RFS Advance Access published November 7, 2014

Learning About Unstable, Publicly Unobservable Payoffs

Elise Payzan-LeNestour
UNSW Australia Business School, University of New South Wales, Sydney

Peter Bossaerts
David Eccles School of Business, University of Utah, Salt Lake City and
Faculty of Business and Economics, University of Melbourne, Melbourne

Neoclassical finance assumes that investors are Bayesian. In many realistic situations, Bayesian learning is challenging. Here, we consider investment opportunities that change randomly, while payoffs are observable only when invested. In a stylized version of the task, we wondered whether performance would be affected if one were to follow reinforcement learning principles instead. The answer is a definite yes. When asked to perform our task, participants overwhelmingly learned in a Bayesian way. They stopped being Bayesians, though, when not nudged into paying attention to contingency shifts. This raises an issue for financial markets: who has the incentive to nudge investors? (*JEL* C91, D83, D87, G02, G11)

We study learning when the agent cannot observe the payoffs on the available investment opportunities unless she actively invests in them. In abstract terms, this problem is known as a multi-armed bandit problem (Rothschild 1974). For instance, the risk/reward data of most structured financial instruments that trade exclusively over the counter are not publicly available in an actionable time frame. The same is true for the vast majority of investment opportunities in asset classes such as private equity, foreign direct investment, and natural resources

An earlier version of this article circulated under the title "Bayesian Learning in Unstable Settings: Experimental Evidence Based on the Bandit Problem." We thank the Editor, David Hirshleifer, for very helpful suggestions. E. Payzan-LeNestour also thanks Renée Adams, Andrew Caplin, Andres Carvajal, Pierre-André Chiappori, Claire Célérier, Giorgio Coricelli, Rajna Gibson, Jacob Goeree, Michel Habib, Peter Hammond, Byoung-Hyoun Hwang, Cami Kuhnen, Michelle Lowry, Andreas Ortmann, Sébastien Pouget, Antonio Rangel, seminar participants at Caltech, the University of Warwick, INSEAD, the Institute for Empirical Research in Economics, the University of Geneva, the University of Melbourne, New York University, as well as job market seminar participants at UCLA Anderson, Kellogg, Chicago Booth, Toulouse School of Economics, Duke Fuqua, McGill, and Rotman, for very useful comments and suggestions. On the practitioner side, very special thanks to Lionnel Pradier, recently retired from his position as Head of Quantitative Research at JP Morgan and currently "in residence" at the UNSW Australia Business School, for countless enlightening discussions on this paper and for his invaluable help in validating the implications of this study for finance practitioners. The usual caveat applies. Chen Feng and Frédéric Ollier provided outstanding research assistance. We gratefully acknowledge financial support from the Swiss Finance Institute, NCCR FINRISK of the Swiss National Science Foundation, and The University of New South Wales. Send correspondence to E. Payzan-LeNestour, School of Banking and Finance, UNSW Business School, 2052 Sydney UNSW; telephone: +61 2 9385 4273. E-mail: elise@unsw.edu.au.

© The Author 2014. Published by Oxford University Press on behalf of The Society for Financial Studies.
All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com.
doi:10.1093/rfs/hhu069

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

exploration licenses. One has to actually invest in the assets to learn about their risk/return tradeoff. Such learning is challenging because the risk/reward profiles of many—if not most—investments are unstable in the sense that they change occasionally and in a sudden way. Prior field research has identified the presence of such jumps or regime shifts as a major feature of the return-generating process (see Ang and Timmermann 2011, for a survey).[1] Market participants are aware that such jumps occur, and have to identify them as they emerge. Formally, this situation is referred to as that of a "restless" bandit (Whittle 1988).

Here we report results from experiments on human learning in a six-armed restless bandit task. Each arm represents an investment opportunity that the agent needs to try out to learn about its expected value. The latter "jumps" regularly through time owing to regime shifts affecting the payoff probabilities.

To model subject behavior, we decomposed the decision problem into two components, information acquisition and information processing or learning, as suggested in Rangel, Camerer, and Montague (2008). Information acquisition relates to the exploitation/exploration trade-off—that is, the desire to exploit the options that are currently deemed best, in competition with the motive to explore new opportunities that might turn out to be better. Learning concerns the process whereby new facts are integrated with old beliefs. Recently, neuroscientists (e.g., O'Doherty et al. (2004) and Behrens et al. (2007)) have shown that the neural pathways underlying these two cognitive processes are dissociable, raising the possibility that their nature is entirely different (e.g., one is based on heuristics, while the other is optimal).

Gans et al. (2007) have provided insight in the type of behavior displayed in the information acquisition stage of a standard bandit problem, where contingencies never change. This study reports that the logit rule describes exploratory behavior. The logit rule, however, is not optimal in the standard bandit problem. This is perhaps not surprising, because the standard bandit problem is rather contrived: the optimal policy requires one to stop exploring at a certain point, way before acquiring perfect information, but that makes sense only if the decision maker is absolutely sure that the contingencies will *never* change. We instead study a more realistic case, the restless bandit. There, logit has provided the best fit as well (Daw et al. 2006).[2] If optimality is understood as "Bellman optimality" (i.e., the policy satisfies the Bellman equations), then logit is unlikely to be optimal, because the parameters of this rule do not change with beliefs. Unfortunately, we do not know what the true optimal policy is for the case of the restless bandit, because the Bellman equations are intractable. Hence, we do not know how inferior the logit rule is. Still, as we discuss in

[1] Such instability is thought to be at the origin of fat tails in return distributions; e.g., Mandelbrot (1957) and Gabaix et al. (2003).

[2] One could conjecture that choice in the standard bandit is the same as in the restless bandit because subjects do not believe that contingencies will never change and, hence, perceive that bandit as possibly restless.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

more detail later, the logit rule is optimal in a weaker sense (i.e., weaker than Bellman optimality). We will also show that, among a number of alternative ways to model choice, the logit rule provides the best fit for our data, confirming the findings in Daw et al. (2006).

Here, we focus on learning rather than information acquisition. We contrasted Bayesian learning, which estimates the outcome probabilities of each arm, with reinforcement (or "adaptive") learning, a far simpler and straightforward learning approach, which captures the idea that one sticks to given choices as long as they generate rewards; otherwise, one switches (Charness and Levin 2005). Reinforcement learning has been at the core of learning in games (Erev and Roth 1998), recently enriched with the idea of counterfactuals based on fictitious play (Camerer and Ho 1999). It has also been popular in experimental psychology and experimental finance (Pouget 2007), and it has solid neurobiological foundations (Schultz et al. 1997). Yet, and despite the complexity of our bandit task, we found strong evidence in favor of Bayesian learning.

Participants were told in the task instructions that the option values would jump regularly during the experiment. Notably, participants fell back to boundedly rational learning in a follow-up experimental treatment in which they were not provided with this information. Taken together, our experimental results thus show that when people know that regime shifts do and will occur, they can detect them as they emerge, and correctly estimate the outcome probabilities in that case. But when the same people are not warned that their environment is unstable, they fail to detect the regime shifts, and their estimation of the option values is correspondingly suboptimal. This suggests that nudging (Thaler and Sunstein 2009) market participants into looking out for temporal swings in payoff distributions is a prerequisite for the emergence of optimal learning.

We explicitly instructed participants about the nonstationarity of the options in the game. One may reasonably wonder whether this makes reinforcement learning a strawman. It does not, as long as one allows the learning rate to be close to 1 (we shall do so). This way, the algorithm discounts observations further in the past, and nonstationarity in the outcome-generating process is accounted for. The difference with Bayesian updating then is only a matter of whether the participants explicitly tracked outcomes to detect changes in the contingencies. Under Bayesian updating, the decision maker has a model of how contingencies change, which she applies to infer from outcomes whether and when changes occur. In contrast, under reinforcement learning, the decision maker is agnostic about how contingencies change, though she still acknowledges that they may change, by fixing the learning rate to be very high (close to 1). A higher learning rate reflects her opinion that changes are more frequent. See Behrens et al. (2007) and Jepma and Nieuwenhuis (2011).

Our findings are important inasmuch as multi-armed bandit problems abound in the practice of finance. The importance of investment in opportunities with

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

returns that are not publicly observable has increased in recent times because of a shift in equity holding from publicly traded shares to privately managed stakes ("private equity"). There, the performance of individual investments is known only to the investors, and remains opaque to those who did not invest. According to the Private Equity Growth Council, in 2011 alone, private equity investment in the United States is estimated to have been over $140 billion. Venture capital falls in the same category: until the public offering of shares, the performance of individual projects is observable only to the venture capitalists who invested in the project. While annual new venture capital activity is rather modest (around $27 billion in 2012, according to the National Venture Capital Association), it is generally considered to be a crucial engine for long-term economic growth.

Bank lending to firms is another instance of the bandit problem in finance: the potential performance of a loan to an individual firm cannot generally be assessed unless the loan is put in place; only then can the lender have access to verifiable data. According to the Board of Governors of the Federal Reserve System, bank lending to industrial and commercial firms stood at over $1.5 trillion at the end of 2012. Thus, financial applications that take on the structure of multi-armed bandit problems are voluminous, and when not voluminous, they nevertheless are important for economic well-being.[3]

There is a certain sense in which learning in multi-armed bandit problems is actually simple. The fact that one cannot observe the outcomes on the unchosen options means that the values of these options cannot be updated, unless payoffs correlate across options. In contrast, when payoffs on all options are publicly observable, the decision maker faces a complex multidimensional learning problem that readily becomes intractable. As a matter of fact, human decision makers resort to deliberately ignoring the outcomes of many, if not most, options. Such rational inattention (Sims 2003, 2006) actually converts problems with fully observable payoffs into bandit problems. Hence the relevance of the study of bandit problems extends even to traditional investments in publicly traded securities such as stocks and bonds.

The restless feature of our bandit task parallels the instability of financial decision-making. For instance, in bank lending, jumps in the payoff probabilities reflect shifts in the creditworthiness of the borrower. In private equity and venture capital, they may correspond to shifts in the demands of the goods or services produced. The occurrence of jumps does complicate learning substantially. When observing a streak of low payoffs, the decision maker has to determine whether she should use this to update her current estimate of the value of the option at hand, or whether the streak signifies a break with the past,

[3] When investors only observe relevant information at the moment they decide to invest, information cascades may ensue. This occurs when the information investors receive concerns the moves of others (whether others invested). See Chari and Kehoe (2004). In our setting, all information will be produced exogenously; the agent's decision only concerns whether to collect the (exogenous) information through investing.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

in which case she should discard her current estimate and start updating from scratch.

Strikingly, it is the occurrence of jumps in the payoff probabilities that makes the question of whether learning is Bayesian crucial. Without the (hidden) jumps at the heart of a restless bandit problem, there is no hidden "state" that the decision maker needs to infer from observed data, and without such inference, Bayesian updating effectively works like reinforcement learning (future predictions depend on past predictions and forecast errors). This is a well-established result (see Aoki 1987) that we confirm here in Monte Carlo simulations of a jump-free variant of the current task. In it, Bayesian and reinforcement learners behave alike. In contrast, the incremental earnings from applying Bayesian learning are substantial in the actual task, because of the presence of jumps. We elaborate in Section 2.2.2.

The superior economic performance of the Bayesian agent in the task comes from the fact that he detects the regime shifts and hence can finely track the payoff probabilities, which allows him to exploit highly skewed arms when they are in "good runs" (i.e., when they have very high reward probability)—and to avoid those arms when they are in "bad runs" (very low reward probability). In contrast, the adaptive agent is slow to adjust to the hidden regime shifts in the payoffs inasmuch as he ignores them and assumes that a hitherto-good arm will remain good in the future, until the value estimate of the arm eventually drops to the level of the other arms he has last visited. At that point, the Bayesian agent will long have figured out that there has been a jump, changed valuations accordingly, and acted on these valuation changes.

The current study adds to the growing literature in behavioral finance. Prior behavioral finance studies mostly emphasize the role of cognitive biases and restrictions, such as limitations in the number of variables agents can keep track of or pay attention to.[4] In contrast, considerably less is known about the domains in which humans fare well. Quite reassuringly, our findings suggest that finance practitioners can learn about unstable and unobservable payoffs in a way that approximates the Bayesian benchmark, so long as they are provided with sufficient information about the stochastic structure underlying the payoffs. As such, the current study contributes to the growing experimental finance literature on learning in financial markets (e.g., Kluger and Wyatt 2004; Kuhnen forthcoming).

There was little evidence available from prior work to indicate how close to the optimal benchmark learning might be in our restless bandit task. Only a few experiments have featured restless bandit tasks, and either they were not designed to answer this question (e.g., Daw et al. 2006), or they could not answer it formally (Yi et al. 2009). One exception is Estes (1984), which provides evidence that participants in a two-armed bandit task managed to

[4] E.g., Simon (1955, 1987), Kahneman (1973), Huberman and Regev (2001), Hirshleifer et al. (2009), DellaVigna and Pollet (2009), and Gao et al. (2011).

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

retain in memory information about a complex pattern of temporal swings in payoff probabilities, and adjusted sensitively and rapidly to that pattern. The current study fleshes out and extends Estes (1984), by documenting how, in our six-armed bandit task, subjects managed to track the changing outcome probabilities of all six arms.

Our study also relates to the literature examining the effect of complexity on financial decision-making (e.g., Brunnermeier and Oehmke 2009; Carlin 2009; Carlin and Manso 2011; Carlin, Kogan, and Lowery forthcoming). In particular, Carlin, Kogan, and Lowery (forthcoming) perform an experimental study to explore the effect of complexity on trading strategies. They find that complexity causes several trade anomalies. The present experiment shows that complexity is not always a decisive obstacle to sophisticated behavior. Berk and Hughson (2009) also noted this, in a nonfinancial setting. (They observed that the same people that did not fare well in a simple task did use the optimal decision rule in a much more complex one.) Together, these findings suggest that what is objectively (computationally) complex is not necessarily the most difficult for the human brain.

The current study has important implications for market participants and regulators. Following the global financial crisis, for example, several European regulators have mandated banks and fund managers to incorporate scenario analysis in the disclosures of their structured retail products.[5] The new disclosure policies are specifically mandating an emphasis on the consequences of regime shifts in the scenario used. Thus, rather than removing those products from the retail market altogether, the regulators chose to give retail investors a nudge, whereby investors are induced to incorporate regime shifts in their investment evaluations. Our findings justify this regulatory choice. Regulators have also encouraged institutional market participants to explicitly incorporate the possibility of regime shifts in their risk management processes—for example, by insisting on incorporating the possibility of such shifts in credit scoring models.[6] The message from regulators here seems to be that nudging market participants into explicitly watching for such shifts is critical for sound investment and risk management practices. Our findings certainly lend support to these regulatory practices.

Our results show that nudges do need to be given to the agents for an optimal learning to occur. As such, they raise an important issue for financial markets: who has the incentives to provide investors with the nudging they need to ensure superior performance?

[5] See *The Official Journal of the European Union*, *Commission Regulation (EU) No 583/2010 of 1 July 2010*, Paragraph 12 (p.2), Paragraph 16 (p.3), and last but not least, Section 5, p. 12.

[6] See, e.g., *OCC Bulletin 97-24*, May 20, 1997.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

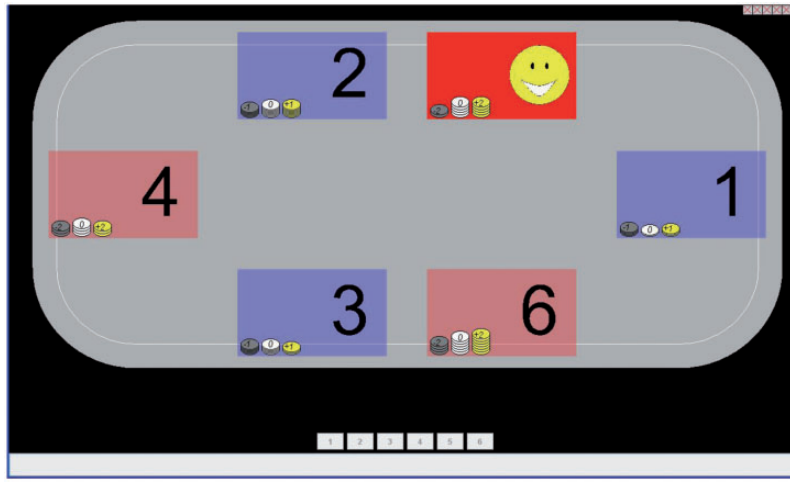


Figure 1
Experimental Design
The six-armed bandit task was implemented as a board game. Six locations correspond to six arms. Locations are color-coded (locations 1, 2, and 3 are blue; 4, 5, and 6 are red). Blue locations pay 1, 0 or −1 CHF (Swiss francs). Red locations pay 2, 0 or −2 CHF. Blue locations change less often than do red locations. The chosen option is highlighted (in this instance, location 5). Participants can freely choose a location at each trial. Histories of outcomes in locations chosen in the past are shown by means of coin piles.

1. Experimental Design

1.1 Six-armed bandit task

To study the nature of learning about payoffs that are both publicly unobservable and unstable, we designed a six-armed bandit task. The principle of the task is that on each trial, each one of six arms (assets) generates one of three possible outcomes; the subject selects one of the arms and immediately receives the outcome returned by the chosen arm, but she is not told the outcomes returned by the other arms. The expected values of the arms jump throughout the experimental session, whereby the bandit is "restless" (more on this below).

Importantly, the arms are not named after investment opportunities in the real world, and instead are presented as "locations" on a board. Three locations are blue and three were red (see Figure 1). Rewards (and losses) are expressed directly in monetary terms. Specifically, each location generates on each trial one of three possible outcomes: 1, −1, or 0 CHF (Swiss franc) for the blue locations; 2, −2, or 0 CHF for the red locations. There is no reference to prices or returns.

The current task is therefore devoid of financial connotation. This is both to avoid framing effects and to minimize confounding factors: given the extent of financial illiteracy that has been documented (e.g., Bernheim 1995; Lusardi and Mitchell 2007), we would expect participants in a standard financial task to depart from Bayesian behavior, but then it would be hard to tell to whether such

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

departure is caused by participants' financial illiteracy, or by limited learning abilities.

At each trial, the time allowed to make a decision (indicated through a timer at the bottom of the screen) averaged four seconds. If the subjects did not answer within the allotted time, they received a penalty. (The task ends automatically after the sixth penalty, and in this case the player does not receive any payoff.) The task was self-paced in the sense that subjects could respond before the end of the allotted time, in which case they proceeded to the next trial.

Note that the pace of the task was quick. Such "time pressure" was a side effect of achieving a large number of samples for each participant—about 500 trials within the 30-minute task, which is much better than in previous studies (e.g., Camerer and Ho 1999; Behrens et al. 2007). Reportedly, such high pace did not hinder subjects' play. (We asked subjects their opinion about the pace of the game in pilot sessions as well as in the after-task debriefings.)

The goal of the subjects was to maximize the outcomes accumulated from sequential play of 30 minutes. At the end of the experiment, the subjects received the accumulated outcomes minus a fixed fee. The fee was fixed before the session but revealed to the subjects only after the task was completed. This design feature was chiefly meant to prevent well-established wealth effects (e.g., the house money effect) from occurring during the task.

1.2 Stochastic structure

The task is Markovian. Each location is a trinomial generator whose current outcome depends only on the current outcome probabilities. The three locations of a given color differ in their skewness. There is one very skewed location (entropy level:[7] 0.3) that yields high (in absolute terms) expected value, one "median" location (entropy: 0.67) that yields moderate (absolute) expected value, and one random location (entropy: 1) that yields expected value zero. The identity of the locations is fixed.

While absolute expected value is constant for each location, the sign of expected value occasionally flips, and locations thus switch from having positive to negative expectation and back. The flips in the outcome probabilities are governed by regime shifts (or "jumps") implemented through two independent Bernoulli processes, one for the blue locations and one for the red. For each process and at each trial, either jump or no jump is realized. When jump occurs for one of the two colors, then at the three locations of this color, the probabilities of two outcomes flip. (As an example, imagine a location returning the good outcome 80% of the time and the bad outcome 20% of the time; then a jump occurs, and after the jump occurs, the location returns

[7] The entropy of a location measures how much its outcome probabilities differ; it is highest when all three outcome probabilities are equal, in which case the location is completely unpredictable or "random."

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

the bad outcome 80% of the time and the good outcome 20% of the time.) The jump intensity is 1/4 for red, 1/16 for blue, and thus the red locations are more restless than the blue ones.

Our motivation for choosing this specific stochastic structure was to have high statistical power in the model comparison of the Bayesian and adaptive models. The logic was that if the Bayesian and adaptive models were behaving too similarly in the task, the model comparison would make little sense and clear regularities would never come through in the experimental data. So we compared the discriminatory power of several possible task designs in Monte Carlo simulations. The discriminatory power of each task design was simply measured by the percentage of trials in which the models made different choices during the task. It appeared that the best discriminatory power was achieved with the current design, for reasons we intuitively explain in Section 1.4.

1.3 Information provided to the subjects

In the task instructions, the subjects were given a detailed explanation of the stochastic structure of the task. Specifically, the subjects were told that each location returns one out of three possible outcomes, whose (fixed) values the subjects also knew. They also knew that jumps occur independently for the two colors, that the (fixed) chance of a jump is higher for the red locations than for the blue ones; that when there is a jump, the probabilities of two out of the three outcomes permute. Finally, subjects knew that within each color group, there is one very skewed location, one relatively skewed (median) location, and one random (neutral) location.[8]

The fact that the subjects were introduced to the task by going into a detailed explanation of the outcome-generation process is a distinctive feature of the current experiment. In previous restless bandit tasks, the task instructions were vague; see, for example, Daw et al. (2006), Behrens et al. (2007), Yi et al. (2009), and Jepma and Nieuwenhuis (2011). We discuss below the importance of providing subjects with sufficiently detailed information about the decision context.

Thus, the subjects did not face model uncertainty (Draper 1995) (also referred to as distribution uncertainty [Kacperczyk and Damien 2011]). However, parameter uncertainty (Pastor and Veronesi 2009) was maximal, inasmuch as the values of the outcome probabilities, jump intensities, and skewness levels were unknown. Furthermore, the identity of the locations—which among the three locations of the same color was the skewed, median, or random one—was also unknown.

[8] We did not use the term "skewed" in the instructions (available in Appendix A.5); we said "biased" instead. We avoided any technical term, so the task was arguably accessible to subjects without scientific background. We also wanted to avoid framing effects.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

1.4 Heuristic presentation of potential approaches to performing the task

1.4.1 Choice. In principle, the subjects should adjudicate between the locations by comparing their Bellman values—that is, the discounted expected payoffs assuming optimal continuation strategies. Absent jumps, subjects could compute these values as Gittins indexes (Gittins and Jones 1974). With restless bandits, this is no longer the case, because the state at each arm is changing whether or not the arm is being sampled (Whittle 1988). Simple analogs to the Gittins indices for restless bandits have yet to be derived. As a matter of fact, in the bandit problem people behave as if they used the logit rule, which makes people explore locations with a frequency proportional to their estimated expected payoffs. We elaborate in Section 3.1.

1.4.2 Learning. To estimate the location values, the subjects could infer the outcome probabilities as described in Section 2.2.1. Alternatively, the subjects could resort to adaptive expectations or reinforcement learning: under this simple approach which we describe in detail in Section 2.2.2, the next outcomes are directly forecasted from the observed outcomes, with disregard for the hidden outcome probabilities.

When combined with the logit rule, the reinforcement learning model[9] closely implements the win-stay-lose-switch heuristic described in Charness and Levin (2005). This heuristic policy stays with a location until the average reward starts declining, at which point it either returns to a location that did best the last time it was visited or randomly tries another one. Reinforcement learning (combined with logit choice) has been popular in computational neuroscience, behavioral game theory, experimental psychology, and, recently, experimental finance (Pouget 2007), not only because of long-running behavioral evidence in its favor, but also because of strong neurobiological foundations. We elaborate in Section 2.2.2.

Even with the same decision rule (logit), Bayesians behave differently, if only because they track outcome probabilities and, hence, have a good sense of which location in each color group is the skewed one. Bayesians use this to opportunistically move to skewed locations when they realize they are in a "good run."[10] When the skewed locations for both color groups are bad, Bayesians typically try the median locations. When the median locations are bad too, as a last resort they "take shelter" at the random locations, which have expected value of zero. In contrast, unaware of the outcome probabilities, reinforcement learners navigate the six locations pretty much indiscriminately. As a result, Bayesians spend more time at the skewed locations than do reinforcement learners. Figure 2 illustrates the effect of skewness. It reports the

[9] In this paper, by "model" we mean a learning algorithm (either Bayesian or reinforcement learning) along with a decision rule, which is always the logit rule unless stated otherwise below.

[10] The expressions in quotation marks are taken from subjects' own reports on their play during the experiment.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

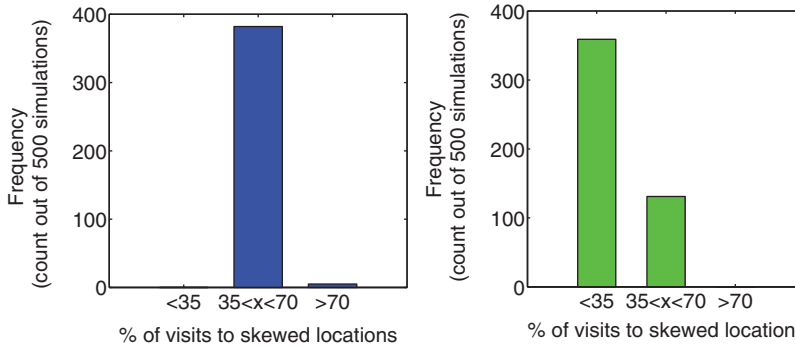


Figure 2
Comparison of the frequency of visits to the skewed locations by Bayesian versus reinforcement learners in simulated data
The left graph shows the empirical distribution of the fraction (percent) of visits to the skewed locations in 500 simulations of 500 trials each of Bayesian decision making. The right graph shows the empirical distribution when the decision maker is a reinforcement learner. The Bayesian agent appears to spend much more time at the skewed locations than the reinforcement learner does.

distribution of the fraction of visits to the skewed locations in 500 simulations of Bayesian decision making (left graph) and of reinforcement learning (right graph). The Bayesian model spent an average of 63% of the time at the skewed locations (standard deviation is 0.9); the reinforcement learning model only 27% (standard deviation is 0.17).

1.5 Why the current experimental design

The foregoing behavioral signature of Bayesian learning, which consists of visiting predominantly the locations identified as the skewed ones, explains why we introduced three levels of skewness for the options within a color group. By making the behaviors of the Bayesian and reinforcement learning models markedly different, this design feature increases the discriminatory power of the design. As an alternative design we envisioned using the four-armed bandit variant featuring only the skewed and random locations, but in Monte Carlo simulations, the choice predictions of the models were more correlated in that design—meaning that design would have less discriminatory power than the current one. Also, we needed six arms because, if we did find evidence for Bayesian learning in the current task, we planned to next investigate the neurobiological underpinnings of this Bayesian learning (more on this in Section 4), and six arms are needed to do this.[11]

[11] Specifically, a key feature of Bayesian learning in a restless bandit task is that it involves the combined perception of three kinds of uncertainty: estimation uncertainty or ambiguity (uncertainty regarding the true values of the payoff probabilities), standard risk (uncertainty left even if the decision maker knows the payoff probabilities), and jump risk (the chance that a jump has occurred, which the agent assesses at each trial). To investigate where and how these three levels of uncertainty are encoded in the human brain, one needs to obtain independent variation of the three levels, and for this six arms are needed.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

2. Formal Models of Behavior

In this section, we formally describe the behavioral models we fitted to subjects' choices.

2.1 Choice

Task participants will not learn if they do not try (they will not see outcomes of options they do not visit). We model exploration using the state of the art in experimental psychology and computer science (e.g., Ishii et al. 2002; Daw et al. 2006)—namely, the logit rule. The probability (propensity) to visit location l ($\forall l = 1, \ldots, 6$) at trial T is

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

where $Q(l,T-1)$ denotes the value estimate of location l after the $T-1$th trial. Estimation is based on either Bayesian learning (Section 2.2.1) or reinforcement learning (Section 2.2.2).

For parsimony, we take the decision maker to be risk neutral, as in prior research (e.g., Camerer and Ho 1999; Gabaix et al. 2006). The assumption of risk-neutrality is not central to our results, though, as we explain in Section 3.

The logit rule has both theoretical and empirical support in bandit tasks: Ishii et al. (2002) establish that it is optimal in a broad, information-theoretic sense,[12] and Daw et al. (2006) provide neural evidence that it is the correct model of human exploration. In contrast, in a simpler setting where subjects would see the outcomes even of unchosen options, we would not know how to model choice. This fact may seem counterintuitive at first, inasmuch as it may seem simpler to avoid the need to explore and provide the decision maker with feedback from all options. One would think that choice then immediately reveals beliefs. But reality is very different: it has been shown that in settings without exploration, choices are notoriously random (e.g., Kable and Glimcher 2009; Rolls et al. 2010) and that the logit rule fails to model such randomness in certain individuals (Frydman et al. 2011). Modeling choice in a pure learning setting is thus problematic.

[12] Specifically, the logit rule maximizes the expected utility from one's actions subject to a constraint on the entropy of the choice rule:

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

Intuitively, if there is a clear winner in the choice among locations, low entropy is allowed for, which means that the decision maker can choose the optimal location with high probability. Conversely, if all locations are estimated to be equally valuable, the policy has to exhibit high entropy—which means all locations will be visited with approximately equal probability. The inverse of the parameter β captures the importance of acquiring information.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Why is modeling of choice in bandit tasks easier than in (supposedly) simpler pure learning tasks? One possibility is that the bandit problem has higher ecological relevance than the pure learning problem. The idea is that humans and their ancestors have been exposed more to bandit-like problems than to situations where they see outcomes independent of whether they tried the option or not. Through evolutionary selection, the behavior in ecologically relevant tasks is likely to be better adapted, and hence, more comprehensible (Brennan and Lo 2011). In this view, it is difficult to make sense of anomalies in human behavior in ecologically irrelevant settings if one does not first study what happens in situations that humans and their ancestors can be expected to have encountered repeatedly during their evolution.

While we argue that the logit specification is the correct model of choice in our task, we nonetheless checked that it was not pivotal to generate the results reported in this paper. Specifically, we performed the analysis assuming other choice models–for example, power, probit, and the "greedy" choice rule, which always selects the location of currently greatest estimated value. The evidence in favor of Bayesian learning prevailed whatever the choice rule, as we report in Section 3.

2.2 Learning

Here, we explain the two competing approaches to processing information in our task: Bayesian learning and reinforcement learning.

2.2.1 Bayesian learning. In each trial T, an option l generates either the loss outcome, denoted by r_1, with probability $p_{l\,1\,T}$; the null outcome (r_2), with probability $p_{l\,2\,T}$; or the reward outcome (r_3), with probability $p_{l\,3\,T}$. The triplet $\mathbf{p_{l\,T}} = (p_{l\,1\,T}, p_{l\,2\,T}, p_{l\,3\,T})$ is in the three-dimensional simplex Θ. $\left(\Theta = \left\{ \mathbf{p} \;\middle|\; p_i \geq 0, i = 1 \ldots 3, \; \sum_{i=1}^{3} p_i = 1 \right\}.\right)$

Bayesians estimate the expected value of each location l after the Tth trial:

$$Q(l, T) = \bar{\mathbf{p}}_{\mathbf{lT}} . \mathbf{r},$$

where $\bar{\mathbf{p}}_{\mathbf{lT}}$, the estimated outcome probability, is the mean of the posterior distribution of the outcome probability (henceforth, "posterior probability distribution" or PPD) $P_{lT}(\mathbf{p_{lT}})$:

$$\bar{\mathbf{p}}_{\mathbf{lT}} = \int_{\Theta} \mathbf{p_{lT}} P_{lT}(\mathbf{p_{lT}}) d\mathbf{p_{lT}}. \tag{1}$$

So the core of Bayesian learning is to infer the PPD $P_{lT}(\mathbf{p_{lT}})$. There are two ways in which such inference could be done. One consists of dynamically adjusting the memory of the process with which one learns the outcome probabilities on each trial, depending on how sure one is that a jump has occurred at the trial. Under this model, the agent exponentially discounts

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

the past, and the discount ("forgetting") factor increases with the strength of evidence for a jump. It has been argued that such an exponential discounting of the past, which we refer to as "Forgetting Bayesian" (FB), is a good formalization of neuronal dynamics in a changing environment (Sugrue et al. 2004; Yu and Cohen 2009), as well as an intuitive, tractable, and optimal solution to learning in a changing world (Kulhavy and Zarrop 1993; Quinn and Karny 2007). So this formulation of the learning process is grounded by a convergence of theoretical and neuroscientific constructs.

The jump detection that the FB agent implements "on the spot" does not require that the agent estimate the jump probabilities of the hidden-state Markov model of the task. An alternative approach, which we refer to as the "Hierarchical Bayes" (HB) model, is to estimate the jump probabilities, learn the outcome probabilities conditional on whether a jump has occurred or not at each trial, and then compute the marginal outcome probabilities by integrating out all the possible values of the jump probability parameter using the estimated jump probabilities. This HB model has solid neuroscientific support (Behrens et al. 2007) and is particularly plausible in the current experiment inasmuch as the task instructions revealed to the subjects the hidden-state Markov model of the task. Thus, the assumption that at least some subjects used that information seems natural.

We next present the essentials of each Bayesian approach; computational details are in the appendix.

2.2.1.1. Forgetting Bayesian (FB) approach. For each of the six locations, the FB agent represents his prior belief about the outcome probability $\mathbf{p}=(p_1, p_2, p_3)$ through the following Dirichlet distribution with center $\hat{\mathbf{p}}_\mathbf{0}$ and precision $\nu_\mathbf{0}=(\nu_0, \nu_0, \nu_0)$:

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

where Γ denotes the gamma function and $\mathbf{1}$ denotes the indicator function. The agent may use, for instance, the uninformative prior $\hat{\mathbf{p}}_\mathbf{0}=(1/3,1/3,1/3)$ and $\nu_0=1/2$.[13]

At each trial T, the agent updates his beliefs about the outcome probability at the visited location l based on the outcome just observed, r_T. The agent encodes that observation through the count vector $\mathbf{c}_{\mathbf{l}\mathbf{T}}=(c_{li\,T}, i=1,\ldots 3)$, where $c_{li\,T}=\delta_{r_i}(r_T)$. ($\delta_x$ denotes point mass at x.) For example, if the outcome generated at location l at trial T is the reward outcome (r_3 as per our notations), then $\mathbf{c}_{\mathbf{l}\mathbf{T}}=(0,0,1)$.

[13] This is the Jeffreys prior, which is commonly used in Bayesian statistics. We tried different values for ν_0, and the results of the horse race between the Bayesian and reinforcement learning models were always the same (see Section 3). So this particular specification of the prior is not pivotal for our purpose in this study. Note that we refrained from fitting ν_0 to the data to minimize the number of free parameters in the estimation procedure.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

In a jump-free world, the posterior belief about the outcome probability—the PPD—would be computed through usual application of Bayes' law, which simply transforms the most recent belief as per the likelihood of the new observation. Upon a jump however, this most recent belief should be forgotten—that is, learning should start anew, which is implemented by setting the posterior belief to the prior P_0.

Since jumps are not observed directly, the agent must weight the two possibilities (his most recent belief versus P_0) based on his estimate of the likelihood that a jump has not occurred at trial T, $\lambda(T)$, a function of the strength of evidence for a jump at time T. The detailed computation of the jump likelihood $\lambda(T)$ can be found in Appendix A.1.

Formally, the posterior distribution of the outcome probability is the geometric mean of the most recent belief and P_0. The weight put on the most recent belief equals $\lambda(T)$ and the one put on P_0 equals $1-\lambda(T)$.

This posterior turns out to be the Dirichlet distribution with center $\hat{\mathbf{p}}_{\mathbf{l}\,\mathbf{T}}$ and precision ν_{lT} defined as follows. Let $\Delta_l(T-1)$ denote the set of trials before (and including) trial $T-1$ when location l was chosen:

$$\hat{\mathbf{p}}_{\mathbf{l}\,\mathbf{T}} = \frac{1}{\nu_{lT}}\left[\nu_0\ \hat{\mathbf{p}}_{\mathbf{0}} + N_l(T)\ {<<}\mathbf{c_l}{>>}(\mathrm{T})\right],$$

$$\nu_{lT} = \nu_0 + N_l(T),$$

where

$$N_l(T) = \sum_{t\in\Delta_l(T-1)} \prod_{s=t+1}^{T} \lambda(s) + 1,$$

$${<<}\mathbf{c_l}{>>}(T) = \frac{\displaystyle\sum_{t\in\Delta_l(T-1)} \left(\prod_{s=t+1}^{T} \lambda(s)\right) \mathbf{c_{lt}} + \mathbf{c_{lT}}}{N_l(T)}.$$

The updating of the PPD at each trial thus boils down to discounting the past data exponentially, with a discount factor that is adjusted "on the spot" at each trial according to the likelihood that the outcome probability has just changed. $N_l(T)$ is the effective number of data used in the updating of the PPD at trial T, and ${<<}\mathbf{c_l}{>>}(\mathrm{T})$ is a sufficient statistic for $\hat{\mathbf{p}}_{\mathbf{l}\,\mathbf{T}}$.

It should be noted that by returning to the prior P_0 after a jump, the model ignores the fact that the entropy level of each location is fixed. To account for this fact, the model should instead return to a Dirichlet distribution whose center has the same entropy level as the most recent estimated outcome probability. We tested the augmented version of the model that returns to such a prior upon a jump. As one would expect, the augmented model fitted the data even better than did the benchmark model. But the latter is simpler and turned out to be sufficient for our purpose in this paper (cf. the results reported in Section 3). Besides, in

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Monte Carlo simulations, the economic performance of the augmented version was not markedly better than the one of the simplified model. Thus, the more complicated version is no further discussed in this paper.[14]

2.2.1.2. Hierarchical Bayesian (HB) approach. To learn the outcome probabilities at the six locations, the HB agent needs to estimate the jump-intensity parameter for each color. To be more specific, let $\tilde{J}_{\text{blue}}$ and $\tilde{J}_{\text{red}}$ denote the Bernoulli processes governing the jump process for the blue locations and the red ones, respectively. At each trial, $\tilde{J}_{\text{blue}}$ takes a value of 1 if the outcome probabilities of the blue locations have jumped at the trial, which occurs with probability α_{blue}; likewise, $\tilde{J}_{\text{red}}$ takes a value of 1 if the red locations have jumped at the trial, which occurs with probability α_{red}. To compute the PPD, the HB agent must assess the posterior distributions of α_{red} and α_{blue}.

To illustrate this and without loss of generality, in the following exposition we take l, the location chosen at trial T, to be a red location. The HB agent recognizes that the outcome probability is unchanged when there is no jump at time T ($J_{\text{red}\,T}=0$) and permuted when there is one ($J_{\text{red}\,T}=1$). She further recognizes that $P(J_{\text{red}\,T}=1)=\alpha_{\text{red}}$ and $P(J_{\text{red}\,T}=0)=1-\alpha_{\text{red}}$, and hence she represents the evolution of the outcome probability $\mathbf{p_{lT-1}}$ through the transition distribution

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

where $P_{0T}(\mathbf{p_{lT}}|\mathbf{p_{lT-1}})$, the distribution of the outcome probability $\mathbf{p_{lT}}$ after a jump at time T, given $\mathbf{p_{lT-1}}$, is a uniform distribution centered around the average outcome probability that summarizes all the possible permutations of the components of $\mathbf{p_{lt-1}}$ (see Section A.3 in the Appendix for details). So the HB agent recognizes that the stochastic structure of the task is hierarchical in the sense that the evolution of the outcome probability is governed by the jump process at the top, as represented in Figure 3.

At each trial, the posterior belief about the unknown parameters (jump and outcome probabilities) is updated in the light of the data history $\underline{c}_{\mathbf{lT}}=(\mathbf{c_{lt}},t\in\Delta_l(T))$, where $\Delta_l(T)$ is the set of trials before (and including) trial T when location l was chosen. The update of the agent's posterior belief is the joint posterior distribution over the set of unknown parameters $\{p_{lT},\alpha_{\text{red}}\}$:

$$P(\mathbf{p_{lt}},\alpha_{\text{red}}|\underline{c}_{\mathbf{lT}})=\frac{P(\mathbf{p_{lt}},\alpha_{\text{red}}|\underline{c}_{\mathbf{lT-1}})l(\mathbf{c_{lT}}|\mathbf{p_{lT}})}{\int_{\Theta}\int_{0}^{1}P(\mathbf{p_{lt}},\alpha_{\text{red}}|\underline{c}_{\mathbf{lT-1}})l(\mathbf{c_{lT}}|\mathbf{p_{lT}})d\mathbf{p_{lT}}d\alpha_{\text{red}}},$$

[14] Mathematical details on the augmented FB model, along with the results of the optimization procedure and Monte Carlo simulations when using it, are available on request.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

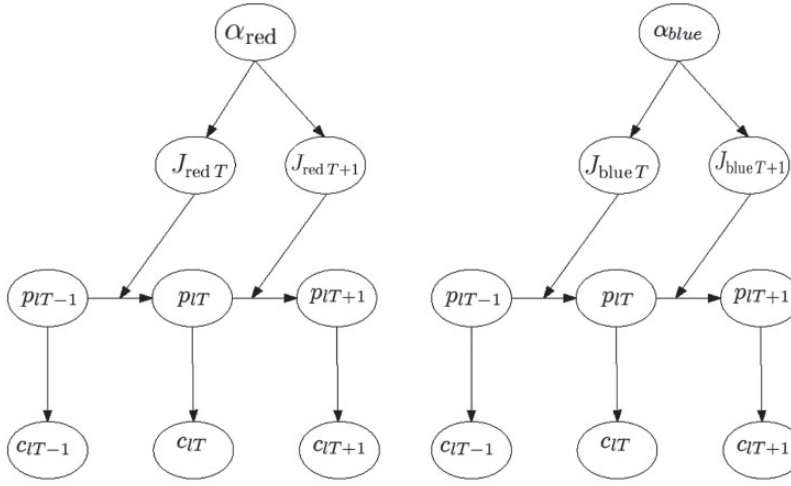


Figure 3
Diagram of the hidden state Markov model representation of the task, for a red location (left) and for a blue location (right)

where $l(\mathbf{c_{lT}}|\mathbf{p_{lT}}) = \prod_{i=1}^{3} p_{liT}^{c_{liT}}$ is the (multinomial) likelihood of the observed outcome, $\mathbf{c_{lT}}$, given the outcome probability $\mathbf{p_{lT}}$.

The HB agent assesses this joint posterior distribution by first decomposing it into the marginal posterior distribution of the jump intensity parameter and the posterior distribution of the outcome probability parameter given the jump intensity parameter:

$$P(\mathbf{p_{lt}}, \alpha_{\text{red}}|\underline{c}_{\mathbf{lT}}) = P(\mathbf{p_{lT}}|\underline{c}_{\mathbf{lT}}, \alpha_{\text{red}})\ f_T(\alpha_{\text{red}}),$$

where $f_T(\alpha_{\text{red}})$ denotes the posterior distribution of α_{red} at trial T.

Both $f_T(\alpha_{\text{red}})$ and $P(\mathbf{p_{lT}}|\underline{c}_{\mathbf{lT}}, \alpha_{\text{red}})$ are computed recursively through sequential application of Bayes' law. The detailed computations, which are standard but tedious, are available on request. Note that the computation of f_T requires the definition of the prior distribution (f_0). A possible prior distribution for α_{blue} (resp α_{red}) is the uniform distribution on the interval $[0, 1/5]$ (resp $[1/5, 1/2]$). Importantly, the results reported in Section 3 are the same across different specifications for the prior.[15]

Once the joint posterior is computed, the PPD readily obtains as the following marginal posterior distribution:

$$P_{lT}(\mathbf{p_{lT}}) = \int_0^1 P(\mathbf{p_{lt}}, \alpha_{\text{red}}|\underline{c}_{\mathbf{lT}}) d\alpha_{\text{red}}.$$

[15] In particular, we used different choices for the intervals of the uniform distribution, and we also replaced the uniform distribution with different functional forms for the prior. In each case the results reported in Section 3 were unchanged.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

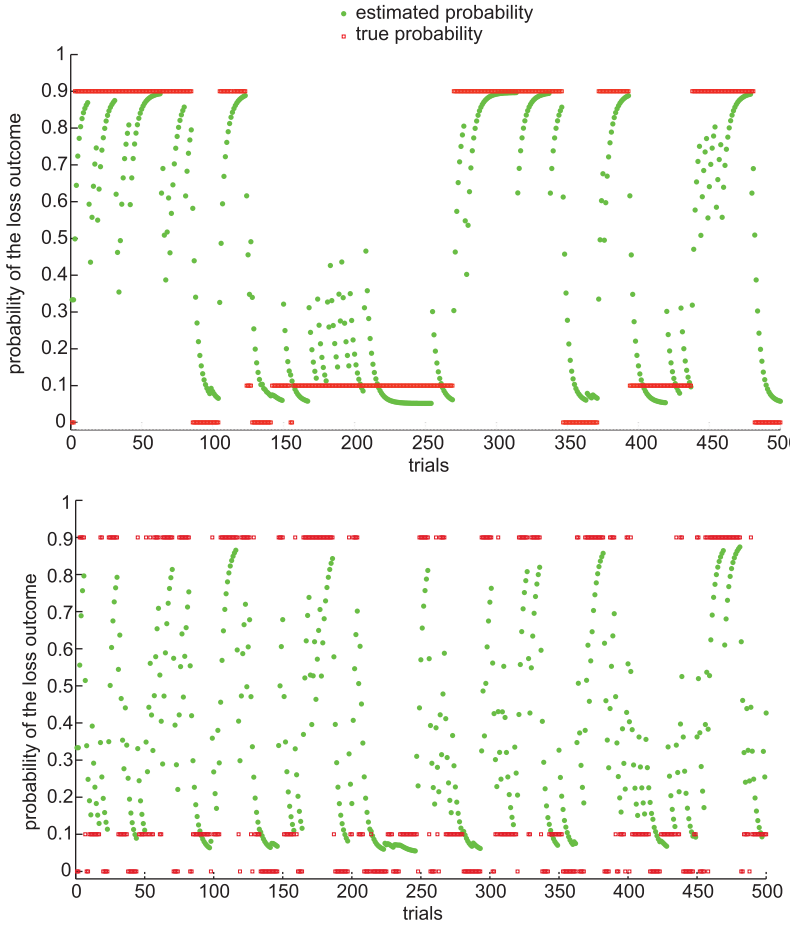


Figure 4
Forgetting Bayesian learning in 500 simulated trials at the skewed blue location (top) and at the skewed red location (bottom)
The top (bottom) graph shows FB learning of the probability to get the loss outcome in one simulation of 500 trials at the skewed blue (red) location. The horizontal axis shows the trials of the simulation. The vertical axis indicates the value of the probability to get the loss outcome. The graph reports both the estimated probability and the true probability.

2.2.1.3. Dynamic learning of the outcome probabilities in simulated data. Once the PPD is assessed, using either the FB or HB approach, the estimated outcome probability (Equation 1) is directly computed as indicated in Appendix A.4.

How accurate are the Bayesian models in their learning of the outcome probabilities in simulations of the task? Figure 4 compares the estimated probability of the bad outcome with the true probability, for 500 simulated trials in which the FB agent stayed at the skewed blue location (top graph)

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

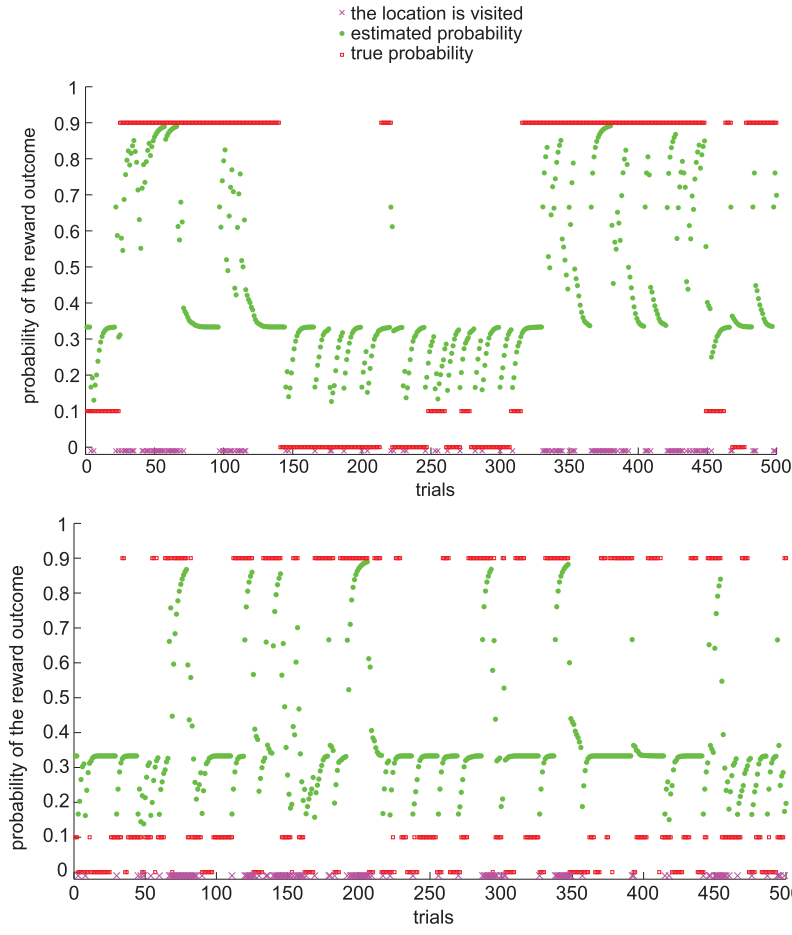


Figure 5
Probabilistic learning by the Forgetting Bayesian model in one simulation of 500 **trials**
The top (bottom) graph shows how the FB model learns the probability to get the reward outcome at the skewed blue (red) location while exploring/exploiting the locations using the logit rule. The horizontal axis shows the trials of the simulation. The vertical axis indicates the value of the probability to get the reward outcome. The graph reports both the estimated probability and the true probability. Crosses at the bottom indicate trials when the FB model visited the location.

and at the skewed red location (bottom graph) throughout. We focus on the skewed location because its value changes most around jumps. The estimated probability generally converges to the true one not only at the blue location, but also at the red one, despite the frequent jumps. For example, in the top graph, observe the 100 first trials, or the period from trial 275 to trial 350. On the bottom graph, see the period from trial 90 till trial 120.

Figure 5 illustrates how the FB model learns the outcome probabilities while exploring/exploiting the locations using the logit rule. The graph reports the

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

estimated and true probabilities of a reward at the skewed blue location (top) and at the skewed red location (bottom), as well as the trials when these locations were visited. Convergence to the true value is often observed, even at the red location.

The counterparts of Figure 4 and Figure 5 for HB learning are similar and available on request. Both models appear to track the outcome probabilities equally well. Both models are also equally coherent inasmuch as their beliefs are martingales: at any time, the current beliefs are the best forecasts of the future beliefs—that is, the models do not expect their beliefs in the future to change in any particular direction (proof available on request).

2.2.2 Reinforcement learning. Reinforcement learners ignore the outcome probabilities (let alone their jumps), and instead derive the expected value $Q(l,T)$ (for each location l and at each trial T) through adaptive expectations (e.g., Evans and Honkapohja 2001). The reinforcement learning rule is as follows:

- If l is sampled at trial T, then $Q(l,T)=Q(l,T-1)+\eta\ \delta(T)$, where $\delta(T)=r_{lT}-Q(l,T-1)$; η (the learning rate) is fixed.
- If l is not sampled at trial T, then $Q(l,T)=Q(l,T-1)$.

This reinforcement learning rule is embedded in the Experience-Weighted Attraction (EWA) framework proposed by Camerer and Ho (1999). Specifically, it is a special case of the EWA rule when the parameter that weights the strength of the outcomes at the unchosen options, relative to the outcome at the chosen one, is null. This modeling choice reflects the fact that in the current task, counterfactual learning is absent by design, since the agent does not see the outcomes at the unchosen locations. If, at each trial, the agent could see the outcomes at locations other than the chosen one, a natural formalization of bounded rationality in the task would be the EWA rule featuring counterfactual learning.

By contrasting reinforcement learning with Bayesian learning, the current study recognizes the long-standing distinction between two orthogonal learning modes, one causal or model-based, which infers the task parameters (rational expectations, or forward-looking reasoning), and the other, model-free, which does not (adaptive expectations, or backward-looking reasoning). Neuroscientists have begun identifying in the human brain distinctive neural substrates for each learning mode, thus supporting this conceptual distinction (see, for example Daw et al. 2005; Glascher et al. 2010). The distinction turns out to be relevant across domains as well as across species.[16]

[16] For instance, Brosnan (2008) documents that capuchin monkeys choose tools after inferring their functional characteristics, rather than merely remembering past success and failure with the tools. Inferring the hidden functional characteristics of tools is akin to inferring the hidden outcome probabilities of options in the current task.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Our reinforcement learning rule has solid foundations, including neurobiological. Specifically, dopamine neurons in the brain stem are known to signal $\delta(T)$ (Schultz et al. 1997) and to transmit this signal to their subcortical and prefrontal, cortical projection areas to guide the learning of outcomes. This has been well documented (e.g., Kuhnen and Knutson 2005). These solid foundations are the reason we chose to formalize model-free learning with the reinforcement learning rule, rather than with heuristics from the adaptive toolbox (Gigerenzer and Selten 2001). The issue with these adaptive heuristics is that they have limited cognitive foundation and correspondingly low empirical support, as Newell (2005) highlights.

It should be noted that the economic performance of reinforcement learners is good in many domains. In a jump-free world in particular, in which no hidden state is involved, reinforcement learning is (behaviorally) indistinguishable from Bayesian learning. The formal proof is in Aoki (1987). We regenerated those findings in Monte Carlo simulations of the task—a batch of 500 simulations of the task, each composed of 500 trials—in which we switched off the jump probability parameter (set it to zero). The Bayesians and reinforcement learners behaved alike, and both models earned on average 225 CHF (standard error: 6.5 for the Bayesian model, 7.4 for the reinforcement learning model).[17]

The fact that reinforcement learning is simple and performs well generally may explain that our brains have evolved to implement it. In other words, for participants in the current task, implementing reinforcement learning was, arguably, the "default option."

The current task puts reinforcement learners at a disadvantage vis-à-vis Bayesians because it involves a hidden state (the hidden jumps). The following analogy illustrates intuitively why Bayesian updating is superior to adaptive learning when a hidden state is involved. In the famous Monty Hall problem (Kluger and Wyatt 2004), a reward is behind one of three closed doors. The agent is asked to pick a door at random, after which the donor (knowing behind which door the reward is located) opens one of the remaining doors behind which there is no reward. At that point, the agent is allowed to switch choices. Whether to switch involves a complex inference problem whereby the agents determines the chances of the hidden "state" of the door at hand ("the reward is or is not behind the door"). Bayes' law immediately suggests that the chance that the reward is behind this door is 2/3, and hence, switching is always beneficial. The adaptive agent eventually could learn to switch, but it will take multiple trials before the agent realizes that it is always beneficial to switch. Meanwhile, the agent may make suboptimal choices, leading to inferior performance relative to the Bayesian.

Despite the inherent superiority of Bayesian learning when hidden jumps are involved, some scholars have argued that more sophisticated reinforcement

[17] In all the simulations reported in this paper, we set the values of the free parameters entering the behavioral models equal to the mean maximum-likelihood estimates from the estimation procedure reported in Section 3.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

models can accommodate nonstationarity in the environment by flexibly modulating the learning rate (Courville et al. 2006). So in our analysis of the empirical results, we tested those versions. Specifically, in addition to the benchmark rule in which the learning rate (η) is constant across locations and across time, we tested a variant that allows for two learning rates, one for the blue and another for the red locations (η_{blue} and η_{red}). We also tested a variant that allows the learning rate to temporarily increase after a jump. Finally, we tested the Pearce-Hall extension of the reinforcement learning rule (Pearce and Hall 1980), in which the learning rate is proportional to the size of the prediction error in the recent past—that is, the learning rate increases with increased levels of outcome volatility. All three variants are in line with recent evidence from neuroscience that the learning rate increases with increased instability (Behrens et al. 2007). The variant with the two learning rates η_{blue} and η_{red} fitted the data best (even after correcting for the difference in number of free parameters). For expositional reasons, we report only the results obtained with that variant—the RL model.[18]

In a batch of 500 simulations of the current task, the Bayesian model earned about 100 CHF on average, whereas the RL model made 65 CHF (standard error of the gain: 3.3 for the Bayesian model, 1.9 for the RL model). For reference, the random play model (which chooses the six locations with equal probability) earned, on average, −16 CHF in those simulations. This shows that reinforcement learning works fairly well in the current task but is nonetheless trumped by the Bayesian model, for the reasons stressed above.

3. Results

3.1 Experimental procedure

Sixty-two undergraduates at the Ecole Polytechnique Fédérale de Lausanne performed the task. Upon arrival in the lab, the participants were seated in front of a personal computer. The participants watched the online instructions (Appendix A.5) for about twenty-five minutes, whereupon the experimenter asked them to fill out a multiple-choice questionnaire that checked their understanding. After fifteen minutes, the participants reviewed the answers with the experimenter.

Subsequently, the participants completed a run of the task lasting for thirty minutes. The length of the time series varied across participants, from about 450 trials to 600 trials, with an average of 500 trials.

In the written and verbal instructions provided in the lab, we reminded the subjects that their remuneration above and beyond their show-up reward of 5 CHF would accumulate during the thirty-minute task. The subjects were also told that they could earn more than 100 CHF, depending on the quality of their

[18] The results obtained with the other models are available on request. The worst model did markedly better than random choice—according to a pseudo-R^2.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

decisions and on chance factors. The payoffs ranged from five CHF to 180 CHF, with a median of 92 CHF (standard deviation was 33). So our subjects were provided with considerable monetary incentives—higher than in similar experiments and quite important relative to their standard of living.

At the end of the experimental session, the participants filled out a debriefing questionnaire that assessed to what extent they attempted to learn the outcome probabilities, and to what extent they succeeded. All the material used in the experimental protocol is available at https://www.dropbox.com/sh/ojti2f8a6lch2rw/srLYIKQC9T/protocol.zip.

3.2 Estimation procedure

For each one of the subjects ($s = 1, \ldots 62$), we fitted the free parameters of each behavioral model by maximizing the log-likelihood compounded over trials:

$$LL_s = \sum_{t=1}^{T_s} \ln P^{\pi}(l_{st}*, t),$$

where $l_{st}*$ denotes the location that was actually chosen by subject s at trial t, and T_s is the length of the time series for player s. In this log-likelihood maximization, the HB and FB models had one free parameter, β. The RL model had two additional parameters: the learning rates η_{blue} and η_{red}.

Our estimation was subject-specific. We also fitted behavior on the basis of fixed parameters across subjects, as in, for example, Camerer and Ho (1999), by maximizing the likelihood of observed choices across subjects s and trials t:

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

We used the Nelder–Mead simplex algorithm to find the optimum. To investigate the robustness of the results, we used a genetic algorithm as an alternative to ensure that we avoided local minima. The results we obtained with the genetic algorithm corroborated the ones from the Nelder-Mead simplex search method.

3.3 Empirical results

We found strong evidence that subject choices reflected Bayesian updating. This was already immediate from the time subjects spent at skewed locations. They spent an average of 50% of their time at the skewed locations (standard deviation is 0.1). When we compared the fractions obtained in the simulations with either of the Bayesian models, FB or HB, to the one observed in the lab, we could not reject the null that the fractions were equal at a threshold of 0.001 of a Welch t-test. In contrast, any threshold of the same test led to the conclusion that our subjects did not choose skewed locations the way the RL model would (despite using learning rates that were fitted to the subjects' actual choices).

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Our experimental design thus allows us to pinpoint sharply what it is that subjects were doing to improve on reinforcement learning. Like the Bayesian agent, subjects managed to discover which location in each color group was the skewed one, which allowed them to opportunistically move to the skewed locations and exploit them when their expected value was very high.

The results of the model comparison formally show that our subjects were overwhelmingly Bayesians. Both Bayesian models match actual behavior markedly better than does the RL model, even according to a pure negative log-likelihood criterion (which does not penalize the RL model for having two additional degrees of freedom). Figure 6 displays the negative log-likelihood under either the FB or HB Bayesian model against the negative log-likelihood under the RL model, for each subject. The HB model fits better than the RL model in 89% of the cases (subjects) and the FB model beats the RL one in 81% of the cases. A paired t-test leads us to reject the null hypothesis that the fits are equal with a p-value lower than 10^{-4}. One hazard with the paired t-test is that we will reject the null while the evidence against it is actually weak (Jeffreys 1961). When we replaced the paired t-test with the more severe test of Berger and Sellke (1987), which is based on the lower bound on the evidence against the null, the conclusions were unchanged. Non-parametric tests (Wilcoxon test and sign test) lead to the same conclusion (p-value is 10^{-5}). The coefficient of variation of the fitted β across subjects is of the same order of magnitude for the three models (it is 0.29 in the HB model, 0.26 in the FB model, and 0.47 in the RL model), indicating that the results cannot be attributed to a fudge factor effect. The results of the fixed-parameter estimation are analogous.

As a qualitative check, we examined the answers to the debriefing questionnaire that each subject filled out at the end of the experiment. Subjects were often able to estimate quite accurately the relative jump probabilities across color groups.[19]

3.3.1 Robustness checks. To check the robustness of the results of the model comparison analysis, we first compared model fits separately for two subperiods: the beginning of the task, defined as the 50 first trials of the task, and the rest of the task. In both subperiods, the Bayesian models fitted the observed behavior better than did the RL model. We found the same when we repeated the exercise defining the beginning of the task as the first 100 or 200 trials.

Second, we replaced the logit rule with the greedy rule (the agent always chooses the location with maximum expected value) and re-performed the model comparison. Specifically, we calculated the percentage of trials in which the model correctly predicted subject choices. The distributions of the

[19] Subject answers are available on request in a hard-copy format.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

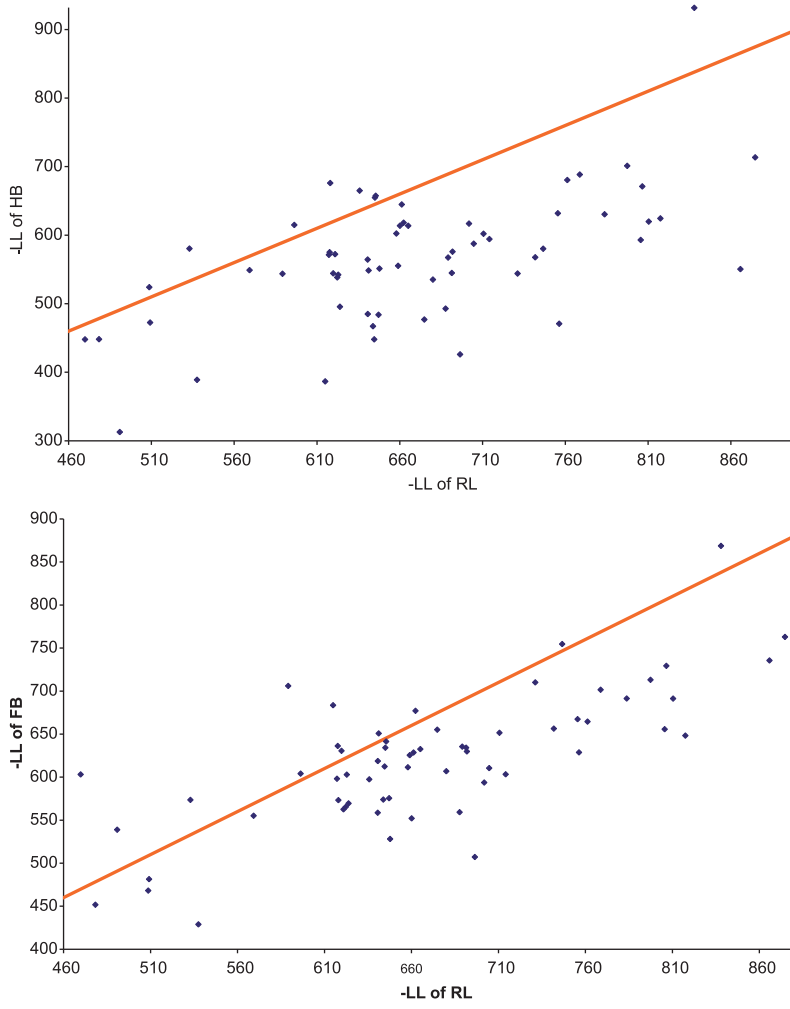


Figure 6
Comparative fits of the Bayesian and reinforcement learning models
The top graph shows the comparative fits of the HB and RL models. The bottom graph shows the comparative fits of the FB and RL models. The comparisons are based on the negative log-likelihood (-LL) criterion. Each data point 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.

prediction accuracy of the Bayesian and RL models differ significantly (chi-square test, p-value < 0.001). Figure 7 shows that the median level of prediction accuracy of both Bayesian models is significantly higher than that of the RL model. Figure 8 further shows that the FB model first-order stochastically dominates the RL model: for any level of choice prediction accuracy, the probability that we observed a level equal to or better than this level is higher

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

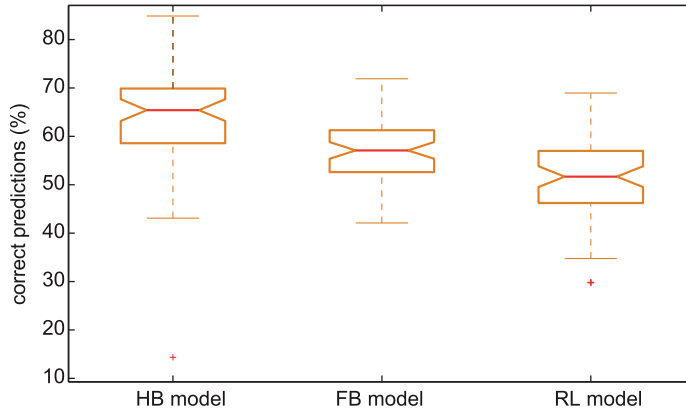


Figure 7
Comparison of the levels of prediction accuracy of the models
The box-and-whisker plots represent the distributions of the prediction accuracies of the models—HB (left), FB (middle), and RL (right)—across subjects. The prediction accuracy of a model is the percentage of trials in which the model correctly predicted choice during the task. Boxes represent the interquartile range ($25th$ to $75th$ percentile), and whiskers indicate the $5th$ and $95th$ percentiles. Crosses beyond the whiskers are outliers (e.g., the HB algorithm has one outlier, for which the prediction accuracy is only 10%). The notch in each box represents confidence interval about the median, represented by the horizontal line at the middle of the notch. Box plots whose notches do not overlap have different medians at the 5% significance level. The difference between the medians of the HB (FB) and RL models is 14% (6%). Since the notches in the box plots do not overlap, we conclude, with 95% confidence, that the medians of the Bayesian and RL models differ.

with the FB model than with the RL model. The HB model does not have first-order stochastic dominance over the RL model because of one outlier, which is apparent in Figure 7. Without this outlier, the HB model would also dominate the RL model.

Third, when we dispensed with the assumption of risk neutrality, and instead fitted parameters of the utility function proposed in Prelec (1998), which allows for risk sensitivity, loss aversion, and probability weighting, the superiority of the Bayesian models was reinforced.

We also re-performed the model comparison using different specifications for the prior distributions used by both Bayesian models. When we set the precision ν_0 to $1/3$ or 1 rather than $1/2$ in the FB model, the result of the horse race between the FB and RL models was unchanged. The results also appeared to be robust to the specification of the prior distribution of the jump intensity parameters in the HB model. We re-performed the model comparison using different intervals for the uniform priors. We also replaced the uniform prior with an exponential prior as well as a beta prior (in particular, the Jeffreys prior: $f_0(\alpha_{\text{red}}) = 1/\sqrt{\alpha_{\text{red}}(1-\alpha_{\text{red}})}$). The results were always qualitatively the same.

Finally, we compared the fits of the current RL model with advanced reinforcement learning rules where the learning rate increases after the outcome contingencies have jumped (see Section 3.2.2). RL was significantly better.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

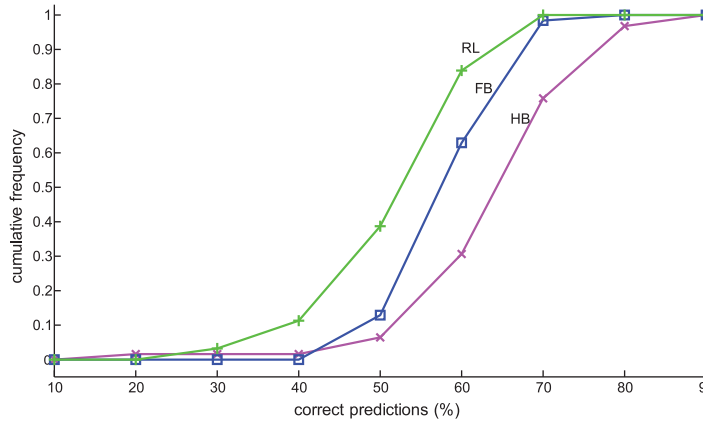


Figure 8
Comparison of the levels of prediction accuracy of the models (cont.)
The horizontal axis represents the percentage of correct predictions made by the models. The vertical axis represents the cumulative frequency: the cumulative frequency of x is the frequency of observation ($n = 62$) of a level of prediction accuracy worse than x. The curve for the FB model is always below the one for the RL model so there is first-order stochastic dominance of the FB model over the RL model. There is not first-order stochastic dominance of the HB model over the RL model owing to one outlier (which is apparent in Figure 7).

3.3.2 Replication of the results. Although 62 subjects is a relatively small sample size, as noted earlier, we recorded many choices for each subject, which explains why we could draw confident statistical conclusions from this sample size. In addition, the original results were replicated in three follow-up experiments—two behavioral and one neuroimaging experiment—with 30, 38, and 18 new subjects, respectively. Importantly, most of the subjects in those follow-up experiments were from institutions that are not more technical than average. Thus, our results seem to be generalizable inasmuch as they are not restricted to one—arguably sophisticated—population of subjects. We elaborate in the discussion section.

4. Discussion

The finding that subjects acted like Bayesians in our restless bandit task is at odds with the ample evidence supporting the idea that agents are boundedly rational in their learning abilities (for example Tversky and Kahneman 1971; Kahneman and Tversky 1972; Grether 1992; Kluger and Wyatt 2004; Charness and Levin 2005). However, our finding is consistent with the substantial evidence that people are Bayesians within important domains of human psychology. These domains include sensorimotor learning, reward learning, word learning, learning of visual chunks, and predictions about everyday events. See, for example, Körding and Wolpert (2004), Griffiths and Tenenbaum (2006), Behrens et al. (2007), Xu and Tenenbaum (2007), and Orbán et al. (2008).

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Thus, whether agents are able to exhibit sophisticated learning seems to depend finely on the nature of the objects to learn about, as well as on the characteristics of the learning environment. Which characteristics of the current task may explain why task participants managed to learn optimally?

Perhaps the current task lends itself to optimal learning because it is "ecologically relevant" (Brennan and Lo 2011). The conjecture here is that human rationality resembles the normative Bayesian model only to the extent that our primitive environments encouraged such adaptive behavior (Gigerenzer and Hoffrage 1995). Sampling the returns of unstable assets is arguably like foraging in primitive environments, where Bayesian learning outperforms simpler heuristics. In the same logic, perhaps humans perform poorly in tasks where the use of Bayes' law boils down to the implementation of declarative logic (Tversky and Kahneman 1971; Kahneman and Tversky 1972; Grether 1992; Kluger and Wyatt 2004) inasmuch as the latter was "useless" in primitive environments.[20]

To better understand the environmental factors that explained the emergence of Bayesian learning in our task, we ran two follow-up experiments. The first was to test the conjecture that providing subjects with high monetary incentives was pivotal for subject learning to be sophisticated in our challenging task. The second was to determine whether telling our subjects, in the instructions, that the payoff probabilities of the options would jump without warning during the task was important for the emergence of Bayesian learning.

We thus ran a couple of sessions of the exact same experiment, except this time no monetary incentives were provided. Subjects were undergraduates from the University of New South Wales. We found a striking reversal of the original findings, with the RL model fitting behavior better for 75% of the subjects (p-value < 0.001).[21] We then replicated the original experiment with 38 new subjects (same pool: undergraduates at the University of New South Wales) that were paid exactly like in the original experiment. For 74% of the subjects, the Bayesian model fitted the choice data better than did the RL model, and the difference in the fits was highly significant (p-value $= 0.0002$).

This finding is consistent with the ample evidence that the introduction of monetary incentives markedly decreases the extent of irrationality in the lab; for example, Siegel and Goldstein (1959), Siegel and Andrews (1962), Hertwig and Ortmann (2001), Parco et al. (2002), and Charness et al. (2010). Wilcox (1993) shows that the effect of providing monetary incentives is maximal when the task is highly complex, like in the current experiment.

[20] Johnson et al. (2002) make similar conjecture to explain that their subjects did not use backward induction in a simple three-round bargaining game. They argue that backward induction was "useless" in primitive environments (inasmuch as games in the real world do not have a definite terminal point).

[21] Importantly, the subjects in those sessions were motivated to do well inasmuch as they performed the task for course credit. So absence of monetary incentives was not merely a proxy for lack of motivation or random choice.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

In another experimental treatment, we exactly replicated the conditions of the original experiment except we did not inform subjects (same cohort as in the original experiment: undergraduates at the Ecole Polytechnique Fédérale de Lausanne) about the presence of jumps in the outcome probabilities. Subjects failed to detect the jumps, and resorted to reinforcement learning (N = 32; p-value = 0.01). But when the same subjects next performed the task under the original instructions, Bayesian learning better explained behavior than did reinforcement learning (N = 30; p-value < 0.001), thus replicating again the results reported in the current paper.[22]

This finding suggests that informing investors about key aspects of the payoff generating process is essential to help them estimate these payoffs. In other words, nudging (Thaler and Sunstein 2009) investors into looking out for regime shifts in asset returns is critical. We further discuss the meaning of such nudging in the conclusion.

An important question concerns whether some individuals are more likely to be Bayesians than others in our task. Like most lab experiments, our pool of participants lacks the heterogeneity in individual characteristics necessary to uncover many of the interesting relationships we may find using field data. However, we ran a follow-up neuroimaging experiment in which we scanned the brains of 18 subjects (undergraduates at Trinity College) while they performed the task. The purpose was to uncover the neurochemical underpinnings of Bayesian learning in the task. The results point to a key function of the neurotransmitter norepinephrine in the implementation of Bayesian learning in the task (paper published in *Neuron*; full reference omitted here to comply with anonymity instructions). Specifically, norepinephrine release in the prefrontal cortex appears to be one way by which the brain signals jumps in the payoff probabilities. Individuals with impaired norepinephrine signaling are thus less likely to be Bayesians in the task.

The foregoing neuroimaging study established that the very components that you need to implement Bayesian learning in the task are very much represented in the human brain. These components comprise three kinds of uncertainty signals—standard risk, jump risk, and estimation uncertainty—which are intrinsically Bayesian—that is, the other learning modes ignore them (reference omitted here to comply with anonymity instructions). If one doubts that task participants are truly Bayesian, one wonders why their brain would bother to compute these uncertainty signals. In that sense, our neural findings strengthen the behavioral evidence reported in the present study.

[22] The prevalence of Bayesian learning in one treatment, and of reinforcement learning in the other, cannot simply be attributed to a framing effect. The instructions were exactly the same in the original and follow-up treatments, except the latter did not explicitly mention that the outcome probabilities would jump.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

5. Conclusion

We designed a restless bandit task to study the nature of learning about payoffs that are both publicly unobservable and unstable. A first finding that emerged from our simulated data is that adhering closely to the Bayesian benchmark turns out to be critical for performing well in our task. In Monte Carlo simulations of the task, the economic performance of the Bayesian agent was significantly higher than that of the adaptive agent. Notably, the large gap between the performances of the two agents evaporated in simulations in which the option values were stationary, meaning the outperformance of the Bayesian model did specifically arise from the regime shifts in the option values. This finding suggests that learning optimally is particularly important in the business world when decision problems more and more resemble a restless bandit task.

Given the difficulty of our task, our second finding, that task participants were overwhelmingly Bayesians, suggests that complexity in itself—measured by the amount of calculations needed—is not always a good predictor of bounded rationality. New complexity metrics could be developed, and we would advocate that they be tied to ecological relevance: the more frequently our ancestors had to cope with a given problem, the more the problem should be easy to solve for our brains. Ecological relevance has not been of interest to finance yet, but given its importance to understanding behavior in other fields, such as the neurosciences, we would advocate that it receive more attention.

We further found that providing task participants with sufficiently detailed information about the stochastic structure of the task was pivotal for the emergence of sophisticated learning. This finding points to the importance of re-emphasizing to agents the occurrence of regime shifts in investment returns. Even though many—if not most—market participants are aware that returns are unstable, when they don't know the exact nature of such instability (regime shifts), they do not attempt to explicitly track jumps in contingencies, and instead resort to reinforcement learning. There, instability is accommodated through a learning rate close to 1 (thereby discounting observations in the farther past). In the context of a restless bandit problem, this approach, which is "agnostic" about why the realized returns changed, leads, however, to inferior performance.

We would argue that even though most sophisticated practitioners are aware that jumps occur, not all of them explicitly incorporate in their actual investment and risk management processes an "always-on" vigilance to detect potential regime shifts. The evidence documented here suggests that it is the failure to achieve such extreme vigilance, rather than a lack of computational capabilities, that underlies the emergence of suboptimal learning.

Our results show that nudges do need to be given to the agents for optimal learning to occur. As discussed in the introduction, regulators seem to have

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

grasped this. In several occasions, they tried to force market participants to explicitly take into account regime shifts in the risk/reward profiles of financial instruments. For example, The Basel II Framework has been revised to introduce a newly defined "stressed VAR" in replacement of the standard VAR used until then to estimate market risk.[23] This redefined VAR has to include a one-year period where significant losses were sustained, thereby explicitly incorporating regime shifts into the VAR market risk indicator.

Institutional investors could also use the current findings to improve existing processes, by actively seeking to account for possible jumps. For instance, to make the occurrence of regime shifts salient on traders' individual dashboards, firms could use new real-time visual clues on individual trading screens. We would further advocate that new firm-wide Key Performance Indicators be specifically designed to signal regime shifts to their users.

Our results raise an important issue for financial markets: who has the incentives to provide investors with the nudging they need to ensure superior performance? Some market participants behave as if they wanted to avoid providing such nudges to their clients. For instance, in most hedge fund marketing material, the emphasis is on average historic performance. This ignores the fact that many hedge funds perform differentially across two regimes: in upturns, return correlation with the market as a whole is low; in downturns, however, it is high. Not to alert investors to these contingency shifts would make them resort to reinforcement learning, we argue.

Even if nudges are provided, it may not be in the interest of some market participants to pay attention. Toxic loans contracted by local governments are a case in point (Pérignon and Vallée 2014). The interest rate on these loans is calculated with a formula that is very sensitive to jumps in the reference assets. Even if given nudges to look out for such jumps, politically motivated clients may ignore them because it is not in their interest to pay attention.

Our findings imply that, in the context of projects with shifting return opportunities and payoffs that are observable only to the investor, it is not innocuous to assume that investors are Bayesian. While humans appear to be perfectly capable of learning in a Bayesian way in this context, it requires that they are nudged into paying explicit attention to the contingency shifts. It is not obvious whether and to what extent such nudging emerges naturally in financial markets. Absent nudging, investors can be expected to resort to reinforcement learning, with a corresponding drop in performance. Theoretical analysis of asset pricing ought to take this into account.

[23] See *Basel Committee on Banking Supervision—Revisions to the Basel II market risk framework*, February 2011, available at www.bis.org.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Appendix A

A.1 Jump Likelihood in the Forgetting Bayesian Model

The core of the Forgetting Bayesian approach is to detect whether a jump has occurred at each trial. Let $\lambda_{\text{blue}}(T)$ (resp $\lambda_{\text{red}}(T)$) denote the subjective probability that no jump has occurred for the blue (red) locations. Without loss of generality, take l, the visited location at trial T, to be red. Formally,

$$\lambda(T)_{\text{red}} = P(J_{\text{red}\,T} = 0 \mid \underline{\mathbf{c}}_{\mathbf{IT}}).$$

To compute $\lambda(T)_{\text{red}}$, let $P_{l\,T}(\mathbf{p})(J_{\text{red}\,T}=0)$ and $P_{l\,T}(\mathbf{p})(J_{\text{red}\,T}=1)$ denote the posterior probability distribution after no jump and after jump, respectively. Absent further information, the Bayesian model sets the prior belief of a jump equal to 0.5. Thus, the subjective jump likelihood at time T, $P(J_{\text{red}\,T}=0 \mid \underline{\mathbf{c}}_{\mathbf{IT}})$, equals to

$$\frac{1/2\int_{\Theta} l(\mathbf{c_{IT}}|\mathbf{p})P_{l\,T}(\mathbf{p})(J_{\text{red}\,T}=0)d\mathbf{p}}{1/2\int_{\Theta} l(\mathbf{c_{IT}}|\mathbf{p})P_{l\,T}(\mathbf{p})(J_{\text{red}T}=0)d\mathbf{p}+1/2\int_{\Theta} l(\mathbf{c_{IT}}|\mathbf{p})P_{l\,T}(\mathbf{p})(J_{\text{red}\,T}=1)d\mathbf{p}}.$$

We can rewrite the previous form as

$$\frac{1}{1+\dfrac{\int_{\Theta} l(\mathbf{c_{IT}}|\mathbf{p})P_{l\,T}(\mathbf{p_{IT}})(J_{\text{red}\,T}=1)d\mathbf{p}}{\int_{\Theta} l(\mathbf{c_{IT}}|\mathbf{p})P_{l\,T}(\mathbf{p})(J_{\text{red}\,T}=0)d\mathbf{p}}},$$

where

$$P_{l\,T}(\mathbf{p})(J_{\text{red}T}=1) = P_0(\mathbf{p}),$$

$$P_{l\,T}(\mathbf{p})(J_{\text{red}T}=0) = P_{l\,T/T}(\mathbf{p}) \quad \left(= \frac{l(\mathbf{c_{IT}}|\mathbf{p})P_{l\,T-1}(\mathbf{p})}{\int_{\Theta} l(\mathbf{c_{IT}}|\mathbf{p})P_{l\,T-1}(\mathbf{p})d\mathbf{p}}\right).$$

Therefore, we have:

$$\lambda_{\text{red}}(T) = \frac{1}{1+\dfrac{AC}{B}},$$

where

$$\begin{cases} A = \int_{\Theta} l(\mathbf{c_{IT}}|\mathbf{p})P_0(\mathbf{p})d\mathbf{p}, \\ B = \int_{\Theta} l(\mathbf{c_{IT}}|\mathbf{p})^2 P_{l\,T-1}(\mathbf{p})d\mathbf{p}, \\ C = \int_{\Theta} l(\mathbf{c_{IT}}|\mathbf{p})P_{l\,T-1}(\mathbf{p})d\mathbf{p}. \end{cases}$$

Since $l(\mathbf{c_{IT}}|\mathbf{p}) = \prod_{i=1}^{3} p_i^{c_{liT}}$, we write A as follows:

$$A = \int_{\Theta} \prod_{i=1}^{3} p_i^{c_{liT}} \frac{\prod_{i=1}^{3} p_{liT}^{v_0\hat{p}_{i0}-1}}{\frac{\prod_{i=1}^{3}\Gamma(v_0\hat{p}_{i0})}{\Gamma(v_0)}} d\mathbf{p}.$$

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Because

$$\int_{\Theta}\prod_{i=1}^{3}p_i^{c_{liT}+\nu_0\hat{p}_{i0}-1}\,d\mathbf{p}=\frac{\prod_{i=1}^{3}\Gamma(c_{liT}+\nu_0\hat{p}_{i0})}{\Gamma(\sum_{i=1}^{3}c_{liT}+\nu_0)},$$

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 $T-1$. (For example, suppose that location l delivered the loss outcome at trial $T-1$; then $c_{lT-1}=(1,0,0)$, and i^* is equal to 1.) Because $\Gamma(x+1)=x\Gamma(x)$, and since $c_{liT}=0,\quad \forall i\neq i^*$ while $c_{li^*T}=1$, we can further simplify this expression:

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

Hence,

$$\begin{aligned}A&=\frac{\Gamma(1+\nu_0\hat{p}_{i^*0})\prod_{i\neq i^*}\Gamma(\nu_0\hat{p}_{i0})\Gamma(1+\nu_0\hat{p}_{i^*0})}{\nu_0\prod_{i=1}^{3}\Gamma(\nu_0\hat{p_{i0}})}\\&=\frac{1}{\nu_0}\frac{\Gamma(1+\nu_0\hat{p}_{i^*0})}{\Gamma(\nu_0\hat{p}_{i^*0})}\\&=\hat{p}_{i^*0}.\end{aligned}$$

The calculation of B is analogous:

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

Since

$$\Gamma(2+\nu_{lT-1})=\Gamma(1+(1+\nu_{lT-1}))=(1+\nu_{lT-1})\Gamma(1+\nu_{lT-1})=(1+\nu_{lT-1})\nu_{lT-1}\Gamma(\nu_{lT-1}),$$

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

and

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

B simplifies to:

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

Using analogous arguments,

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

Consequently,

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

Thus, $\lambda(T)$ depends on $\frac{\hat{p}_{i^*0}}{\hat{p}_{i^*lT-1}}$, the odds ratio (i.e., strength of evidence) for the hypothesis that a jump has occurred at time T.

A.2 Posterior Probability Distribution in the Forgetting Bayesian Model

Here we drop the color reference to avoid any unnecessary notational burden: $\lambda(T)$ stands for $\lambda_{\text{blue}}(T)$ if the location visited at trial T, l, is blue, and $\lambda_{\text{red}}(T)$ if it is red.

To update his belief about the outcome probability at trial T (the posterior outcome probability distribution or PPD, $P_{l\,T}$) , the decision maker starts by assessing $P_{l\,T/T}$, the PPD absent jumps. This PPD comes from applying Bayes' law in the usual way, by combining the most recent PPD available at the beginning of trial T ($P_{l\,T-1}$, the PPD carried over from trial $T-1$), with the (multinomial) likelihood of trial T observation $\mathbf{c_{lT}}$:

$$P_{l\,T/T}(\mathbf{p_{lT}})\propto P_{l\,T-1}(\mathbf{p_{lT}})\;l(\mathbf{c_{lT}}|\mathbf{p_{lT}}),\quad\text{where}\quad l(\mathbf{c_{lT}}|\mathbf{p_{lT}})=\prod_{i=1}^{3}p_{liT}^{c_{liT}}.$$

In principle, the PPD should either be $P_{l\,T/T}$, when there is no jump, or P_0, when there is one. Since the decision maker does not observe jumps directly, she must weight the two cases based on her estimate of the likelihood that a jump has not occurred at trial T, $\lambda(T)$, which is a function of the strength of evidence for a jump at time T.

The PPD minimizes Bayes risk (Berger 1980) in the following sense:

$$P_{lT}=\underset{P\in\,\Upsilon}{\arg\min}\left\{\lambda(T)\;KL(P,P_{l\,T/T})+(1-\lambda(T))\;KL(P,P_0)\right\},$$

where Υ denotes the probability space on Θ and $KL(.,.)$ denotes the Kullback-Leibler divergence measure (an intuitive measure of the distance between two distributions):

$$P_{lT}=\underset{P\in\,\Upsilon}{\arg\min}\left\{\lambda(T)\int_{\Theta}P\;(\mathbf{p})\ln\frac{P(\mathbf{p})}{P_{l\,T/T}(\mathbf{p})}\;d\mathbf{p}+(1-\lambda(T))\int_{\Theta}P(\mathbf{p})\;\ln\frac{P(\mathbf{p})}{P_0(\mathbf{p})}\;d\mathbf{p}\right\}.$$

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Under the assumption that the product between $P_{l\,T/T}$ and P_0 is not zero everywhere,

$$P_{lT} = \left(P_{l\,T/T}\right)^{\lambda(T)} (P_0)^{1-\lambda(T)}. \tag{A1}$$

For large T, the posterior probability distribution (PPD) is well approximated by the Dirichlet distribution defined in the main text. The proof which is quite tedious is available on request.

A.3 Conditional Probability Distribution After a Jump in the Hierarchical Bayesian Model

P_{01} is the uniform distribution on Θ. For T strictly greater than 1, let $P_{0T}(\mathbf{p_{IT}}|\mathbf{p_{IT-1}})$ denote the distribution of the outcome probability after a jump at time T, given the most recent outcome probability triplet $\mathbf{p_{IT-1}}$. For concreteness, take $\mathbf{p_{IT-1}}$ to be $(0.6, 0.4, 0)$. The possible outcome probabilities after $\mathbf{p_{IT-1}}$ has flipped are $(0.4, 0.6, 0)$, $(0, 0.4, 0.6)$, and $(0.6, 0, 0.4)$, with equal probability. The average triplet summarizing the possible permutations is therefore $1/3\ (0.4, 0.6, 0) + 1/3\ (0, 0.4, 0.6) + 1/3\ (0.6, 0, 0.4)$ which turns out to be equal to $(1/3, 1/3, 1/3)$. $P_{0T}(\mathbf{p_{IT}}|(0.6, 0.4, 0))$ is a two-dimensional uniform distribution centered around (the first two components of) $(1/3, 1/3, 1/3)$:

$$\mathrm{U}(\ [1/3 - 0.1; 1/3 + 0.1] \times [1/3 - 0.1; 1/3 + 0.1]\ ).$$

While the above reflects the true environment of the task, there is a hazard that the instructions (Appendix A.5) deluded some subjects that after a jump, the largest of the three probability components (0.6 in the previous example) would flip with one of the two others, in which case the possible outcome probabilities after the flip would be $(0.4, 0.6, 0)$ and $(0, 0.4, 0.6)$, and hence the average triplet summarizing the possible flips would be $1/2\ (0.4, 0.6, 0) + 1/2\ (0, 0.4, 0.6) = (0.2, 0.5, 0.3)$. In this case, $P_{0T}(\mathbf{p_{IT}}|(0.6, 0.4, 0))$ is a two-dimensional uniform distribution centered around $(0.2, 0.5, 0.3)$:

$$\mathrm{U}(\ [0.2 - 0.01; 0.2 + 0.01] \times [0.5 - 0.01; 0.5 + 0.01]\ ).$$

Fortunately, all the results involving the HB model were qualitatively the same with both versions of the HB model.

A.4 Estimated Outcome Probabilities

Under the FB approach, the estimated outcome probability—or the posterior mean outcome probability—equals the center $\hat{\mathbf{p}}_{\mathbf{IT}}$ of the posterior probability distribution as defined in the main text. Under HB learning, Equation 1 is assessed numerically, after computing the PPD as indicated in the main text, as a Riemann sum based on a two-dimensional grid with steps of 0.01 on each coordinate p_1 and p_2.

A.5 Instructions of the Task

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

The Boardgame consists of a board composed of **six locations, three blue and three red**.

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.

Each round, every location delivers **one outcome out of three possible**:

- A **blue** location returns 1 CHF or -1 CHF or otherwise 0 CHF.

- A **red** location returns 2 CHF or -2 CHF or otherwise 0 CHF.

At each round, you choose to visit one of the six locations and receive the outcome generated by the chosen location at this round.

You *don't* see the outcomes returned by the other locations at this trial.

Initially you don't know the chances to win and lose on each one of the 6 locations.

At each round, you cannot predict for sure the next outcome - like in *the Roulette*.

The three locations of a given color differ in their degree of bias: one location is **very biased** - one outcome is much likely to occur; another location is **not biased at all;** the third location is **biased but to a lesser extent** than the very biased location.

To understand what "biased" means, consider the location in Fig. 1: on it, 1 CHF occurs 10% of the time and -1 CHF the rest of the time. Such a location is biased towards -1 CHF in the sense that -1 CHF is more likely to occur than the two other outcomes.

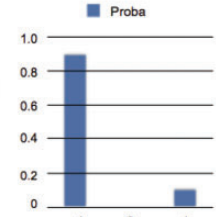


Fig. 1

Now, consider the symmetric case of the location in Fig. 2: on it, the three scenarios are equally likely to occur (-1 CHF is as likely to occur as +1 CHF or 0 CHF), whereby the location is not biased at all.

The three degrees of bias are the same across the two colors.

CAVEAT! What makes this game challenging is that the locations **change** throughout the game: specifically, the chances to give the three outcomes **switch** over time, meaning that a hitherto-good location is gonna suddenly turn bad, and this at any point in time!

To understand the sense in which the locations change, imagine a red location that gave in the past 2 CHF 75% of the time, and never returned -2 CHF. Suddenly, the chance to get 2 CHF switches with that of getting -2 CHF, whereby upon the change the location delivers -2 CHF 75% of the time!

You are not warned in advance when changes occur.

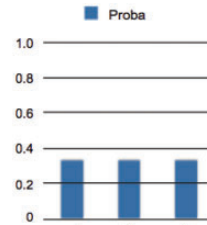
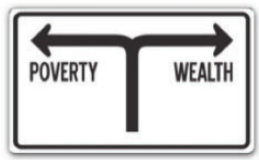


Fig. 2



Changes are color-specific:

- When a change occurs for the red (blue) color, all three red (blue) locations change at the same time.

- Changes occur independently for blue and red.

Note: A change may occur at any time. A change on blue (red) may well happen while you are playing with a red (blue) location.

The **red locations are relatively unstable**: the chance that a change will occur is higher at the red locations than at the blue ones.

The chance that a change will occur is **fixed** throughout the game, for each color.

The degree of bias of each location is **fixed** throughout the game.

You play for 30 minutes (on average 500 rounds) and accumulate the rewards and losses received throughout the game.

You go home with the accumulated outcomes minus a fixed fee.

Figure A1
Instructions of the task

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

References

Ang, A., and A. G. Timmermann. 2011. Regime changes and financial markets. Netspar Discussion Paper No. 06/2011-068.

Aoki, M. 1987. *State space modeling of time series*. Berlin: Springer-Verlag.

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

Berger, J. O. 1980. *Statistical decision theory and Bayesian analysis*. New York: Springer.

Berger, J. O., and T. Sellke. 1987. Testing a point null hypothesis: the irreconciliability of p values and evidence. *Journal of the American Statistical Association* 82:112–22.

Berk, J., and E. N. Hughson. 2009. Can boundedly rational agents make optimal decisions? A natural experiment. Robert Day School of Economics and Finance Research Paper No. 2008-7.

Bernheim, D. 1995. Do households appreciate their financial vulnerabilities? An analysis of actions, perceptions, and public policy. In *Tax policy and economic growth: Proceedings of a symposium sponsored by the American Council for Capital Formation, Center for Policy Research* (pp. 1–30). Washington, DC: Center for Policy Research.

Brennan, T. J., and A. W. Lo. 2011. The origin of behavior. *Quarterly Journal of Finance* 1:55–108.

Brosnan, S. F. 2008. Animal behavior: The right tool for the job. *Current Biology* 19:R124–25.

Brunnermeier, M. K., and M. Oehmke. 2009. *Complexity in financial markets*. Princeton University.

Camerer, C. F., and T.-H. Ho. 1999. Experience-weighted attraction learning in normal form games. *Econometrica* 67:827–74.

Carlin, B. I. 2009. Strategic price complexity in retail financial markets. *Journal of Financial Economics* 91:278–87.

Carlin, B. I., S. Kogan, and R. Lowery. Forthcoming. Trading complex assets. *Journal of Finance*.

Carlin, B. I., and G. Manso. 2011. Obfuscation, learning, and the evolution of investor sophistication. *Review of Financial Studies* 24:754–85.

Chari, V. V., and P. J. Kehoe. 2004. Financial crisis as herds: Overturning the critiques. *Journal of Economic Theory* 119:128–50.

Charness, G., E. Karni, and D. Levin. 2010. On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda. *Games and Economic Behavior* 68:551–56.

Charness, G., and D. Levin. 2005. When optimal choices feel wrong: A laboratory study of Bayesian updating, complexity and affect. *American Economic Review* 95:1300–1309.

Courville, A. C., N. D. Daw, and D. S. Touretzky. 2006. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences* 10:294–300.

Daw, N. D., Y. Niv, and P. Dayan. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8:1704–11.

Daw, N. D., J. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan. 2006. Cortical substrates for exploratory decisions in humans. *Nature* 441:876–79.

DellaVigna, S., and J. M. Pollet. 2009. Investor inattention and Friday earnings announcements. *Journal of Finance* 64:709–49.

Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:45–97.

Erev, I., and A. E. Roth. 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* 88:848–81.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Estes, W. K. 1984. Global and local control of choice behavior by cyclically varying outcome probabilities. *Journal of Experimental Psychology* 10:258–70.

Evans, G. W., and S. Honkapohja. 2001. *Learning and expectations in macroeconomics*. Princeton, NJ: Princeton University Press.

Frydman, C., C. Camerer, P. Bossaerts, and A. Rangel. 2011. MAOA-L carriers are better at making optimal financial decisions under risk. *Proceedings of the Royal Society B* 1714:2053–59.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H. E. Stanley. 2003. A theory of power-law distributions in financial market fluctuations. *Nature* 423:267–70.

Gabaix, X., D. Laibson, G. Moloche, and S. Weinberg. 2006. Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review* 96:1043–68.

Gans, N., G. Knox, and R. Croson. 2007. Simple models of discrete choice and their performance in bandit experiments. *Manufacturing and Service Operations Management* 9:383–408.

Gao, P., Z. Da, and J. Engelberg. 2011. In search of attention. *Journal of Finance* 66:1466–91.

Gigerenzer, G., and U. Hoffrage. 1995. How to improve Bayesian reasoning without instructions: Frequency formats. *Psychological Review* 102:684–704.

Gigerenzer, G., and R. Selten (eds.). 2001. *Bounded rationality: The adaptive toolbox*. Cambridge, MA: MIT Press.

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

Glascher, J., N. Daw, P. Dayan, and J. P. O'Doherty. 2010. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66:585–95.

Grether, D. M. 1992. Testing Bayes rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics* 95:537–57.

Griffiths, T. L., and J. B. Tenenbaum. 2006. Optimal predictions in everyday cognition. *Psychological Science* 17:767–73.

Hertwig, R., and A. Ortmann. 2001. Experimental practices in economics: A methodological challenge for psychologists. *Behavioral and Brain Sciences* 24:383–451.

Hirshleifer, D., S. S. Lim, and S. H. Teoh. 2009. Driven to distraction: Extraneous events and underreaction to earnings news. *Journal of Finance* 64:2289–325.

Huberman, G., and T. Regev. 2001. Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. *Journal of Finance* 56:387–96.

Ishii, S., W. Yoshida, and J. Yoshimoto. 2002. Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural Networks* 15:665–87.

Jeffreys, H. 1961. *Theory of probability*. Oxford: Oxford University Press.

Jepma, M., and S. Nieuwenhuis. 2011. Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience* 23:1587–96.

Johnson, E. J., C. Camerer, S. Sen, and T. Rymon. 2002. Detecting failures of backward induction: Monitoring information search in sequential bargaining. *Journal of Economic Theory* 104:16–47.

Kable, J. W., and P. W. Glimcher. 2009. The neurobiology of decision: Consensus and controversy. *Neuron* 63:733–45.

Kacperczyk, M., and P. Damien. 2011. Asset allocation under distribution uncertainty. McCombs Research Paper Series No. IROM-01-11.

Kahneman, D. 1973. *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Kahneman, D., and A. Tversky. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology* 3:430–54.

Kluger, B. D., and S. B. Wyatt. 2004. Are judgment errors reflected in market prices and allocations? Experimental evidence based on the Monty Hall problem. *Journal of Finance* 59:969–97.

Körding, K. P., and D. M. Wolpert. 2004. Bayesian integration in sensorimotor learning. *Nature* 427:244–47.

Kuhnen, C. M. Forthcoming. Asymmetric learning from financial information. *Journal of Finance*.

Kuhnen, C. M., and B. Knutson. 2005. The neural basis of financial risk taking. *Neuron* 47:763–70.

Kulhavy, R., and M. B. Zarrop. 1993. On a general concept of forgetting. *International Journal of Control* 58:905–24.

Lusardi, A., and O. S. Mitchell. 2007. Baby boomer retirement security: The role of planning, financial literacy, and housing wealth. *Journal of Monetary Economics* 54:205–24.

Mandelbrot, B. B. 1957. *Fractales, Hasard et Finance*. Paris: Flammarion.

Newell, B. R. 2005. Re-visions of rationality. *Trends in Cognitive Sciences* 9:11–15.

O'Doherty, J. P., P. Dayan, J. Schultz, R. Deichmann, K. Friston, and R. J. Dolan. 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304:452–54.

Orbán, G., J. Fiser, R. N. Aslin, and M. Lengyel. 2008. Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America* 105:2745–50.

Parco, J. E., A. Rapoport, and W. E. Stein. 2002. Effects of financial incentives on the breakdown of mutual trust. *Psychological Science* 13:292–97.

Pastor, L., and P. Veronesi. 2009. Learning in financial markets. National Bureau of Economic Research Working Paper 14646.

Pearce, J. M., and G. Hall. 1980. A model for pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review* 87:532–52.

Pérignon, C., and B. Vallée. 2014. Political incentives and financial innovation: The strategic use of toxic loans by local governments. HEC Paris Research Paper No. FIN-2013-1017.

Pouget, S. 2007. Adaptive traders and the design of financial markets. *Journal of Finance* 62:2835–63.

Prelec, D. 1998. The probability weighting function. *Econometrica* 66:497–527.

Quinn, A., and M. Karny. 2007. Learning for non-stationary Dirichlet processes. *International Journal of Adaptive Control and Signal Processing* 21:827–55.

Rangel, A., C. Camerer, and P. R. Montague. 2008. A framework for studying the neurobiology of value-based decision making. *Nature Neuroscience* 9:545–56.

Rolls, E. T., F. Grabenhorst, and G. Deco. 2010. Decision-making, errors, and confidence in the brain. *Journal of Neurophysiology* 104:2359–74.

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

Schultz, W., P. Dayan, and P. R. Montague. 1997. A neural substrate of prediction and reward. *Science* 275:1593–99.

Siegel, S., and J. M. Andrews. 1962. Magnitude of reinforcement and choice behavior in children. *Journal of Experimental Psychology* 63:337–41.

Siegel, S., and D. A. Goldstein. 1959. Decision-making behavior in a two-choice uncertain outcome situation. *Journal of Experimental Psychology* 57:37–42.

Simon, H. A. 1955. Behavioral model of rational choice. *Quarterly Journal of Economics* 49:99–118.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014

Simon, H. A. 1987. Bounded rationality. In J. Eatwell, M. Milgate, and P. Newman (eds.), *The new Palgrave: A dictionary of economics* (pp. 266–68). London: Macmillan.

Sims, C. A. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50:258–70.

Sims, C. A. 2006. Rational inattention: Beyond the linear-quadratic case. *American Economic Review* 96:158–63.

Sugrue, L. P., G. S. Corrado, and W. T. Newsome. 2004. Matching behavior and the representation of value in the parietal cortex. *Science* 304:1782–87.

Thaler, R. H., and C. R. Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. New York: Penguin.

Tversky, A., and D. Kahneman. 1971. Belief in the law of small numbers. *Psychological Bulletin* 76:105–10.

Whittle, P. 1988. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability* 25:287–98.

Wilcox, N. T. 1993. Lottery choice: Incentives, complexity and decision time. *Economic Journal* 103:1397–417.

Xu, F., and J. B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review* 114:245–72.

Yi, S. K. M., M. Steyvers, and M. Lee. 2009. Modeling human performance in restless bandits with particle filters. *Journal of Problem Solving* 2:81–101.

Yu, A. J., and J. D. Cohen. 2009. Sequential effects: Superstition or rational behavior. In M. I. Jordan, Y. LeCun, and S. A. Solla (eds.), *Advances in neural information processing systems,* vol.21 (pp. 1873–80). Cambridge, MA: MIT Press.

Downloaded from http://rfs.oxfordjournals.org/ at University of New South Wales on November 19, 2014