ &\qquad\qquad \text{(by (20e) and (11))} \\ &\geq \frac{1}{[\![\pi]\!]^{\mathtt{G}}} \int_{\Sigma^1} \big[p^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + p^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}\big] \pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \int_{\Sigma^1} p^{\mathtt{G}} \pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big) \end{aligned}$$

because the random variables $\mathbf{C} : \Sigma^1 \ni (p^{\mathtt{B}}, p^{\mathtt{G}}) \mapsto p^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + p^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}$ and $\mathbf{D} : \Sigma^1 \ni (p^{\mathtt{B}}, p^{\mathtt{G}}) \mapsto p^{\mathtt{G}}$ are comonotonic, since the function $p^{\mathtt{G}} \mapsto p^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + p^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} = p^{\mathtt{G}}(\mathcal{U}^{\mathtt{G}} - \mathcal{U}^{\mathtt{B}}) + \mathcal{U}^{\mathtt{B}}$ is increasing as a consequence of $\mathcal{U}^{\mathtt{G}} > \mathcal{U}^{\mathtt{B}}$ by (7)

$$= \int_{\Sigma^1} \left[p^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + p^{\mathtt{G}}\mathcal{U}^{\mathtt{G}}\right]\pi\big(d(p^{\mathtt{B}}, p^{\mathtt{G}})\big)$$

(by the definition (10) of $[\![\pi]\!]^{\mathtt{G}}$)

$$= \int_{\Sigma^1} L(x, \varepsilon)\pi(dx) \qquad \text{(by (20e).)}$$

We can prove in the same way that $V \geq V \circ \theta^{\mathtt{B}}$, so that we have obtained

$$V \circ \theta^{\mathtt{G}} \geq V \geq V \circ \theta^{\mathtt{B}} \; . \tag{30}$$

The rest of the proof follows from Proposition 1. In particular, once the optimal DM selects the "avoid" option, the DM will never more experiment. Indeed, the optimal rule of Proposition 1 states that, once the optimal DM selects the "avoid" option, the DM does not observe the random outcomes, hence the DM no longer updates the posterior π_t because of the dynamics (13) so that the DM keeps avoiding.

This ends the proof. □

A.4 Monotonicity property w.r.t. the probability $p^{\mathtt{B}}$

We consider the set of functions

$$\mathcal{Z} = \left\{\varphi : \Delta(\Sigma^1) \to \mathbb{R}_+ \,\middle|\, \varphi \text{ measurable and } \varphi \circ \theta^{\mathtt{G}} \geq \varphi \circ \theta^{\mathtt{B}}\right\} , \tag{31}$$

where the shift mappings $\theta^{\mathtt{B}}, \theta^{\mathtt{G}} : \Delta(\Sigma^1) \to \Delta(\Sigma^1)$ have been defined in (11).

Proposition 6. *For any sequence $\{\varphi_s\}_{s=0,\ldots,t}$ of functions in $\mathcal{Z}$, the function $[0,1] \ni p^{\mathtt{B}} \mapsto \mathbb{E}_{\mathcal{B}(p^{\mathtt{B}}, 1-p^{\mathtt{B}})}\left[\prod_{s=0}^{t} \varphi_s(\pi_s)\right]$ is nonincreasing, where, for any $(p^{\mathtt{B}}, p^{\mathtt{G}}) \in \Sigma^1$, the probability distribution $\mathcal{B}(p^{\mathtt{B}}, p^{\mathtt{G}})$ on the sample space $\mathbb{H}_\infty$ in (1) is given in (5), and where the sequence $\{\pi_s\}_{s=0,\ldots,t}$ of posteriors is given by*

$$\pi_0 = \pi_0 \;\; \textit{and} \;\; \pi_{s+1} = f(\pi_s, \mathbf{W}_{s+1}) = \begin{cases} \theta^{\mathtt{B}}\pi_s & \textit{if } \mathbf{W}_{s+1} = \mathtt{B} \; , \\ \theta^{\mathtt{G}}\pi_s & \textit{if } \mathbf{W}_{s+1} = \mathtt{G} \; . \end{cases} \tag{32}$$

Proof. Let $(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})$ and $(\overline{\overline{p}}^{\mathtt{B}}, \overline{\overline{p}}^{\mathtt{G}})$ in Σ^1 be such that $\overline{p}^{\mathtt{B}} \leq \overline{\overline{p}}^{\mathtt{B}}$ (or, equivalently, that $\overline{p}^{\mathtt{G}} \geq \overline{\overline{p}}^{\mathtt{G}}$). We will show, by induction on $t \in \mathbb{N}$, that, for any sequence $\{\varphi_s\}_{s=0,\dots,t}$ of functions in $\mathcal{Z}$, as in (31), we have the inequality

$$\mathbb{E}_{\mathcal{B}(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}\big[\prod_{s=0}^{t} \varphi_s(\pi_s)\big] \geq \mathbb{E}_{\mathcal{B}(\overline{\overline{p}}^{\mathtt{B}}, \overline{\overline{p}}^{\mathtt{G}})}\big[\prod_{s=0}^{t} \varphi_s(\pi_s)\big] \ . \tag{33}$$

Before that, we need one notation and two preliminary results. For any function $\varphi : \Delta(\Sigma^1) \to \mathbb{R}$, we put

$$\overline{P}\varphi = \overline{p}^{\mathtt{B}}(\varphi \circ \theta^{\mathtt{B}}) + \overline{p}^{\mathtt{G}}(\varphi \circ \theta^{\mathtt{G}}) \ , \ \ \overline{\overline{P}}\varphi = \overline{\overline{p}}^{\mathtt{B}}(\varphi \circ \theta^{\mathtt{B}}) + \overline{\overline{p}}^{\mathtt{G}}(\varphi \circ \theta^{\mathtt{G}}) \ .$$

On the one hand, from the equalities

$$\overline{P}\varphi - \overline{\overline{P}}\varphi = (\overline{p}^{\mathtt{B}} - \overline{\overline{p}}^{\mathtt{B}})(\varphi \circ \theta^{\mathtt{B}}) + (\overline{p}^{\mathtt{G}} - \overline{\overline{p}}^{\mathtt{G}})(\varphi \circ \theta^{\mathtt{G}}) = (\overline{p}^{\mathtt{B}} - \overline{\overline{p}}^{\mathtt{B}})(\varphi \circ \theta^{\mathtt{B}} - \varphi \circ \theta^{\mathtt{G}}) \ ,$$

we readily get that, as $\overline{p}^{\mathtt{B}} - \overline{\overline{p}}^{\mathtt{B}} \leq 0$,

$$\varphi \in \mathcal{Z} \implies (\varphi \circ \theta^{\mathtt{B}} - \varphi \circ \theta^{\mathtt{G}}) \leq 0 \implies \overline{P}\varphi \geq \overline{\overline{P}}\varphi \ . \tag{34}$$

On the other hand, we have that

$$\varphi \in \mathcal{Z} \implies \overline{P}\varphi \in \mathcal{Z} \text{ and } \overline{\overline{P}}\varphi \in \mathcal{Z} \ , \tag{35}$$

because, if $\varphi \in \mathcal{Z}$, one has

$$\begin{aligned} (\overline{P}\varphi) \circ \theta^{\mathtt{B}} &= \overline{p}^{\mathtt{B}}(\varphi \circ \theta^{\mathtt{B}} \circ \theta^{\mathtt{B}}) + \overline{p}^{\mathtt{G}}(\varphi \circ \theta^{\mathtt{G}} \circ \theta^{\mathtt{B}}) \quad \text{(by definition of } (\overline{P}\varphi) \circ \theta^{\mathtt{B}}) \\ &= \overline{p}^{\mathtt{B}}(\varphi \circ \theta^{\mathtt{B}} \circ \theta^{\mathtt{B}}) + \overline{p}^{\mathtt{G}}(\varphi \circ \theta^{\mathtt{B}} \circ \theta^{\mathtt{G}}) \end{aligned}$$

because $\theta^{\mathtt{B}} \circ \theta^{\mathtt{G}} = \theta^{\mathtt{G}} \circ \theta^{\mathtt{B}}$ as easily seen from the definitions (11)

$$\begin{aligned} &\leq \overline{p}^{\mathtt{B}}(\varphi \circ \theta^{\mathtt{G}} \circ \theta^{\mathtt{B}}) + \overline{p}^{\mathtt{G}}(\varphi \circ \theta^{\mathtt{G}} \circ \theta^{\mathtt{G}}) \\ &\qquad\qquad \text{(as } \varphi \in \mathcal{Z} \text{ and by definition (31) of } \mathcal{Z}) \\ &= (\overline{P}\varphi) \circ \theta^{\mathtt{G}} \qquad\qquad \text{(by definition of } (\overline{P}\varphi) \circ \theta^{\mathtt{G}}.) \end{aligned}$$

The same inequality holds true for $\overline{\overline{P}}\varphi$.

Now, we can prove the inequality (33) by induction. For $t = 0$, the inequality (33) is true as it is the trivial equality $\varphi_0(\pi_0) = \varphi_0(\pi_0)$. Let us

suppose that the inequality (33) holds true, for any sequence $\{\varphi_s\}_{s=0,\dots,t}$ of functions in $\mathcal{Z}$. We consider a sequence $\{\varphi_s\}_{s=0,\dots,t,t+1}$ of functions in $\mathcal{Z}$, and we calculate

$$\begin{aligned}\mathbb{E}_{\mathcal{B}(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\big[\prod_{s=0}^{t+1}\varphi_s(\pi_s)\big] &= \mathbb{E}_{\mathcal{B}(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\Big[\prod_{s=0}^{t}\varphi_s(\pi_s)\mathbb{E}_{\mathcal{B}(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\big[\varphi_{t+1}(\pi_{t+1})\mid \pi_s\ ,\ \ s=0,\dots,t\big]\Big]\\ &= \mathbb{E}_{\mathcal{B}(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\Big[\prod_{s=0}^{t}\varphi_s(\pi_s)(\overline{P}\varphi_{t+1})(\pi_t)\Big]\end{aligned}$$

by (32) in which the sequence $\{W_s\}_{s=0,\dots,t+1}$ is i.i.d. under the probability distribution $\mathcal{B}(p^{\mathtt{B}},p^{\mathtt{G}})$ in (1) with probability $p^{\mathtt{B}}$ (resp. $p^{\mathtt{G}}$) to take the value $\mathtt{B}$ (resp. $\mathtt{G}$)

$$\begin{aligned}&\geq \mathbb{E}_{\mathcal{B}(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\Big[\prod_{s=0}^{t}\varphi_s(\pi_s)(\overline{\overline{P}}\varphi_{t+1})(\pi_t)\Big] \qquad \text{(by (34))}\\ &= \mathbb{E}_{\mathcal{B}(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\Big[\prod_{s=0}^{t-1}\varphi_s(\pi_s)\times\big(\varphi_t(\overline{\overline{P}}\varphi_{t+1})\big)(\pi_t)\Big]\\ &\geq \mathbb{E}_{\mathcal{B}(\overline{\overline{p}}^{\mathtt{B}},\overline{\overline{p}}^{\mathtt{G}})}\Big[\prod_{s=0}^{t-1}\varphi_s(\pi_s)\times\big(\varphi_t(\overline{\overline{P}}\varphi_{t+1})\big)(\pi_t)\Big]\end{aligned}$$

by the induction inequality (33) because $\varphi_s\in\mathcal{Z}$ for $s=0,\dots,t$ by assumption, that $\overline{\overline{P}}\varphi_{t+1}\in\mathcal{Z}$ by (35) as $\varphi_{t+1}\in\mathcal{Z}$ by assumption, and that a product of nonnegative functions in $\mathcal{Z}$ is also in $\mathcal{Z}$

$$\begin{aligned}&= \mathbb{E}_{\mathcal{B}(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\big[\prod_{s=0}^{t+1}\varphi_s(\pi_s)\big]\\ &\qquad\qquad\text{(by going backward in the same way.)}\end{aligned}$$

This ends the proof. □

A.5 Proposition 7

The following Proposition 7 details what are the estimates $[\![\pi_t]\!]^{\mathtt{B}}$ of the objective probability value $\overline{p}^{\mathtt{B}}$ that the optimal DM is forming during the course of learning, and how she/he assesses the environment. It also establishes

how the probability of relevant events monotonically depends upon objective probabilities. To our knowledge, these results are new.

Proposition 7. *Let* $(\mathbf{W}_1, \mathbf{W}_2, \ldots)$ *be a sequence of independent Bernoulli trials governed by the objective probability distribution* $\mathbb{P}^{\delta(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})} = \mathcal{B}(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})$*, as in* (5)*, where* $(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}}) \in \Sigma^1$*. Suppose that the DM adopts the corresponding strategy* $\mathcal{S}^*$ *of Proposition 1, based on the observations* $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots)$ *inductively given by* (3) *(for* $\mathcal{S} = \mathcal{S}^*$*) and on the sequence of posteriors* π_t *given by the dynamics* (13)*.*

Suppose also that the DM holds the prior beta distribution $\pi_0 = \beta(n_0^{\mathtt{B}}, n_0^{\mathtt{G}})$*, where* $n_0^{\mathtt{B}} > 0$ *and* $n_0^{\mathtt{G}} > 0$ *are two positive scalars. Then, if we define the numbers* $N_t^{\mathtt{B}}$ *and* $N_t^{\mathtt{G}}$ *of bad and good outcomes up to stage* t *by*

$$N_0^{\mathtt{B}} = N_0^{\mathtt{G}} = 0 \ , \ \ N_t^{\mathtt{B}} = \sum_{s=1}^{t} \mathbf{1}_{\{\mathbf{Y}_s = \mathtt{B}\}} \ , \ \ N_t^{\mathtt{G}} = \sum_{s=1}^{t} \mathbf{1}_{\{\mathbf{Y}_s = \mathtt{G}\}} \ , \ \ t = 1, 2, \ldots \tag{36}$$

the posterior π_t *in Proposition 1 is the beta distribution*

$$\pi_t = \beta(n_0^{\mathtt{B}} + N_t^{\mathtt{B}}, n_0^{\mathtt{G}} + N_t^{\mathtt{G}}) \ , \tag{37}$$

on the simplex Σ^1*, whose conditional expectation* (10) *is given by the statistics*

$$[\![\pi_0]\!]^{\mathtt{B}} = \frac{n_0^{\mathtt{B}}}{n_0^{\mathtt{B}} + n_0^{\mathtt{G}}} \ \text{ and } \ [\![\pi_t]\!]^{\mathtt{B}} = \frac{n_0^{\mathtt{B}} + N_t^{\mathtt{B}}}{n_0^{\mathtt{B}} + n_0^{\mathtt{G}} + t} \ , \ \ t = 1, 2 \ldots \ . \tag{38}$$

Moreover, here are the assessments of the objective probability value $\overline{p}^{\mathtt{B}}$ *and of the environment made by the above optimal DM.*

a) Infinite learning $\tau = +\infty$*. Infinite learning can only happen when* $\pi_0 \in \Pi_\varepsilon$*. When* $\tau = +\infty$*, the optimal DM experiments forever and, the statistics* $[\![\pi_t]\!]^{\mathtt{B}}$ *in* (38) *asymptotically reaches the objective probability value* $\overline{p}^{\mathtt{B}}$*, almost surely under the objective probability distribution* $\mathbb{P}^{\delta(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}$*, that is,*

$$\lim_{t \to +\infty} [\![\pi_t]\!]^{\mathtt{B}} = \overline{p}^{\mathtt{B}} \ , \tag{39a}$$

or, in more precise terms,

$$\mathbb{P}^{\delta(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})} \Big\{ \lim_{t \to +\infty} [\![\pi_t]\!]^{\mathtt{B}} = \overline{p}^{\mathtt{B}} \ , \ \ \tau = +\infty \Big\} = \mathbb{P}^{\delta(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})} \big\{ \tau = +\infty \big\} \ . \tag{39b}$$

Then, the optimal DM asymptotically makes an accurate assessment of the objective best option as $\lim_{t \to +\infty} \big([\![\pi_t]\!]^{\mathtt{B}} \mathcal{U}^{\mathtt{B}} + [\![\pi_t]\!]^{\mathtt{G}} \mathcal{U}^{\mathtt{G}} \big) = \overline{p}^{\mathtt{B}} \mathcal{U}^{\mathtt{B}} + \overline{p}^{\mathtt{G}} \mathcal{U}^{\mathtt{G}}$*.*

Infinite learning happens with probability $\mathbb{P}^{\delta(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\big(\pi_t \in \Pi_\varepsilon\ ,\ \ \forall t = 0,1,2\ldots\big)$, *which goes up to 1 when the objective probability* $\overline{p}^{\mathtt{B}}$ *of the bad outcome* $\mathtt{B}$ *goes down to 0. As a consequence, an accurate estimation of the objective probability of a rare bad outcome is likely.*

b) No learning $\tau = 0$. *No learning happens if and only if* $\pi_0 \in \Pi_\alpha$. *When* $\tau = 0$, *the optimal DM never experiments and the DM initial estimate* $[\![\pi_0]\!]^{\mathtt{B}}$ *of the objective probability value* $\overline{p}^{\mathtt{B}}$ *satisfies*

$$p_c \leq [\![\pi_0]\!]^{\mathtt{B}}\ , \tag{40}$$

where the critical probability p_c *is defined in* (16a). *From the start, the optimal DM assesses that the environment is prone to prudence, as* $[\![\pi_0]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_0]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} \leq \mathcal{U}_\alpha$.

c) Finite learning $1 \leq \tau < +\infty$. *Finite learning can only happen when* $\pi_0 \in \Pi_\varepsilon$. *When* $1 \leq \tau < +\infty$, *the optimal DM experiments till stage* τ *and the DM stops her/his estimation of the objective probability value* $\overline{p}^{\mathtt{B}}$ *at a value* $[\![\pi_\tau]\!]^{\mathtt{B}}$ *which satisfies*

$$p_c \leq [\![\pi_\tau]\!]^{\mathtt{B}}\ , \tag{41a}$$

or, in more precise terms,

$$\mathbb{P}^{\delta(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\Big\{p_c \leq [\![\pi_\tau]\!]^{\mathtt{B}}\ ,\ \ 1 \leq \tau < +\infty\Big\} = \mathbb{P}^{\delta(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\big\{1 \leq \tau < +\infty\big\}\ . \tag{41b}$$

When the optimal DM stops experimenting, she/he assesses that the environment is prone to prudence, as $[\![\pi_\tau]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_\tau]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} \leq \mathcal{U}_\alpha$.

Finite learning happens with probability $\mathbb{P}^{\delta(\overline{p}^{\mathtt{B}},\overline{p}^{\mathtt{G}})}\big(\exists t = 1,2,\ldots\ ,\ \ \pi_t \in \Pi_\alpha\big)$, *which goes down to 0 when the objective probability* $\overline{p}^{\mathtt{B}}$ *of the bad outcome* $\mathtt{B}$ *goes down to 0. As a consequence, if the objective probability* $\overline{p}^{\mathtt{B}}$ *of the* bad *outcome* $\mathtt{B}$ *is low enough, in the sense that* $\overline{p}^{\mathtt{B}} \leq p_c$, *when the experiment phase ends at* $\tau < +\infty$, *we have*

$$\overline{p}^{\mathtt{B}} \leq p_c \leq [\![\pi_\tau]\!]^{\mathtt{B}}\ , \tag{42}$$

hence the DM will overerestimate *the objective probability* $\overline{p}^{\mathtt{B}}$ *of the bad outcome* $\mathtt{B}$, *but this with a vanishing probability as* $\overline{p}^{\mathtt{B}} \downarrow 0$.

Proof. By the dynamics (13), we easily establish that (37) holds true. Equation (38) follows from property of beta distributions (17).

a) By Proposition 1, when $\tau = +\infty$ it is optimal to select decision ε and experiment forever. Thus, the observations $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots)$ in (3) coincide with $(\mathbf{W}_1, \mathbf{W}_2, \ldots)$, and we get that $N_t^{\mathtt{B}} = \sum_{s=1}^t \mathbf{1}_{\{\mathbf{W}_s = \mathtt{B}\}}$ and $N_t^{\mathtt{G}} = \sum_{s=1}^t \mathbf{1}_{\{\mathbf{W}_s = \mathtt{G}\}}$, for all $t = 1, 2, \ldots$, by (36). As the random variables $(\mathbf{W}_1, \mathbf{W}_2, \ldots)$ in (2) are i.i.d. under $\mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}}$, by the Law of large numbers we have that

$$\frac{n_0^{\mathtt{B}} + N_t^{\mathtt{B}}}{n_0^{\mathtt{B}} + N_t^{\mathtt{B}} + n_0^{\mathtt{G}} + N_t^{\mathtt{G}}} = \frac{n_0^{\mathtt{B}} + \sum_{s=1}^t \mathbf{1}_{\{\mathbf{W}_s = \mathtt{B}\}}}{n_0^{\mathtt{B}} + n_0^{\mathtt{G}} + t} \to_{t \to +\infty} \overline{p}^{\mathtt{B}} \ , \ \ \mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}} - \text{p.s.}$$

By (38), asymptotically the statistics $[\![\pi_t]\!]^{\mathtt{B}}$ reaches the objective probability value $\overline{p}^{\mathtt{B}}$ almost surely under the probability $\mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}}$.

Now, we show that the function $[0,1] \ni \overline{p}^{\mathtt{B}} \mapsto \mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}}\big(\pi_t \in \Pi_\varepsilon \ , \ \ \forall t = 0, 1, 2 \ldots\big)$ is nonincreasing. For this purpose, it suffices to prove that, for any stage t, the function

$$[0,1] \ni \overline{p}^{\mathtt{B}} \mapsto \mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}}\big(\pi_s \in \Pi_\varepsilon \ , \ \ \forall s = 0, 1, 2 \ldots, t\big) \tag{43}$$

is nonincreasing, and then let $t \to +\infty$. Now, the functions $\varphi_s(\pi) = \mathbf{1}_{\{\pi \in \Pi_\varepsilon\}}$ (the same function for all $s = 0, 1, 2 \ldots$) satisfy the assumptions of Proposition 6 because

$$\begin{aligned}
(\varphi_s \circ \theta^{\mathtt{G}})(\pi) &= \mathbf{1}_{\{\theta^{\mathtt{G}}\pi \in \Pi_\varepsilon\}} \\
&= \mathbf{1}_{\{(V \circ \theta^{\mathtt{G}})(\pi) > \frac{\mathcal{U}_\alpha}{1-\rho}\}} && \text{(by (14))} \\
&\geq \mathbf{1}_{\{(V \circ \theta^{\mathtt{B}})(\pi) > \frac{\mathcal{U}_\alpha}{1-\rho}\}} && \text{(as } V \circ \theta^{\mathtt{G}} \geq V \circ \theta^{\mathtt{B}} \text{ by (30))} \\
&= \mathbf{1}_{\{\theta^{\mathtt{B}}\pi \in \Pi_\varepsilon\}} && \text{(by (14))} \\
&= (\varphi_s \circ \theta^{\mathtt{B}})(\pi) \ .
\end{aligned}$$

We conclude, using Proposition 6, that the function (43) is nonincreasing.

Finally, we easily establish that $\mathbb{P}^{\delta_{(0,1)}}\big(\pi_t \in \Pi_\varepsilon \ , \ \ \forall t = 0, 1, 2 \ldots\big) = 1$. Indeed, under the probability $\mathbb{P}^{\delta_{(0,1)}}$, we have $\mathbf{W}_t = \mathtt{G}$ for all stage t almost-

surely, hence $\pi_{t+1} = \theta^{\mathtt{G}}\pi_t$ by the dynamics (13). Therefore, we get that

$$\begin{aligned}
\pi_t \in \Pi_\varepsilon &\implies V(\pi_t) > \frac{\mathcal{U}_\alpha}{1-\rho} && \text{(by (14))}\\
&\implies V(\theta^{\mathtt{G}}\pi_t) \geq V(\pi_t) > \frac{\mathcal{U}_\alpha}{1-\rho} && \text{(by (30))}\\
&\implies V(\pi_{t+1}) > \frac{\mathcal{U}_\alpha}{1-\rho} && \text{(since } \pi_{t+1} = \theta^{\mathtt{G}}\pi_t\text{)}\\
&\implies \pi_{t+1} \in \Pi_\varepsilon\ . && \text{(by (14))}
\end{aligned}$$

Since $\pi_0 \in \Pi_\varepsilon$, we deduce that $\mathbb{P}^{\delta_{(0,1)}}\big(\pi_t \in \Pi_\varepsilon\ ,\ \ \forall t = 0, 1, 2 \ldots\big) = 1$.

b) See the proof below.

c) Let us suppose that $\tau < +\infty$. We have that

$$\begin{aligned}
\mathcal{U}_\alpha &= (1-\rho)V(\pi_\tau) && \text{(by definition (15) of } \tau \text{ and since } \tau < +\infty\text{)}\\
&\geq [\![\pi_\tau]\!]^{\mathtt{B}}\mathcal{U}^{\mathtt{B}} + [\![\pi_\tau]\!]^{\mathtt{G}}\mathcal{U}^{\mathtt{G}} && \text{(by the inequality (27b))}\\
&= -[\![\pi_\tau]\!]^{\mathtt{B}}\big(\mathcal{U}^{\mathtt{G}} - \mathcal{U}^{\mathtt{B}}\big) + \mathcal{U}^{\mathtt{G}} && \text{(since } [\![\pi_\tau]\!]^{\mathtt{B}} + [\![\pi_\tau]\!]^{\mathtt{G}} = 1 \text{ by (10).)}
\end{aligned}$$

Rearranging the terms, and using (16a), we obtain that $[\![\pi_\tau]\!]^{\mathtt{B}} \geq p_c = \frac{\mathcal{U}^{\mathtt{G}} - \mathcal{U}_\alpha}{\mathcal{U}^{\mathtt{G}} - \mathcal{U}^{\mathtt{B}}}$.

The rest of the proof follows using the property that, if $\pi_0 \in \Pi_\varepsilon$, then

$$\mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}}\big(\exists t = 1, 2, \ldots\ ,\ \ \pi_t \in \Pi_\alpha\big) = 1 - \mathbb{P}^{\delta_{(\overline{p}^{\mathtt{B}}, \overline{p}^{\mathtt{G}})}}\big(\pi_t \in \Pi_\varepsilon\ ,\ \ \forall t = 0, 1, 2 \ldots\big)\ .$$

This ends the proof. □

References

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

[2] D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case.* Athena Scientific, Belmont, Massachusetts, 1996.

[3] 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.

[4] J. Boutang and M. De Lara. *Les Biais de l'esprit : Comment l'évolution a forgé notre psychologie*. Odile Jacob, Paris, 2019.

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

[6] 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.

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

[8] 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.

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

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

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

[12] 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.

[13] 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.

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

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

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

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

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

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

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