Decision-making under uncertainty in dynamic settings: an experimental study

Élise Payzan*, Peter Bossaerts†

April 8, 2009

Abstract

In modern financial markets, uncertainty is dynamic. Agents attempt to infer assets' underlying return-generating processes, but these processes jump randomly over time. In such nonstationary contexts, optimal Bayesian updating is remarkably complex. It calls for explicit learning of both outcome and jump probability (the "Hierarchical Bayes model"), or, if jumps are accounted for only implicitly, through discounting of the old data, optimal learning of the discount parameter (the "Forgetting Bayes model"). A Bayesian thus deals not only with the riskiness of the returns, but also with parameter uncertainty and jump risk. Parameter uncertainty stems from the fact that a Bayesian does not merely consider a single-point estimate but all the possible values of the unknown probability of earning excess returns. Jump risk comes from the abrupt changes in this probability. If optimal Bayesian inference appears too complicated in such dynamic settings, one can completely avoid the multiple layers of uncertainty, by learning the assets' values through a simple *win-keep lose-switch* heuristic (the "Reinforcement Learning" model). We ask to what extent people can account for these jumps and update their probability estimates rationally (i.e., according to Bayes rule). We propose a novel experimental task with which to study decision-making under uncertainty in such dynamic problems. We found that both Bayesian models explained subjects' choices uniformly better than the Reinforcement Learning model. Our result suggests that people are capable of processing information rationally in the face of the most complex situations, as long as the latter are sufficiently compelling.

Keywords: Decision Making Under Uncertainty, Risk, Jumps, Bayesian Learning, Reinforcement Learning

École Polytechnique Fédérale de Lausanne and *Swiss Finance Institute*. Correspondence should be sent at elise.payzan@epfl.ch

†*École Polytechnique Fédérale de Lausanne*, *Swiss Finance Institute*, and *California Institute of Technology*

1 Optimal Vs Heuristic Decision-Making In A Complex Task

Animals foraging for food, traders picking stocks, oil men exploiting oil wells, people playing bandits in a casino, all have to choose between various reward-generating-processes (food sources, assets, oil wells, bandits) with unknown reward probabilities. Contrary to casino bandits which are fixed, the reward-generating-processes encountered in some real-world situations are affected by abrupt changes (jumps), as when a market event causes a huge discontinuity in assets' returns, when oil unexpectedly dries up at a hitherto-productive place, etc. Day-traders and oil men need to detect such unexpected jumps and adapt behavior to the new contexts (e.g., start investing in new assets). Such jumps are routinely encountered in modern financial markets – see, e.g., Barndorff-Nielsen and Shephard (2006), Huang and Tauchen (2005), Tauchen and Zhou (2006).

The present experimental study is an attempt to learn whether individuals can process information rationally – and hence do optimal allocation decisions – in such dynamic settings. The answer may be negative. The type of uncertainty human beings are optimized for is the one generated by Nature, not the one generated through modern financial markets. Most of the ecologically relevant changes that animals encounter in their natural environments are steady (see, e.g., the slow diffusion of the reward rate provided by a source of food); abrupt changes are rare events.[1] So it is an open question whether people are sufficiently sophisticated that they can fare well in the face of jumps.

To examine the nature of learning in such dynamic settings, we designed the "Boardgame," a six-armed bandit task in the form of a board with six locations, three blue and three red. Each trial, every single location delivers one outcome among three possibilities. At a blue location, the possible outcomes are 1 CHF, −1 CHF, 0 CHF; at a red one, 2 CHF, −2 CHF, 0 CHF. At the beginning of the game, the nature of each location – i.e., the probabilities of the three outcomes – is unknown. Furthermore, the locations jump randomly throughout play. That a location jumps means two of its outcome probabilities swap. For example, consider a location which returns the reward outcome with probability 0.8; the loss outcome, with probability 0.2. This location jumps and it now returns the reward outcome with probability 0.2; the loss outcome, with probability 0.8. In this instance of a jump, the reward probability swaps with the loss probability. There are two independent jump processes, one for the red locations, one for the blue ones. When a jump occurs for one of the two colors, the three locations of this color jump at the same time. The two jump intensities are fixed. The jump intensity for the red locations is larger then the jump intensity for the blue ones, whereby jumps are more frequent at the red locations than at the blue ones. Given a color, the three locations differ in their riskiness. The Boardgame is a high-frequency sampling task, so the riskiness of a location relates to the ability to predict next outcome while sampling this location. Given a color, one location has high risk in

[1] Jumps in the rates of reward of natural sources are encountered only after rare events such as a vulcan, etc.

that it is unpredictable (say, "random"): even though one knows perfectly its probabilities, one cannot predict next outcome (as is the case when the outcome probabilities are close to 1/3). One location has low risk in that it is predictable (say, "biased"): if one knows its probabilities, one may want to bet on the nature of next outcome (as is the case when the outcome probabilities are very different i.e., one outcome is much more likely than the others). One location is in between (say, "median") i.e., less predictable than the biased one, and more predictable than the random one. These risk levels are fixed. Each trial, the player (a she) selects one location. She immediately receives the outcome generated by the chosen location, and does not see any outcome elsewhere. She accumulates rewards and losses throughout the game. The goal is to maximize the cumulative earnings during the game. The player knows that jumps occur independently for the two colors, that the red locations are more unstable than the blue ones, and that the jump intensities are fixed. The player also understands what it means for a location to jump (the swap feature). Besides, she knows that for each color, the three locations differ in their level of risk. However, the player knows neither the absolute values of the two jump intensities, nor those of the three levels of risk. Additionally, among the three locations of a same color, the player does not know which is the biased one, which is the median one, and which is the random one. To examine the nature of learning behind choice in this task, we had 62 subjects play during 30 minutes, and recorded their choice at each trial of play (500 trials on average).

The Boardgame was meant to be the simplest possible setting to study decision making in a relevant multi-armed bandit task with jumps. Given the difficulty of this class of changepoint problems, it was critical to make the game engaging, transparent, and well-structured. For otherwise we would measure anything but noise in our experimental task. This led us to introduce the following features.

Firstly, we introduced six locations, three blue and three red, instead of only two, to make the game engaging and facilitate learning. First, the jump intensities for blue vs red were sufficiently different that the red locations should be perceived as really unstable compared to the blue ones, stable by contrast. This "counterpoint" effect was meant to facilitate learning, by helping the player be sensitive to the uncertainty coming from the jumps. Further, given a color, we used contrasts of entropies.[2] We contrasted one minimal-entropy location, very predictable when generating returns, with one maximal-entropy location, highly unpredictable, and with one median-entropy location, quite predictable, but to a lesser degree than the minimal-entropy location (more on this below, in the detailed presentation of the task). The instructions are deliberately vague regarding the three entropy levels.[3] The fact that the player knows neither the absolute levels of entropy nor which location is biased, which is median, and which is random, is meant to render the game engaging:

[2]The entropy of a location measures how much the probabilities differ across the three possible outcomes. Entropy is highest when all probabilities are equal, whereby it relates to the unpredictability of a location (how much the generated outcomes surprise one).

[3]We refrained from using the term "entropy" in the instructions. Rather, we used pictures to explain intuitively what it means for a location to be "predictable"/biased.

a sophisticated player may first attempt to pin the nature of each location down, then do strategic allocations (e.g., time visits of the locations perceived to be biased, to catch the good runs). The majority of our subjects reportedly did so. However, the use of *six* bandits – a relatively large number for this class of problems – could cause an excessive memory-load for the player.[4] We thus provide the player with the history of past received outcomes, in the form of cues displayed on the board. Hence our subjects did not have to remember everything, and could play effectively. So overall, the contrast between the six locations made the game absorbing.

Secondly, we wanted the dynamics of the locations to be well-structured and transparent, for otherwise learning the *nonstationary* returns distributions would be hampered. Indeed, Epstein and Schneider (2007) has argued that probabilistic learning is implausible when the probabilities are nonstationary. But the situations envisioned in claims like this are more unstructured than the one in our Boardgame, where i) jump detection is facilitated, as the probabilities change in a very specific sense, through simple swaps; and ii) the instructions are very transparent regarding the nature of such jumps. So, the player can in principle accommodate her learning to the presence of the jumps.

We believe the resulting Boardgame balances reasonably well degree of complexity and level of induced engagement, as our subjects could plausibly learn something about the locations throughout play, despite the complexity of the problem. They were given strong incentives to do so.[5]

Our goal was to infer something about the learning processes behind our subjects' choices. We tested two competing theories of learning in the Boardgame. The first, which we refer to as the "full-rationality hypothesis" or "Bayesian hypothesis," posits sophisticated players who process information rationally (i.e., according to Bayes rule) to estimate at each trial the outcome probabilities of the locations. The second theory, which we call the "bounded-rationality hypothesis," states that players are unsophisticated "reinforcement learners," because Bayesian learning is too demanding in this task. While Bayesians understand that the outcomes returned by the locations are caused by the hidden probabilities, reinforcement learners, in contrast, ignore probabilities. At the start of the game, they arbitrarily forecast next outcome for each location. Then, at each trial, they update the forecast attached to the visited location according to the prediction error (the gap between the returned outcome and the forecast) they observe. That is, reinforcement learners have adaptive expectations of the values of the six locations. Normatively, reinforcement learning is ad-hoc, and one could cook other heuristics that may fit better with our task. Nonetheless, we know from prior work in decision neuroscience that reinforcement learning is so ingrained in the mammal brain that it appears to govern reward learning in various experimental tasks. In the light of these results, one may posit Boardgame players to be reinforcement learners as well.

[4] To alleviate the complexity of the task, we may have suppressed the median-entropy location and just keep the two extreme ones. We tested this alternative setting with a few subjects, but it was reported to be much less engaging.

[5] The aim of the game is to maximize the accumulated earnings during the game. The subjects knew that a good player earns a lot of money in this game (more than 150 CHF), while a mediocre player gets back home with the show-up fee only (5 CHF).

We thus took reinforcement learning to represent bounded rationality in our task.

Since the present paper emphasizes learning, both theories are deliberately general on how the player selects one location at each trial, on the basis of the values she has learnt. Specifically, both theories posit that in this game, each subject was capable of weighing optimally the merits of exploitation against those of exploration, in accordance with her idiosyncratic "exploratory tendency." (Exploitation refers to the motive to select the location with the highest estimated value; exploration, to the opposing motive to visit the other locations, to get information about their value – after all, one only has estimates of the values.)

To be able to compare the two theories, we described them with "models." A model is a learning algorithm (to estimate the values of the locations) along with a decision rule (to link the estimated values to choice). As hinted above, the Bayesian and reinforcement learning models share the same decision rule: we used the logit rule to model the foregoing general assumption about choice. The models differ in their learning algorithms. There are two possible candidates to formulate the full-rationality hypothesis. Full Bayesian learning calls for explicit learning of both outcome and jump probabilities (the "Hierarchical Bayes model"). Nonetheless, given the richness of the information to be processed in our complicated task, a sophisticated player may refrain from learning explicitly the two jump intensities. After all, what really matters is to infer outcome probability. A second kind of sophisticated player infers outcome probability with a natural sampling scheme that accommodates its sample size to the strength of evidence in favor of a jump at each trial (the "Forgetting Bayes model"). Both kinds of sophisticated players process information rationally. The competing "Reinforcement Learning model," in contrast, is composed of a suboptimal reinforcement learning heuristic.

We examined which model best explained the choices we recorded from each of the 62 subjects. We set up the stochastic structure of the game so that we were able to discriminate between the Bayesian models and the Reinforcement Learning model, because they prescribed different courses of action. The models were fit to the data with maximum likelihood, using the Nelder-Mead simplex method and a genetic algorithm to find the maxima. We found that the Bayesian models explained subjects' decisions almost uniformly better than the Reinforcement Learning model, with slight superiority of the Hierarchical Bayes model. This means our subjects acted more like Bayesians.

To our knowledge, the present paper is a first experimental attempt to learn whether individuals can process information rationally in dynamic situations of uncertainty. The optimal behavior in our task is conceptually and computationally very difficult. So, from the bounded rationality perspective, it is hard to believe that subjects perform such calculations. One may expect them to play heuristically, if not randomly. Still, sophisticated thinking prevailed in our experiment.

Relating our finding to prior work leads to the following fact: In somewhat difficult problems, heuristic rules outperform the normative prediction – see, e.g., Tversky and Kahneman (1971), Kahneman and Tversky (1972), Grether

(1992), Charness and Levin (2005), whereas in our *very* difficult decision problem, the normative rule appeared to be a much better prediction. Even more strikingly, our finding contrasts with recent examples that point to significant bounds to human rationality even when computational effort would be minimal – see, e.g., Johnson et al. (2002).

Standard theories of bounded rationality don't produce this fact, which suggests that task complexity is not necessarily an obstacle for rationality to emerge. On the contrary, we believe the complexity of our task, together with its compelling nature – because of both its deeply engaging game play and its high monetary incentives, to explain the prevalence of the optimal decision plan. As such, our result may prompt a reevaluation of the scope of bounded rationality.

We will first present the Boardgame. We then propose two classes of behavior in the Boardgame. In the first, people process information rationally in the face of our task. In the second, they are guided by a simple heuristic. We show in detail how these two candidate plans proceed in our task. Last and foremost, we compare the two theories, and show that the first is a much more plausible explanation of our subjects' behavior.

2 Experimental Task

Our experimental task is a six-armed bandit task in which the bandits are six locations displayed on a board. Each trial, every single location returns one outcome. There are three possible outcomes ("states"): the location returns either a reward, or a loss, or nothing. The probabilities associated with each state are hidden. Further, they change abruptly (jump) over time. Each trial, the player selects one location. She then immediately receives the outcome generated by the chosen location (and does not see any outcome elsewhere). Rewards and losses are accumulated throughout play. The aim is to make as much money as possible.

How do you think players behave in a problem like this? Owing to the jumps, this class of decision problems is very difficult to solve, and hence not very relevant, as people are expected to play randomly. However, we attempted to create a design sufficiently engaging, well-structured, and transparent, that its complexity would not hamper effective learning – and informed choice. This led us to create the so called Boardgame.

Engaging. In the Boardgame, the locations contrast in their levels of risk (entropy) and instability (jump frequency). Specifically, there are three blue locations and three red (see Fig. 1). The two different colors point to two Bernoulli jump processes, one that concerns the blue locations, and one that concerns the red ones. Each jump process is a sequence of iid random variables with two possible outcomes at each trial, *Jump* or *No jump*. The occurrence of a jump for red (resp blue) thus means all the red (resp blue) locations change at the same time. Jump intensity is 1/4 for red, 1/16 for blue, whereby the red locations are very unstable compared to the blue ones, experienced as stable by contrast. Given a color, the three locations differ in their level of risk/entropy. The minimal-entropy location, median-entropy location, and

maximal-entropy location, have an entropy level equal to 0.3, 0.65, and 1.1, respectively. An instance of a minimal-entropy location is a location delivering the reward outcome with probability 0.8 and the loss outcome with probability 0.2. If one knows the underlying probability of such a location, one expects next outcome to be good. As such, the minimal-entropy location is predictable. Conversely, each outcome is equally likely when sampling the maximal-entropy location, so one does not expect the realization of a particular outcome there. The median-entropy location is relatively predictable, but to a lesser extent than the minimal-entropy location: see, e.g., a location generating the three outcomes with probabilities 0.6, 0.3, and 0.1. We used such "counterpoint effects" with the two dimensions (instability and risk) in order to render the game engaging and help the players be sensitive to the two dimensions. The cost of doing that is that there are *six* locations, a relatively large number. To alleviate the memory-load for our subjects, we thus provided them with the history of past outcomes throughout play (see Fig. 1).

The resulting task is still a very difficult game, which requires one to be extremely focused. So we gave the player huge monetary incentives to play well. Each trial, 1 CHF can be earned at a blue location (2 CHF at a red one), and there are on average 500 trials per play. Before the start of the game, the player knows she will receive the accumulated earnings minus a fixed price – we ensured that every single subject understood the *call* nature of the payoff.[6] Therefore, before starting the game, our subjects knew that they would earn a lot if they could accumulate a lot of rewards during the game, and that they would get back home with nothing but the show-up fee of 5 CHF in the case they would underperform during the game.

Well-structured. The Boardgame is Markovian (more on this below), which means that for each location, the generated outcome at a trial depends on the outcome probability at this trial, which in turn depends on the outcome probability at the previous trial through the jump process. Besides, the nature of the changes in the probabilities is very specific: a jump at a location means two of its outcome probabilities swap. Jump detection is thus possible, as the structure of the game is transparent to the player.

Transparent. The player can intuitively understand the hierarchical structure of the task, because she knows that the probabilities are governed by the jump process. She is aware of the meaning of the two colors, of the nature of the changes (the swap feature), and of the presence of the three types of location (in terms of their level of unpredictability) for each color. However, the player is not told the absolute levels of the two jump intensities, neither those of the entropies. Besides, among the three locations of the same color, she does not know which is the very predictable one, which is the median one, and which is the unpredictable one. The goal was to balance non-triviality of the task and sufficient transparency. To judge to what extent such a target was reached, the reader should think like an actual player and attempt to go through the game rules as actually presented in the experiment (see Fig. 2

[6]However, we deliberately refrained from telling the subjects the amount of the fixed price, because the knowledge of the fixed price would possibly influence their choices during the game – and we did not want to have to model such influences.

and Fig. 3).[7]

We thus posited our subjects to be capable of playing effectively – at least, not randomly – in the face of this complex and compelling game. What does it mean for a subject to "play effectively"? We had two candidates in mind. The first decision plan is the one of a probabilistically sophisticated Bayesian player; the second, the one of an unsophisticated player, called reinforcement learner, who uses a simple win-stay lose-move heuristic. The present paper answers the question of whether actual players acted more like Bayesians or like reinforcement learners. Before presenting the horse race between the two hypotheses, we have to be explicit about their nature.

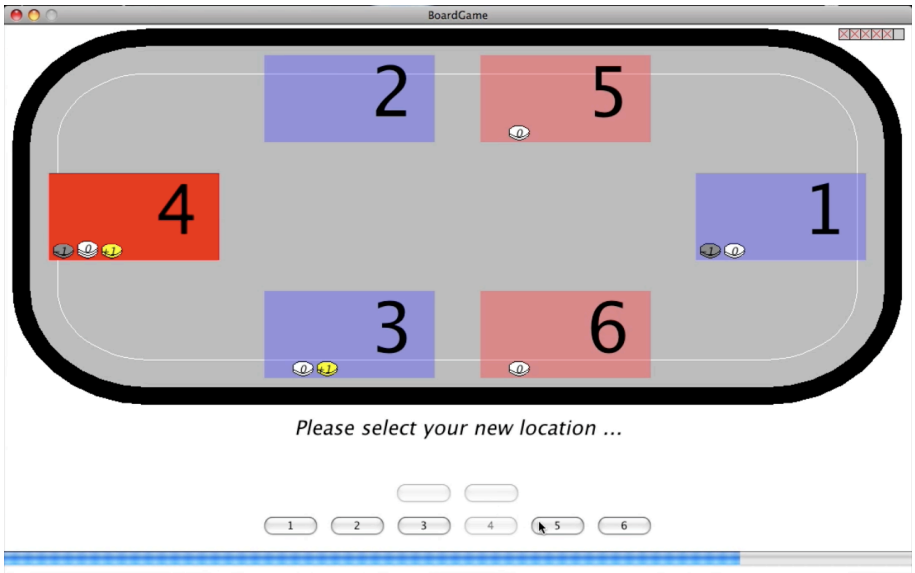


Figure 1: **User Interface of the Boardgame**

[7] All our subjects were French-speakers. They consulted the French version of the Boardgame website. See "Experimental Protocol" in the Appendix.

Home Sign up **The game** How to play (instructions) Demo GO TO THE GAME

Please read carefully this section, which explains the rules of the Boardgame.

It will take you about 10 to 15 minutes to read this section, as the Boardgame is quite complicated. Please be patient.

A play of the boardgame lasts around 30 minutes.

At the beginning, you'll see on the screen a board composed of 6 locations. This board will remain on screen throughout play. Each location has a number - the symbols 1 through 6 are used. These numbers don't mean anything; they just serve to distinguish between the six locations on the board.

There are three blue locations and three red locations.

Each round, every location delivers one outcome. There are three possible outcomes: a blue (resp red) location returns either 1 CHF (resp 2 CHF), or - 1 CHF (resp - 2 CHF), or 0 CHF. That is, three possible scenarios each round: win, lose, no change.

At the start of each round, when you've selected one location (see the ``Instructions'' page), you never know for sure whether you're gonna win or lose or get 0 CHF.

The chances to be in each scenario differ across the locations.

So there may be ``good'' and ``bad'' locations, but at the beginning of the game, you don't know which ones are good and which ones are bad. It's up to you to discover that.

Even though you knew the chances to be in each scenario, you still would not know for sure the outcome you're gonna receive before you actually see it (it's like in the Roulette). However, knowing the chances may allow you to anticipate somehow which outcome you're gonna receive.

For instance, if you know that the location you've just selected gives 2 CHF 80% of the time and -2 CHF the rest of the time, you expect more to receive 2 CHF.

Furthermore, knowing the chances of the three scenarios gives you an idea of the degree of uncertainty/unpredictability of a location.

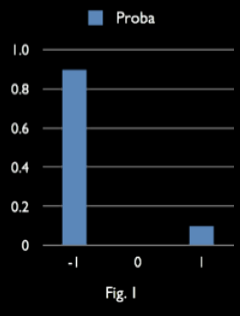


Fig. 1

For instance, consider (Fig. 1) a blue location giving 1 CHF 10% of the time and -1 CHF the rest of the time - i.e., the probability to win 1 CHF is 0.1, the probability to lose 1 CHF is 0.9, and the probability to get 0 CHF is 0.

This location is very "biased" because the bad scenario is much more likely than the two others. If asked, you would bet more on the occurrence of the bad scenario than on the occurrence of the good scenario - which is unlikely. And you are sure that 0 CHF is not gonna happen.
In that sense, this location is predictable.

Figure 2: **Rules of the Boardgame**

Now, consider (Fig. 2) instead a blue location for which the three scenarios are equally likely.

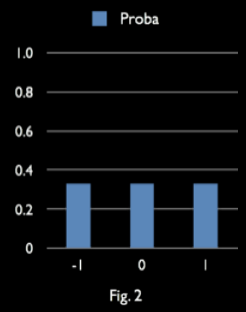


Fig. 2

This location has maximal uncertainty / unpredictability, in that you cannot anticipate at all the outcome that such a location is gonna return: everything is possible - there is no reason why -1 CHF should happen more rather than +1 CHF or 0 CHF.

The locations are more or less biased: among the three blue locations, there is one location that is very biased, another one that is less biased, and the third one is not biased at all - and hence very unpredictable. As for the three red locations, you have exactly the same three degrees of uncertainty/unpredictability.

The uncertainty/unpredictability level of each location is fixed throughout.

You have to test a given location several times to discover something about its nature.

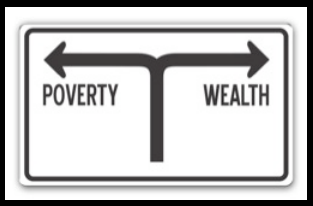
For example, imagine you've decided to visit location 2 (a blue one), which you select ten rounds in a row. It appears that you've obtained: 1 CHF, 1 CHF, 0 CHF, -1 CHF, 1 CHF, 1 CHF, 1 CHF, 1 CHF, 1 CHF, 1 CHF.

If you continue to select location 2 for a while, do you think you will receive 1 CHF more than half of the time, or less than half of the time?

CAVEAT! What makes this game challenging, and hopefully engaging, is that the locations change throughout play: there are unexpected changes in the chances to be in each scenario, which means that a hitherto-good location may well suddenly turn into a bad one, and this at any point in time!

For example, location 6 (a red one) has appeared to be a good location, delivering 2 CHF 75% of the time, and never giving -2 CHF. Then, unexpectedly, the location changes: now, 2 CHF never occurs and -2 CHF occurs 75% of the time!

You are not warned in advance when such changes occur.



When a change occurs for the red color, all the red locations change at the same time; when a change occurs for blue, all the blue locations change at the same time. However, the changes concerning the blue locations and the changes concerning the red ones occur independently.

We don't tell you how often the blue locations change; same thing for red.

However, keep in mind that the red locations are unstable (the blue ones, stable): changes for red are frequent compared to changes for blue.

Also, keep in mind that a change at the red locations may well happen even though you are currently visiting a blue location, and similarly blue locations may well change while you are visiting a red location.

In principle, a change may occur at each round, although it is very unlikely that this will happen.

The color of each location is fixed throughout.

Figure 3: **Rules of the Boardgame (cont'd)**

3 Analysis of the decision problem

If you were a Boardgame player, how would you play? The goal of the player is to maximize her accumulated earnings. If she knew the expected values of the locations, at any time she would pick the location with the highest expected value.[8] But these values are not known. At each trial, the player has to select one location on the basis of her estimates of the values. Optimal choice thus needs to trade-off the desire to exploit the location deemed best at a particular time, and the motive to explore the other ones, to get information about them. How to solve this trade-off between exploitation and exploration in the Boardgame?

3.1 Choice

If the locations were stationary, the player would maximize her cumulative earnings by selecting the location with the greatest *Gittins index* (the expected total future returns at a particular time) (Gittins and Jones, 1974).[9] In the Boardgame however, the locations are nonstationary, hence one has to forgo the Gittins index.

We formalized the exploitation/exploration trade-off as follows. Assume that the player uses a stationary stochastic policy π. Let $Q(l,T)$ denote the estimated value of location l after the Tth trial. π is a function from the vector Q to the probability vector $P^\pi = (P^\pi(l), l = 1, \ldots, 6)$, under the constraint that $\sum_{l=1}^{6} P^\pi(l) = 1$. To solve the trade-off between exploitation and exploration, we suggest maximizing w.r.t P^π the following criterion from *information theory*:

$$\sum_{l=1}^{6} Q(l,T)\ P^\pi(l,T) - \frac{1}{\beta}\sum_{l=1}^{6} P^\pi(l,T)\ lnP^\pi(l,T).$$

This criterion function weights the merits of exploitation against those of exploration: exploitation maximizes the estimated expected reward (the first term), whereas exploration is captured by the randomness (entropy) of P^π (the second term). The inverse of β captures the "willingness to explore" of the player, as the larger β, the smaller the weight of exploration (equivalently, the larger the weight of exploitation). We assume heterogeneity between subjects: β is subject-specific. The solution of the maximization of this criterion under the constraint that $\sum_{l=1}^{6} P^\pi(l,T) = 1$ is the logit rule, whereby the logit specification has a cognitive foundation here, aside from latent randomness.[10]

[8] Without loss, we assume risk-neutrality. Had risk-aversion/risk-loving entered our models, this would not have changed the result of our horse race between the Bayesian model(s) and the Reinforcement Learning model.

[9] This is true provided the player discounts exponentially the value of each reward over an infinite horizon of play.

[10] Usually, the usage of the logit rule to model stochastic choice is just part of the estimation of the utility models (McFadden, 1974). In the present study, β is a *fudge factor* as usual, but it also has a cognitive interpretation.

Logit decision rule.

$$\forall l = 1, \ldots, 6 \quad P^{\pi}(l,T) = \frac{\exp\left(\beta Q(l,T)\right)}{\sum_{l'=1}^{6} \exp\left(\beta Q(l',T)\right)},$$

where $Q(l,T)$ denotes the estimated value of location l after the Tth trial.

To examine to what extent the choice of the logit rule drove the result of the horse race, we also compared the predictions of the different models when choice was modeled with a purely *greedy* rule (β tends to ∞ in the criterion above), and also when choice was modeled with this purely greedy rule augmented with annealing (fixed noise is added, so that the resulting rule generates experimentation independent of valuation). It appeared that i) the fits were best with the logit rule,[11] and ii) the ranking of the models was invariant with the usage of a particular choice rule (more on this in the Results Section). Therefore, the difference in the fits came from the way the subjects processed information during the game.

3.2 Learning

We conjectured two types of learners, one fully-rational, and one bounded-rational.

Fully-rational players are sophisticated both in the way they set their beliefs, and in the way they update them. Firstly, they are *probabilistically sophisticated* (Machina and Schmeidler, 1992). This means their subjective probability of a state does not depend on the outcome they get in that state. They thus follow *Savage principle* and won't fall victim to a *Dutch Book* or, in the language of finance, they won't provide an arbitrage opportunity. Secondly, fully-rational players update their beliefs optimally. This means they can learn the six outcome probabilities through the most efficient information-processing method, Bayes rule.

It is likely that a less rational model using sets of priors – see, e.g., Gilboa and Schmeidler (1989), Ghirardato and Marinacci (2002), Epstein and Schneider (2003), Klibanoff et al. – instead of probabilistic beliefs would outperform our model. We by no means suggest our Bayesian model being the absolute truth. Our goal was to study whether people's behavior was close to being optimal in our task, and we precisely exploited the fact that our Bayesian formulation is behaviorally false – in the sense that it entails extreme levels of sophistication. Indeed, would our sophisticated model outperform the bounded rationality alternative, this would strongly suggest that our subjects were close to learning optimally during the game. However, would we find the reverse (i.e., the bounded rationality hypothesis to be more plausible), we would do the horse race again, this time between a multiple-priors model, which assumes a lesser degree of sophistication, and our reinforcement learning model. That is,

[11] The fit of each model was better under the logit rule, even after penalizing the latter for having one additional degree of freedom (β is a free parameter in the estimation).

we would suppress the hypothesis of probabilistic sophistication, while keeping the second chief aspect of rationality, which relates to the way people update their beliefs.

In our complicated task, Bayesian updating falls into two categories. The first, the Hierarchical Bayes model – henceforth, HB model – is the optimal learning protocol. It uses the Markovian structure of the game (more on this below) to learn all the unknown parameters, outcome and jump probabilities. The second, called the Forgetting Bayes model – henceforth FB model – can accommodate its learning rate to the strength of evidence in favor of a jump at each trial. As such, it is tractable and excellent at learning outcome probability. Contrary to the HB model, the FB model does not make any assumption about the nature of the jumps.[12] As such, the FB model describes rational updating when one cannot reasonably process all available information on how parameters vary.[13] In our complicated task, it is thus an excellent method of probabilistic learning, very close to the optimal HB approach.

In contrast to the fully-rational (HB or FB) players, bounded-rational players try to predict the absolute occurrence of incoming outcomes, rather than their probability. That is, such players merely backward-forecast next outcome. We call such learning by adaptive expectations "reinforcement learning," to point to its solid cognitive foundations (more on this below).[14] Together with the logit decision rule, reinforcement learning entails a *win-stay lose-switch* heuristic, also referred to as *Matching Law* – see, e.g., Herrnstein (1970) and Duffy (2006). It prescribes that the locations that yielded good outcomes in the past should be visited more often in the future.

We are now more explicit about the nature of Bayesian updating in our task, then contrast it with the reinforcement learning approach. We start with the normative model of learning in the Boardgame, the Hierarchical Bayes method.

4 Competing models of learning in the Boardgame

4.1 Hierarchical Bayes model

4.1.1 Hidden Markov model of the environment

HB players have in mind the true Markovian structure of the game, so they can track outcome probability without having to store the entire history of

[12] The FB model employs a forgetting operator as an alternative for the proper transition-probability operator.

[13] Before each experimental play, we checked in the lab through an MCQ questionnaire that the subjects had well understood the stochastic structure of the game – the nature of the jump processes, the different nature of the three locations for each color, etc. Nevertheless, it could be, arguably, that the subjects could not make use of all these sources of information during the game.

[14] Our use of the label "reinforcement learning" is also to stress that the forecasted values derived from the non-Bayesian heuristic are not probabilistic – i.e., they are not proper "expectations."

estimated probabilities and outcomes. To clarify what it means for the task to be Markovian, we need some notation. Each location $l = 1 \ldots 6$ can be characterized as follows.

- Let Θ denote the "3-simplex."[15] $\mathbf{p_{lt}} = (p_{l\,1\,t}, p_{l\,2\,t}, p_{l\,3\,t}) \in \Theta$ is the probability vector (triplet) for location l at time t. Each location l is multinomial $\tilde{\mathbf{r}}_{\mathbf{lt}} \sim \text{Multi}(\mathbf{p_{lt}})$. There are three possible outcomes: for l blue, $r_{l1} = -1$ CHF, $r_{l2} = 0$ CHF, $r_{l3} = 1$ CHF; for l red, $r_{l1} = -2$ CHF, $r_{l2} = 0$ CHF, $r_{l3} = 2$ CHF. To infer the hidden two-dimensional[16] probability parameter, it is thus unnecessary to explicitly consider states of the remote past, because the outcome returned at time t is independent of past outcomes, and depends only on the current probability triplet $\mathbf{p_{lt}}$.

- The transition from $\mathbf{p_{lt}}$ to $\mathbf{p_{lt+1}}$ is controlled by a Bernoulli jump process. There are two independent jump processes, one for the red locations, $\widetilde{J_{\text{red}}} \sim Bern(\alpha_{\text{red}})$, and one for the blue ones, $\widetilde{J_{\text{blue}}} \sim Bern(\alpha_{\text{blue}})$. $J_{\text{red}\,t}$ is equal to 1 when a jump occurred for red at time t (in which case all the red locations change at time t), and 0 otherwise. $J_{\text{blue}\,t}$ is equal to 1 when a jump occurred for blue at time t (in which case all the blue locations change at time t), and 0 otherwise. While the values of α_{red} and α_{blue} are unknown, it is known that $\alpha_{\text{red}} > \alpha_{\text{blue}}$. The prior distribution of α_{red}, denoted by $f_0(\alpha_{\text{red}})$, is the uniform distribution within the interval $[1/5, 1/2]$; the one of α_{blue}, $f_0(\alpha_{\text{blue}})$, the uniform distribution within the interval $[0, 1/5]$.[17]

- The evolution of the probabilities is as follows. Assume without loss that l is a red location.[18] The changeability of the probability vector is represented by the transition probability distribution $P_l(\mathbf{p_{lt}}|\mathbf{p_{lt-1}})$. Let $\delta_{\mathbf{p_{lt-1}}}$ denote point mass at $\mathbf{p_{lt-1}}$. For the moment, take for granted $P_{0t}(\mathbf{p_{lt}}|\mathbf{p_{lt-1}})$, the probability distribution at location l after a jump at time t, given $\mathbf{p_{lt-1}}$. Note that

  $$P_l(\mathbf{p_{lt}}|\mathbf{p_{lt-1}}, \alpha_{\text{red}}, J_{\text{red}\,t}) = \delta_{\mathbf{p_{lt-1}}}(\mathbf{p_{lt}})$$
  $$\text{unless } J_{\text{red}\,t} = 1, \text{ in which case}$$
  $$P_l(\mathbf{p_{lt}}|\mathbf{p_{lt-1}}, \alpha_{\text{red}}, J_{\text{red}\,t}) = P_{0t}(\mathbf{p_{lt}}|\mathbf{p_{lt-1}}).$$

  Further note that $P(J_{\text{red}\,t} = 1) = \alpha_{\text{red}}$ and $P(J_{\text{red}\,t} = 0) = 1 - \alpha_{\text{red}}$. So, the transition distribution of the probability is

  $$P_l(\mathbf{p_{lt}}|\mathbf{p_{lt-1}}, \alpha_{\text{red}}) = (1 - \alpha_{\text{red}})\delta_{\mathbf{p_{lt-1}}}(\mathbf{p_{lt}}) + \alpha_{\text{red}} P_{0t}(\mathbf{p_{lt}}|\mathbf{p_{lt-1}}).$$

At any location, P_{01} is the uniform distribution on Θ; for t strictly greater than 1, $P_{0t}(\mathbf{p_{lt}}|\mathbf{p_{lt-1}})$ is a uniform distribution that is centered around a triplet,

[15] $\Theta = \left\{ \mathbf{p} \;\middle|\; p_i \geq 0, i = 1 \ldots 3, \; \sum_{i=1}^{3} p_i = 1 \right\}$.

[16] $p_1 + p_2 + p_3$ is equal to 1, so knowing two components is tantamount to knowing $\mathbf{p}$.

[17] We should write f_{red} and f_{blue} to refer to the respective distributions (since they differ). For notational convenience we don't.

[18] In what follows, for l blue, replace α_{red} by α_{blue} and J_{red} by J_{blue}.

$\mathbf{perm}(\mathbf{p_{lt-1}})$, which represents all the possible new triplets after p_{lt-1} has jumped. The formal definition of $P_{0t}(\mathbf{p_{lt}}|\mathbf{p_{lt-1}})$ is delegated to the Appendix, because it is rather involved. Here we give a heuristic definition, based on one example. Suppose that $\mathbf{p_{lt-1}}$ is $(0.8, 0.2, 0)$. What does it mean for $\mathbf{p_{lt-1}}$ to jump? By design, it means that its first component swaps either with the middle one, or with the third one.[19] So starting from $(0.8, 0.2, 0)$, the possible permuted triplets are

$(0.2, 0.8, 0)$, $(0, 0.8, 0.2)$, $(0.2, 0, 0.8)$, $(0, 0.2, 0.8)$.

$\mathbf{perm}((0.8, 0.2, 0))$, the average permuted triplet, is

$$1/4(0.2, 0.8, 0) + 1/4(0, 0.8, 0.2) + 1/4(0.2, 0, 0.8) + 1/4(0, 0.2, 0.8).$$

So $\mathbf{perm}((0.8, 0.2, 0)) = (0.1, 0.45, 0.45)$ and $P_{0t}(\mathbf{p_{lt}}|(0.8, 0.2, 0))$ is a two-dimensional uniform distribution that is centered around the first two[20] components of (0.1,0.45,0.45):

$$\mathrm{U}\left([0.1 - 0.1; 0.1 + 0.1] \times [0.45 - 0.1; 0.45 + 0.1] \right).$$

Fig. 4 provides a diagram of the Markovian structure of the game.

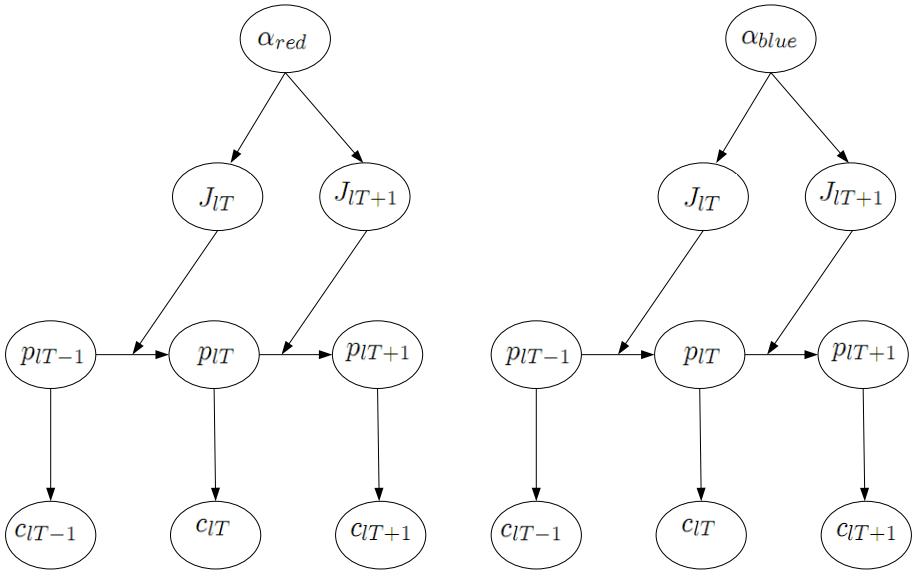


Figure 4: **Diagram of the underlying structure of the Boardgame, for a red location (left) and for a blue location (right).**

[19] These are the only admissible swaps, be design. The other ones would not constitute relevant jumps because the experienced outcomes after such jumps are essentially the same as before the jump. To see this, consider in the foregoing example a jump leading to a swap between probability 0.2 and probability 0, whereby the new underlying probability for location l is $(0.8, 0, 0.2)$. When sampling several times in a raw location l just after such a jump, it still appears that on average, the bad scenario crops up almost all the time and that the two other ones almost never obtain, exactly as before the jump.

[20] Since $p_1 + p_2 + p_3$ is equal to 1, once two components are specified the remaining one is known.

As we show now, the Markovian structure of the task allows to learn both outcome and jump probabilities. We break the solution process into steps to suggest how it works in a relatively transparent way. Full derivation of the solution, along with proofs, are in the Appendix.

4.1.2 Algorithm

Let $\delta_{r_{li}}$ denote point mass at r_{li}. Time t observation is the count vector $\mathbf{c_{lt}} = (c_{lit}, i = 1 \dots 3)$, where $c_{lit} = \delta_{r_{li}}(r_{lit})$. $\underline{c}_{\mathbf{lT}}$ denotes the data for location l at time T:

$\underline{c}_{\mathbf{lT}} = (\mathbf{c_{lt}}, t \in \Delta_l(T))$, with $\Delta_l(T) = \{\, t |\ l \textit{ is visited at time } t, t \leq T\}$.

For any location l, the likelihood function at trial T is, using the Markov structure of the environment:

$$l(\underline{c}_{\mathbf{lT}} | \underline{p}_{\mathbf{lT}}) = \prod_{t \in \Delta_l(T)} \prod_{i=1}^{3} p_{lit}^{c_{lit}} = \exp\left(\sum_{i=1}^{3} \sum_{t \in \Delta_l(T)} c_{lit} \ln p_{lit} \right). \tag{1}$$

Goal At any time T and for each location l, the goal is to compute the value $Q(l, T)$, which in a Bayesian framework is the expected outcome at the Tth trial. As such, the posterior probability distribution $P_{lT}(\mathbf{p_{lT}})$ – with which to get the posterior mean of $\mathbf{p_{lT}}$ – needs to be computed. The first step is to compute the posterior distribution of the jump parameter α_{red},[21] denoted $f_T(\alpha_{\text{red}}) \equiv P(\alpha_{\text{red}} | \underline{c}_{\mathbf{lT}})$.

Posterior distribution of the jump parameter For expositional clarity, throughout in this part we will skip the index l that refers to a specific location, although the learning process is location-specific. So $P_T(\mathbf{p_T})$, $\underline{c}_{\mathbf{T}}$ and α stand for $P_{lT}(\mathbf{p_{lT}})$, $\underline{c}_{\mathbf{lT}}$, and α_{red},[22] respectively, with $\underline{c}_{\mathbf{T}} = (\mathbf{c_1}, \mathbf{c_2}, \dots, \mathbf{c_T})$.

At time T , the posterior distribution of the jump parameter is, by definition,

$$f_T(\alpha) = \int_\Theta P(\mathbf{p_T}, J_T = 1, \alpha | \underline{c}_{\mathbf{T}})\, d\mathbf{p_T} + \int_\Theta P(\mathbf{p_T}, J_T = 0, \alpha | \underline{c}_{\mathbf{T}})\, d\mathbf{p_T}. \tag{2}$$

The calculation of the joint likelihoods $P(\mathbf{p_T}, J_T = 1, \alpha | \underline{c}_{\mathbf{T}})$ and $P(\mathbf{p_T}, J_T = 0, \alpha | \underline{c}_{\mathbf{T}})$ involves multidimensional integrals. We used the Markovian structure of the task to derive a relatively tractable recursive form, with which to compute the posterior probability of the jump parameter at each trial:

$$f_T(\alpha) = f_{T-1}(\alpha) \int_\Theta l(\mathbf{c_T} | \mathbf{p_T}) P(\mathbf{p_T} | \underline{c}_{\mathbf{T-1}}, \alpha) \left[\alpha P_{0T}(\mathbf{p_T} | \mathbf{p_{T-1}}^*) + (1 - \alpha) \right] d\mathbf{p_T},$$

where $\mathbf{p_{T-1}}^*$ is the mode of the posterior probability distribution P_{T-1}. See the Appendix for a proof. After deriving a tractable recursive equation on

[21] α_{blue} if l is blue.

[22] α_{blue} if l is blue.

$P(\mathbf{p_T}|\underline{c}_{\mathbf{T-1}}, \alpha)$ (details are given in the Appendix), we used a three-dimensional meshgrid to assess $f_T(\alpha)$ as a Riemann sum with steps of 0.01 on each coordinate p_1, p_2, and α.

Posterior probability distribution Armed with the posterior distribution of the jump parameter, the HB learner can infer the posterior probability distribution. Marginalization over all the possible values of the jump parameter[23] generates the posterior distribution of p_T, the hidden probability parameter:

$$P_T(\mathbf{p_T}) = \int_0^1 P(\mathbf{p_T}|\underline{c}_{\mathbf{T}}, \alpha)\ f_T(\alpha)\, d\alpha.$$

We used the aforementioned three-dimensional grid to compute $P_T(\mathbf{p_T})$ as a Riemann sum with steps of 0.01 on the coordinate α.

Posterior mean probability The HB player then estimates the probability by computing the mean of $P_T(\mathbf{p_T})$:

$$\bar{\mathbf{p}}_{\mathbf{T}} = \int_\Theta \mathbf{p_T} P_T(\mathbf{p_T})\, d\mathbf{p_T}. \tag{3}$$

We used a two-dimensional grid to assess the integral in Equation (3) as a Riemann sum with steps of 0.01 on every single coordinate p_1 and p_2.

4.1.3 Learning by the HB model

Learning of the probabilities We did simulations to test the quality of learning by the HB algorithm. Each simulation was of 500 trials in length. The underlying stochastic structure was identical to that in the experimental sessions.[24] We compared the *learned* outcome probability (see Equation (3), p. 17) with the *true* outcome probability. We first ignored the decision aspect and studied the learning of the probabilities by the HB algorithm, when the player was forced to stay at the same location throughout the game. In Fig. 5 (p. 18), the top graph (bottom graph) shows the estimated probability of the loss outcome[25] at the minimal-entropy blue (red) location for each trial in one simulation. These graphs suggest that the HB algorithm is very good at learning the underlying outcome probabilities – convergence of the estimated probability to the true probability is quick. Nonetheless, it appears on the bottom graph that the FB player did not have sufficient time to fully learn the probability at the red location, which jumped very often.

[23] An alternative route is to derive from $f_T(\alpha)$ the posterior mean of α (denoted $\bar{\alpha}$), from which the posterior probability distribution is obtained as $P_T(\mathbf{p_T}) = P(\mathbf{p_T}|\underline{c}_{\mathbf{T-1}}, \bar{\alpha})$. As pointed to in MacKay (2003), marginalization over all the possible values of α leads to more robust estimates.

[24] The low entropy level is 0.3, the median one, 0.65, and the high one, 1.1; jump frequency is 1/16 at the blue locations, 1/4 at the red locations.

[25] The counterparts of Fig. 5 for the probabilities to be in the two other states (win and no change) are similar.

The top graph of Fig. 6 (p. 19) shows the probability of the reward outcome[26] at the minimal-entropy blue location, as estimated by the HB model. The bottom graph is the counterpart for the minimal-entropy red location. Remember that the HB model is a HB player who, at each trial, learns the underlying probabilities of the locations through the HB algorithm, then feeds the expected values into the logit rule, to select one location. The β coefficient of our simulated HB player is the average maximum likelihood estimate we got from the subjects (more on this in the next Section). On this graph, the estimated probability is stable and converges well at the blue location. By contrast, it is erratic at the red location, which reflects the high jump frequency there.

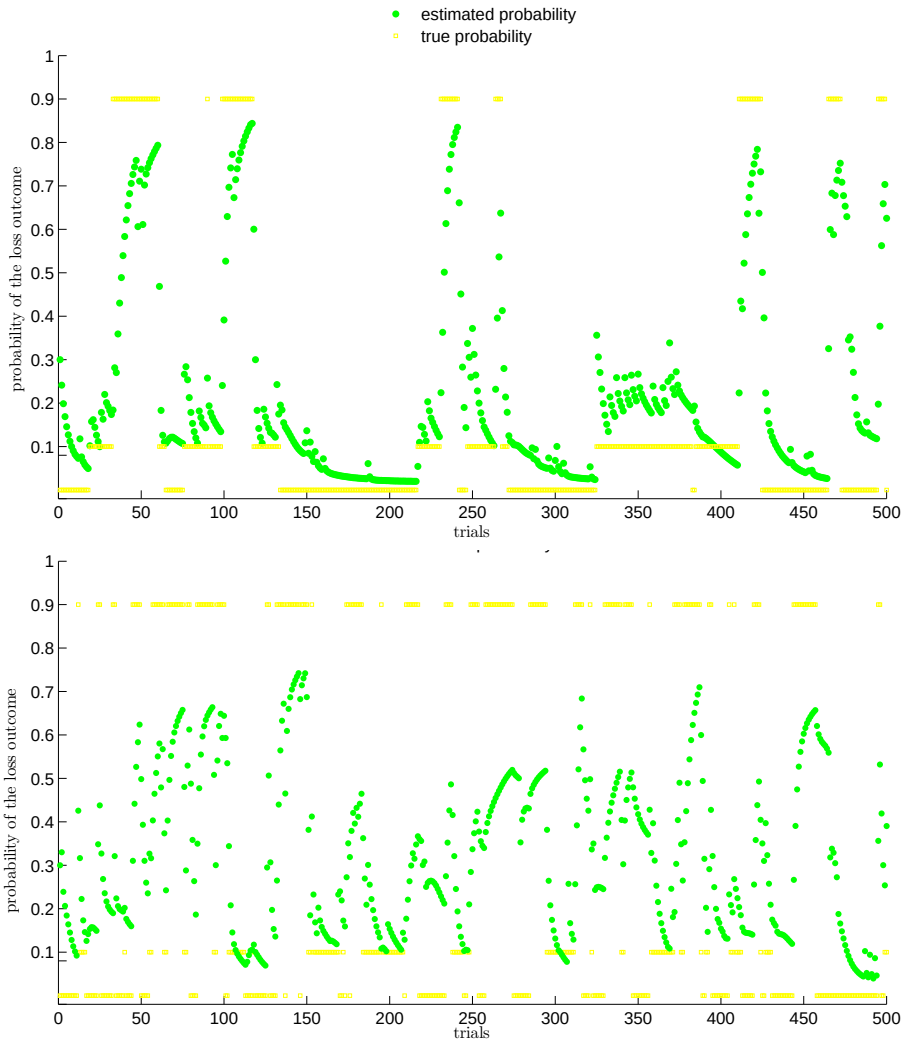


Figure 5: **The top (bottom) graph shows the learning by a simulated player of the probability of the loss outcome at the minimal-entropy blue (red) location. The player was forced to stay at this location throughout the game. He learnt the probabilities using the HB algorithm at each trial.**

[26] The counterparts of Fig. 6 for the probabilities to be in the two other states (lose and no change) are similar.

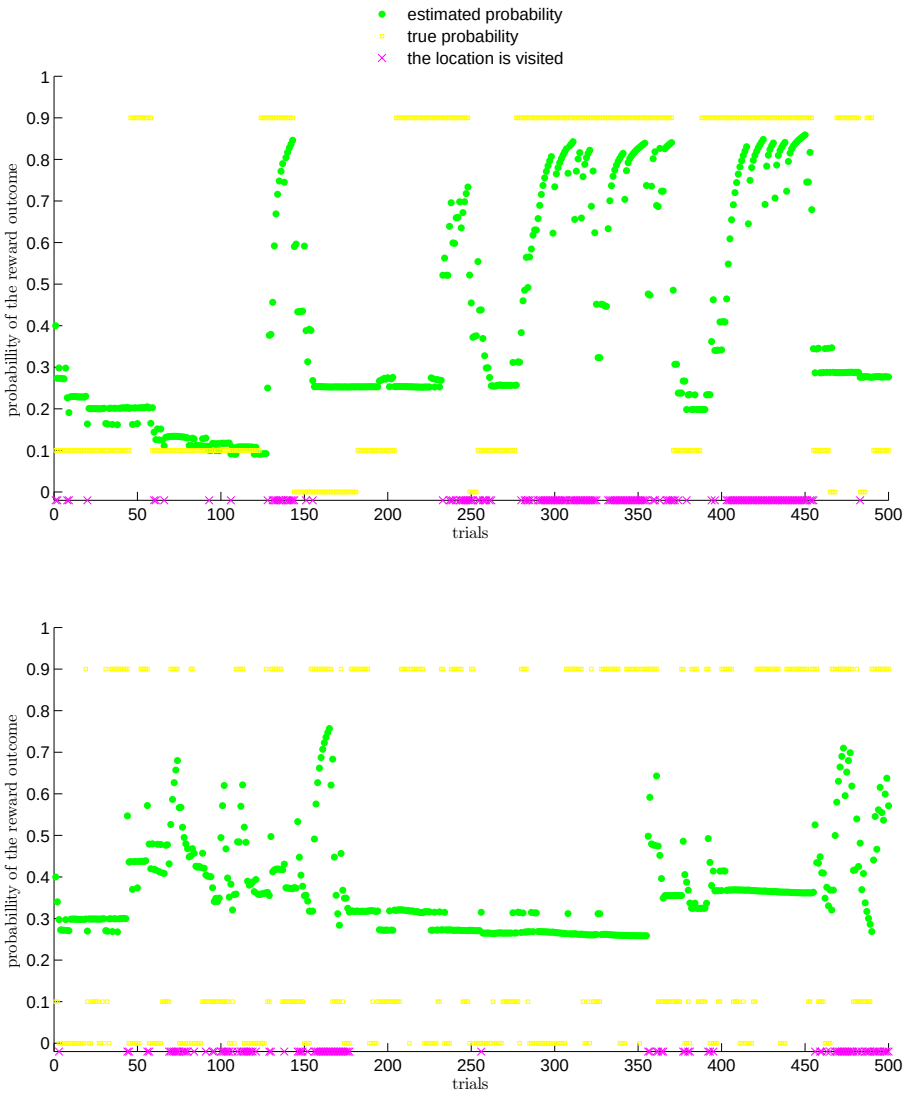


Figure 6: **Learning by the HB model of the reward probability at the minimal-entropy blue location (top graph) and at the low entropy red location (bottom graph), at each trial of one simulated play.**

Learning of the jumps We further studied how well the HB model learns the jump intensities. At time T, the estimated jump frequency is

$$\bar{\alpha}_T = \int_0^1 \alpha f_T(\alpha)\, d\alpha.$$

Fig. 7 shows the jump frequencies (for red locations and blue locations) as estimated by the HB model at each trial in one simulation. The jump frequency estimates appear to be inconsistent.[27] An explanation of such inconsistency is that a Bayesian may not want to learn the truth in the Boardgame, owing to the aforementioned trade-off between exploitation and exploration. Rothschild (1974) showed that even in a *stationary* multi-armed bandit problem, optimal behavior leads to incomplete learning. In our *nonstationary* multi-armed bandit task, it may be even more so that a full learning is uncalled for. From this perspective, it could be, arguably, that learning explicitly the two jump intensities merely adds an extra layer of complication. We now present a method that allows one to estimate outcome probability rationally – as the HB algorithm does – while avoiding to learn a full posterior probability distribution of the jump intensities.

4.2 Forgetting Bayes Learning

Now to the presentation of the FB approach. A statistical justification of its usage, as well as formal derivations of all the forms we introduce below, are in the Appendix. Here we sketch the FB method heuristically, and try to explain intuitively how the ensuing FB algorithm works. Our goal is twofold. Firstly, we show that FB players fully account for the jumps. Like HB players, FB players learn the nonstationary outcome probabilities, They solve the same changepoint problem differently. While HB players learn the distribution of the two jump frequencies, FB players do not. They detect jumps "on the spot," and such jump detection drives their learning rate at each point in time. Secondly, we want to stress that the FB algorithm is tractable and particularly well adapted to our highly nonstationary task.

4.2.1 Updating rule

The FB algorithm is a natural sampling scheme under Dirichlet prior. Remember that FB players are probabilistically sophisticated – like HB players. We take their prior probability distribution, denoted by P_0,[28] to be Dirichlet with center $\hat{\mathbf{p}}_\mathbf{0}$ and precision $\nu_\mathbf{0} = (\nu_0, \nu_0, \nu_0)$:

[27] On the bottom graph of Fig. 7, which shows the estimated jump frequency for the red locations, $f_T(\alpha_{red})$ does not tend to 1 for every neighborhood of the true value of α_{red} (1/4), even after 400 trials. The counterpart for the blue locations (top graph) hints at a possible (slow) convergence to the true value of α_{blue} though.

[28] P_{l0}, the prior probability distribution at location l, is the same for all the locations, so $P_{l0} \equiv P_0$.

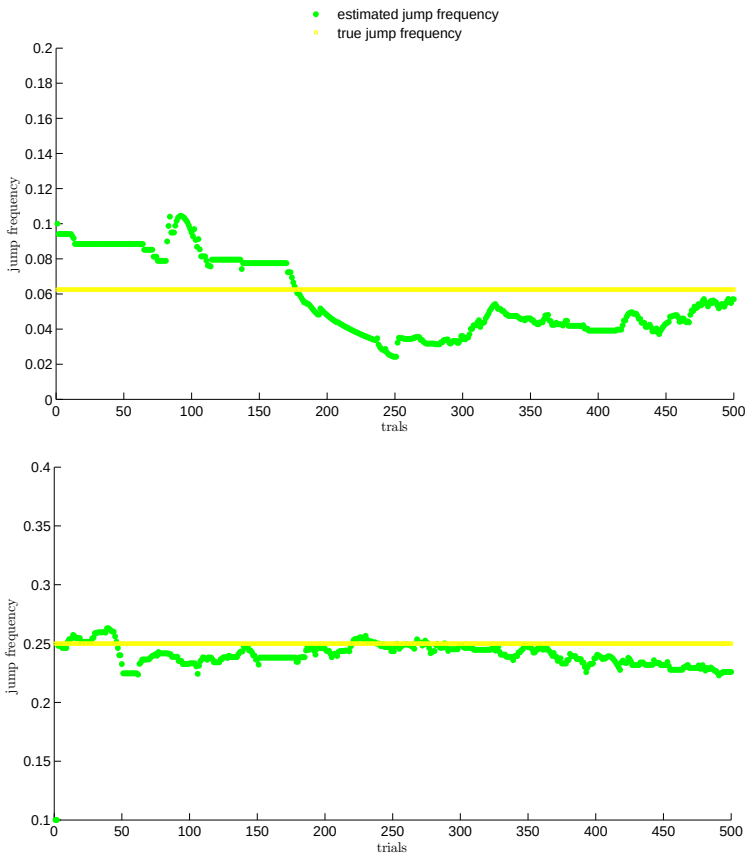


Figure 7: **Learning by the HB model of α_{blue} (top) and α_{red} (bottom), at each trial of one simulation.**

$$P_0(\mathbf{p}) = \left[\frac{\prod_{i=1}^{3}\Gamma(\nu_0\hat{p}_{i0})}{\Gamma(\nu_0)}\right]^{-1} \prod_{i=1}^{3} p_i^{(\nu_0\hat{p}_{i0}-1)} \quad \delta_\Theta(\mathbf{p}), \tag{4}$$

with $\mathbf{p} = (p_1, p_2, p_3)'$,

$$\Gamma(\nu_0\hat{p}_{i0}) = \int_\Theta x^{\nu_0\hat{p}_{i0}-1} \ \mathrm{e}^{-x}\, dx.$$

The center $\hat{\mathbf{p}}_\mathbf{0}$ is the base measure: $\hat{\mathbf{p}}_\mathbf{0} = E_{\mathrm{Dir}(\hat{\mathbf{p}}_\mathbf{0},\nu_\mathbf{0})}[\mathbf{p}]$. We formalized absence of prior knowledge related to outcome probability by setting $\hat{p}_{i0} = 1/3$, for $i = 1 \ldots 3$. The precision parameter $\nu_\mathbf{0}$ controls the extent to which the probability mass is localized around the center $\hat{\mathbf{p}}_\mathbf{0}$. We set it equal to $(1, 1, 1)$. Intuitively, $\nu_0\hat{p}_{i0}$ is tantamount to the "prior observation counts" for outcome i, thereby measuring (in units of i.i.d samples) the weight of the prior in the inference.

The FB method accounts for the occurrence of jumps in the probabilities, albeit it avoids learning explicitly the two jump frequency parameters. Specifically, at any point in time T, FB learners ask themselves whether a jump has just occurred. Using the available evidence in favor of a jump, they set their subjective probability that a jump has not occurred, and then accommodate the sample size to the strength of this belief, that we denote $\lambda(T)$. The formal computation of $\lambda(T)$ is in the Appendix. If a FB player is strongly convinced that a jump has occurred ($\lambda(T)$ close to 0), he restarts estimation from the prior P_0 i.e., forgets most information gathered from previous observations. Conversely, if he is strongly convinced that no jump occurred ($\lambda(T)$ close to 1), he keeps the latest posterior probability distribution (PPD) of outcome probability, and updates it, using Bayes rule, after observing the new outcome obtained at time T. Mixed evidence in favor of a jump ($\lambda(T)$ lies somewhere between 0 and 1) leads to some degree of discounting of the past data, whereby the ensuing PPD is a geometric mean between the latest PPD and the prior P_0.

The return to the prior, after a jump has been detected, may seem like an inefficient way to process information: after a location has jumped, should not FB players retain in memory what they have learnt about the entropy, since the latter is known to be fixed? Contrary to the HB model, the FB model ignores the entropy dimension. As we show below, the FB model is capable of learning the nonstationary probabilities very well precisely because it does not try to "do too much" and merely focuses on jump detection. The return to the prior is stabilizing and helps learning outcome probability in the face of the numerous jumps.

For large T, at any location l, the PPD of the unknown triplet at time T is well approximated by

$$
\begin{aligned}
&\mathbf{p_{l\,T}} \sim \mathrm{Dir}(\hat{\mathbf{p}}_{\mathbf{l\,T}}, \nu_{\mathbf{lT}}),\\
&\hat{p}_{i\,l\,T} = \frac{1}{\nu_{lT}}\left[\nu_0\,\hat{p}_{i\,0} + N^{\lambda}{}_l(T)\text{«}c_{li}(\mathrm{T})\text{»}\right],\\
&\nu_{lT} = \nu_0 + N^{\lambda}{}_l(T),
\end{aligned}
$$

$$
\text{where } \text{«}c_{li}(\mathrm{T})\text{»} = \frac{\sum_{t\in\Delta_l(T)}\left(\prod_{s=t}^{T}\lambda(s)\right)c_{lit}}{N^{\lambda}{}_l(T)},
$$

$$
N^{\lambda}{}_l(T) = \sum_{t\in\Delta_l(T)}\prod_{s=t}^{T}\lambda(s).
$$

The computation of the PPD is readily obtained, because «c_{li}(T)» is a *sufficient statistic* for $\hat{p}_{i\,l\,T}$. A statistic is Bayes sufficient if, for any prior, the posterior distribution only depends on the data through it. In other words, FB players only have to keep track of «c_{li}(T)» . Then, as the previous expression suggests, they take $N^{\lambda}{}_l(T)$, the effective number of data, to be the weight given to the sufficient statistic in the inference. We derived recursive forms (see the Appendix for a proof), whereby «c_{li}(T)» – and hence, the PPD – is readily computed at each trial through the following updating rule:

$$
\text{«}c_{li}(\mathrm{T})\text{»} = \text{«}c_{li}(\mathrm{T\text{-}1})\text{»}\,(1-\eta_l(T)) + \eta_l(T)c_{liT},
$$

where the learning rate $\eta_l(T)$, formally defined as the inverse of $N^{\lambda}{}_l(T)$, is computed recursively too:

$$
\eta_l(T) = \frac{1}{1+\frac{\lambda(T)}{\eta_l(T-1)}}.
$$

The forgetting mechanism is at the heart of the inference of the PPD. Such forgetting effect is twofold. First, as a location l is not visited, $N^{\lambda}{}_l(T)$ decreases (equivalently, $\eta_l(T)$ increases). So, in the inference of the probability for this location, more weight is put on the prior and less on the likelihood. This is to account for the possibility that jumps occurred since the last visit at the location. Second, after a jump has been detected (λ is minimal), the effective number of data drops sharply (equivalently, the learning rate soars), so that estimation restarts from the prior. These two effects are apparent in Fig. 8, which shows the effective number of data in the inference of outcome probability at the minimal-entropy red location, at each of the first 100 trials of one simulation with the FB model. From trial 19 to trial 27, the increase of the sample size is concave. In a stationary environment, the increase would be linear. This concavity is caused by the fact that the belief that a jump has just occurred is rarely null, whereby the FB algorithm is going to discount somehow the past data at each trial. Now consider the subinterval from trial 53 to trial 83: the effective sample size, starting from a value of five observations, is quickly reduced to zero as the location is not visited any more. To see how

jump detection drives the effective sample size, consider trial 27 on Fig. 8, p. 24. The drop of the effective sample size is not caused by the FB player's selecting a different location after trial 27,[29] but to $\lambda(27)$'s being minimal: the FB player detected a jump at trial 27. Accounting for the jumps is thus the essence of probabilistic learning by the FB algorithm.

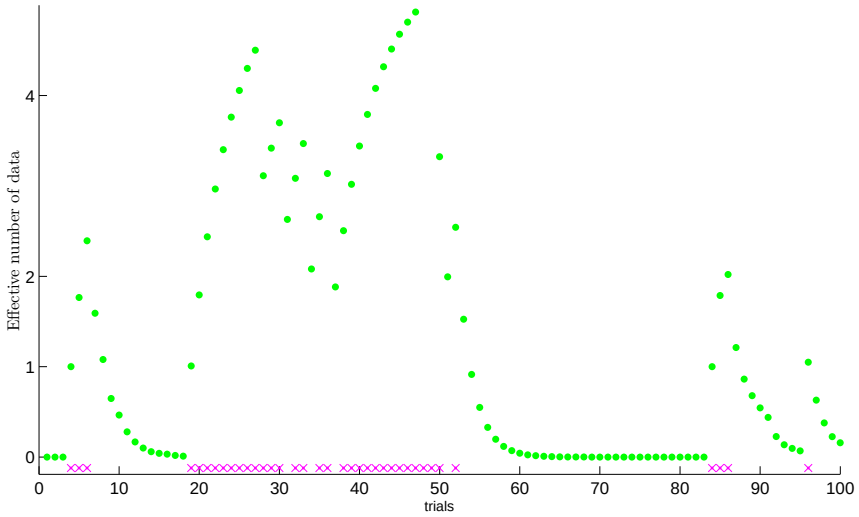


Figure 8: **Effective sample size in the inference of outcome probability of the minimal-entropy red location, at each one of** 100 **trials during one simulated play** (500 **trials) by the FB model. Crosses at the bottom indicate that the low entropy red location was visited at the corresponding trials.**

[29] Indeed, the cross at trial 27 indicates that the location under study is visited at trial 27.

4.2.2 Estimated probability

As the HB learner, the FB learner derives from the PPD the estimated outcome probability, from which he directly gets $Q(l,T)$, the expected value of location l at trial T (for $l = 1 \ldots 6$). The estimated outcome probability is

$$\bar{p}_{l\,i\,T} = \frac{N^{\lambda}{}_{l}(T)\text{«}c_{li}(\mathrm{T})\text{»} + 1/3}{N^{\lambda}{}_{l}(T) + 1}. \tag{5}$$

See the Appendix for a proof.

4.2.3 Learning performance of the FB algorithm

As in the study of the HB model, we performed simulations of 500 trials in length to compare the *learned* outcome probability (see Equation (5), p. 25) with the true probability.

Fig. 9 shows the learned probability of the loss outcome by a simulated FB player who was forced to stay at the same location (the minimal-entropy location) throughout the game. The learned outcome probability converges to the true value at the blue location (on the top graph, see, e.g., the value of the learned probability for the first 100 trials, and from trial 275 to trial 350). Convergence is also observed at the red location (on the bottom graph consider, e.g., the period from trial 90 till trial 120, approximately) despite its large jump frequency. Such probabilistic learning is facilitated because our FB algorithm focuses on jump detection. The quality of jump detection is apparent. For instance, note the quick and accurate adjustment to a jump in probability at about time 90 and time 350 in the top graph.

Fig. 10 shows the learning of the probability of the reward outcome by the FB model. As in the simulated data with the HB model, the value of β (the free parameter of the logit rule) is the maximum likelihood estimate we got from the subjects (more on this in the next Section). Convergence to the true value 0.9 is often observed, even at the red location (see, e.g., the estimated probabilities at times 80, 200 and 350 in the bottom graph). Notice that convergence to low probabilities (0; 0.1) never obtains: when the location is "bad" (which is what the low probabilities mean), the FB algorithm leaves it, learning stops, and beliefs return to the prior $[1/3, 1/3, 1/3]$. Such a return to the prior, which reflects the forgetting effect (see, e.g., around trial 120 on the top graph, around trial 360 on the bottom graph), helps stabilize the learning of outcome probability.[30]

It thus appears in these simulated data that the FB model is very good at learning outcome probability. As hinted above, one important characteristic

[30] We also tested an augmented FB model. The latter is a FB model that learns the entropies explicitly, whereby it does not return to the uninformative prior P_0 after a jump has been detected at a location. Rather, it returns to a prior having the estimated level of entropy of the location. The quality of learning of this model was not better. On the contrary, it appeared to be less stable than the FB model. (This may be due the fact that if the estimated entropies are far from their true level, one would better return to the uninformative prior P_0.)

of the FB model is that it processes information rationally (i.e., according to Bayes rule), while deliberately avoiding to process *all* available information. We suggest such a selective use of the available information being particularly adaptive in our complicated task, where a full learning may not be optimal, as already noted, with reference to Rothschild (1974). Therefore, while some may argue that after all, the FB model is a heuristic "in between" the sophisticated HB model and the crude RL model, we argue the opposite. The FB model falls into the full-rationality hypothesis, because our view on rationality has to do with how people process the information they consider using – whence the suggested dichotomy between Bayesian learning vs reinforcement learning. However, we are agnostic about the optimal way to select the relevant information in a complicated task like ours. However, the high quality of probabilistic learning by the FB model suggests that the FB model does this selection optimally. We now present the alternative way to learn in the Boardgame, through simple reinforcement learning.

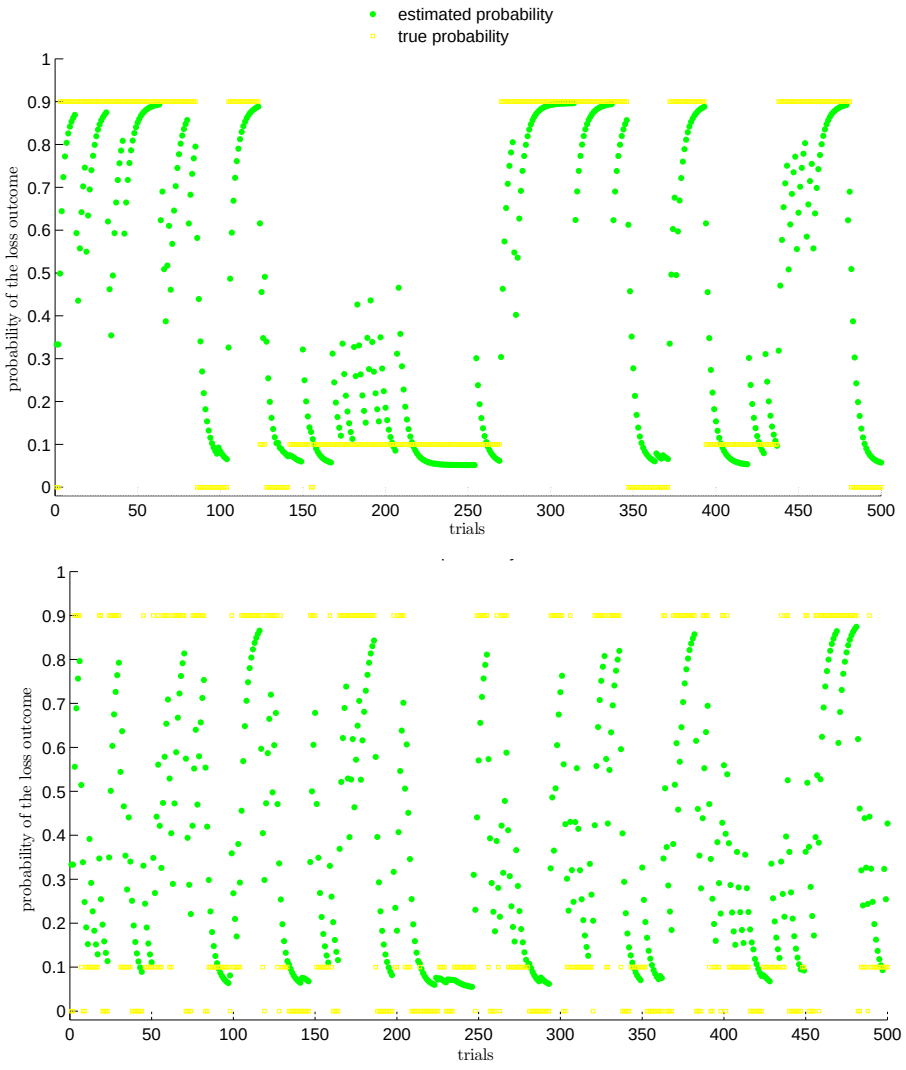


Figure 9: **The top (bottom) graph shows the learning by a simulated FB player of the probability of the loss outcome at the minimal-entropy blue (red) location at each trial. The player was forced to stay at this location throughout play. He learnt the probabilities using the FB algorithm.**

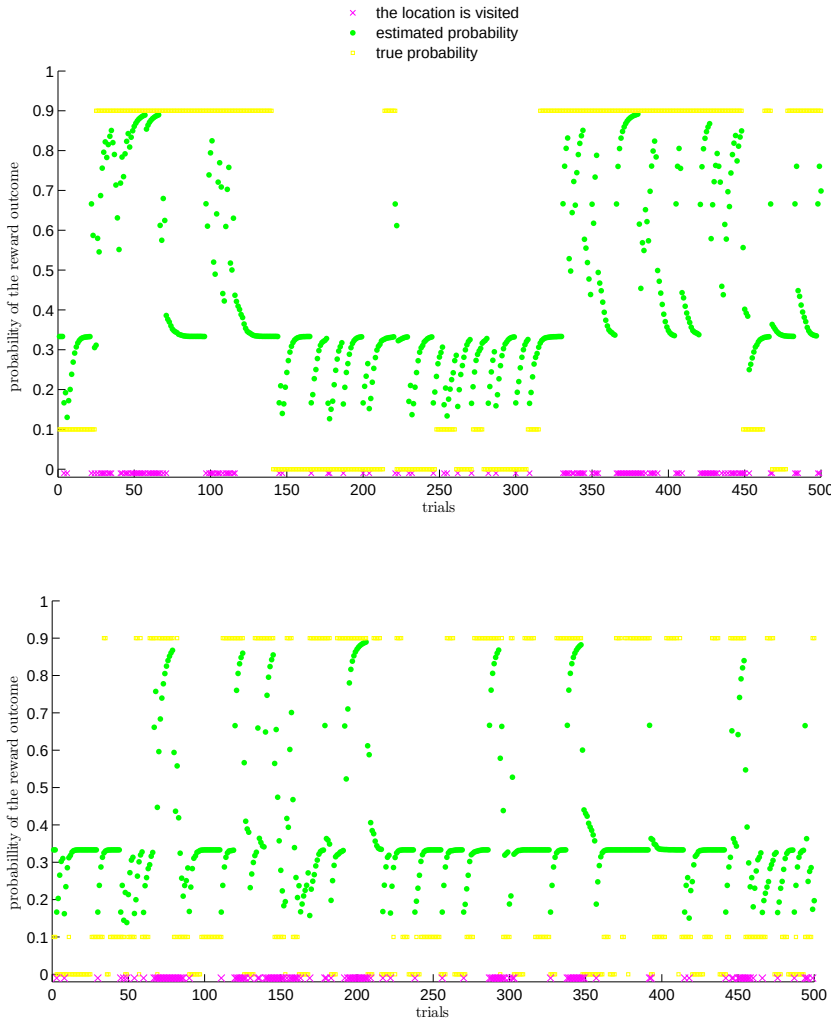


Figure 10: **Learning by the FB model of the probability of the reward outcome at the minimal-entropy blue location (top graph) and at the minimal-entropy red location (bottom graph), at each trial of one simulated play.**

4.3 Reinforcement Learning

4.3.1 The RL model

For each location $l = 1, \ldots, 6$, from an initial value denoted $Q(l, 0)$ (that we set equal to 0), the RL model updates at each trial T the six estimated values according to an error-correction principle:

- If l is sampled at trial T,

$$\begin{cases} Q(l, T+1) = Q(l, T) + \eta_{\text{blue}}\, \delta(T) & \text{if } l \text{ is blue,} \\ \\ Q(l, T+1) = Q(l, T) + \eta_{\text{red}}\, \delta(T) & \text{if } l \text{ is red,} \end{cases} \tag{6}$$

where $\delta(T) = r_{lT} - Q(l, T)$ is the *prediction error* in trial T.

- If l is not visited at trial T, then Q(l,T+1)= Q(l,T).

Note that η_{blue} drives the rate of learning at the blue locations; η_{red}, at the red ones. These parameters are exogenous.

4.3.2 Why use this heuristic?

The rule belongs to the class of *Temporal Difference* models (Sutton and Barto, 1998), where learning is driven by the prediction error. Prior work in decision neuroscience has shown that such a prediction error is instantiated in neural mechanisms, such as the firing of dopaminergic neurons (Schultz et al., 1997). It appears that reinforcement learning characterizes reward learning in various contexts, in monkeys, rats, and humans. Humans are sophisticated creatures that may use optimal Bayesian learning in some conditions. But reinforcement learning is pervasive in humans too. So, we took the RL model to be the most plausible definition of "bounded rationality" in the Boardgame.

Our RL model is not the only candidate within the reinforcement learning class though. In particular, a variant of the RL model with a single learning rate would have provided a more parsimonious account of the data. It appeared that the RL model fitted the data better than the simple version.[31]

One may still be skeptical about our usage of the RL model to compete with the Bayesian hypothesis. The RL model may be judged as excessively unsophisticated, as RL players seem to be insensitive to the realized volatility[32] in their environment. We could have conjectured a lesser degree of unsophistication. In the Boardgame, the most sophisticated form of reinforcement learning accommodates the speed of learning to the realized volatility of the locations, whereby the learning rate is higher when the location is experienced as more volatile. A simple modification of the RL model produces this feature. Instead of having the two exogenous learning rates η_{blue} and η_{red}, compute the learning rate at each trial, and for each location, as the (scaled) level of "surprise" (the absolute value of the prediction error last time the location was sampled).[33] This sophisticated version of reinforcement learning appeared not to fit the data well compared to the RL model.[34] So, in the horse race between the fully-rational and bounded-rational models, we believed the RL model to be the most credible competitor of the Bayesian model (HB or FB).

[31] This is after penalizing the RL model for one additional degree of freedom. The results are available upon request.

[32] We use the term "realized volatility" here – rather than "jumps" – because for a reinforcement learner, there are no such things as jumps. Remember that reinforcement learners ignore the underlying causes behind observables.

[33] In the machine learning literature, this model is called the *Pearce-Hall* model (Pearce and Hall, 1980).

[34] The results are available upon request.

5 Model comparison

We assessed the comparative fits of the three models by exploiting data series of 500 trials in average length,[35] from each of 62 subjects (all students at the EPFL, in Lausanne).

5.1 Method

We compared players' choice with predictions made by the competing models. The idea is as follows. Anthropomorphize each model and think of it as a back-seat driver (a "he") commenting on the choices of the player (a "she") at every single trial. The back-seat driver, armed with the actual sequence of decisions of the player and her outcomes, develops his own estimates of the values of the different locations, and makes "recommendations" (predictions) for subsequent choices. To measure to what extent these recommendations match actual choice, we used the likelihood of actual choice under each model. Specifically, for each one of the 62 subjects, we fitted the free parameters of the models[36] to the subject's choice data by maximizing the log-likelihood of observed choice compounded over trials:

$$LL = \sum_{t=1}^{T_{\max}} \ln P^{\pi}(l_t*, t),$$

where l_t* is the location that is actually chosen by the player at trial t, and $T_{\max}$ is the length of the time series.

Note that the approach is *subject-specific* here. For completeness, in the reference to the learning literature, we also fitted behavior on the basis of *fixed* parameters. This was not necessary in our case because unlike in, e.g., Camerer and Ho (1999), samples per subject are reasonably large. The fixed-parameter regression maximizes the likelihood of observed choices across subjects s and trials t:

$$LL_{\textit{fixed}} = \sum_{s=1}^{62} \sum_{t=1}^{T_s} \ln P^{\pi}(l_{st}*, t),$$

where $l_{st}*$ is the location that is actually chosen by player s at trial t, and T_s is the length of the time series for player s.

The Nelder–Mead simplex algorithm was used to optimize the parameter fits. To investigate the robustness of our results, we alternatively used a genetic algorithm to ensure that we avoided local minima.[37] The results that we got with the genetic algorithm are fully consistent with the ones from the Nelder-

[35] The Boardgame is self-paced. Each trial, players have at most 5 seconds to choose their next location, and move to it thereafter. Hence reaction time varies between subjects and the number of trials goes from 450 to 600, with an average of about 500.

[36] In the HB and FB models, there is only one free parameter, β. The RL model adds two parameters: the learning rates η_{blue} and η_{red}.

[37] Simulated annealing has the same advantage, but a genetic algorithm maintains a pool of solutions rather than just one, and new candidate solutions are generated not only by *mutation* but also by *combination* of two solutions from the pool.

Mead simplex search method.[38]

Before doing the horse race between the models, we checked that the predictions of the RL model were sufficiently de-correlated from those of the Bayesian models. For otherwise, the model comparison would have little meaning if any. Having six possible actions rather than simply two helped achieve a reasonably high de-correlation between the predictions of the competing models. Indeed, the larger the choice set, the more likely that the models disagree.[39] Also, as explained earlier, we set up the stochastic structure of the game so that jump detection is facilitated in our task. This was important not only to make the game engaging, but also to have the competing models predict different courses of action. Indeed, because accounting for the jumps worked reasonably well – with the HB or FB method – the Bayesian models often moved to a new location after a jump. The RL model was less reactive, because its two learning rates are exogenous.[40] We setup a test which confirmed that the competing models predicted different courses of action. See the Appendix for the presentation and results of this test.

5.2 Results

5.2.1 Estimation results

We compared the HB, FB, and RL models in terms of their ability to explain the courses of action we recorded from our subjects. Table 1 compares the negative log likelihoods based on fixed-parameter estimation. It reveals behavior to be fit best by the HB model, followed by the FB model. The RL model is worst. To compare the models at the individual subject level, we plotted, for

Table 1: Model fits (negative log likelihood $-LL$) to about 32000 choices, based on fixed parameter estimation

Model	number parameters	- LL_{fixed}
HB model	1	**35407**
FB model	1	**39178**
RL model	3	**43079**

each subject, the HB model's negative log likelihood against the RL model's negative log likelihood (see Fig. 11, top graph), and similarly we plotted, for each subject, the FB model's negative log likelihood against the RL model's negative log likelihood (Fig. 11, bottom graph). When an observation is below the 45 degree line, the Bayesian model fits better than the RL model. Overall, the HB model fits better in 89% of the cases (subjects), while the FB model beats the RL model in 81% of the cases.

[38] The results obtained with the genetic algorithm are available upon request.

[39] Another advantage of having 6 locations rather than 2 is that it is harder to be accurate for any model – "guessings" (the policy that randomly chooses one location among the six possible ones at each trial) leads to only 16.6% of correct predictions on average, against 50% when there are only two possible choices.

[40] Another reason is that under the RL algorithm, the estimated Q-values of the locations that are not currently visited are "frozen" at their most recent value, which is likely to

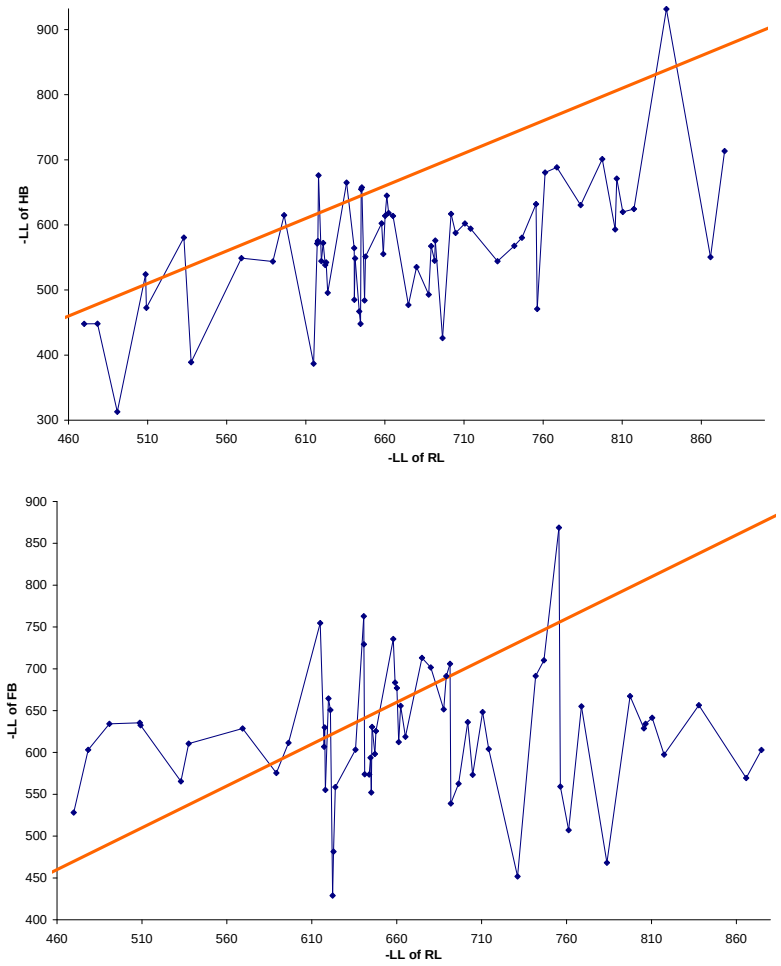


Figure 11: **Comparative fits of the HB and RL models (top graph), and of the FB and RL models (bottom graph), based on the negative log likelihood criterion.** Each data point corresponds to one subject (500 samples on average per subject). The Bayesian model fits better when the data point is below the 45 degree line.

So, the Bayesian models comfortably outperformed the RL model with a pure negative log likelihood criterion, even though they have no free parameter for the inference model (whereas the RL model has two). In principle, the RL model should be penalized for these additional degrees of freedom. So we also evaluated the models on the basis of Bayesian model comparison techniques, that report negative log likelihood penalized for additional degrees of freedom, and the superiority of the Bayesian models is strengthened. Details of this method and the results are provided in the Appendix.

Finally, note that the HB model outperformed the FB model for 75% of the subjects (for spaces reasons, we do not show the corresponding graph), which is congruent with the result we got from the fixed-parameter estimation. As such, subjects' choices reflected optimal learning in the Boardgame. It is doubtful that subjects did so consciously; rather, we expect them to learn intuitively. Bayesian learning may indeed be happening at a low level, as is the case with sensory processing, known to be best explained by sophisticated algorithms, such as Bayesian integration of auditory and visual processes (Shams et al., 2005). Such processes are subconscious.

It is unlikely that the relative fits of the models were driven by our assuming the logit rule to model choice. For we found the variability of the fitted inverse temperature parameter β to be of the same order of magnitude across the models,[41] indicating that the logit model generated equal *fudge factor* effect. Further, the ranking of the three models was confirmed when using a deterministic predictor of choice, as we show now.

5.2.2 Prediction accuracy

We switched-off any source of randomness coming from the choice rule, and did the horse race again. Instead of using the logit rule, we took our subjects to be purely greedy. The "HB algorithm" refers to a back-seat driver behaving as a purely greedy HB learner; the "FB algorithm," to a back-seat driver behaving as a purely greedy FB learner; and the "RL algorithm," to a back-seat driver behaving as a purely greedy reinforcement learner. We compared these three back-seat drivers in terms of their ability to mimic our subjects' choices. Specifically, for each subject, at every single trial of play, we examined whether the choice prescribed by each algorithm was the one that the subject made. We looked at the percentage of trials for which the prediction matched the actual choice. The predictive performance of an algorithm in predicting the subject's course of action relates to this percentage. We assessed the performance of each algorithm for each one of the 62 subjects, whereby we derived the cross-subject distribution of the predictive performance of each algorithm. Fig. 12 displays the distribution of the performance of each algorithm in the form of a box-and-whisker plot. Boxes represent the interquartile range ($25th$ to $75th$ percentile), and whiskers indicate the $5th$ and $95th$ percentiles. Crosses beyond the whiskers are outliers; note that the HB algorithm has one outlier,

be bad – unless exploration is high.

[41] The coefficient of variation of the fitted β is 0.29 in the HB model, 0.26 in the FB model, and 0.47 in the RL model.

for which the predictive performance is only 10%. The notch in each box represents confidence interval about the median (the horizontal line at the middle of the notch). It appears that the median performance of both Bayesian algorithms is significantly higher than the one of the RL algorithm.[42] The HB and FB algorithms thus beat the RL algorithm.

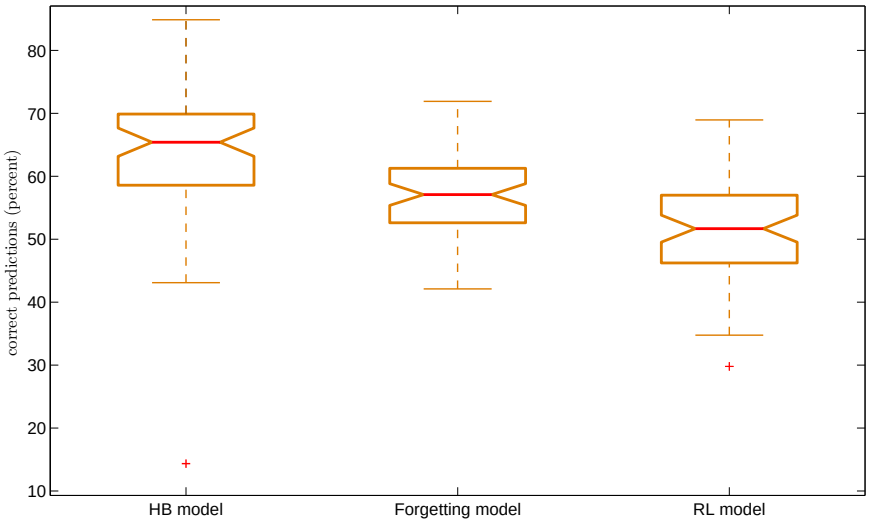


Figure 12: **Box-and-whisker plots representing the distributions of the performance of the HB, FB, and RL algorithms, across the** 62 **subjects. The performance of an algorithm relates to the percentage of trials in which it predicted actual choice during a play.** Boxes represent the interquartile range ($25th$ to $75th$ percentile), and whiskers indicate the $5th$ and $95th$ percentiles; crosses beyond the whiskers are outliers. The notch in each box represents confidence interval about the median, represented by horizontal line at the middle of the notch. The width of a notch is computed so that box plots whose notches do not overlap have different medians at the 5% significance level. The difference between the medians of the HB (FB) algorithm and the RL algorithm is 14% (6%). Since the notches in the box plot do not overlap, we conclude, with 95% confidence, that the true medians do differ in each case.

[42] We also performed a chi-square test that confirmed the distribution of the performance of both HB and FB to differ significantly from the one of RL ($p < 0.001$). These data are available upon request.

Further, Fig. 13 shows that the FB algorithm has *first-order stochastic dominance* (henceforth FOSD) over the RL algorithm: for any percent level of correct predictions x, the probability of observing a percentage of correct predictions equal to or better than x is higher with FB than with RL. The HB algorithm has "almost-FOSD" over the RL algorithm: it does not have FOSD over the RL algorithm because of the outlier mentioned before (for which the prediction accuracy was only 10%), but once this single outlier is excluded, HB outperforms RL according to the FOSD criterion.

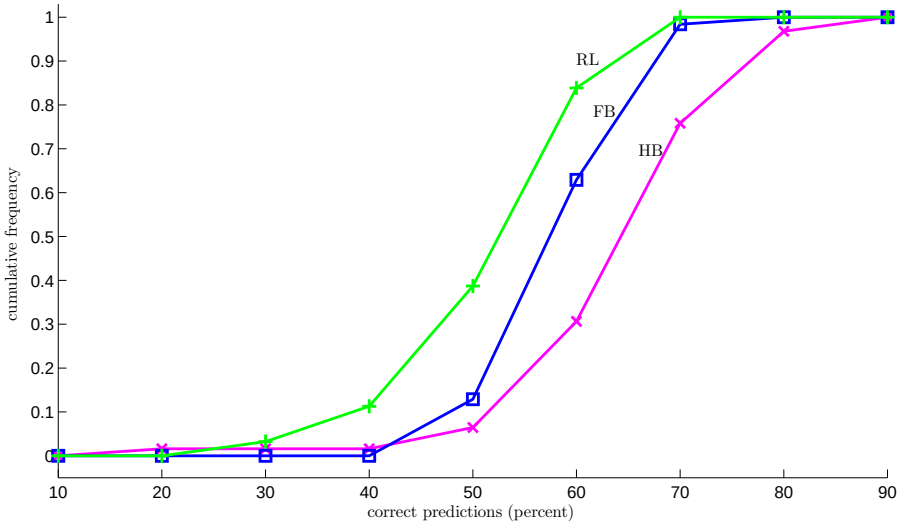


Figure 13: **First-order stochastic dominance (FOSD) of the Bayesian algorithms over the RL algorithm.** The performance level (percentage of correct predictions) is on the x-axis, and the cumulative frequency of these performance levels is on the y-axis. The cumulative frequency of x is the frequency with which performance levels worse than x were obtained. Note that the FB curve is always below the RL curve, which means the FB algorithm has FOSD over the RL algorithm.

5.2.3 Interpretation of our result

To an extent, our result is strong evidence in favor of the full-rationality hypothesis. Our Bayesian models are much more *complex* (in Kolmogorov's sense) than the simple RL model, as it is much easier to write a computer program that simulates a RL learner, compared to a computer program that simulates HB or FB. The more complex an explanation, the more detailed/specific, so the more evidence you need to find it in the space of all possible explanations. We thus expect the simple (general) RL description to fit our observations exactly as well as it would fit other behavioral data. In contrast, the Bayesian formulation describes all the aspects of learning in full detail. Under this complex formulation, the chance of falsifying our claim that people are apt at processing information rationally in our task is higher. Hence, the superiority of the complex explanation over the more general one makes the full-rationality hypothesis much more credible than the bounded-rationality one.

We wrap up this section with two cautionary notes. Firstly, we have not shown that the Bayesian models are reasonable in a stand-alone sense. We do not claim their absolute veracity. In particular, it is unlikely that people are that sophisticated in the way they set their beliefs. We simply did a horse race between two theories. Our objective was to examine whether our subjects processed information rationally in our task. Our finding that *even* our extremely sophisticated Bayesian model comfortably outperformed the bounded-rational model gives some credit to the idea that subjects did process information rationally.[43]

Secondly, the implicit assumption behind this horse race is that choices are relatively consistent across the trials of play. If choices were changing continually during play, it would not make sense to say that a given player "acted more like a Bayesian" or "acted more like a reinforcement learner." It could be, arguably, that people change their learning style over time. Prior work suggests that within-subjects, there is variability in the nature of updating during a same session – see, e.g., Charness and Levin (2005) and M.Kuhnen and Knutson. These studies point to inefficient deviations from Bayesian updating, these deviations being caused by strong affective states. However, such departures from Bayesian learning may be an optimal thing to do at times. Intuitively, a Bayesian player who feels very unsure about her probability estimates at some point during the game, may switch to reinforcement learning. This points to a competition between the two competing learning methods (Bayesian learning vs reinforcement learning). This idea that people may forego the usage of the optimal Bayesian method when they don't trust their estimates has been formalized in the computational neuroscience literature (Daw et al., 2006b).

[43] We found consistent evidence in favor of the Bayesian hypothesis in the debriefing questionnaires. After play, we asked our subjects to do an entropy-based ranking of the locations (e.g., "*Can you tell which one of the blue locations was the most unpredictable? The most predictable?*"). The majority of the subjects was surprisingly good at ranking the three locations. Remarkably, they were capable of dissociating even the median-entropy and minimal-entropy locations, albeit these two were not that different (the minimal-entropy location was very biased when generating outcomes, the median-one was biased too, but to a lesser extent). We believe that such an accuracy reflects sophisticated play.

6 Conclusion

The Bayesian hypothesis thus appeared to be the best predictor of actual behavior in the Boardgame. This finding indicates that people may be capable of processing information rationally in the face of extremely complex environments, such as modern financial markets.

Bayesian updating is difficult in the Boardgame. Viewed from the bounded rationality paradigm, our finding is puzzling. Some may allege that Boardgame players cannot and do not even attempt to learn outcome probabilities, and are reinforcement learners instead. In fact, in our very difficult decision problem, the normative rule did appear to be a much better prediction. This finding contrasts with experimental studies of choice in simple situations such as three-round alternating-offer bargaining over a shrinking pie (Johnson et al., 2002), which show extreme lack of sophistication in decision making (lack of backward induction). Sophisticated thinking thus appears when it is computationally complex, as in our Boardgame, and does not emerge when it is simple. This fact may prompt a reevaluation of the scope of the bounded rationality paradigm. One may conjecture that backward induction in the context of *one-shot* bargaining has low relevance (after all, is not the hallmark of social interactions their repetition?). In contrast, our sampling task, albeit very difficult, is natural. Our brain may thus be well adapted to implement the most efficient learning protocol available. This may explain why people appear to be sophisticated in our task.

Our finding may also surprise one because it points to a positive role of emotions in Bayesian learning. Given the complexity of our task, we had to render it compelling. So, we gave huge monetary incentives to the players, and emotions were reportedly strong during our subjects' play.[44] In the light of prior work, one could expect the strong affect induced by the game to hinder Bayesian learning in it– see Charness and Levin (2005) and M.Kuhnen and Knutson. Our finding that our subjects acted more like Bayesians suggests that in some cases, the key point may be to integrate emotions properly, not to neutralize them.[45]

[44] The majority of our subjects reported the game to be "stressful" in the debriefing questionnaire after play.

[45] The idea that emotions play a chief role in learning is supported by prior studies in neuroeconomics. For instance, amygdala, the "fear center" of the brain, has been associated with ambiguity (Hsu et al., 2005). Such a fear signal may embody the need to start learning, whereby it is critical for Bayesian learning. Likewise, hormones such as cortisol have been shown to influence the behavior of professional traders (Coates and Herbert, 2008), presumably causing risk signals in limbic structures such as the insula, which in turn may cause the heart-beat and other bodily signals with which humans signal risk to themselves (Lo and Repin, 2002).

7 Appendix

Definition of the conditional probability distribution after a jump in the HB model

$P_{0t}(\mathbf{p_{lt}}|\mathbf{p_{lt-1}})$ denotes the time t distribution at location l after a jump, conditional on $\mathbf{p_{lt-1}}$. This distribution is centered on a triplet, $\mathbf{perm}(\mathbf{p_{lt-1}})$, which represents the "swaps" between the probability components:

$$\mathbf{perm}(\mathbf{p_{lt-1}}) = (\mathrm{perm}_1(\mathbf{p_{lt-1}}), \mathrm{perm}_2(\mathbf{p_{lt-1}}), 1 - \mathrm{perm}_1(\mathbf{p_{lt-1}}) - \mathrm{perm}_2(\mathbf{p_{lt-1}}))\,.$$

$\mathbf{perm}(\mathbf{p_{lt-1}})$ is obtained from a simple permutation function f.
Let $\mathbf{p}_{\text{reordered l t-1}} = (p_{\max\ l\ t-1}, p_{\text{mid } l\ t-1}, p_{\min\ l\ t-1})$. f is defined as follows:

$$\begin{aligned}\mathbf{perm}(\mathbf{p_{lt-1}}) \equiv f(\mathbf{p}_{\text{reordered l t-1}}) = \\ 1/4(p_{\min\ l\ t-1}, p_{\text{mid } l\ t-1}, p_{\max\ l\ t-1}) + \\ 1/4(p_{\text{mid } l\ t-1}, p_{\min\ l\ t-1}, p_{\max\ l\ t-1}) + \\ 1/4(p_{\min\ l\ t-1}, p_{\max\ l\ t-1}, p_{\text{mid } l\ t-1}) + \\ 1/4(p_{\text{mid } l\ t-1}, p_{\max\ l\ t-1}, p_{\min\ l\ t-1}).\end{aligned}$$

The distribution after a jump conditional on $\mathbf{p_{lt-1}}$ is the uniform distribution

$$\begin{aligned}&\mathrm{U}\,(\ [\mathrm{perm}_1(\mathbf{p_{lt-1}}) - 0.1; \mathrm{perm}_1(\mathbf{p_{lt-1}}) + 0.1] \\ &\times[\mathrm{perm}_2(\mathbf{p_{lt-1}}) - 0.1; \mathrm{perm}_2(\mathbf{p_{lt-1}}) + 0.1]\ )\,.\end{aligned}$$

Derivation of the posterior distribution of the jump parameter in the HB algorithm

At time T , the posterior distribution on the jump parameter is

$$f_T(\alpha) = \int_\Theta P(\mathbf{p_T}, J_T = 1, \alpha|\underline{c}_\mathbf{T})\, d\mathbf{p_T} + \int_\Theta P(\mathbf{p_T}, J_T = 0, \alpha|\underline{c}_\mathbf{T})\, d\mathbf{p_T}, \qquad (7)$$

so we need to calculate the joint likelihoods $P(\mathbf{p_T}, J_T = 1, \alpha|\underline{c}_\mathbf{T})$ and $P(\mathbf{p_T}, J_T = 0, \alpha|\underline{c}_\mathbf{T})$. Take first $P(\mathbf{p_T}, J_T = 1, \alpha|\underline{c}_\mathbf{T})$, an untractable multidimensional integral:

$$\begin{aligned}&P(\mathbf{p_T}, J_T = 1, \alpha|\underline{c}_\mathbf{T}) \propto \\ &\sum_{J_{T-1}} \sum_{J_{T-2}} \cdots \int_\Theta \cdots \int_\Theta P(\mathbf{p_T}, J_T = 1, (\underline{p}_{\mathbf{T-1}}, \underline{J}_{T-1}, \alpha, \underline{c}_\mathbf{T})\, d\underline{p}_{\mathbf{T-1}}.\end{aligned}$$

The Markovian structure of the game – both the probabilities and the jumps are independent over time – permits us to reduce the dimensionality of the

problem to a (still complicated) multidimensional integral over all the possible $T-1$ probabilities:

$$P(\mathbf{p_T}, J_T = 1, \alpha|\underline{\mathbf{c}}_\mathbf{T}) \propto$$
$$\sum_{J_{T-1}} \int_\Theta P(\mathbf{p_{T-1}}, J_{T-1}, \alpha|\underline{\mathbf{c}}_\mathbf{T-1}) l(\mathbf{c_T}|\mathbf{p_T})$$
$$\times P(\mathbf{p_T}|\mathbf{p_{T-1}}, J_T = 1) P(J_T = 1|\alpha)\, d\mathbf{p_{T-1}}.$$

The joint likelihood is proportional to

$$\sum_{J_{T-1}} \int_\Theta P(\mathbf{p_{T-1}}, J_{T-1}, \alpha|\underline{\mathbf{c}}_\mathbf{T-1}) l(\mathbf{c_T}|\mathbf{p_T}) P(\mathbf{p_T}|\mathbf{p_{T-1}}, J_T = 1) P(J_T = 1|\alpha)\, d\mathbf{p_{T-1}}.$$

Proof. To assess $P(\mathbf{p_T}, J_T = 1, \alpha|\underline{\mathbf{c}}_\mathbf{T})$, first note that it is proportional to $P(\mathbf{p_T}, J_T = 1, \alpha, \underline{\mathbf{c}}_\mathbf{T})$:

$$P(\mathbf{p_T}, J_T = 1, \alpha, \underline{\mathbf{c}}_\mathbf{T}) =$$
$$\sum_{J_{T-1}} \sum_{J_{T-2}} \cdots \int_\Theta \cdots \int_\Theta P(\mathbf{p_T}, J_T = 1, \underline{\mathrm{p}}_\mathbf{T-1}, \underline{\mathrm{J}}_{T-1}, \alpha, \underline{\mathbf{c}}_\mathbf{T})\, d\underline{\mathrm{p}}_\mathbf{T-1},$$

Using the Markov property, consider

$$P(\mathbf{p_T}, J_T = 1, \alpha, \underline{\mathrm{p}}_\mathbf{T-1}, \underline{\mathrm{J}}_{T-1}, \underline{\mathbf{c}}_\mathbf{T}) =$$
$$\left[\prod_{t=1}^{T} l(\mathbf{c_t}|\mathbf{p_t}) P(\mathbf{p_t}|\mathbf{p_{t-1}}, J_t) P(J_t|\alpha)\right] \alpha_0(\alpha) P_0(\mathbf{p_0}).$$

Develop the previous expression to get

$$P(\mathbf{p_T}, J_T = 1, \alpha, \underline{\mathbf{c}}_\mathbf{T}) =$$
$$\sum_{J_{T-1}} \sum_{J_{T-2}} \cdots \int_\Theta \cdots \int_\Theta \prod_{t=1}^{T-1} l(\mathbf{c_t}|\mathbf{p_t}) P(\mathbf{p_t}|\mathbf{p_{t-1}}, J_t) P(J_t|\alpha) \alpha_0(\alpha) P_0(\mathbf{p_0})$$
$$\times P(\mathbf{p_T}|\mathbf{p_{T-1}}, J_T = 1) l(\mathbf{c_T}|\mathbf{p_T}) P(J_T = 1|\alpha)\, d\underline{\mathrm{p}}_\mathbf{T-1}.$$

The ensuing form gives the result:

$$P(\mathbf{p_T}, J_T = 1, \alpha|\underline{\mathbf{c}}_\mathbf{T}) \propto$$
$$\sum_{J_{T-1}} \int_\Theta P(\mathbf{p_{T-1}}, J_{T-1}, \alpha|\underline{\mathbf{c}}_\mathbf{T-1}) l(\mathbf{c_T}|\mathbf{p_T}) P(\mathbf{p_T}|\mathbf{p_{T-1}}, J_T = 1) P(J_T = 1|\alpha)\, d\mathbf{p_{T-1}}.$$

To have a more tractable expression, suppose that HB learners do not marginalize over all the possible $T-1$ probabilities, but only consider $\mathbf{p_{T-1}}^*$,

their most preferred hypothesis at trial $T-1$ in that it is (one of) the mode(s) of the posterior:[46]

$$\mathbf{p_{T-1}}^* = \underset{\mathbf{p_{T-1}} \in \urcorner_{T-1}}{\arg\max}\ P_{T-1}(\mathbf{p_{T-1}}),$$

where $\urcorner_{T-1}$ denotes the set of admissible triplets $\mathbf{p_{T-1}}$ of the contour.[47]

This leads to the following simplified expression for $P(\mathbf{p_T}, J_T = 1, \alpha|\underline{c}_\mathbf{T})$:

$$P(\mathbf{p_T}, J_T = 1, \alpha|\underline{c}_\mathbf{T}) \propto$$

$$\sum_{J_{T-1}} P(\mathbf{p_{T-1}}^*, J_{T-1}, \alpha|\underline{c}_\mathbf{T-1}) l(\mathbf{c_T}|\mathbf{p_T}) P(\mathbf{p_T}|\mathbf{p_{T-1}}^*, J_T = 1) P(J_T = 1|\alpha).$$

Using the fact that $J_T \sim Bern(\alpha)$, so that $P(J_T = 1|\alpha) = \alpha$, we have

$$\begin{aligned} &P(\mathbf{p_T}, J_T = 1, \alpha|\underline{c}_\mathbf{T}) \propto \\ &P(\mathbf{p_{T-1}}^*, J_{T-1} = 0, \alpha|\underline{c}_\mathbf{T-1}) l(\mathbf{c_T}|\mathbf{p_T}) P(\mathbf{p_T}|\mathbf{p_{T-1}}^*, J_T = 1)\alpha \\ +&P(\mathbf{p_{T-1}}^*, J_{T-1} = 1, \alpha|\underline{c}_\mathbf{T-1}) l(\mathbf{c_T}|\mathbf{p_T}) P(\mathbf{p_T}|\mathbf{p_{T-1}}^*, J_T = 1)\alpha, \end{aligned}$$

where

$$\begin{aligned} &P(\mathbf{p_{T-1}}^*, J_{T-1} = 0, \alpha|\underline{c}_\mathbf{T-1}) = (1-\alpha) P(\mathbf{p_{T-1}}^*|J_{T-1} = 0, \alpha, \underline{c}_\mathbf{T-1})\ f_{T-1}(\alpha), \\ &P(\mathbf{p_{T-1}}^*, J_{T-1} = 1, \alpha|\underline{c}_\mathbf{T-1}) = \alpha P(\mathbf{p_{T-1}}^*|J_{T-1} = 1, \alpha, \underline{c}_\mathbf{T-1}) \quad f_{T-1}(\alpha), \\ &\text{and } P(\mathbf{p_T}|\mathbf{p_{T-1}}^*, J_T = 1) = P_{0T}(\mathbf{p_T}|\mathbf{p_{T-1}}^*). \end{aligned}$$

The implication is

$$\begin{aligned} &P(\mathbf{p_T}, J_T = 1, \alpha|\underline{c}_\mathbf{T}) \propto \\ &\alpha \left[(1-\alpha)\, P(\mathbf{p_{T-1}}^*|J_{T-1} = 0, \alpha, \underline{c}_\mathbf{T-1}) l(\mathbf{c_T}|\mathbf{p_T}) P_{0T}(\mathbf{p_T}|\mathbf{p_{T-1}}^*)\right. \\ &\left. + \alpha P(\mathbf{p_{T-1}}^*|J_{T-1} = 1, \alpha, \underline{c}_\mathbf{T-1}) l(\mathbf{c_T}|\mathbf{p_T}) P_{0T}(\mathbf{p_T}|\mathbf{p_{T-1}}^*)\right] \\ &\times f_{T-1}(\alpha). \end{aligned}$$

Simple rewriting leads to

$$P(\mathbf{p_T}, J_T = 1, \alpha|\underline{c}_\mathbf{T}) \propto$$

[46] Such approximation is usual practice in the machine learning literature; see, e.g., Yu and Dayan (2005).

[47] The contour here refers to the set of triplets with an entropy level equal to that of the generic location.

$$
\begin{aligned}
&\alpha \; l(\mathbf{c_T}|\mathbf{p_T}) \\
&\times \underbrace{\left[(1-\alpha)\, P(\mathbf{p_{T-1}}^*|J_{T-1}=0,\alpha,\underline{c}_{\mathbf{T-1}}) + \alpha P(\mathbf{p_{T-1}}^*|J_{T-1}=1,\alpha,\underline{c}_{\mathbf{T-1}})\right]}_{*} \\
&\times P_{0T}(\mathbf{p_T}|\mathbf{p_{T-1}}^*) \;\; f_{T-1}(\alpha).
\end{aligned}
$$

We proceeded in an analogous way to calculate $P(\mathbf{p_T}, J_T = 0, \alpha|\underline{c}_{\mathbf{T}})$. Since $P(\mathbf{p_T}|\mathbf{p_{T-1}}, J_T = 0) = \delta_{\mathbf{p_{T-1}}}(\mathbf{p_T}) = 1$ if $\mathbf{p_T} = \mathbf{p_{T-1}}$ and 0 otherwise, the expression for the joint probability $P(\mathbf{p_T}, J_T = 0, \alpha|\underline{c}_{\mathbf{T}})$ boils down to

$$
\begin{aligned}
&P(\mathbf{p_T}, J_T = 0, \alpha|\underline{c}_{\mathbf{T}}) \propto \\
&\\
&\quad (1-\alpha) l(\mathbf{c_T}|\mathbf{p_T}) \\
&\times \underbrace{\left[(1-\alpha)\, P(\mathbf{p_{T-1}}^*|J_{T-1}=0,\alpha,\underline{c}_{\mathbf{T-1}}) + \alpha P(\mathbf{p_{T-1}}^*|J_{T-1}=1,\alpha,\underline{c}_{\mathbf{T-1}})\right]}_{*} \\
&\times f_{T-1}(\alpha).
\end{aligned}
$$

Now Identify $*$ as $P(\mathbf{p_T}|\underline{c}_{\mathbf{T-1}}, \alpha)$ and revert to Equation (7) (p. 38) to get the main result:

$$
f_T(\alpha) = f_{T-1}(\alpha) \int_\Theta l(\mathbf{c_T}|\mathbf{p_T}) P(\mathbf{p_T}|\underline{c}_{\mathbf{T-1}}, \alpha) \left[\alpha P_{0T}(\mathbf{p_T}|\mathbf{p_{T-1}}^*) + (1-\alpha)\right] d\mathbf{p_T}.
$$

□

Derivation of a recursive equation for $P(\mathbf{p_T}|\underline{c}_{\mathbf{T-1}}, \alpha)$ in the HB model

Starting from

$$
P(\mathbf{p_T}|\underline{c}_{\mathbf{T-1}}, \alpha) = P(\mathbf{p_T}, J_{T-1} = 1|\underline{c}_{\mathbf{T-1}}, \alpha) + P(\mathbf{p_T}, J_{T-1} = 0|\underline{c}_{\mathbf{T-1}}, \alpha),
$$

use Bayes law to derive the following recursive equation:

$$
\begin{aligned}
&P(\mathbf{p_T}|\underline{c}_{\mathbf{T-1}}, \alpha) \propto \\
&\\
&\quad \alpha P_{0T-1}(\mathbf{p_T}|\mathbf{p_{T-2}}^*)\, l(\mathbf{c_{T-1}}|\mathbf{p_T}) + (1-\alpha)\, P(\mathbf{p_T}|\underline{c}_{\mathbf{T-2}}, \alpha)\, l(\mathbf{c_{T-1}}|\mathbf{p_T})\,.
\end{aligned}
$$

Derivation of the posterior probability distribution in the FB model

Because it will help the reader to proceed incrementally, we will start by ignoring the jumps ("Benchmark"). Then discounting will be introduced with the jumps. In the final part of this section ("Metalearning"), we will explain what the appropriate discounting rate is.

7.0.4 Benchmark

The prior probability distribution, denoted by P_0,[48] is Dirichlet with center $\hat{\mathbf{p}}_\mathbf{0}$ and precision $\nu_\mathbf{0} = (\nu_0, \nu_0, \nu_0)$:

$$P_0(\mathbf{p}) = \left[\frac{\prod_{i=1}^{3}\Gamma(\nu_0\hat{p}_{i0})}{\Gamma(\nu_0)}\right]^{-1} \prod_{i=1}^{3} p_i^{(\nu_0\hat{p}_{i0}-1)} \quad \delta_\Theta(\mathbf{p}), \tag{8}$$

with $\mathbf{p} = (p_1, p_2, p_3)'$,

$$\Gamma(\nu_0\hat{p}_{i0}) = \int_\Theta x^{\nu_0\hat{p}_{i0}-1} \ \mathrm{e}^{-x}\, dx.$$

The center $\hat{\mathbf{p}}_\mathbf{0}$ is the base measure: $\hat{\mathbf{p}}_\mathbf{0} = E_{\mathrm{Dir}(\hat{\mathbf{p}}_\mathbf{0},\nu_\mathbf{0})}[\mathbf{p}]$. Absence of prior knowledge related to outcome probability is formalized by $\hat{p}_{i0} = 1/3$, for $i = 1 \ldots 3$. The precision parameter $\nu_\mathbf{0}$ controls the extent to which the probability mass is localized around the center $\hat{\mathbf{p}}_\mathbf{0}$. We set it equal to $(1, 1, 1)$. Intuitively, $\nu_0\hat{p}_{i0}$ is tantamount to the "prior observation counts" for outcome i, thereby measuring (in units of i.i.d samples) the weight of the prior in the inference.

If outcome probability is fixed, it has long been known that starting from P_0, the posterior probability distribution is also Dirichlet:[49]

$$p_{l\,T} \sim \mathrm{Dir}(\hat{\mathbf{p}}_{\mathbf{l}\,\mathbf{T}}, \nu_{\mathbf{l}\mathbf{T}}), \tag{9}$$

$$\hat{p}_{i\,l\,T} = \frac{1}{\nu_{T_l}}\left[\nu_0\,\hat{p}_{i\,0} + T_l <c_{li}>(\mathrm{T})\right], \tag{10}$$

$$\nu_{lT} = \nu_0 + T_l, \tag{11}$$

where T_l denotes the cardinal of $\Delta_l(T)$[50] and $<c_{li}>(\mathrm{T}) = \frac{1}{T_l}\sum_{t\in\Delta_l(T)} c_{lit}$.[51]

7.0.5 Natural sampling in the context of jumps

For nonstationary sampling, the inference is more complicated. We used a *stabilized forgetting* technique (Quinn and Kárný, 2007) for the learning of the nonstationary outcome probability. At each trial the algorithm applies the usual Bayes operator first and a stabilized forgetting operator thereafter. By "usual Bayes operator" at time T – denoted by B_T, we mean the operator transforming a given prior into the likelihood times this prior:

$$\forall \mathbf{p_T},\ B_T\,P(\mathbf{p_T}) \propto l(\mathbf{C_T}|\mathbf{p_T})P(\mathbf{p_T}).$$

[48] P_{l0}, the prior probability distribution at location l, is the same for all the locations, so $P_{l0} \equiv P_0$.

[49] This closure of the prior for multinomial sampling characterizes the exponential family of models with hidden variables (Sato, 2001), of which the Dirichlet distribution is a member.

[50] Remember that $\Delta_l(T) = \{\ t|\ \ l \textit{ is visited at time } t, t \leq T\}$.

[51] As such, $<c_{li}>(\mathrm{T})$ is the empirical distribution corresponding to the unknown probability $p_{i\,l}$.

We will turn now to the definition of the forgetting operator.

The forgetting operator The forgetting operator is a substitute for the transition-probability operator $P(\mathbf{p_{l\,t+1}}|\mathbf{p_{l\,T}})$ when a complete model of parameter changes is not available. To define things we need some notation. Let $P_{l\,T+1/T}$ denote the prediction of outcome probability at time $T+1$ when at time T. $P_{l\,T+1/T}$ is also the prior on outcome probability at time $T+1$. Take $P_{l\,T/T} = P_l(\mathbf{p_T}|\underline{c}_{\mathbf{lT}})$ to be the posterior probability distribution at trial T:

$$P_{l\,T/T} \propto B_T\, P_{l\,T/T-1},$$

where $P_{l\,T/T-1}$ is the prior at time T. Viewed from time T, $P_{l\,T/T}$ is the best guess about the unknown outcome probability, based on available evidence at time T.

Let us now introduce the forgetting operator, which we will denote F_T. What F_T does is to derive the prediction $P_{l\,T+1/T}$ from the posterior probability distribution $P_{l\,T/T}$, by assessing the likelihood of a jump between $T-1$ and T. Specifically:

- After a jump, the prediction should not be $P_{l\,T/T}$ but another reference probability distribution $P^*_{l\,T+1/T}$, which describes available knowledge of $\mathbf{p_{lT+1}}$ based on prior information and information contained in $\underline{c}_{\mathbf{lT}}$ (this is the minimal guaranteed information about the future value $\mathbf{p_{lT+1}}$). Here $P^*_{l\,T+1/T} = P_0$, where P_0 is defined by Equation (8), p. 42. To take $P_{l\,T+1/T}$ equal to the reference prior means to restart the parameter estimation by forgetting all information gathered from previous observations. This is optimal after a jump (at which point past data brings little information about the current probability parameter). Notice that, by re-starting from P_0, our decision maker ignores crucial structural information, namely, that entropies are not equal across locations. Alternative assumptions did not improve the behavioral fit, however. One may conjecture that there are trade-offs between accurate jump detection and entropy learning: successful entropy estimation makes it harder to identify jumps and vice versa.

- Conversely, in absence of a jump, one should carry over the latest posterior probability distribution $P_{l\,T/T}$.

Thus, either $\mathbf{p_{lT+1}} \sim P_{l\,T/T}$, or $\mathbf{p_{lT+1}} \sim P_0$.

Had jump detection been perfect, at any time either the reference or the latest posterior probability distribution would be considered. F_T mixes both kinds of information, thereby representing uncertainty about the occurrence of a jump:

$$P_{l\,T+1/T} = F_T(P_{l\,T/T}, P_0).$$

Specifically, F_T derives $P_{l\,T+1/T}$ as a *geometric mean* value between P_0 and $P_{l\,T/T}$; we will explain the nature of this mean now. Let $\lambda(T)$[52] denote the

[52] $\lambda(T)$ stands for $\lambda(T)_{\text{blue}}$ if the location that is visited at T is blue, and for $\lambda(T)_{\text{red}}$ if it is red.

subjective probability that no jump occurred at time T. Formally, $\lambda(T) = P(J_{lT} = 0 | \underline{c}_{\mathbf{lT}})$. $\lambda(T)$ equals to 0 means certainty about the occurrence of a jump at time T; conversely, $\lambda(T)$ equals to 1 means certainty about non-occurrence. A value of the subjective probability between 0 and 1 formalizes belief uncertainty related to jump occurrence. See the next subsection for a formal calculation of the value of $\lambda(T)$.

F_T is solution of the following Bayes risk minimization program:

$$\min_{P \in \Upsilon} \left\{ \lambda(T)\, KL(P, P_{lT/T}) + (1 - \lambda(T))\, KL(P, P_0) \right\},$$

where Υ denotes the probability space on Θ and $KL(.,.)$ stands for the *Kullback-Leibler divergence* measure. So the criterion can be rewritten

$$\lambda(T) \int_\Theta P\,(\mathbf{p}) \ln \frac{P(\mathbf{p})}{P_{lT/T}(\mathbf{p})}\; d\mathbf{p} + (1 - \lambda(T)) \int_\Theta P(\mathbf{p})\; \ln \frac{P(\mathbf{p})}{P_0(\mathbf{p})}\; d\mathbf{p}. \qquad (12)$$

Note that the Kullback-Leibler divergence used in the criterion is not $KL(P_{lT/T}, P)$ – the one that would be used in a usual Bayes risk minimization program, but $KL(P, P_{lT/T})$. We used such "reverse KL" because under the assumption that the product between $P_{lT/T}$ and P_0 is not 0 everywhere,

$$P_{lT+1/T} = F_T\,(P_{lT/T}, P_0) = \arg\min_{P \in \Upsilon} \{\, (12) \,\} = \left(P_{lT/T}\right)^{\lambda(T)} (P_0)^{1-\lambda(T)}.$$

Thus, F_T is the *geometric* mean between the latest posterior probability distribution and the reference prior. Had we used the usual Bayes risk decision criterion, we would have ended up with the *arithmetic* mean. The sense of these two minimization programs is strictly equivalent, but the geometric mean is more tractable, in the following sense.

Lemma. *F_T is closed for Dirichlet probability distributions $P_{lT/T}$ and P_0. Thus, $P_{lT+1/T}$ is Dirichlet.*

Proof. Getting back to probability distribution function (p.d.f) of a Dirichlet defined by (8), we write explicitly the p.d.f $P_{lT+1/T}(\mathbf{p})$:

$$P_{lT+1/T}(\mathbf{p}) =$$

$$\left[\prod_{i=1}^{3} \frac{\Gamma\left(\left((1-\lambda)\nu_0 + \lambda\nu_{T/T}\right) \left(\frac{1}{(1-\lambda)\nu_0 + \lambda\nu_{T/T}} \left((1-\lambda)\hat{p_0} + \lambda\hat{p}_{lT/T}\right)\right)\right)}{\Gamma(\nu_0)} \right]^{-1}$$

$$\times \prod_{i=1}^{3} p_i^{\left((1-\lambda)\nu_0 \hat{p_{0i}} + \lambda\nu_{T/T}\hat{p}_{T/T\,i} - 1\right)} \quad \delta_\Theta(p).$$

Developing the previous expression and regrouping leads to the desired form:

$$\mathbf{p_{lT+1/T}} \sim \mathrm{Dir}(\hat{\mathbf{p}}_{\mathbf{lT+1/T}}, \nu_{\mathbf{T+1/T}}),$$

$$\text{with} \quad \begin{cases} \hat{\mathbf{p}}_{\mathbf{l}\,\mathbf{T+1/T}} = \frac{1}{\nu_{T+1/T}} \left((1-\lambda)\, \nu_0 \hat{\mathbf{p_0}} + \lambda\, \nu_{T/T} \hat{\mathbf{p}}_{\mathbf{l}\,\mathbf{T/T}} \right), \\ \\ \nu_{T+1/T} = (1-\lambda)\nu_0 + \lambda \nu_{T/T}. \end{cases} \tag{13}$$

□

With a weighted sum of two Dirichlet distributions, Lemma 7.0.5 does not hold, whereby the "reverse KL" is critical to make $P_{l\,T+1/T}$ conjugate.

Updating of the probabilities Equipped with Lemma 7.0.5, we will now state the main result: the Dirichlet measure is also closed for the sequential use of the two operators (Bayes operator and forgetting operator). In other terms, the space of the nonstationary posterior probability distributions is conjugate under the sequence of i.i.d sampling *and* stabilized forgetting.

Proposition. *For large T, at any location l, the posterior probability of the unknown triplet at time T is well approximated by*

$$\begin{aligned} &\mathbf{p}_{\mathbf{l}\,\mathbf{T}} \sim Dir(\hat{\mathbf{p}}_{\mathbf{l}\,\mathbf{T}}, \nu_{\mathbf{l}\mathbf{T}}), \\ &\hat{p}_{i\,l\,T} = \frac{1}{\nu_{lT}} \left[\nu_0\, \hat{p}_{i\,0} + N^{\lambda}{}_l(T)\, «c_{li}(T)» \right], \\ &\nu_{lT} = \nu_0 + N^{\lambda}{}_l(T), \\ \textit{where } «c_{li}(T)» &= \frac{\sum_{t \in \Delta_l(T)} \left(\prod_{s=t}^{T} \lambda(s) \right) c_{lit}}{N^{\lambda}{}_l(T)}, \\ N^{\lambda}{}_l(T) &= \sum_{t \in \Delta_l(T)} \prod_{s=t}^{T} \lambda(s). \end{aligned}$$

Proof. For expositional clarity we provide a proof for $\lambda(T)$ equals to λ at each trial T. The principle of the proof is exactly the same under time-dependent λ(T). Besides, to simplify and without loss, assume that the location is visited at each trial from $t = 1$ up to $t = T$ included (so that $T_l = T$).

Consider Equation (13), p. 45, and note that $P_{l\,T/T-1} = F_{T-1}(P_{lT-1/T-1}, P_0)$. This leads to

$$\hat{\mathbf{p}}_{\mathbf{lT/T-1}} = \frac{1}{\nu_{lT/T-1}} \left[(1-\lambda)\nu_0 \hat{\mathbf{p}}_{\mathbf{0}} + \lambda \nu_{T-1/T-1} \hat{\mathbf{p}}_{\mathbf{lT-1/T-1}} \right].$$

On the other hand, using (9), p. 42,

$$\nu_{lT-1/T-1} \hat{\mathbf{p}}_{\mathbf{lT-1/T-1}} = \nu_{lT-1/T-2} \hat{\mathbf{p}}_{\mathbf{lT-1/T-2}} + \mathbf{c_{lT-1}},$$

$$\text{so} \quad \hat{\mathbf{p}}_{\mathbf{lT/T-1}} = \frac{1}{\nu_{lT/T-1}} \left[(1-\lambda)\nu_0 \hat{\mathbf{p}}_{\mathbf{0}} + \lambda (\nu_{lT-1/T-2} \hat{\mathbf{p}}_{\mathbf{lT-1/T-2}} + \mathbf{c_{lT-1}}) \right].$$

Proceeding recursively, apply the two operators successively: first the forgetting operator to replace $\nu_{lT-1/T-2} \hat{p}_{lT-1/T-2}$, then the Bayes operator, and

so on:

$$\hat{\mathbf{p}}_{\mathbf{lT/T-1}} = \frac{1}{\nu_{lT/T-1}}\left[(1-\lambda)\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \lambda\left(\nu_{lT-2/T-2}\hat{\mathbf{p}}_{\mathbf{lT-2/T-2}} + \mathbf{c}_{\mathbf{lT-1}}\right)\right]$$

$$= \frac{1}{\nu_{lT/T-1}}\left[(1-\lambda)\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \lambda\left((1-\lambda)\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \lambda\nu_{lT-1/T-2}\hat{\mathbf{p}}_{\mathbf{lT-1/T-2}} + \mathbf{c}_{\mathbf{lT-1}}\right)\right]$$

(Using (13), p. 45)

$$= \frac{1}{\nu_{lT/T-1}}\left[(1-\lambda)\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \right.$$

$$\left.\lambda\left((1-\lambda)\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \lambda\left(\nu_{lT-2/T-3}\hat{\mathbf{p}}_{\mathbf{lT-2/T-3}} + \mathbf{c}_{\mathbf{lT-2}}\right) + \mathbf{c}_{\mathbf{lT-1}}\right)\right]$$

(Using (9), p. 42)

$$\vdots$$

$$= \frac{1}{\nu_{lT/T-1}}\left[(1-\lambda)\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \lambda(1-\lambda)\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \lambda^2(1-\lambda)\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \ldots\right.$$

$$\left.+\lambda^{T-2}(1-\lambda)\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \lambda^{T-1}\nu_{1/0}\hat{\mathbf{p}}_{\mathbf{l1/0}} + \lambda\mathbf{c}_{\mathbf{lT-1}} + \lambda^2\mathbf{c}_{\mathbf{lT-2}} + \ldots + \lambda^{T-1}\mathbf{c}_{\mathbf{l1}}\right].$$

Apply the forgetting operator again: $\nu_{1/0}\hat{\mathbf{p}}_{\mathbf{l1/0}} = [(1-\lambda)\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \lambda\nu_0\hat{\mathbf{p}}_{\mathbf{0}}] = \nu_0\hat{\mathbf{p}}_{\mathbf{0}}$. Since $\lim_{T\to\infty}\lambda^T\nu_0\hat{\mathbf{p}}_{\mathbf{0}} = 0$, we have that

$$\hat{\mathbf{p}}_{\mathbf{lT/T-1}} \approxeq \frac{1}{\nu_{lT/T-1}}\left[(1-\lambda)\sum_{k=0}^{T-2}\lambda^k\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \sum_{t=1}^{T-1}\lambda^{T-t}\mathbf{c}_{\mathbf{lt}}\right].$$

We also know that

$$\lim_{T\to\infty}(1-\lambda)\sum_{k=0}^{T-2}\lambda^k = 1.$$

Combining the two leads to

$$\lim_{T\to\infty}\hat{\mathbf{p}}_{\mathbf{lT/T-1}} = \frac{1}{\nu_{lT/T-1}}\left[\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \sum_{t=1}^{T-1}\lambda^{T-t}\mathbf{c}_{\mathbf{lt}}\right].$$

So for T large enough, a valid approximation of $\nu_{lT/T-1}\hat{\mathbf{p}}_{\mathbf{lT/T-1}}$ is $\left[\nu_0\hat{\mathbf{p}}_{\mathbf{0}} + \sum_{t=1}^{T-1}\lambda^{T-t}\mathbf{c}_{\mathbf{lt}}\right]$, which implies that

$$\begin{cases} \hat{\mathbf{p}}_{\mathbf{lT/T-1}} \approxeq \frac{1}{\nu_{lT/T-1}} \left[\nu_0 \hat{\mathbf{p}}_{\mathbf{0}} + \sum_{t=1}^{T-1} \lambda^{T-t} \mathbf{c_{lt}} \right], \\ \\ \nu_{lT/T-1} = \nu_0 + \sum_{t=1}^{T-1} \lambda^{T-t}. \end{cases}$$

Applying the Bayes operator one more time, at trial T, leads to

$$\hat{\mathbf{p}}_{\mathbf{lT/T}} = \frac{1}{\nu_{lT/T}} \left[\nu_{T/T-1} \hat{\mathbf{p}}_{\mathbf{lT/T-1}} + \mathbf{c_{lT}} \right].$$

Replace for the expression of $\hat{\mathbf{p}}_{\mathbf{lT/T-1}}$, then conclude:

$$\hat{\mathbf{p}}_{\mathbf{lT/T}} \approxeq \frac{1}{\nu_{lT/T}} \left[\nu_0 \hat{\mathbf{p}}_{\mathbf{0}} + \sum_{t=1}^{T} \lambda^{T-t} \mathbf{c_{lt}} \right],$$

$$\nu_{lT/T} = \nu_0 + \sum_{t=1}^{T} \lambda^{T-t}.$$

In the presentation of the model, $\hat{\mathbf{p}}_{\mathbf{lT}}$ stands for $\hat{\mathbf{p}}_{\mathbf{lT/T}}$, to simplify notations. □

From the Proposition, we can derive equivalent formulations in terms of recursions, as follows.

Corollary. *Consider* $\eta_l(T) = \left[\sum_{t \in \Delta_l(T)} \left(\prod_{s=t}^{T} \lambda(s) \right) \right]^{-1}$. *For* $T = \sup \Delta_l(T)$,

$$\text{«}c_{li}(T)\text{»} = \text{«}c_{li}(T\text{-}1)\text{»} \, (1 - \eta_l(T)) + \eta_l(T) c_{liT}.$$

Further,

$$\eta_l(T) = \frac{1}{1 + \frac{\lambda(T)}{\eta_l(T-1)}}.$$

Equivalently,

$$N^{\lambda}{}_l(T) = 1 + \lambda(T) N^{\lambda}(T-1).$$

Proof. To simplify notation, the proof is done for a constant $\lambda(T)$, denoted by λ. Note that the proof is exactly the same under a time-dependent parameter.

Starting from

$$\begin{aligned}
\text{«}c_{li}(\text{T})\text{»} &= \eta_l(T)\sum_{k=1}^{T}\lambda^{T-k}c_{lik} \quad \text{with } \eta_l(T) = \left[\sum_{k=1}^{T}\lambda^{T-k}\right]^{-1} \\
&= \eta_l(T-1)\left[\lambda\sum_{k=1}^{T-1}\lambda^{T-1-k}c_{lik} + c_{liT}\right]\frac{\eta_l(T)}{\eta_l(T-1)} \\
&= \lambda\left[\eta_l(T-1)\sum_{k=1}^{T-1}\lambda^{T-1-k}c_{lik}\right]\frac{\eta_l(T)}{\eta_l(T-1)} + \eta_l(T)c_{liT},
\end{aligned}$$

we derive

$$\text{«}c_{li}(\text{T})\text{»} = \lambda\text{«}c_{li}(\text{T-1})\text{»}\frac{\eta_l(T)}{\eta_l(T-1)} + \eta_l(T)c_{liT}. \tag{14}$$

On the other hand we have

$$\begin{aligned}
\eta_l(T-1)^{-1} &= \left[\sum_{k=1}^{T-1}\lambda^{T-1-k}\right] \\
&= \left[\sum_{k=1}^{T}\lambda^{T-k} - 1\right]\frac{1}{\lambda} \\
&= \left[\eta_l(T)^{-1} - 1\right]\frac{1}{\lambda},
\end{aligned}$$

so the first implication is

$$\eta_l(T) = \frac{1}{\frac{\lambda}{\eta_l(T-1)+1}}.$$

Exploiting the previous equality, rewrite (14) as

$$\text{«}c_{li}(\text{T})\text{»} = \text{«}c_{li}(\text{T-1})\text{»}(1-\eta_l(T)) + \eta_l(T)c_{liT}.$$

□

For $T \neq \sup\Delta_l(T)$, we set

$$\begin{aligned}
\text{«}c_{li}(\text{T})\text{»} &= \text{«}c_{li}(\text{T-1})\text{»}, \\
\eta_l(T) &= \frac{\eta_l(T-1)}{\lambda(T)}, \\
{N^{\lambda}}_l(T) &= \lambda(T){N^{\lambda}}_l(T-1).
\end{aligned}$$

Computation of $\lambda(T)$ in the FB model

$$\lambda(T) = \frac{1}{1 + \dfrac{AC}{B}}, \text{ with } \quad \frac{AC}{B} = \frac{\hat{p}_{i^*0}(\nu_{lT-1} + 1)}{\nu_{lT-1}\hat{p}_{i^*lT-1} + 1}.$$

Proof. $\lambda(T)$ is defined as

$$\lambda(T) = P(J_T = 0 \,|\, \underline{\mathbf{c}}_{\mathbf{lT}}).$$

Let $P_{lT}(\mathbf{p_{lT}})(J_T = 0)$ and $P_{lT}(\mathbf{p_{lT}})(J_T = 1)$ denote the posterior probability distribution after *No jump* and after *Jump*, respectively. We take the prior probability that a jump occurred to be 1/2. This leads to

$$P(J_T = 0 | \underline{\mathbf{c}}_{\mathbf{lT}}) =$$

$$\frac{1/2 \displaystyle\int_\Theta l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) P_{lT}(\mathbf{p_{lT}})(J_T = 0)\, d\mathbf{p_{lT}}}{1/2 \displaystyle\int_\Theta l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) P_{lT}(\mathbf{p_{lT}})(J_T = 0)\, d\mathbf{p_{lT}} + 1/2 \displaystyle\int_\Theta l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) P_{lT}(\mathbf{p_{lT}})(J_T = 1)\, d\mathbf{p_{lT}}}$$

$$= \frac{1}{1 + \dfrac{\displaystyle\int_\Theta l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) P_{lT}(\mathbf{p_{lT}})(J_T = 1)\, d\mathbf{p_{lT}}}{\displaystyle\int_\Theta l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) P_{lT}(\mathbf{p_{lT}})(J_T = 0)\, d\mathbf{p_{lT}}}},$$

$$\text{with} \quad P_{lT}(\mathbf{p_{lT}})(J_T = 1) = P_0(\mathbf{p_{lT}}),$$
$$P_{lT}(\mathbf{p_{lT}})(J_T = 0) = P_{lT/T}(\mathbf{p_{lT}}) = \frac{l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) P_{lT-1}(\mathbf{p_{lT}})}{\int_\Theta l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) P_{lT-1}(\mathbf{p_{lT}})\, d\mathbf{p_{lT}}}.$$

Therefore we have that

$$\lambda(T) = \frac{1}{1 + \dfrac{AC}{B}}, \quad \text{with} \quad \begin{cases} A = \displaystyle\int_\Theta l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) P_0(\mathbf{p_{lT}})\, d\mathbf{p_{lT}}, \\ B = \displaystyle\int_\Theta l(\mathbf{c_{lT}}|\mathbf{p_{lT}})^2 P_{lT-1}(\mathbf{p_{lT}})\, d\mathbf{p_{lT}}, \\ C = \displaystyle\int_\Theta l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) P_{lT-1}(\mathbf{p_{lT}})\, d\mathbf{p_{lT}}. \end{cases}$$

Now note the following.

- $l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) = \displaystyle\prod_{i=1}^{3} p_{liT}^{c_{liT}}$;
- the density $P_0(\mathbf{p_{lT}})$ is known from (8), p 42;

- the posterior belief $P_{l\,T}(\mathbf{p_{l\,T}})$ is

$$P_{l\,T-1}(\mathbf{p}) = \left[\frac{\prod_{i=1}^{3}\Gamma(\nu_{l\,T-1}\hat{p}_{i\,l\,T-1})}{\Gamma(\nu_{l\,T-1})}\right]^{-1} \prod_{i=1}^{3} p_i^{(\nu_{l\,T-1}\hat{p}_{i\,l\,T-1}-1)} \quad \delta_{\Theta}(\mathbf{p}),$$

with $\mathbf{p} = (p_1, p_2, p_3)'$, $\Gamma(\nu_{l\,T-1}\hat{p}_{i\,l\,T-1}) = \int_{\Theta} x^{\nu_{l\,T-1}\hat{p}_{i\,l\,T-1}} \mathrm{e}^{-x}\, dx$,

and the value of $\nu_{l\,T-1}$ and $\hat{p}_{i\,l}$ is given in the Proposition (p. 47).

Develop the expression for A; this gives

$$A = \int_{\Theta} \prod_{i=1}^{3} p_{liT}^{c_{liT}} \frac{\prod_{i=1}^{3} p_{liT}^{\nu_0\hat{p}_{i0}-1}}{\frac{\prod_{i=1}^{3}\Gamma(\nu_0\hat{p}_{i0})}{\Gamma(\nu_0)}}\, d\mathbf{p_{lT}}.$$

Then noting that $\int_{\Theta} \prod_{i=1}^{3} p_{liT}^{c_{liT}+\nu_0\hat{p}_{i0}-1}\, d\mathbf{p_{lT}} = \frac{\prod_{i=1}^{3}\Gamma(c_{liT}+\nu_0\hat{p}_{i0})}{\Gamma(\sum_{i=1}^{3} c_{liT}+\nu_0)}$, we can rewrite A as follows:

$$A = \frac{\Gamma(\nu_0)\prod_{i=1}^{3}\Gamma(c_{liT}+\nu_0\hat{p}_{i0})}{\Gamma(\underbrace{\sum_{i=1}^{3} c_{liT}}_{1}+\nu_0)\prod_{i=1}^{3}\Gamma(\nu_0\hat{p}_{i0})}.$$

Let i^* refer to the realized component of the count vector at time $T-1$.[53] We can further simplify the previous expression, using the equality

[53] For example, suppose that location l delivered the loss outcome at time $T-1$; then $c_{l\,T-1} = (1, 0, 0)$, and i^* is equal to 1.

$\Gamma(x+1) = x\Gamma(x)$, and the fact that $c_{liT} = 0,\ \forall i \neq i^*$, while $c_{li^*T} = 1$:

$$A = \frac{\Gamma(\nu_0)\prod_{i=1}^{3}\Gamma(c_{liT}+\nu_0\hat{p}_{i0})}{\Gamma(1+\nu_0)\prod_{i=1}^{3}\Gamma(\nu_0\hat{p}_{i0})} = \frac{\prod_{i=1}^{3}\Gamma(c_{liT}+\nu_0\hat{p}_{i0})}{\nu_0\prod_{i=1}^{3}\Gamma(\nu_0\hat{p}_{i0})}.$$

$$\text{Hence}\quad A = \frac{\Gamma(1+\nu_0\hat{p}_{i^*0})\prod_{i\neq i^*}\Gamma(\nu_0\hat{p}_{i0})\Gamma(1+\nu_0\hat{p}_{i^*0})}{\nu_0\prod_{i=1}^{3}\Gamma(\nu_0\hat{p_{i0}})}$$

$$= \frac{1}{\nu_0}\frac{\Gamma(1+\nu_0\hat{p}_{i^*0})}{\Gamma(\nu_0\hat{p}_{i^*0})}$$

$$= \hat{p}_{i^*0}.$$

C is calculated in exactly the same fashion, which gives

$$C = \frac{\Gamma(\nu_{lT-1})\prod_{i=1}^{3}\Gamma(c_{liT}+\nu_{lT-1}\hat{p}_{ilT-1})}{\Gamma(\underbrace{\sum_{i=1}^{3}c_{liT}}_{1}+\nu_{lT-1})\prod_{i=1}^{3}\Gamma(\nu_{lT-1}\hat{p}_{ilT-1})} = \hat{p}_{i^*lT-1}.$$

The calculation of B is similar:

$$B = \frac{\Gamma(\nu_{lT-1})\prod_{i=1}^{3}\Gamma(2c_{liT}+\nu_{lT-1}\hat{p}_{ilT-1})}{\Gamma(2\underbrace{\sum_{i=1}^{3}c_{liT}}_{1}+\nu_{lT-1})\prod_{i=1}^{3}\Gamma(\nu_{lT-1}\hat{p}_{ilT-1})}.$$

$$\text{Also,}\quad \Gamma(2+\nu_{lT-1}) = \Gamma(1+(1+\nu_{lT-1})) = (1+\nu_{lT-1})\Gamma(1+\nu_{lT-1})$$
$$= (1+\nu_{lT-1})\nu_{lT-1}\Gamma(\nu_{lT-1}).$$

$$\text{And}\quad \prod_{i=1}^{3}\Gamma(2c_{liT}+\nu_{lT-1}p_{il\hat{T}-1}) = \prod_{i\neq i^*}\Gamma(\nu_{lT-1}\hat{p}_{ilT-1})\Gamma(2+\nu_{lT-1}\hat{p}_{i^*lT-1}).$$

Therefore B simplifies to

$$B = \frac{1}{(1+\nu_{lT-1})\nu_{lT-1}}\frac{\Gamma(2+\nu_{lT-1}\hat{p}_{ilT-1})}{\Gamma(\nu_{lT-1}\hat{p}_{i^*lT-1})}.$$

$$\text{So} \quad \frac{AC}{B} = \frac{\hat{p}_{i^*0}(\nu_{lT-1}+1)}{\nu_{lT-1}p_{i^*lT-1}+1}.$$

□

Estimated outcome probability in the FB model

$$\bar{\mathbf{p}}_{\mathbf{lT}} = \int_{\Theta} \mathbf{p_{lT}}\, P_{lT}(\mathbf{p_{lT}})\, d\mathbf{p_{lT}}$$

$$= E_{P_{lT}}(\mathbf{p_{lT}})$$

$$= E_{\mathrm{Dir}(\hat{\mathbf{p}}_{\mathbf{lT}},\nu_{\mathbf{lT}})}(\mathbf{p_{lT}})$$

$$= \hat{\mathbf{p}}_{\mathbf{lT}}$$

$$= \frac{N^{\lambda}(T)\langle\!\langle \mathbf{c_l}(\mathbf{T})\rangle\!\rangle + \nu_0\, \hat{p}_0}{N^{\lambda}(T)+\nu_0}.$$

Under the specification $\nu_{\mathbf{0}} = (1,1,1)$ and $\hat{p}_{i\,0} = 1/3 \;\; \forall i = 1,2,3$, the expression provided in the core text then obtains.

Discriminatory power of our design

To test whether the courses of action prescribed by a Bayesian disagreed with the ones prescribed by a reinforcement learner, we used the purely greedy rule to generate choice. That is, for each subject, we compared the choice made by a purely greedy Bayesian learner (referred to as "Bayesian algorithm") to the one made by a purely greedy reinforcement learner (referred to as "RL algorithm"), and this at each trial. If the Bayesian algorithm (HB or FB) prescribes the same action as the RL algorithm too often, making a horse race between the two makes little sense. We reasoned that the predictions of the algorithms would be more correlated than those of the models – because the logit rule is stochastic, so even though the models perfectly agree in that they have the same choice probabilities, it is still expected that the choices they generate differ. In other words, we wanted the disagreement between the back-seat drivers to come from learning, not from the latent randomness of the choice rule (see Section 5.1 for an explanation of the concept of a back-seat driver). So, each trial, before the real player made her choice, there was

one choice prediction/recommandation coming from each algorithm. These trials in which the algorithms selected a different location were classified as "conflicting" (the algorithms "disagreed"). For each subject, we looked at the percentage of conflicting trials during her play.[54]

The top graph in Fig. 14 shows the distribution across subjects of the fraction of conflicting trials between the HB and RL algorithms. Specifically, this graph shows that Bayesian (HB or FB) learning and reinforcement learning lead to different choices. Indeed, the HB and RL algorithms often disagreed: for as many as 50 subjects (out of 62), the percentage of trials where HB and RL disagreed is at least 40%. Similarly, for 57 subjects (out of 62), the percentage of trials in which FB and RL disagreed is at least 20%. So, for almost all the subjects, the HB and RL algorithms disagreed about which location had the highest value at least 40% of the time; the FB and RL algorithms disagreed at least 20% of the time.

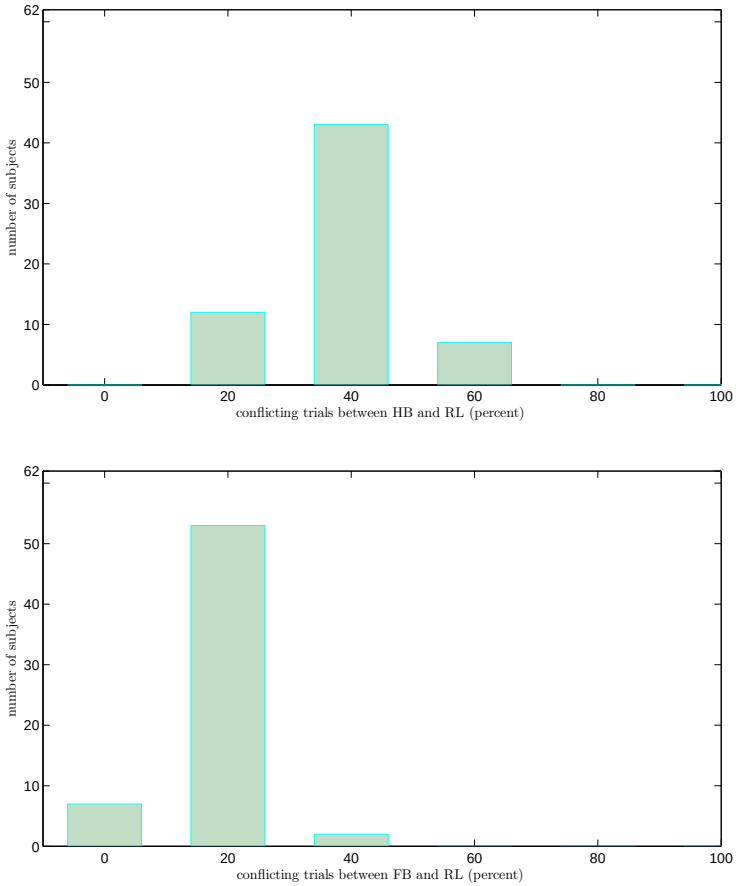


Figure 14: **The top graph shows the distribution (62 subjects) of the percentage of trials in which the HB and RL algorithms disagreed about the location the subject should select next. The bottom graph shows similar disagreement, this time between the FB and RL algorithms.**

[54] Remember that the number of trials per play varied between subjects from 450 to 600, with an average value of 500.

Bayesian model comparison

Here we report negative log likelihoods penalized for model complexity. The BIC value of a model is $-2\ln LL + \ d\ln(n)$, where LL stands for the log likelihood of the model with the fitted parameter(s), d is the number of free parameters, and n is the sample size. Specifically, we looked at the difference between the BIC value of the competing models. Indeed, when the sample size is large enough, this difference is a good approximation of (twice times) the logarithm of the *Bayes factor* (Kass and Raftery, 1995), which measures the relative success of each model at predicting the data (the posterior odds). To see this, let $P(D|H_{Bayes})$ be the probability of seeing the actual data, under the hypothesis that the Bayesian model is true. $P(D|H_{RL})$ is the similar probability under the hypothesis that the RL model is true. The Bayes factor is

$$\frac{P(D|H_{Bayes})}{P(D|H_{RL})} = \underbrace{\frac{P(H_{Bayes}|D)}{P(H_{RL}|D)}}_{posterior\ odds} * \frac{P_{\text{prior}}(H_{Bayes})}{P_{\text{prior}}(H_{RL})}.$$

So the Bayes factor is equal to the *posterior odds* because here the prior probability for each model is the same.

In Fig. 15 (p. 55) we plotted, for every single subject, the HB model's BIC against the RL model's BIC (top graph), and similarly we plotted, for each subject, the FB model's BIC against the RL model's BIC (bottom graph). When the data are below the 45 degree line, this means that BIC is higher for the RL model, so the Bayes factor of the Bayesian model against the RL model is positive – that is, there is positive evidence for the Bayesian model against the RL model. There is positive evidence for the FB model against the RL model 81% of the time, and for the HB model against the RL model 89% of the time.

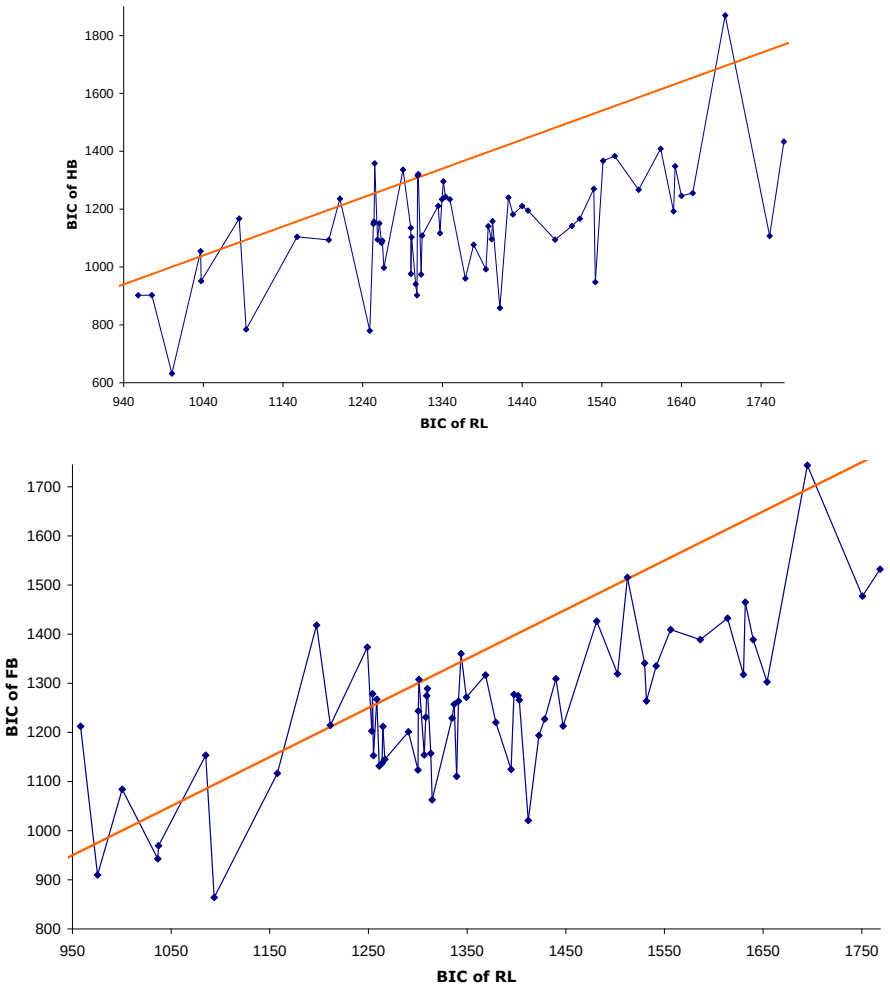


Figure 15: Comparative BIC criteria for the HB and RL models (top graph), and for the FB and RL models (bottom graph). Each data point provides the estimated BIC for one subject (500 trials on average per subject). When a data point is below the 45 degree line, the Bayesian model fits better.

Experimental protocol

Each experimental session started with 30 minutes of instructions in the lab. We invited the subjects to read carefully the pages "The Game" and "How to play" on the Boardgame website (http://decisions.epfl.ch/Boardgame1/) – login: boardgame; password: brdgmesecure. We ensured that the subjects had well understood the task by having them fill an MCQ questionnaire about the rules of the game before they started playing. They played during 30 minutes. After play, they filled a debriefing questionnaire that allowed us to understand how they approached the task.

References

Masanao Aoki. *State Space Modeling of Time series*. Springer-Verlag, 1987.

W. Brian Arthur. Designing economic agents that act like human agents: A behavioral approach to bounded rationality. *American Economic Review*, 81:353–359, 1991.

Robert Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. (Princeton, NJ: Princeton University Press, 1997.

Ole E. Barndorff-Nielsen and Neil Shephard. Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, vol. 4(1):pp. 1–30, 2006. 2

T. Behrens, M. W. Woolrich, M. E. Walton, and M. Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9): 1214–21, 2007.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics, 1980.

Gary Bishop and Greg Welch. An introduction to the kalman filter. In *ACM SIGGRAPH*, 2001.

Peter Boasserts. *The Paradox of Asset Pricing*. Princeton University Press, 2002.

Peter Bossaerts and Kerstin Preuschoff. Adding prediction risk to the theory of reward learning. *Annals of the New York Academy of Sciences*, 1104: 135–146, 2007.

Colin Camerer and Teck-Hua Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 67, 1999. 30

Colin Camerer and Teck-Hua Ho. Ewa learning in games: Probability form, heterogeneity, and time variation. *Journal of Mathematical Psychology*, 42: 305–326, 1998.

Gary Charness and Dan Levin. When optimal choices feel wrong: A laboratory study of bayesian updating, complexity, and affect. *The American Economic Review*, 95:1300–1309, 2005. 6, 36, 37

J. M. Coates and J. Herbert. Endogenous steroids and financial risk taking on a london trading floor. *Proc Natl Acad Sci U S A*, 105(16):6167–6172, Apr 2008. doi: 10.1073/pnas.0704025105. URL http://dx.doi.org/10.1073/pnas.0704025105. 37

Aaron C Courville, Nathaniel D Daw, and David S Touretzky. Bayesian theories of conditioning in a changing world. *Trends Cogn Sci*, 10(7):294–300, Jul 2006. doi: 10.1016/j.tics.2006.05.004. URL http://dx.doi.org/10.1016/j.tics.2006.05.004.

D. Dacunha-Castelle. *Chemins de l'aléatoire*. Flammarion, 1996.

Nathaniel D Daw and Kenji Doya. The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16:199–204, 2006.

Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8:1704–1711, 2005.

Nathaniel D Daw, John P O'Doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, Jun 2006a. doi: 10.1038/nature04766. URL http://dx.doi.org/10.1038/nature04766.

Nathaniel D. Daw, John P. O'Doherty, Peter Dayan, Ben Seymour, and Raymond J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441:876–879, 2006b. 36

Peter Dayan and Theresa Long. Statistical models of conditioning. In *Proceedings of the conference on Advances in neural information processing systems 10*, 1997.

Bruno de Finetti. Bayeisanism: Its unifying role for both the foundations and applications of statistics. *International Statistical Review*, 42:117–130, 1974.

P. Diaconis and D. Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986.

Kenji Doya. Metalearning and neuromodulation. *Neural Networks*, 15:495–506, 2002.

John Duffy. *Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics*, chapter Agent-Based Models and Human Subject Experiments. North-Holland, 2006. 13

R. Elliott and R. J. Dolan. Activation of different anterior cingulate foci in association with hypothesis testing and response selection. *Neuroimage*, 8 (1):17–29, Jul 1998. doi: 10.1006/nimg.1998.0344. URL http://dx.doi.org/10.1006/nimg.1998.0344.

D. Ellsberg. Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, 75:643–669, 1961.

Larry G. Epstein and Martin Schneider. Recursive multiple-priors. *Journal of Economic Theory*, 113:1–31, 2003. 12

Larry G. Epstein and Martin Schneider. Learning under ambiguity. *Review of Economic Studies*, 74:1275–1303, 2007. 4

Leon Festinger. *A Theory of Cognitive Dissonance*. Evanston, 1957.

C. R. Gallistel, Terence A. Mark, Adam Philip King, and P.E. Latham. The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *Journal of Experimental Psychology*, 27:354–372, 2001.

Paolo Ghirardato and Massimo Marinacci. Ambiguity made precise: A comparative foundation. *Journal of Economic Theory*, 102:251–289, 2002. 12

Itzhak Gilboa and David Schmeidler. Maxmin expected utlity with non-unique prior. *Journal of Mathematical Economics*, 18:141–153, 1989. 12

J Gittins and D Jones. *Progress in Statistics*. Amsterdam: North-Holland, 1974. 11

D Grether. Testing bayes rule as a descriptive model: The representativeness heuristic. *Quaterly Journal of Economics*, 95:537–557, 1992. 5

Alan N Hampton, Peter Bossaerts, and John P O'Doherty. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J Neurosci*, 26(32):8360–8367, Aug 2006. doi: 10.1523/JNEUROSCI.1010-06.2006. URL http://dx.doi.org/10.1523/JNEUROSCI.1010-06.2006.

Michael E. Hasselmo. Neuromodulation: acetylcholine and memory consolidation. *Trends in Cognitive Sciences*, 3:351–359, 1999.

Chip Heath and Amos Tversky. Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4:5–28, 1991.

R. J. Herrnstein. On the law of effect. *Journal of Experimental Analysis of Behavior*, 13:243–266, 1970. 13

Jaakko Hintikka. Unknown probabilities, bayesianism, and de finetti's representation theorem. *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1970:325–341, 1970.

Junichiro Hirayama, Junichiro Yoshimoto, and Shin Ishii. Bayesian representation learning in the cortex regulated by acetylcholine. *Neural Netw*, 17(10):1391–1400, Dec 2004. doi: 10.1016/j.neunet.2004.06.006. URL http://dx.doi.org/10.1016/j.neunet.2004.06.006.

Ming Hsu, Meghana Bhatt, Ralph Adolphs, Daniel Tranel, and Colin F Camerer. Neural systems responding to degrees of uncertainty in human decision-making. *Science*, 310(5754):1680–1683, Dec 2005. doi: 10.1126/science.1115327. URL http://dx.doi.org/10.1126/science.1115327. 37

Xin Huang and George Tauchen. The relative contribution of jumps to total price variance. *Journal of Financial Econometrics*, vol. 3(4),:pp. 456–499, 2005. 2

Shin Ishii, Wako Yoshida, and Junichiro Yoshimoto. Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural Netw*, 15(4-6): 665–687, 2002.

Eric J. Johnson, Colin Camerer, Sankar Sen, and Talia Rymond. Detecting failures of backward induction: Monitoring information search in sequential bargaining. *Journal of Economic Theory*, 104:16–47, 2002. 6, 37

Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3:430–454, 1972. 5

Sham Kakade and Peter Dayan. Dopamine: generalization and bonuses. *Neural Networks*, 15:549–559, 2002.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of The American Statistical Association*, 90:773–795, 1995. 54

P. Klibanoff, M. Marinacci, and S. Mukerji. Recursive smooth ambiguity preferences. working paper, MEDS, Kellogg School of Management. 12

Etienne Koechlin, Chrystèle Ody, and Frédérique Kouneiher. The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648): 1181–1185, Nov 2003. doi: 10.1126/science.1088545. URL http://dx.doi.org/10.1126/science.1088545.

Camelia M Kuhnen and Brian Knutson. The neural basis of financial risk taking. *Neuron*, 47(5):763–770, Sep 2005. doi: 10.1016/j.neuron.2005.08.008. URL http://dx.doi.org/10.1016/j.neuron.2005.08.008.

Andrew W Lo and Dmitry V Repin. The psychophysiology of real-time financial risk processing. *Journal of Cognitive Neuroscience*, 14:323–339, 2002. 37

Mark J Machina and David Schmeidler. A more robust definition of subjective probability. *Econometrica*, 60:745–780, 1992. 12

David J. C. MacKay. *Information Theory, Inference, and learning algorithms*. 2003. 17

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman, 1982.

D. McFadden. *Frontiers of Econometrics*, chapter Conditional Logit Analysis of Qualitative Choice Behavior. Academic Press, 1974. 11

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10:6–38, 1995.

Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49:267–290, 2002.

Camelia M.Kuhnen and Brian Knutson. The influence of affect on beliefs, preferences and financial decisions. 36, 37

John F. Muth. Rational expectations and the theory of price movements. *Econometrica*, 29:315–335, 1961.

Yael Niv, Jeffrey A. Edlund, Peter Dayan, and John P. O'Doherty. Neural prediction errors reveal risk-sensitive learning.

John O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–454, Apr 2004. doi: 10.1126/science.1094285. URL http://dx.doi.org/10.1126/science.1094285.

Elise Payzan and Peter Bossaerts. Hiearchical versus forgetting bayes: probabilistic learning and choice under paramount uncertainty.

J. M. Pearce and G. Hall. A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol Rev*, 87(6):532–552, Nov 1980. 29

Russell A. Poldrack. Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10, 2006.

Kerstin Preuschoff, Steven R Quartz, and Peter Bossaerts. Human insula activation reflects risk prediction errors as well as risk. *J Neurosci*, 28(11): 2745–2752, Mar 2008. doi: 10.1523/JNEUROSCI.4286-07.2008. URL http://dx.doi.org/10.1523/JNEUROSCI.4286-07.2008.

A. Quinn and M. Kárný. Learning for non-stationary dirichlet processes. *International Journal of Adaptive Control and Signal Processing*, 21:827–855, 2007. 42

M. B. Zarrop R. Kulhavý. On a general concept of forgetting. *International Journal of Control*, 58:905–924, 1993.

Michael Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9:185–202, 1974. 20, 26

Ueli Rutishauser, Adam N Mamelak, and Erin M Schuman. Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex. *Neuron*, 49(6):805–813, Mar 2006. doi: 10.1016/j.neuron.2006.02.015. URL http://dx.doi.org/10.1016/j.neuron.2006.02.015.

M. Sato. Online model selection based on the variational bayes. *Neural Computation*, 13:1649–1681, 2001. 42

David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57:571–587, 1989.

W . Schultz. Getting formal with dopamine and reward. *Neuron*, 36:241 – 263, 2002.

Wolfram Schultz, Peter Dayan, and P. Read Montague. A neural substrate of prediction and reward. *Science*, 275:1593 – 1599, 1997. 29

Reinhard Selten. Evolution, learning, and economic behavior. *Games and Economic Behavior*, 3:3–24, 1991.

Ben Seymour, John P O'Doherty, Peter Dayan, Martin Koltzenburg, Anthony K Jones, Raymond J Dolan, Karl J Friston, and Richard S Frackowiak. Temporal difference models describe higher-order learning in humans. *Nature*, 429(6992):664–667, Jun 2004. doi: 10.1038/nature02581. URL http://dx.doi.org/10.1038/nature02581.

Ben Seymour, Nathaniel Daw, Peter Dayan, Tania Singer, and Ray Dolan. Differential encoding of losses and gains in the human striatum. *J Neurosci*, 27 (18):4826–4831, May 2007. doi: 10.1523/JNEUROSCI.0400-07.2007. URL http://dx.doi.org/10.1523/JNEUROSCI.0400-07.2007.

Ladan Shams, Wei Ji Ma, and Ulrik Beierholm. Sound-induced flash illusion as an optimal percept. *Neuroreport*, 16(17):1923–1927, 2005. 33

Herbert Simon. Behavioral model of rational choice. *The Quaterly Journal of Economics*, LXIX, 1955.

Bryan A. Strange, Andrew Duggins, William Penny, Raymond J. Dolan, and Karl J. Friston. Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Networks*, 18:225–230, 2005.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press, 1998. 29

George Tauchen and Hao Zhou. Realized jumps on financial markets and predicting credit spreads. Technical report, Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series, 2006. 2

Edward Thorndike. Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review Mdddh Supplement*, 4: 1–109, 1898.

L. Thurstone. A law of comparative judgment. *Psychological Review*, 34: 273–286, 1927.

Amos Tversky and Daniel Kahneman. Belief in the law of small numbers. *Psychological Bulletin*, 76:105–110, 1971. 5

AR Wagner and RA Rescorla. *Inhibition and Learning*, chapter Inhibition in Pavlovian conditioning: application of a theory. 1972.

Wako Yoshida and Shin Ishii. Resolution of uncertainty in prefrontal cortex. *Neuron*, 50(5):781–789, Jun 2006. doi: 10.1016/j.neuron.2006.05.006. URL http://dx.doi.org/10.1016/j.neuron.2006.05.006.

Angela J Yu and Peter Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692, May 2005. doi: 10.1016/j.neuron.2005.04.026. URL http://dx.doi.org/10.1016/j.neuron.2005.04.026. 40