Journal of Experimental Psychology: General
2013, Vol. 142, No. 2, 476–488

0096-3445/13/$12.00 DOI: 10.1037/a0029550

The Role of First Impression in Operant Learning

Hanan Shteingart, Tal Neiman, and Yonatan Loewenstein
The Hebrew University of Jerusalem

We quantified the effect of first experience on behavior in operant learning and studied its underlying computational principles. To that goal, we analyzed more than 200,000 choices in a repeated-choice experiment. We found that the outcome of the first experience has a substantial and lasting effect on participants' subsequent behavior, which we term *outcome primacy*. We found that this outcome primacy can account for much of the underweighting of rare events, where participants apparently underestimate small probabilities. We modeled behavior in this task using a standard, model-free reinforcement learning algorithm. In this model, the values of the different actions are learned over time and are used to determine the next action according to a predefined action-selection rule. We used a novel nonparametric method to characterize this action-selection rule and showed that the substantial effect of first experience on behavior is consistent with the reinforcment learning model if we assume that the outcome of first experience resets the values of the experienced actions, but not if we assume arbitrary initial conditions. Moreover, the predictive power of our resetting model outperforms previouly published models regarding the aggregate choice behavior. These findings suggest that first experience has a disproportionately large effect on subsequent actions, similar to primacy effects in other fields of cognitive psychology. The mechanism of resetting of the initial conditions that underlies outcome primacy may thus also account for other forms of primacy.

Keywords: reinforcement learning, operant conditioning, underweighting of rare events, risk aversion, primacy

First impressions, you know, often go a long way, and last a long time.
—Dickens, *The Life and Adventures of Martin Chuzzlewit*

Operant Learning

According to the *law of effect* formulated by Thorndike over a century ago, actions that are closely followed by satisfaction are more likely to recur, whereas actions followed by discomfort are less likely to reoccur in that situation (Lattal, 1998; Thorndike, 1911). *Operant learning*, in which behavior is a function of the consequences of past behavior, is based on this principle. The computational principles underlying operant learning are a subject of debate. Some neurophysiological evidence supports the view that operant learning is achieved through the synergy of two processes. First, the *values* of the different actions (or more generally, state actions) are learned from past actions and their subsequent rewards. Second, these learned values are used to choose among different actions such that actions associated with a higher value are more likely to be chosen (Doya, 2007; Glimcher, 2009). By contrast, there are alternative views on operant learning that are not based on a valuation system (Dayan & Niv, 2008; Erev & Barron, 2005; Gallistel, Mark, King, & Latham, 2001; Law & Gold, 2009; Loewenstein, 2010; Loewenstein & Seung, 2006; Sugrue, Corrado, & Newsome, 2004).

Reinforcement Learning (RL)

Operant learning is typically modeled quantitatively using *reinforcement learning* (RL) algorithms (Sutton & Barto, 1998), which describe how behavior should adapt to rewards and punishments (Dayan & Niv, 2008). In this framework, the *Q-learning* algorithm (Watkins, 1989; Watkins & Dayan, 1992) is particularly noteworthy, as it has been widely used to model sequential decision-making behavior in humans and animals (Barto, Sutton, & Watkins, 1989; Daw, 2011; Neiman & Loewenstein, 2011; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006). Here we used Q-learning to quantitatively model human behavior in a *repeated choice experiment* in which in every trial t, the participant chooses an action a_t from a finite set of actions and receives a reward r_t. Q-learning describes how the expected average reward (action value), of each action a in trial t, denoted by $Q_t(a)$, changes in response to that trial's action and the resultant reward. The value of the chosen action $Q_t(a_t)$ is updated by

This article was published Online First August 27, 2012.

Hanan Shteingart, Tal Neiman, and Yonatan Loewenstein, Department of Neurobiology, The Alexander Silberman Institute of Life Sciences, Edmond and Lily Safra Center for Brain Sciences, Interdisciplinary Center for Neural Computation and the Center for the Study of Rationality, The Hebrew University of Jerusalem, Israel.

This work was supported by Israel Science Foundation Grant 868/08 from the Ministry of Science and Technology, Israel, and the Ministry of Foreign and European Affairs, the Ministry of Higher Education and Research, France, and the Gatsby Charitable Foundation. The work was conducted in part in the framework of the France-Israel Laboratory of Neuroscience. We thank Ido Erev for kindly sharing with us the Technion Prediction Tournament data set and for discussions.

Correspondence concerning this article should be addressed to Yonatan Loewenstein, Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem 91904, Israel. E-mail: yonatan@huji.ac.il

$$Q_{t+1}(a_t) = Q_t(a_t) + \eta(r_t - Q_t(a_t)), \quad (1)$$

where $0 \leq \eta \leq 1$ is the *learning rate*, which determines the relative contribution of the most recent reward to the expected average reward. The value of the nonchosen actions Q_t ($a \neq a_t$) remains unchanged. The smaller the magnitude of η, the smaller is the contribution of the most recent reward to the value of the action. If $\eta = 1$, the value of action a_t following the value update is simply r_t. If the reward r_t is larger than the estimated action value ($r_t - Q_t(a_t) > 0$), the action value increases, which in turn increases the likelihood that this action will be chosen again in the future. The reverse occurs if the reward is smaller than the action value.

Equation 1 describes how the action values adapt over trials but does not specify how these action values are used to select actions. Several *action selection rules*, which determine the mapping between action values and the policy, have been previously proposed. Two of these, ε-*greedy* and *softmax* are noteworthy, as they are commonly used for modeling behavior (Sutton & Barto, 1998). According to the ε-greedy action selection rule, the alternative associated with the highest estimated action value is chosen, with probability $1 - \varepsilon$ ($0 < \varepsilon < 1$). The other alternatives are chosen randomly with a probability ε. The value of the parameter ε determines the balance between *exploration* and *exploitation* (Cohen, McClure, & Angela, 2007). The larger the value of ε, the more likely that actions associated with a low action value will be chosen (exploration). By contrast, the smaller the value of ε, the more likely that the action with the highest estimated value will be chosen (exploitation).

An alternative action selection rule is the softmax rule. According to this rule, the probability of choosing an action a is proportional to $e^{\beta Q_t}(a)$, where parameter β controls the exploration–exploitation trade-off. The lower the value of β, the more likely that an action associated with a relatively low action value will be selected. In contrast to the ε-greedy action selection rule, the softmax action selection rule has a graded sensitivity to the values of actions. Typically, the empirical trade-off between exploration and exploitation (controlled by ε or β) is estimated by fitting one of these action-selection rules to the empirical data (Daw, 2011).

However, to the best of our knowledge, the shape of the action selection rule has never been estimated nonparametrically. In the Results section, we describe a novel method for estimating the action selection rule.

Initial Conditions in RL

A model of value adaptation and action selection is not fully determined without specifying the initial conditions of the value adaptation rule, Equation 1. This is because the value adaptation rule in Equation 1 is a *difference equation*, in which the current value depends on the value of the previous trial. Therefore, the values of the actions before the first trial need to be specified. The common practice when modeling empirical behavioral data using RL models is to initialize all action values to the same value Q_0 (Daw, 2011). The value of Q_0 is determined either arbitrarily (e.g., $Q_0 = 0$) or by fitting to the empirical data (Daw, 2011; Sutton & Barto, 1998). Theoretical studies have shown that under general conditions, the choice of initial conditions has no effect on the asymptotic learning behavior. In other words, the behavior of the model after a sufficiently large number of trials is independent of Q_0 because the contribution of Q_0 to the value of action a diminishes exponentially with the number of trials in which action a is chosen (Sutton & Barto, 1998). Following these theoretical considerations, little attention has been directed to determining how the initial values of Equation 1 are specified.

Although the asymptotic behavior may be independent of the initial conditions, it is not clear to what extent this asymptotic behavior describes participants' behavior in standard experiments composed of a finite number of trials. There are two reasons why the initial conditions may play an important role in explaining the nonasymptotic experimentally observed behavior. First, the learning rate may be low, leading to a slow adaptation and a prolonged contribution of the initial conditions to behavior. Second, the action selection rule dictates that actions that are associated with a relatively low value would be less often selected than those associated with a relatively high value. This *sampling bias* is also known as *adaptive sampling* or the *hot stove effect* (Denrell, 2005, 2007; Denrell & March, 2001). As a result, more trials would be needed to update the values of actions that are associated with the lower estimated value, potentially prolonging the effect of initial conditions on behavior.

Reset of Initial Conditions Hypothesis

This article explores how the initial conditions of action values are determined and to what extent these initial conditions shape behavior in humans in the first 100 trials of repeated choice experiments. We hypothesize that the initial conditions are not arbitrarily set. Rather, we posit that the initial condition of each action value is "optimistic" (formally $Q_0 = \infty$ for all action values). Moreover, we posit that these initial values are reset to the value of the reward in the first trial in which that action was chosen. As a result, the outcome of the first action is expected to have a disproportionately large effect on subsequent actions, similar to *primacy effects* in other fields of cognitive psychology (Hogarth & Einhorn, 1992; Mantonakis, Rodero, Lesschaeve, & Hastie, 2009). The idea of resetting of the initial conditions can apply to other forms of learning that are not associated with actions or rewards. Thus, we posit that the resetting of initial conditions may also help explain the primacy effect in belief updating (Asch, 1946; Hogarth & Einhorn, 1992).

Predicting Aggregate Behavior

If the initial action values are indeed reset by the outcome of the first choice, a model that incorporates *reset of initial conditions* (RIC) is expected to predict participants' behavior better than a model that assumes any *arbitrary initial condition* (AIC). We test this prediction by comparing the predictive power of several previously proposed models and the one proposed here. Finally, we show that much of the *underweighting of rare events*, in which participants tend to be more risk aversive when the probability for a successful risky attempt is low (Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004), can be attributed to RIC.

The Experiment

To address our questions and test our hypotheses and predictions, we analyzed the results of an experiment by Erev et al. (2010). In this experiment, participants repeatedly chose between two unmarked alternatives in blocks composed of 100 trials. One alternative, denoted as *risky*, yielded either a high or low monetary reward with a fixed probability. The other alternative, denoted as *safe*, yielded a deterministic reward that was approximately equal to the mean reward of the risky alternative. The first experience is expected to be most pronounced if expected rewards are approximately equal for the two alternatives, as is explained in the Discussion section.

Method

The full details of the experimental procedures and methods have been described elsewhere (Erev, Ert, & Roth, 2008; Erev et al., 2010). A relevant summary of these methods is described here.

Participants and Instructions

Two hundred students (Technion, Israel) participated in the experiment; half were in the "estimation" session and the other half in the "competition" session (see Experiment Design section below). Participants were paid 40 Israeli Shekels (ILS) (about U.S. $11.40) for showing up, and could earn more money or lose part of the show-up fee during the experiment. The procedure lasted about 40 minutes on average per participant.

Participants were told that the experiment would include several independent blocks and that in each they would be asked to repeatedly select one of two unmarked buttons that appeared on a computer screen for an unspecified number of trials. Each selection was followed by a presentation of its outcome (in ILS currency). The payoff from the unselected button (the forgone payoff) was not presented. At the end of the experiment, one choice was randomly selected, and the participant's payoff for this choice was added to the show-up fee to determine the final payoff. The instructions (translated from Hebrew) were as follows:

> This experiment includes several games. Each game includes several trials. You will receive a message before the beginning of each game. In each trial, you will be asked to select one of two buttons. Each press will result in a payoff that will be presented on the selected button. At the end of the experiment, one of the trials will be randomly drawn (all the trials are equally likely to be drawn). Your payoff for the experiment will be the outcome (in Sheqels) of this trial. Good luck! (Erev et al., 2010).

Experiment Design

In each trial, pressing the risky button resulted in the delivery of a high monetary payoff (H) with probability P_H, or a low payoff (L) with probability $1 - P_H$. Pressing the safe button resulted in a medium payoff (M). There were 100 choice trials in each block. Different blocks differed in reward schedule parameters, namely H, L, M, and P_H. The location of the buttons changed between sections randomly, so there was no association between button type and location.

There were two experimental sessions: an "estimation" session and a "competition" session. The two sessions used the same methods and examined similar (but not identical) decision problems, as is described below. Both sessions consisted of different collections of 60 problem sets, and the exact problem sets were determined by a random selection of the parameters (rewards and probabilities) L, M, H, and P_H according to a predefined algorithm (Erev et al., 2010). In each session, participants were randomly assigned to one of five different subgroups. Each subgroup contained 20 participants who were presented with the same 12 problem sets. The distribution of P_H across problems is depicted in Figure 1A. In approximately one third of the problems, P_H was relatively small, $P_H < .15$ (denoted as low-P_H problems; black in Figure 1A), in approximately one third it was relatively high, $P_H >$ .85 (denoted as high-P_H problems; white in Figure 1A), and in approximately one third it had an intermediate value (gray in Figure 1A). As shown in Figure 1B, the medium prize M was chosen from a narrow distribution whose mean was equal to the expected value of the risky alternative $\langle r \rangle = P_H \cdot H + (1 - P_H) \cdot L$.

Results

Outcome Primacy

The RIC hypothesis predicts that the outcome of the first trial should have a disproportionately large effect on subsequent choice behavior. To study this prediction, we quantified the extent to which the outcome of the first risky choice, L or H, affects subsequent choices. We separated the blocks of each problem set into two groups, according to the outcome of the first risky choice, L or H. We focused our attention on behavior in 73% of the problem sets (88/120), in which there was at least one block associated with each of the two groups. For each group in these problem sets, we computed the frequency of choosing the risky choice in all trials subsequent to the first risky choice. These two frequencies are an estimate of the probabilities of choosing the risky action, conditioned on the outcome of the first risky choice for the corresponding problem set.

Averaging over the problem sets, we found that the probability of choosing the risky choice, provided that the outcome of the first risky choice was L, is $A_L = 31 \pm 3\%$ (see Figure 2A top, red). This

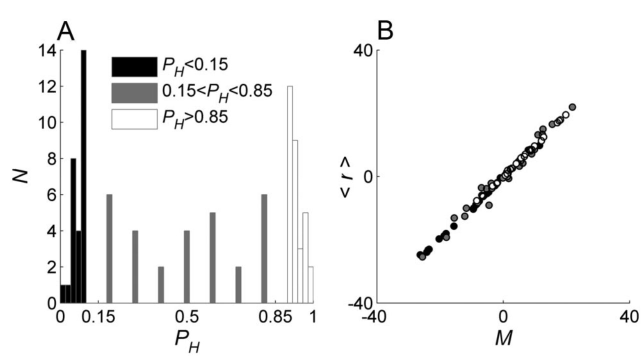

Figure 1. The experimental reward schedule. A: The distribution (N denotes counts) of P_H in the problem sets. B: The expected returns from the "risky" alternatives ($<r>$) as a function of the safe payoff. M: Black, gray, and white correspond to problem sets in which the value of P_H was relatively low ($P_H < .15$), intermediate ($.15 < P_H < .85$), and high ($P_H >$ .85), respectively.

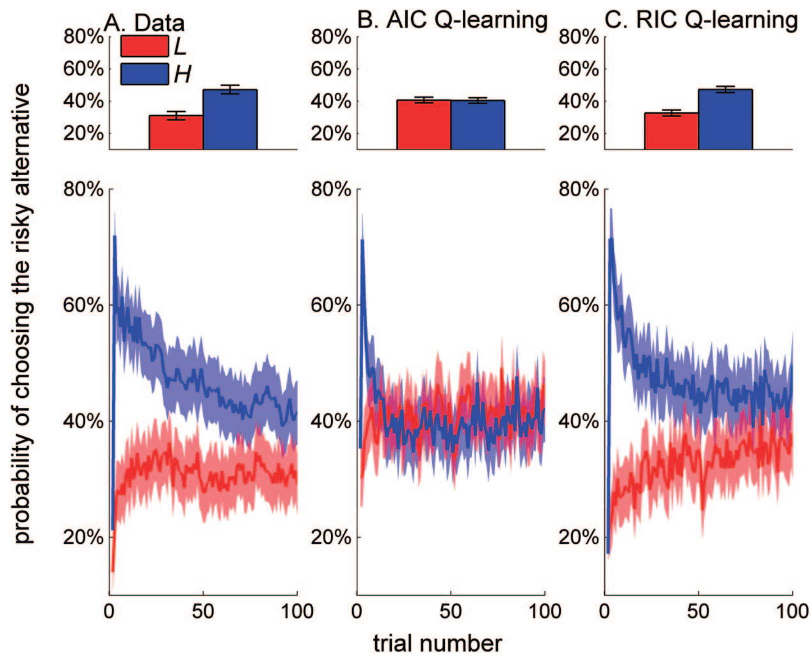

Figure 2. Outcome primacy effect: The average (over problem sets) probability of choosing the risky alternative, conditioned on the outcomes of the first risky choice. Red, low reward (*L*); blue, high reward (*H*). Top panel represents the average probability of choosing the risky alternative, averaged over all subsequent trials; the bottom panel represents the average probability of choosing the risky alternative in a trial. A: The empirical data. B: Simulation of the arbitrary initial conditions (AIC) Q-learning model. C: Simulation of the resetting of initial conditions (RIC) Q-learning model. Bars (top panel) and shaded area (bottom panel) represent the standard error of the mean.

number is substantially smaller than that probability, provided that the outcome of the first risky choice was H, $A_H = 47 \pm 3\%$ (see Figure 2A, top panel, blue), $t(174) = 4.96$, $p = 2 \times 10^{-6}$, 95% CI [9.7%, 22.5%], $g = 0.84$. This result shows that the outcome of the first risky trial has a substantial effect on subsequent choice behavior. Note that A_L and A_H are based on choices made throughout a block of 100 trials.

To further quantify the time scale associated with the effect of the first trial on behavior, we computed, for each of the problems in the 88-problem subset (see above), the probabilities of choosing the risky choice in all trials t, conditioned on the outcome of the first risky choice in that block. These conditional probabilities, averaged over the different problem sets, are depicted in Figure 2A (bottom panel), where the blue and red lines indicate the probability of choosing the risky choice given that the first risk outcome was H and L, respectively. In 92.3% of the blocks, the first risky choice was on either the first or the second trial. Thus, the trial number is approximately equal to the number of trials elapsed from the first risky choice. Therefore, the difference between the blue and red curves is a measure of the effect of the outcome of the first risky choice on behavior in subsequent trials.

We found that even in the last trial, $t = 100$, there was a statistically significant difference between the two curves, $t(206) = 3.397$, $p = 8 \times 10^{-4}$, 95% CI [5.7%, 21.5%]. Similarly, a statistically significant difference between the two curves was observed for each of the trials in Figure 2A, bottom panel ($p < .05$). This result is a demonstration that the outcome of the first risky choice affects behavior for at least 100 trials. This long-lasting effect of the first experience is reminiscent of the primacy effect in other fields of psychology in which the first stimulus is particularly salient (Hogarth & Einhorn, 1992; Mantonakis et al., 2009). Therefore, we denote the effect of the first risky reward on subsequent behavior as *outcome primacy*. In the Discussion section, we elaborate on the similarities between outcome primacy and other forms of primacy.

Modeling Outcome Primacy

Arbitrary initial conditions. The outcome of the first risky choice has a significant and long-lasting effect on choice behavior (see Figure 2A, top and bottom panels). However, this outcome primacy does not necessarily indicate a reset of the initial conditions (the RIC hypothesis). As mentioned in the introduction, a low learning rate and adaptive sampling, which naturally emerges in standard RL algorithms, might give rise to a long time scale (Denrell, 2005, 2007; Denrell & March, 2001). In order to test whether the RL framework can account for outcome primacy, we considered a standard AIC Q-learning algorithm with the following action selection rule, which is motivated by the experimental data (see below):

$$\Pr[a] = (1 - 2\varepsilon)\frac{e^{\beta Q_t(a)}}{\sum_{a'} e^{\beta Q_t(a')}} + \varepsilon. \qquad (2)$$

We term the action selection rule in Equation 2 ε-*softmax* because it is a hybrid of the ε-greedy and softmax action selection rules. If $\varepsilon = 0$, then the ε-softmax is simply the softmax action selection rule. The ε-softmax becomes ε-greedy if $\beta = \infty$. Note that the ε-softmax action selection rule has a graded sensitivity to action values like the softmax action selection rules, and like the ε-greedy, it maintains exploration even when the value of one of the actions is much larger than that of the other action.

The AIC Q-learning model with the ε-softmax action selection rule is characterized by four parameters: (a) the initial conditions Q_0, (b) the learning rate η (see Equation 1) and two parameters of the action selection rule, (c) ε, and (d) β. We found the set of parameters that best fit the sequences of actions of each participant in the experiment by maximizing the likelihood of the sequence. We then used these parameters to simulate the behavior of the AIC Q-learning model such that each simulated participant was tested on the same problem sets as the corresponding human participant.

The results of these simulations are depicted in Figure 2B, which shows that in the AIC Q-learning model, the probability of choosing the risky choice, provided that the outcome of the first risky choice was *L*, is $A_L^{AIC} = 40 \pm 2\%$, which is not statistically different from that number, provided that the outcome of the first risky choice was *H*, $A_L^{AIC} = 40 \pm 2\%$, $t(170) = 0.12$, $p = .91$, 95% CI [−4.2%, 4.7%], $g = 0.25$. Thus, the AIC Q-learning model with the parameters extracted from the behavior of the participants in the experiment is inconsistent with the finding that the outcome of the first risky choice has a substantial effect on the aggregate the probability of choosing the risky alternative (see Figure 2B, top panel).

Moreover, considering the conditional probabilities of choosing the risky alternative over trials (see Figure 2B, bottom panel), we found that in the AIC Q-learning model, these conditional probabilities became statistically indistinguishable from Trial 13 onward, $t(202) = 0.64$, $p = .52$, 95% CI [−5.1%, 10.0%], $g = 0.09$ for Trial 13. These results indicate that the AIC Q-learning cannot account for the outcome primacy effect observed in the behavior of the participants (compare Figure 2A with Figure 2B).

Reset of initial condition. The failure of the AIC Q-learning model to account for the observed outcome primacy prompted us to test the effect of incorporating a reset of the initial conditions into the Q-learning model. In this model, the initial values of the two alternatives are "optimistic": $Q_0 = \infty$ for all action values (Sutton & Barto, 1998). Moreover, these initial values are reset to the value of the immediate reward after the first experience of each alternative (see RIC hypothesis in the introduction). In subsequent trials, these values are updated according to Equation 1. Similar to the analysis of the AIC Q-learning model, we used the method of maximum likelihood to estimate the parameters of the RIC Q-learning model with the ε-softmax action selection rule that best fit the behavior of the participants. Note that the number of parameters that characterize the RIC Q-learning model is smaller than that of the AIC Q-learning model because the initial values are not a free parameter. We then used these parameters to simulate the behavior of the RIC Q-learning model such that each simulated participant was tested on the same problem sets as the corresponding human participant.

The results of these simulations are depicted in Figure 2C, which shows that the probability of choosing the risky alternative in the RIC model, provided that the outcome of the first risky choice was *L*, is $A_L^{RIC} = 32 \pm 2\%$, that is, significantly lower than that probability, provided that the outcome of the first risky choice was *H*, $A_H^{RIC} = 47 \pm 2\%$, $t(164) = 6.02$, $p = 1 \times 10^{-8}$, 95% CI [9.7%, 19.2%], $g = 1.12$. Moreover, the predictions of the RIC model are statistically indistinguishable from the experimentally measured aggregate data: The pairs (A_L, A_L^{RIC}) and (A_H, A_H^{RIC}) are not statistically different, $t(204) = 0.42$, $p = .67$, 95% CI [−3.7%, 5.7%], $g = 0.06$ and, $t(203) = 0.06$, $p = .94$, 95% CI [−5.3%, 5.1%], $g = 0.01$, respectively (see Figure 2C, top panel).

Similarly, when considering the probabilities of choosing the risky alternative over trials conditioned on the outcome of the first risky choice (see Figure 2C, bottom panel), we found that the dynamics of the RIC model were qualitative similar to that of the empirical data (see Figure 2A, bottom panel). Moreover, in the RIC simulation, as in the empirical data, even in the last trial, $t = 100$, there was a statistically significant difference between the two conditional probabilities, $t(201) = 4.34$, $p = 2 \times 10^{-5}$, 95% CI [8.7%, 23.2%], $g = 0.61$.

Short-Term Consequences of the RIC Hypothesis

The RIC hypothesis was also supported by the short-term effect of the outcome of the first risky choice on subsequent behavior: the initial rate of alternations, regardless of action or outcome and the phasic (steplike) change in choice preference according to the outcome of the first risky action.

Initial rate of alternations. In 84% of the blocks (2,006 blocks out of 2,400), the first choice was different from the second, indicating that the probability of alternation in the second trial was significantly larger than chance (binomial, $p = 1 \times 10^{-237}$, 95% CI [82.0%, 85.0%]). Moreover, the probability of alternation to the safe alternative in the second trial after a risky choice in the first trial was higher than chance if the outcome of the first risky choice was either *H* or *L* as depicted in the second trial in Figure 2A, bottom panel (516 blocks out of 645, binomial, $p = 1 \times 10^{-52}$, 95% CI [76.6%, 83.0%], in case that the first risky choice was *H* and 492 blocks out of 569, binomial, $p = 1 \times 10^{-68}$, 95% CI [83.4%, 89.2%], in case that the first risky choice was *L*). In the framework of AIC Q-learning, such alternation can result from *optimistic initial conditions*, that is, initial values higher than typical values of reward on the task (Sutton & Barto, 1998). However, optimistic initial conditions are expected to result, in general, in several trials of a high probability of alternation between the choices, depending on the magnitude of the learning rate. This is because independent of the action outcome, its action value is reduced. By contrast, the probability of alternation in the empirical data already drops below chance in the third transition (1,017 blocks out of 2,400, binomial, $p = 4 \times 10^{-14}$, 95% CI [40.0%, 44.4%]). In contrast to the AIC Q-learning model, the RIC Q-learning model predicts a high rate of alternation in the second trial and a lower-than-chance rate of alternation after both alternatives are chosen, as observed in the behavioral data. Specifically, the alternation rate during the first two trials in the RIC Q-learning model was 83%, which is not significantly different form the empirical alternation rate, $t(4798) = -0.89$, $p = .37$, 95% CI [−0.03, 0.01], $g = 0.026$.

Phasic change in choice preference. The dynamics of the probability of choosing the risky alternative conditioned on the outcome of the first risky choice (see Figure 2A, bottom panel) is characterized by a large phasic response, followed by a slow decay of the difference between the two conditional probabilities. The co-occurrence of the two phenomena, namely, a large phasic response and a slow decay, is difficult to account for in the framework of AIC. The reason is that a trade-off between the two phenomena is expected: A low learning rate would enable a slow decay, but the phasic response would be small. By contrast, a high learning rate that can account for the considerable phasic difference between the two conditional probabilities would result, in general, in fast decay. The latter was observed in the simulation of the AIC model based on participants' estimated parameters (see Figure 2B, bottom panel). By contrast, in the RIC model, these two phenomena are decoupled: The reset of initial conditions results in a large phasic response, independent of the value of the learning rate parameter. Indeed, both a large phasic response and a slow decay are observed in the simulation of the RIC model (see Figure 2C, bottom panel).

Predicting Aggregate Behavior

In the previous subsections, we showed that the RIC model can account for the outcome primacy effect as well as the alternation rate in the second trials and the phasic response. In order to further test the predictive power of the RIC model, we compared it with alternative models of operant learning. As described in the Method section, the behavioral data analyzed in this article were used in a competition (Erev et al., 2010), in which models were compared according to their ability to predict the probability of choosing the risky alternative, averaged over all trials and participants, given the parameters of the problem set (M, H, L, and P_H; see the Method section).

The competition consisted of two sessions, an estimation session and a competition session, each containing 100 participants and 60 problem sets (see the Method section). The estimation session was used to optimize the parameters of the candidate models, and their performance was tested by comparing their predictions with humans' behavior in the competition session. The aggregate probability of choosing the risky alternative was predicted by each model (P_{predict}) for each problem set, and was compared with the empirically measured probability, averaged over all participants for that problem set (P_{empiric}).

The predictive power of the different models was evaluated using three measures: (a) the fraction of problems, in which both P_{predict} and P_{empiric} were either above or below 50% (p_{agree}); (b) the Pearson's normalized correlation (ρ) between P_{predict} and P_{empiric}; (c) the mean square difference (MSD) between P_{predict} and P_{empiric}, averaged over all problem sets (see Table 1). An additional measure was the Equivalent Number of Observation (Erev, Roth, Slonim, & Barron, 2007). However, because this measure is a monotonic function of the MSD, it was not used here.

In order to evaluate the RIC Q-learning model, we estimated the three parameters of the RIC Q-learning model, η, β, and ε, that best fit the trial-by-trial behavior of each of the participants in the estimation session (similar to Figure 2C). The 100 triplets of parameters, one triplet for every participant, were regarded as representatives of the distribution of parameters across the population of participants. Then, for every problem set in the competition session, we estimated the expected aggregate probability of choosing the risky alternative, P_{predict}, by simulating the RIC Q-learning separately for each triplet of parameters and averaging the aggregate probability of choosing the risky over all simulations. As can be seen in Table 1, this *heterogeneous RIC Q-learning model* that takes into account the population heterogeneity outperformed all previously proposed models with respect to MSD and P_{agree} and was performing as well as the best baseline model (explorative sampler with recency) with respect to correlation measurement ρ.

To study the contribution of the population heterogeneity to the predictive power of the RIC Q-learning model, we considered a *homogenous RIC Q-learning model*, which is characterized by the same triplet of parameters for all simulated participants. The single triplet of parameters was found by simulating the model and choosing the triplet that minimized the MSD between P_{predict} and P_{empiric}, averaged over all problem sets in the estimation session, using the Nelder-Mead simplex (direct search) method (Lagarias, Reeds, Wright, & Wright, 1998). Simulating the model with the

Table 1
Performance Comparison Between Models in the Aggregate Risk Aversion Prediction Competition

Model name	No. of parameters	Estimation			Competition		
		P_{agree} (%)	ρ	MSD·10^3	P_{agree} (%)	ρ	MSD·10^3
Basic RL[a]	2	56	0.67	22.4	66	0.51	26.3
Normalized RL[a]	2	76	0.83	9.2	84	0.84	8.7
Normalized RL with inertia[a]	4	75	0.86	8.0	86	0.85	8.4
Two stage sampler[a]	7	80	0.90	6.5	83	0.87	8.4
ACT-R[a]	2	77	0.88	9.4	87	0.89	7.5
Homogenous AIC Q-learning[b]	**4**	**80**	**0.92**	**7.2**	**87**	**0.90**	**7.0**
Explorative sampler with recency[a]	4	82	0.88	7.5	86	0.89	6.6
Heterogeneous RIC Q-learning[c]	**300**	**78**	**0.93**	**5.2**	**88**	**0.89**	**6.4**
Homogenous RIC Q-learning[d]	**3**	**77**	**0.91**	**5.8**	**88**	**0.90**	**6.4**

Note. RL = reinforcement learning; AIC = arbitrary initial conditions; RIC = resetting of initial conditions; MSD = mean square error.
[a] Taken from the competition results (Erev et al., 2010) and ordered by competition session MSD (models proposed in this article are marked by bold font). [b] $\beta = 52$, $\varepsilon = 0.2$, $\eta = 0.4$, $Q_0 = 1$ were chosen by gradient descent optimization of the MSD on the estimation set. [c] The heterogeneous population values were chosen by maximizing the likelihood per participant and are summarized here by their mean, STD and the median in brackets respectively: $\beta = (370, 470, 22)$, $\varepsilon = (0.16, 0.1, 0.17)$, $\eta = (0.5, 0.5, 0.5)$. [d] $\beta = 52$, $\varepsilon = 0.2$, $\eta = 0.4$ were chosen by gradient descent optimization of the MSD on the estimation set.

resultant triplet of parameters over the problems in the competition session, we found that the predictive power of the homogenous RIC Q-learning model is comparable to the heterogeneous RIC Q-learning model (see Table 1). However, in contrast to the heterogeneous RIC Q-learning model, the homogeneous RIC Q-learning model predicts outcome primacy, which is substantially smaller than the experimentally observed outcome primacy (not shown).

Repeating the same analysis for the AIC Q-learning model, we found that the predictive power of a *homogeneous AIC Q-learning model* is lower than that of the RIC Q-learning model, further strengthening the RIC hypothesis. Note that the better descriptive power is despite the fact that the number of parameters that characterize the AIC Q-learning model is larger than that of the RIC Q-learning model (4 and 3 parameters, respectively). Nonetheless, it should be noted that the AIC Q-learning model outperforms previously proposed RL models (compare with Basic RL, Normalized RL, and Normalized RL with inertia rows in Table 1). The primary difference between those models and the AIC Q-learning model is the action selection function used (softmax vs. ε-softmax), which demonstrates the importance of choosing an accurate action selection function when modeling choice behavior.

The Action Selection Rule

In order to model learning behavior in the framework of Q-learning, as was described in the previous sections, the action selection function should be specified. Previous studies have typically assumed a particular functional form of the action selection function and estimated its parameters from the data (Daw, 2011). However, to the best of our knowledge, the action selection rule has not been estimated nonparametrically. The reason is that there is no direct access to the arguments of the action selection function, the action values, and to the output, the probability of choice.

By contrast, here we develop a novel procedure to characterize the shape of the action selection function nonparametrically. This method is based on the behavior of the participants in the third trial of the blocks, in which both the safe and the risky alternatives had been selected in the first two trials (2,006 blocks out of 2,400 blocks). These trials were selected for analysis because they provide an opportunity to estimate the shape of the action selection function nonparametrically. To see this, consider the AIC Q-learning model in blocks in which both the safe and the risky alternatives were selected in the first two trials. According to Equation 1, the values of the risky action Q_3(risky) and the safe action Q_3(safe) in the third trial of these blocks are given by $Q_3(a) = (1 - \eta)Q_0 + \eta r_{t_a}$, where $t_a = \{1, 2\}$ is the trial number in which action a was selected. The difference between the values of the two alternatives $\Delta Q_3 = Q_3(\text{risky}) - Q_3(\text{safe})$ is independent of the initial conditions Q_0, and is linear in the reward difference $\Delta r = r_{t_{\text{risky}}} - r_{t_{\text{safe}}}$. The resulting linear relation $\Delta Q_3 = \eta \Delta r$ enables a direct estimation of the average action selection rule with a scale factor η. Similarly, in the framework of the RIC Q-learning model, the above derivation will result in the relation $\Delta Q_3 = \Delta r$.

Figure 3 depicts the probability of choosing the risky alternative in the third trial as a function of the difference in the rewards Δr. Note that in contrast to the ε-greedy action selection, the probability of choice is graded with the value of Δr even when $\Delta r \approx 0$. Moreover, in contrast to the softmax action selection rule, the probability of choice does not converge to a deterministic policy even when the absolute value of Δr is large. Thus, we chose to model the action selection rule of the participants with the ε-softmax rule (Equation 2), which manifests graded sensitivity to Δr while maintaining exploration even when the absolute difference between the action values is large. This ε-softmax rule was used during all the simulation conducted in this article.

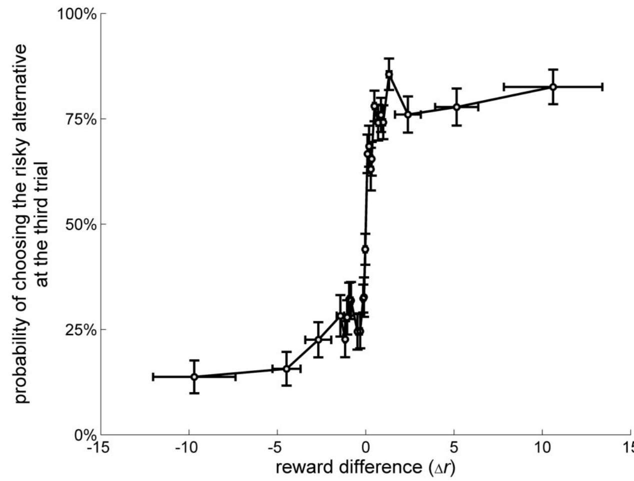

Figure 3. The action selection rule. The probability of choosing the risky alternative in the third trial as a function of the difference in the rewards between the risky and safe alternative in the first two trials averaged over 2,006 blocks in which both alternatives were sampled in the first two trials. The different blocks were grouped according to the value of Δr into 25 bins of approximately equal size. For each bin, the fraction of trials in which the risky alternative was chosen is plotted as a function of the average value of Δr. Error bars correspond to the standard error of the mean.

Underweighting of Rare Events

When learning from experience, participants are more risk aversive the smaller the probability of the high-outcome P_H, a phenomenon that has been termed *underweighting of rare events* (Barron & Erev, 2003; Hertwig et al., 2004) because the participants behave as if they underestimate the probability of the low-probability outcome. In order to quantify the magnitude of the underweighting of rare events in the experiment, we considered the aggregate probability of choosing the risky choice in the low- ($P_H < .15$) and high- ($P_H > .85$) P_H problems (see the Method section) separately. We found that the value of P_H had a substantial effect on participants' choices: in the high-P_H blocks, participants chose the risky alternative in 50 ± 3% of the trials (white in Figure 4A, top panel). By contrast, participants made a risky choice only in 27 ± 3% of the trials in the low-P_H blocks (black in Figure 4A, top panel). The significant difference in the two probabilities of choice, 23 ± 4%, is a measure of the magnitude of the underweighting of rare events effect, $t(89) = 9.1$, $p = 2 \times 10^{-14}$, 95% CI [18.4% 28.6%], $g = 1.91$. Note that this substantial difference in behavior occurred despite the fact that in both cases, the return of the risky alternative was approximately equal to that of the safe alternative (see Figure 1B).

The probability of a high reward (H) in the first risky trial (as in any risky trial) is P_H. Therefore, on average, there will be more H

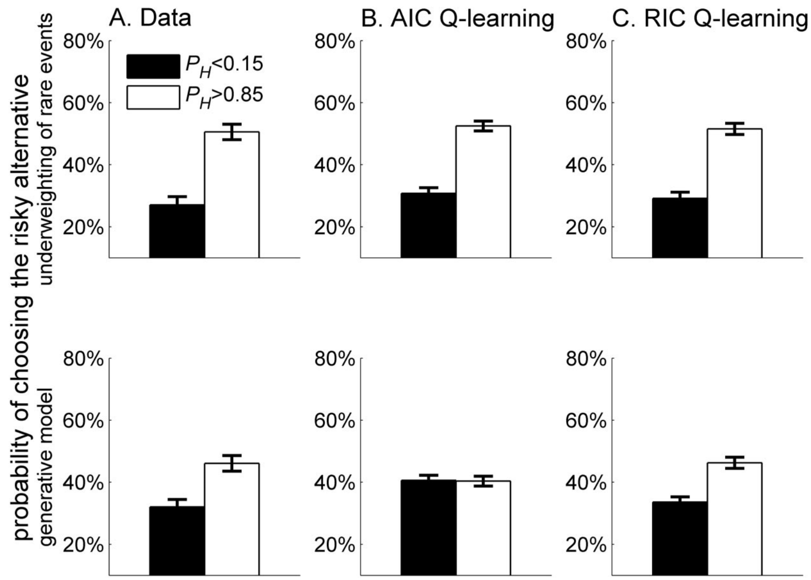

Figure 4. The underweighting of rare events and the generative model. Top panel: The probability of choosing the risky alternative averaged over the low-P_H blocks (black) and the high-P_H blocks (white). Bottom panel: The probability of choosing the risky alternative as predicted by the generative model based on the outcome of the first risky choice. A: The empirical data. B: Simulation of the arbitrary initial conditions (AIC) Q-learning model. C: Simulation of the resetting of initial conditions (RIC) Q-learning model. Error bars correspond to the standard error of the mean.

outcomes for the first risky choice in high-P_H blocks than in low-P_H blocks. Therefore, outcome primacy predicts that this excess of H outcomes in the high-P_H blocks should bias choice in favor of the risky alternative in those blocks, compared with behavior in the low-P_H blocks. Therefore, outcome primacy predicts underweighting of rare events. In order to quantify the contribution of outcome primacy to the underweighting of rare events, we constructed a generative model that predicts the effect of P_H on aggregate choice based on the two conditional probabilities A_L and A_H, which measure the effect of the first risky choice outcome on aggregate behavior (see Figure 2A, top panel). This generative model posits that the probability of a choosing the risky alternative in a block is determined solely by the binary outcome of the first risky choice, H or L. If that outcome is H, the model predicts that the participants would choose the risky alternative in A_H of the trials (see the Outcome Primacy section). If it is L, the risky alternative would be chosen in A_L of the trials. Consequently, according to this generative model, the probability of choosing the risky alternative in a trial in a problem characterized by P_H is:

$$\Pr\left[a = \text{‘}risky\text{’}\,;\, P_H\right] = A_H \cdot P_H + A_L \cdot (1 - P_H). \qquad (3)$$

In order to relate Equation 3, which predicts behavior for a given problem set to average behavior in the low- and high-P_H blocks (see Figure 4A, top panel), we averaged Equation 3 over the different problems, separately for the low- and high-P_H problem sets. The predictions of the generative model for the low- and high-P_H problems are depicted in Figure 4A (bottom panel) in black and white, respectively. The generative model predicts that the magnitude of the underweighting of rare events should be 14 ± 3%, approximately 60 ± 17% of the magnitude of the empirically measured underweighting of rare events (23 ± 4%). This result indicates that outcome primacy contributes substantially to the experimentally observed underweighting of rare events.

Whereas outcome primacy implies underweighting of rare events, the opposite case, namely that underweighting of rare events implies primacy, is not true. To see this, we analyzed the results of the simulations of the AIC Q-learning model and found significant underweighting of rare events: In the high-P_H blocks, the simulated participants chose the risky alternative in 52 ± 2% of the trials (white in Figure 4B, top panel). In contrast, the simulated participants chose “risky” only in 30 ± 2% of the trials in the low-P_H blocks (black in Figure 4B, top panel), $t(89) = 13.8$, $p = 9 \times 10^{-24}$, 95% CI [18.5%, 24.8%], $g = 2.89$. The underweighting of rare events in the AIC Q-learning model is in line with previous studies showing that the underweighting of rare events naturally emerges from RL models (see the Discussion section). Nevertheless, there is no outcome primacy in the AIC Q-learning model, ($A_H^{\mathrm{AIC}} \approx A_L^{\mathrm{AIC}}$), and therefore the generative model cannot explain the underweighting of rare events predicted by the AIC Q-learning model (0 ± 3% out of 22 ± 3%; see Figure 4B, bottom panel).

Similar to the behavioral data and to the AIC Q-learning model, there was a significant underweighting of rare events in the simulations of the RIC Q-learning model: Simulated participants chose the risky alternative in 51 ± 2% of the trials in the high-P_H blocks (white in Figure 4C, top panel) and in 29 ± 2% of the trials in the low-P_H blocks (black in Figure 4C, top panel), $t(89) = 11.8$, $p = 7 \times 10^{-20}$, 95% CI [18.6%, 26.1%], $g = 2.47$. This underweighting of rare events in the simulations is not statistically different from the experimentally observed effect, $t(84) = 0.86$, $p = .39$, 95% CI [−7.0%, 2.7%], $g = 0.18$; and, $t(84) = 0.46$, $p = .65$, 95% CI [−5.2%, −3.2%], $g = 0.09$, for the low and high $P_{H,}$

respectively. Similar to the behavioral data and in contrast to the AIC Q-learning model, outcome primacy accounts for 56 ± 13% of the magnitude of underweighting of rare events in the simulation of the RIC Q-learning model (12 ± 2% out of 22 ± 3%; see Figure 4C, bottom panel).

Another way of demonstrating the contribution of outcome primacy to the underweighting of rare events is to compare the average aggregate choice in the low- and high-P_H blocks, conditioned on the outcome of the first risky choice. We denote these averages by $A_{r_1}^{P_H}$, where $r_1 \in \{L, H\}$ is the outcome of the first risky choice and $P_H \in \{\uparrow, \downarrow\}$ is the P_H block type (low or high, respectively). If participants' aggregate choice behavior is dominated by the primacy effect, it is expected that the P_H block type will have a negligible effect on behavior once conditioned on the first risky outcome, formally, $A_H^{\downarrow} \approx A_H^{\uparrow}$ and $A_L^{\downarrow} \approx A_L^{\uparrow}$. In contrast, if participants' sensitivity to the value of P_H is not mediated by the outcome of the first risky choice, it is expected that within a P_H block type, this outcome will have only a negligible effect on behavior, $A_H^{\downarrow} \approx A_L^{\downarrow}$ and $A_H^{\uparrow} \approx A_L^{\uparrow}$.

Figure 5A depicts the values of $A_{r_1}^{P_H}$, where blue and red hues denote H and L, and dark and light brightness denote low- and high-P_H block type, respectively. We found that the contribution of block type to aggregate behavior was smaller than the contribution of the outcome of the first risky choice. To quantify this result, we used a two-way analysis of variance that showed that the outcome of the first reward effect was statistically significant, $F(1, 149) = 36.13$, $MSE = 1.56$, $\rho = 0.46$, $p = 1.4 \times 10^{-8}$. By contrast, the contribution of the P_H block type and its interaction with the outcome of the first risky choice were not statistically significant, $F(1, 149) = 2.08$, $MSE = 0.09$, $\rho = 0.2$, $p = .15$; and, $F(1, 149) = 1.28$, $MSE = 0.05$, $p = .26$, respectively. These results indicate that the outcome of the first risky choice is the major contributor to the underweighting of rare events and further support the hypothesis that the outcome primacy effect plays an important role in aggregate choice behavior.

Repeating the same analysis for the AIC Q-learning model (see Figure 5B) revealed that in this model, the P_H block type dominates choice behavior, $F(1, 147) = 136.74$, $MSE = 1.67$, $\rho = 0.71$, $p = 1 \times 10^{-22}$, and not the outcome of the first risky choice, $F(1, 147) = 0.2$, $MSE = 2.4 \times 10^{-3}$, $\rho = 0.2$, $p = .66$. By contrast, in the RIC Q-learning model (see Figure 5C), similar to the behavior of the participants, the outcome of the first risky choice affected choice behavior more strongly than the P_H block type, $F(1, 144) = 45.56$, $MSE = 0.99$, $\rho = 0.54$, $p = 3 \times 10^{-10}$; and, $F(1, 144) = 9.31$, $MSE = 0.202$, $\rho = 0.34$, $p = 3 \times 10^{-3}$, respectively.

Discussion

The primary objective of this study was to test our hypothesis that first experience resets the initial conditions in operant learning. We showed that, indeed, the outcome of the first risky choice has a long-lasting effect on subsequent choice behavior, a phenomenon we termed outcome primacy (see Figure 2A). To the best of our knowledge, the question of primacy in operant learning has never been addressed. To test our hypothesis, we estimated the action selection function nonparametrically, modeled it using a ε-softmax function, and implemented it in a Q-learning model (see Figure 3). In line with our hypothesis, this standard RL model is consistent with the effect of the outcome of the first choice on behavior if we assume that the outcome of the first choice resets the value of the action (see Figure 2C), but not if we assume arbitrary initial conditions (see Figure 2B). Our hypothesis is

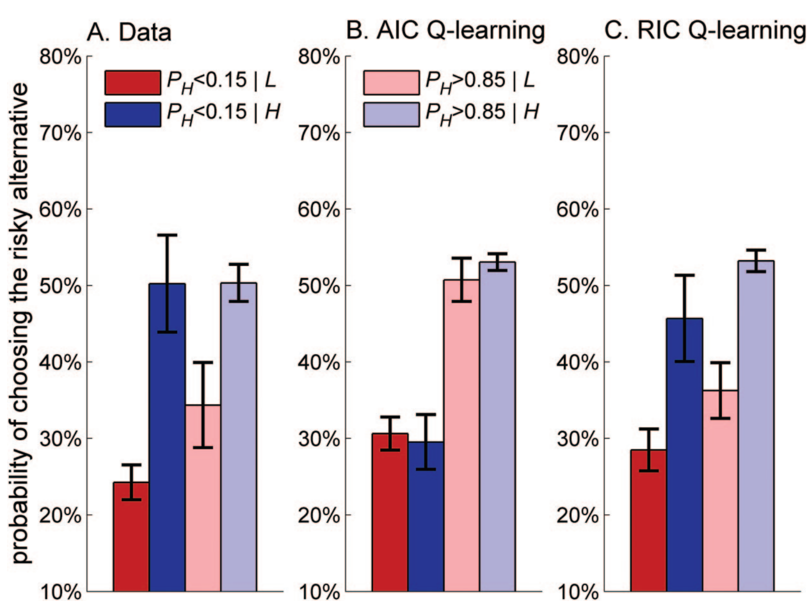

Figure 5. The underweighting of rare events, conditioned on the outcome of the first risky choice. The probability of choosing the risky alternative for the low-P_H (dark) and high-P_H (bright) blocks, conditioned on the outcome of the first choice: L (low, red) and H (high, blue). A: The empirical data. B: Simulation of the arbitrary initial conditions (AIC) Q-learning model. C: Simulation of the resetting of initial conditions (RIC) Q-learning model. Error bars correspond to the standard error of the mean.

further supported by the fact that our model predicts aggregate probability of choice in operant learning more accurately than other previously proposed models (see Table 1). Finally, our results indicate that outcome primacy substantially contributes to the underweighting of rare events (see Figures 4 and 5). These results strongly suggest that outcome primacy plays an important role in shaping behavior in operant-learning tasks.

The RIC Hypothesis and the Underweighting of Rare Events

Previous studies have suggested that the underweighting of rare events can result from *estimation bias*, which is enhanced by adaptive sampling (Denrell, 2005, 2007), also known as the hot stove effect (Denrell & March, 2001). The idea behind estimation bias is that if P_H is sufficiently small, the empirical average of the past outcomes of the risky choices is typically lower than the true (ensemble) average. The opposite effect is expected in problem sets in which P_H is sufficiently large. This effect is particularly pronounced if participants rely on a relatively small sample, due to either limited memory or overweighting of recent samples (Barron & Erev, 2003; Erev, Ert, & Yechiam, 2008; Hertwig et al., 2004). Researchers have hypothesized that the finite number of samples in the experiment is sufficient to account for the estimation bias and the underweighting of rare events (Fox & Hadar, 2006). However, this hypothesis has been contested by findings that rare events are underweighted even when the sample is representative (Hau, Pleskac, Kiefer, & Hertwig, 2008; Hertwig & Erev, 2009; Ungemach, Chater, & Stewart, 2009). It should be noted that *recency*, in which more recent samples are more influential than other samples (Hogarth & Einhorn, 1992), would result in a biased estimation even in representative examples (Hertwig et al., 2004). Such recency naturally emerges in Q-learning (both AIC and RIC) because of the adaptation rule (Equation 1). Similarly, the resetting of initial conditions results in more weight being given to a single experience, the first experience, which yields a similar estimation bias.

Adaptive sampling enhances the estimation bias by the following asymmetry: If the decision maker temporarily underestimates the value of the risky alternative, she or he will tend to avoid it. By contrast, an overestimation of the value of the risky alternative will motivate additional choices of the risky alternative and hence reduce the bias. Adaptive sampling affects choice behavior in two ways. First, it biases participants against the risky alternative, resulting in risk-aversion behavior. Second, it amplifies the underweighting of rare events caused by the estimation bias (Denrell, 2005, 2007). Estimation bias and hot stove effects are implicitly incorporated in the AIC and RIC Q-learning models. The value adaptation results in estimated action values based primarily on the most recent trials. Adaptive sampling is a natural consequence of the action selection rule. In fact, substantial underweighting of rare events was observed in our simulations of the AIC Q-learning model (see Figure 4B, top panel), consistent with previous studies (Denrell, 2005, 2007).

Our analysis focused on the contribution of the first experience, through the outcome primacy, to the underweighting of rare events. The analysis of the empirical data showed that outcome primacy accounts for a substantial part of the underweighting of rate events (see Figures 4A and 5A), which is consistent with the RIC Q-learning model (see Figures 4C and 5C). According to the RIC Q-learning model, the outcome of the first choices makes a disproportionately large contribution to the action values. This overweighting of the first experience effectively decreases the sample used for estimating the action values and thus enhances the estimation bias, and, consequently, the underweighting of rare events.

The underweighting of rare events depicted in Figures 4 and 5 is quantified as an average over the entire block of 100 trials. Because the contribution of the outcome of the first risky choice to behavior decreases with trial number and because outcome primacy contributes substantially to the underweighting of rare events, the magnitude of the underweighting of rare events is expected to decrease with trial number as well. To test this, we computed the magnitude of the underweighting of rare events for each trial individually by computing the difference in the probabilities of choosing the risky alternative in the high- and low-P_H blocks.

As depicted in Figure 6 (magenta), the magnitude of the underweighting of rare events increases within several trials and decreases gradually throughout the block. The phasic increase can be attributed to the resetting of the initial conditions, whereas the decrease can be attributed to an effective increase in the number of samples in the action value estimation, which in turn decreases the sampling bias. This dynamics of the underweighting of rare events is consistent with the simulations of the RIC Q-learning model (see Figure 6, black, first 100 trials).

The simulations of the RIC Q-learning model can also be used to predict the magnitude of underweighting of rare events in a longer experiment. As depicted in Figure 6, black, the magnitude of the underweighting of rare events is expected to plateau at a positive value in longer experiments. This residual underweighting of rare events in the steady state is independent of the reset of initial conditions.

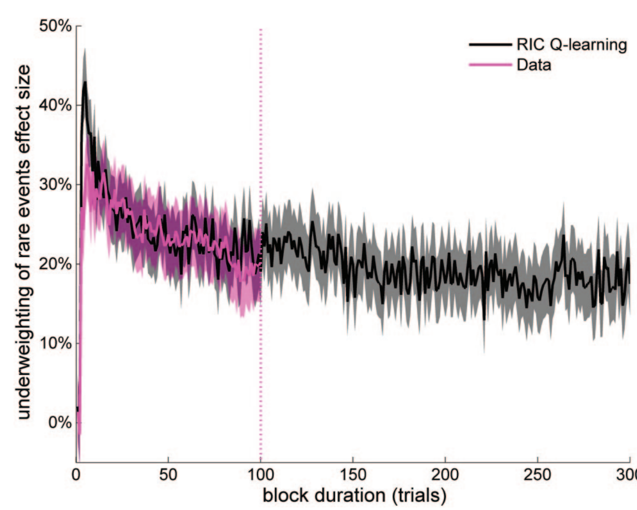

Figure 6. The magnitude of the underweighting of the rare event effect (difference between the probabilities of choosing the risky alternative in high- and low-P_H blocks) for each trial computed for the empirical data set (magenta) and for the resetting of initial conditions (RIC) Q-learning simulation with parameters estimated for each participant. Shadowed margins correspond to the standard error of the mean. The dotted vertical line marks the 100th trial in the block.

Predicting Outcome Primacy in Different Experimental Paradigms

The participants in the experiment exhibited outcome primacy, whose magnitude can be quantified as the difference between the probabilities of choosing the risky choice when the outcome of the first risky choice is H and that probability when the outcome of the first risky choice is L ($\Delta A^{\mathrm{data}} = A_H - A_L = 16 \pm 4\%$). In this section, we consider the contributions of two main characteristics of the experimental schedule to the outcome primacy: (a) the fact that in each trial, only the payoff of the chosen alternative was known to the participant, also known as *obtained payoff*, and (b) the fact that the expected returns from the two alternatives were approximately equal.

In order to estimate the contribution of the obtained payoff paradigm to outcome primacy, we simulated the RIC Q-learning model in a *forgone payoff* paradigm in which both the obtained outcome from the chosen alternative and the foregone outcome from the nonchosen alternative are known to the participant after each trial. Averaging over the problem sets, we found that the magnitude of outcome primacy in the simulation of the forgone payoff paradigm is $\Delta A^{\mathrm{forgone}} = 5 \pm 3\%$, which is significantly smaller than ΔA^{data}, $t(175) = 3.3$, $p = 1 \cdot 10^{-3}$, 95% CI [4.6%, 18.2%], $g = 0.50$. Thus, the contribution of adaptive sampling to outcome primacy is substantial, and we predict that the magnitude of the outcome primacy in a forgone payoff paradigm would be substantially lower than in an obtained payoff paradigm.

To test for the contribution of equal expected rewards to outcome primacy, we repeated the simulations of the RIC Q-learning model for each of the participants, while varying the value of the safe alternative, M, according to $M' = M + q|M|$. The parameter q is a measure of the deviation of the reward schedule from equal returns. The original reward schedule of approximately equal returns corresponds to $q = 0$, whereas a positive (negative) value of q indicates that the value of the safe alternative is larger (smaller) than the expected average reward of the risky alternative.

The top panel in Figure 7 depicts the probability of choosing the risky choice given that the outcome of the first trial was high (H, blue) or low (L, red) as a function of q. The lower panel depicts the difference between these two curves. The solid circles in both plots denote the empirical values, A_L (red in top panel), A_H (blue in top panel), and ΔA^{data} (black in bottom panel). The results of these simulations predict that the magnitude of outcome primacy should be maximal when the two alternatives have approximately the same return. Nevertheless, substantial outcome primacy is expected in all the values of q that we studied ($-1 < q < 1$).

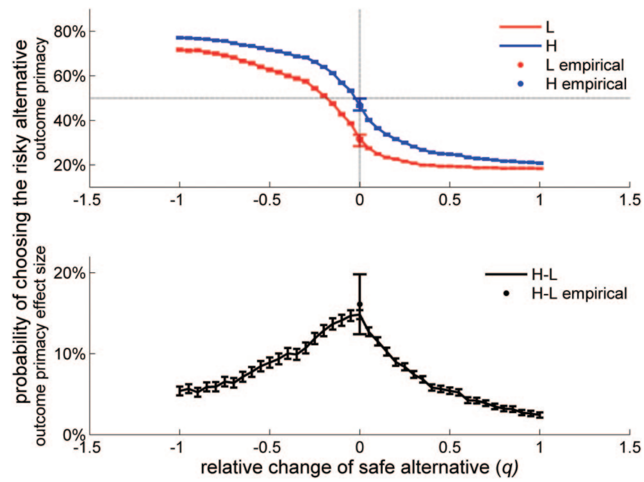

Figure 7. The predicted dependency of the outcome primacy effect on reward schedule according to the simulation of the resetting of initial conditions Q-learning model with parameters of each participant estimated. Top panel depicts the probability of choosing the risky alternative given that the outcome of the first risky choice was either high (H, blue) or low (L, red), as a function of the value of the parameter q, which controls the value of the safe alternative according to $M' = M + q|M|$, where M is the original value in the empirical data, and M' is the safe value used in the simulation. Horizontal and vertical gray lines correspond to choice indifference ($p = .5$) and equal returns ($q = 0$), respectively. Bottom panel depicts the difference between the two curves in the top panel. The solid circles denote the empirical values of A_L, A_H (red and blue in top panel) and ΔA^{data} (black in bottom panel). Simulation was conducted over 20 repetitions of the original experiment (200 participants, 12 blocks each) with the parameter q varying between -1 and 1 in steps of 0.05 (total 41 values). The error bars corresponds to the simulation and data standard error of the mean.

RIC Model as Nonstationary Learning

The magnitude of the learning rate determines the speed–accuracy trade-off in learning. Therefore, the time-dependent learning rate, in which the rate is initially high and later low, is common in machine learning in general and reinforcement learning in particular (Sutton & Barto, 1998). In line with this framework, the RIC model is mathematically equivalent to an AIC model, in which the learning rate changes according to the following rule: $\eta_1(a) = 1$ and $\eta_t = \eta$ for $t > 1$, where η is a constant and $\eta_t(a)$ is the learning rate after t choices of alternative a. Consistent with this idea, the resetting of initial conditions ensures that after a single trial, the estimated action values are in the ballpark of the true values, enabling fast convergence to the true values. By contrast, an arbitrary initial value may be far from the true value, resulting in a slow convergence of the algorithm. We postulate that this might be the rationale behind this cognitive strategy. Furthermore, for a deterministic action–outcome relation, resetting would be the optimal policy for estimating the action value correctly and quickly.

To further test the validity of the RIC Q-learning model, we tested whether other models incorporating a time-dependent learning rate could explain the behavioral data better. In particular, we focused on power-law learning of the form $\eta_t = 1/t^{\alpha}$ because it guarantees convergence of the estimated action value to its true value under general conditions if $1 \leq \alpha < 2$ (Sutton & Barto, 1998). We found that the likelihood of the power-law model is lower than that of the RIC Q-learning model, and qualitatively, the resulting behavior does not capture the primacy effect (not shown).

Nevertheless, it is likely that the RIC Q-learning model is at best a coarse approximation of the true learning strategy. Therefore, more accurate models should take into account time-dependent changes in the adaptation rule as well as in the action selection rule. However, an accurate description of the dynamics of these rules is difficult because of the heterogeneity in learning between different participants, because our only access to the subjective values is via their binary choices and because these rules could be task dependent.

Beyond the RIC Q-Learning Model

One limitation of RIC Q-learning is that it implicitly assumes that consecutive blocks are independent and that prior expectations play no role in the model. However, this is only an approximation of the behavior. To see this, we computed the probabilities of choosing the risky alternative in the second trial following a risky choice in the first trial, conditioned on the outcome of the first trial (*H* or *L*). According to the RIC model, these probabilities are determined by the parameter ε in the action selection rule and are independent of the outcome of the first trial. We found that these probabilities are statistically different: 21 $\pm$ 4% and 14 $\pm$ 4%, after *H* and *L*, respectively, $t(186) = 2.03$, $p = .043$, 95% CI [0%, 12.6%], $g = 0.30$. This result might indicate that prior expectations of the participants also influence their choice behavior in a way that is not predicted by the RIC Q-learning model.

Outcome Primacy and Other Forms of Primacy

The long-lasting effect of the first outcome, which we denoted as outcome primacy, is reminiscent of other forms of primacy in psychology (Mantonakis et al., 2009), where "earlier data have more impact [on behavior] than later data" (Peterson & DuCharme, 1967, p. 1). For example, in memory recall tasks, the probability of recalling the first item in a list is higher than the probability of recalling subsequent items (Murdock, 1962). Similarly, in multiple-choice tasks, in which opinion is based on one-shot experience per option, such as in wine tasting, the first option is more likely to be chosen (Mantonakis et al., 2009). Although the relation between the above examples of primacy and outcome primacy is unclear, we hypothesize that outcome primacy and primacy in *belief-updating tasks*, such as jurors' decision after a sequence of argumentative speeches or the stating of a personality impression after a sequence of words describing personality traits (Asch, 1946; Cromwell, 1950; Lund, 1925; Peterson & DuCharme, 1967; Stone, 1969), can be explained using a similar theoretical framework.

Belief-updating tasks resemble repeated choice tasks in the fact that participants respond after being provided with a sequence of evidence. However, in contrast to the quantitative nature of the sequence of rewards in repeated choice tasks, the evidence in belief-updating tasks can be qualitative and not easily comparable. Order effects in the belief-updating tasks have been previously modeled using the *belief-adjustment model*, in which evidence, despite its qualitative nature, is converted to a numerical reinforcement and is used to update the value associated with the evidence's source, in a manner very similar to Q-learning (Hogarth & Einhorn, 1992). An important difference between the belief-adjustment model and the RIC Q-learning model is that in the former model, the representation of the first experience is nondecaying, whereas in the latter model, first experience resets the initial conditions. This difference in the models manifests in a different prediction: primacy in the belief-adjustment model is predicted to be everlasting, whereas primacy in the RIC Q-learning model is predicted to be a transient, albeit possibly long-lasting, phenomenon. We are unaware of studies of primacy in long belief-updating tasks (we demonstrated outcome primacy in a task, in which the two sequences of evidence are composed of tens of trials). However, in a memory recall task, the magnitude of primacy has been shown to decrease with the length of the list (Murdock, 1962).

Researchers have also suggested primacy emerges because participants pay less attention to successive items of evidence (Anderson, 1981). In the framework of the Q-learning model, this attention decrement can be modeled as a decrease in the learning rate. As discussed above, the RIC hypothesis is a simple example of a time-dependent learning rate, in which the learning rate is initially high and is lower in successive trials.

Conclusion

Learning from experience is one of the most compelling aspects of human cognition. Reinforcement learning provides a computational framework for studying learning from experience by using past actions and their outcome to estimate action values, which in turn are used to direct future actions. Nevertheless, when learning starts, neither previous actions nor outcome are available, and thus initial conditions should be defined. In this article, we described the long-lasting contribution of the first experience to behavior, a phenomenon we termed outcome primacy. The long time scale associated with this effect indicates that behavior does not converge to a steady state within 100 trials, and thus the aggregate behavior reported in experiments may not reflect the asymptotic expected behavior. Outcome primacy can be understood in the framework of RL if we assume that initial conditions are reset by the outcome of first experience. We suggest that the resetting of the initial condition is a general trait of human and animal operant learning, which may be related to other forms of primacy and should not be overlooked when modeling and predicting learning from experience.

References

Anderson, N. (1981). *Foundations of information integration theory*. Boston, MA: Academic Press.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41,* 258–290. doi:10.1037/h0055756

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making, 16,* 215–233. doi:10.1002/bdm.443

Barto, A., Sutton, R., & Watkins, C. (1989). Learning and sequential decision making. In M. Gabriel & J. W. Moore (Eds.), *Learning and computational neuroscience* (pp. 539–602). Cambridge, MA: MIT Press.

Cohen, J., McClure, S., & Angela, J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362,* 933–942. doi:10.1098/rstb.2007.2098

Cromwell, H. (1950). The relative effect on audience attitude of the first versus the second argumentative speech of a series. *Communication Monographs, 17,* 105–122. doi:10.1080/03637755009375004

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning: Attention and performance XXIII*. New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780199600434.003.0001

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, The bad and the ugly. *Current Opinion in Neurobiology, 18,* 185–196. doi:10.1016/j.conb.2008.08.003

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review, 112,* 951–978. doi:10.1037/0033-295X.112.4.951

Denrell, J. (2007). Adaptive learning and risk taking. *Psychological Review, 114,* 177–187. doi:10.1037/0033-295X.114.1.177

Denrell, J., & March, J. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science, 12,* 523–538. doi:10.1287/orsc.12.5.523.10092

Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *HFSP, 30,* 30–40. doi:10.2976/1.2732246/10.2976/1

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112,* 912–931. doi:10.1037/0033-295X.112.4.912

Erev, I., Ert, E., & Roth, A. E. (2008). *Motivation and the basic idea*. Retrieved from http://tx.technion.ac.il/~erev/Comp/Comp.html

Erev, I., Ert, E., Roth, A., Haruvy, E., Herzog, S., Hau, R., . . . Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making, 23,* 15–47. doi:10.1002/bdm.683

Erev, I., Ert, E., & Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *Journal of Behavioral Decision Making, 21,* 575–597. doi:10.1002/bdm.602

Erev, I., Roth, A., Slonim, R., & Barron, G. (2007). Learning and equilibrium as useful approximations: Accuracy of prediction on randomly selected constant sum games. *Economic Theory, 33,* 29–51. doi:10.1007/s00199-007-0214-y

Fox, C., & Hadar, L. (2006). Decisions from experience = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making, 1,* 159–161.

Gallistel, C., Mark, T., King, A., & Latham, P. (2001). The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *Journal of Experimental Psychology: Animal Behavior Processes, 27,* 354–372. doi:10.1037//0097-7403.27.4.354

Glimcher, P. (2009). Choice: Towards a standard back-pocket model. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 503–521). London, England: Academic Press.

Hau, R., Pleskac, T., Kiefer, J., & Hertwig, R. (2008). The description-experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making, 21,* 493–518. doi:10.1002/bdm.598

Hertwig, R., Barron, G., Weber, E., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15,* 534–539. doi:10.1111/j.0956-7976.2004.00715.x

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13,* 517–523. doi:10.1016/j.tics.2009.09.004

Hogarth, R., & Einhorn, H. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24,* 1–55. doi:10.1016/0010-0285(92)90002-J

Lagarias, J., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization, 9,* 112–147. doi:10.1137/S1052623496303470

Lattal, K. (1998). A century of effect: Legacies of E. L. Thorndike's Animal Intelligence Monograph. *Journal of the Experimental Analysis of Behavior, 70,* 325–336. doi:10.1901/jeab.1998.70-325

Law, C., & Gold, J. (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature Neuroscience, 12,* 655–663. doi:10.1038/nn.2304

Loewenstein, Y. (2010). Synaptic theory of replicator-like melioration. *Frontiers in Computational Neuroscience, 4,* 1–12. doi:10.3389/fncom.2010.00017

Loewenstein, Y., & Seung, H. (2006). Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. *Proceedings of the National Academy of Sciences, 103,* 15224–15229. doi:10.1073/pnas.0505220103

Lund, F. (1925). The psychology of belief. *Journal of Abnormal and Social Psychology, 20,* 63–81. doi:10.1037/h0076047

Mantonakis, A., Rodero, P., Lesschaeve, I., & Hastie, R. (2009). Order in choice: Effects of serial position on preferences. *Psychological Science, 20,* 1309–1312. doi:10.1111/j.1467-9280.2009.02453.x

Murdock, B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64,* 482–488. doi:10.1037/h0045106

Neiman, T., & Loewenstein, Y. (2011). Reinforcement learning in professional basketball players. *Nature Communications, 2,* 1–7. doi:10.1038/ncomms1580

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R., & Frith, C. (2006, July 7). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature, 442,* 1042–1045. doi:10.1038/nature05051

Peterson, C., & DuCharme, W. (1967). A primacy effect in subjective probability revision. *Journal of Experimental Psychology, 73,* 61–65. doi:10.1037/h0024139

Stone, V. (1969). A primacy effect in decision-making by jurors. *Journal of Communication, 19,* 239–247. doi:10.1111/j.1460-2466.1969.tb00846.x

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004, June 18). Matching behavior and the representation of value in the parietal cortex. *Science, 304,* 1782–1787. doi:10.1126/science.1094765

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Thorndike, E. (1911). Animal intelligence. In E. Thorndike (Ed.), *Animal intelligence* (p. 244). New York, NY: Macmillan.

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science, 20,* 473–479. doi:10.1111/j.1467-9280.2009.02319.x

Watkins, C. (1989). *Learning from delayed rewards* (Doctoral dissertation). University of Cambridge, Cambridge, England. Retrieved from http://www.cs.rhul.ac.uk/home/chrisw/thesis.html

Watkins, C., & Dayan, P. (1992). Q-learning. *Machine Learning, 8,* 279–292. doi:10.1007/BF00992698

Received August 23, 2011
Revision received June 27, 2012
Accepted July 3, 2012 ■