Cogn Affect Behav Neurosci
DOI 10.3758/s13415-014-0260-4

Physiological and behavioral signatures of reflective exploratory choice

A. Ross Otto · W. Bradley Knox · Arthur B. Markman · Bradley C. Love

Abstract Physiological arousal, a marker of emotional response, has been demonstrated to accompany human decision making under uncertainty. Anticipatory emotions have been portrayed as basic and rapid evaluations of chosen actions. Instead, could these arousal signals stem from a "cognitive" assessment of value that utilizes the full environment structure, as opposed to merely signaling a coarse, reflexive assessment of the possible consequences of choices? Combining an exploration–exploitation task, computational modeling, and skin conductance measurements, we find that physiological arousal manifests a reflective assessment of the benefit of the chosen action, mirroring observed behavior. Consistent with the level of computational sophistication evident in these signals, a follow-up experiment demonstrates that anticipatory arousal is modulated by current environment volatility, in accordance with the predictions of our computational account. Finally, we examine the cognitive costs of the exploratory choice behavior these arousal signals accompany by manipulating concurrent cognitive demand. Taken together, these results demonstrate that the arousal that accompanies choice under uncertainty arises from a more reflective and "cognitive" assessment of the chosen action's consequences than has been revealed previously.

Keywords Decision-making · Reward · Reinforcement learning · Emotion · Arousal

A. R. Otto (✉)
Center for Neural Science, New York University,
4 Washington Place, New York, NY 10003, USA
e-mail: rotto@nyu.edu

W. B. Knox
Massachusetts Institute of Technology, Cambridge, MA, USA

A. B. Markman
University of Texas at Austin, Austin, TX, USA

B. C. Love
University College London, London, UK

Published online: 25 March 2014

Introduction

Emotional response and its concomitant peripheral autonomic response play a central role in the way people manage decisions under uncertainty in a variety of task contexts (Critchley, 2005; Dolan, 2002; Figner, Mackinlay, Wilkening, & Weber, 2009; Mellers, Schwartz, Ho, & Ritov, 1997). For example, in the context of decision making, people exhibit arousal, measured by skin conductance responses (SCRs; Öhman & Soares, 1994) just prior to choices carrying potential monetary losses (Bechara, Tranel, Damasio, & Damasio, 1996; Suzuki, Hirota, Takasawa, & Shigemasu, 2003) or future cognitive costs (Botvinick & Rosen, 2009), suggesting that these arousal signals reflect, in some form, an evaluation of a chosen action's goodness.

Although previous work has considered the causal role of autonomic arousal in choice under uncertainty (Bechara et al., 1996; Damasio, 1994; but see Dunn, Dalgleish, & Lawrence, 2006; Tomb, Hauser, Deldin, & Caramazza, 2002; Whitney, Hinson, Wirick, & Holben, 2007), little work has endeavored to characterize the processes generating these arousal signals. In this report, we utilize the framework of reinforcement learning (RL; Sutton & Barto, 1998) to provide a computationally informed examination of how insightful these signals are.

Affective research has often portrayed these anticipatory emotions as basic and rapid evaluations of the options facing a decision-maker—possibly facilitating rapid action—in contrast to a "cognitive" evaluation of a course of action (Ledoux, 1996; Loewenstein, Weber, Hsee, & Welch, 2001; Zajonc, 1984). Could these emotional responses instead stem from a reflective and intelligent decision-making system, as opposed to a simpler, reflexive decision process? Indeed, work in the domain of aversive Pavlovian conditioning has demonstrated how causal knowledge and more explicit, "cognitive" information can produce anticipatory arousal responses (Lovibond, 2003; Olsson & Phelps, 2004). Here, we evaluate the possibility that SCRs accompanying choice register planning-oriented value calculations utilizing environment structure and evolving uncertainty,

in contrast to merely signaling a coarse calculation of the possible consequences resulting from the choice made.

Answering these questions is critical to understanding the nature of interactions between cognition and emotion. For example, patients with ventromedial prefrontal cortex (vmPFC) lesions fail to manifest anticipatory SCRs for actions carrying large potential monetary losses in putatively risky decision making, and moreover, their choices appear insensitive to these negative consequences (Bechara, Damasio, Damasio, & Lee, 1999; Bechara et al., 1996). It is unclear to what extent these apparent physiological and behavioral anomalies stem from the breakdown of a computationally sophisticated, reflective choice system versus a simpler reflexive system.

A key challenge in characterizing these signals lies in constructing a computational account of choice in paradigms examining autonomic response to risky decision making (Busemeyer & Stout, 2002; Studer & Clark, 2011; Worthy et al., 2013). Furthermore, in the Iowa gambling task—ubiquitously used to examine emotional arousal accompanying risky choice—it is well documented that healthy participants exhibit pronounced SCRs when they make selections to "disadvantageous" actions with negative (experienced) expected utility. Because actions in this paradigm are represented by stimuli that remain constant throughout the task, it is unclear whether arousal signals are merely tied to specific stimuli themselves or, instead, reflect a deeper assessment of the chosen action's goodness.

Here, we utilize a well-understood choice task called the *leapfrog* task (Knox, Otto, Stone, & Love, 2012), in which decision-makers must continually adapt their choice behavior in response to changing payoffs. The task is sufficiently constrained such that our computational modeling approach allows us to clearly delineate between competing accounts of behavior (Blanco, Otto, Maddox, Beevers, & Love, 2013; Knox et al., 2012) and their relationship to arousal responses accompanying choice.

The leapfrog task leverages the tension, found in many real-world decision-making situations, between exploitative choice (choosing a known option that is believed to have yielded the best outcome in the past) and exploratory choice (choosing a possibly inferior option with the hope that it will yield an even better result). This trade-off is a nontrivial problem, and the way in which people negotiate this balance is the subject of a spate of recent cognitive neuroscience research concerning its neural and physiological signatures (Badre, Doll, Long, & Frank, 2012; Cohen, McClure, & Yu, 2007; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Jepma & Nieuwenhuis, 2011).

Consider the choice task depicted in Fig. 1a, termed the *leapfrog* task, in which the decision-maker repeatedly makes choices among options A and B, each time observing the obtained payoff. Although one option is always superior to the other by 10 points, the payoffs associated with the two options change over time in a constrained manner: With some fixed probability, option B (which is initially inferior) increases in value by 20 points and, thereby, becomes superior to option A, and with this same probability, option A can subsequently overtake option B as the superior option. Because the relative superiority of the options changes over time, decision-makers must negotiate the competing demands of exploration and exploitation: An exploitative decision-maker will miss jumps in payoff levels and persist in choosing options that have become inferior, while an overly exploratory decision-maker will incur large opportunity costs associated with sampling the observed inferior option too frequently, thus forgoing the higher payoffs associated with the superior option. An example participant's sequence of choices, denoted with Xs and Os, is superimposed on the payoffs in Fig 1a.

Importantly, this constrained "bandit" task allows us to identify whether people approach exploration in a reflective versus reflexive fashion (Blanco et al., 2013; Knox et al., 2012). A reflexive strategy is informed only by directly observing payoffs and, thus, relies upon occasional, undirected random choices to the observed inferior action in order to explore. A reflective strategy, by contrast, leverages knowledge of the full structure of the environment to maintain a belief about the currently superior option. This evolving belief—which incorporates predictions of unobserved changes in the payoffs of the two options—guides the choice between an exploratory or an exploitative action at each decision. With each successive exploitive choice, the probability that the relative value of the options has flipped increases, making the state of the environment less certain. In this way, exploratory behavior is directed by uncertainty about the state of the environment; as uncertainty increases, exploration becomes more valuable. Critically, the behavioral signatures of the two strategies can be identified on the basis of sequential dependence in exploratory choice: Reflexive choice produces unconditionally and equiprobable exploration over time, while a reflective strategy entails that longer periods of consecutive exploitative choice necessitate more exploration.

We formulated two computational models of choice to verify whether participants negotiated exploration in a reflective manner, informed by predictions of unobserved changes in the environment, or in a reflexive manner, informed only by direct payoff observations. And more interestingly, we used these models to elucidate the reflective signature of the anticipatory SCRs accompanying this choice behavior. We provide qualitative descriptions of each model below; full algorithmic details are given in the Appendix.

The Naïve RL model is reflexive model that assumes that action-values are updated in a reactive fashion to directly experienced rewards. This model reflexively maintains beliefs about payoffs based only on directly observed payoffs. In other words, its estimated payoffs for each action (called *Q*-values in RL) are those most recently observed for the two actions A and B. The crucial feature of its predicted behavior—in the leapfrog task—is a constant probability of making an exploratory choice. This predicted behavior and the contrasting predictions of the reflective model are described in Experiment 1.

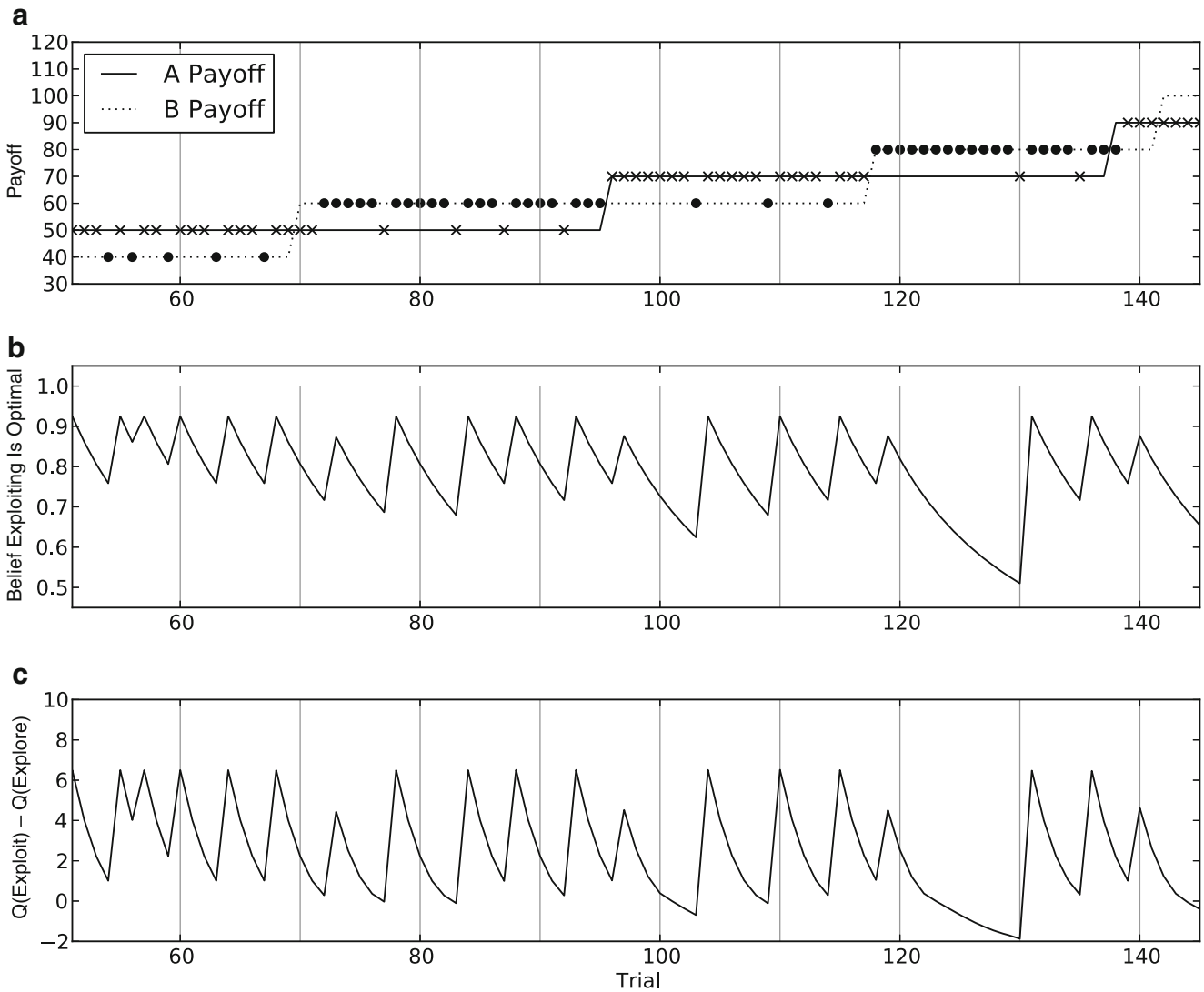

Fig. 1 Choice task and model-inferred variables. **a** An example instantiation of the leapfrog task, in which participants make repeated choices between two options, A and B, each time observing the payoff for the selected option. An example participant's sequence of choices and payoff observations are overlaid as Xs and Os. **b** The Ideal Actor's belief at each choice, based on choices and observations made by the participant up to that point. **c** The Ideal Actor's calculated relative value, based on the beliefs in panel b, of taking the exploitative action, expressed as Q(exploit) − Q(explore). Taking the exploitative action is optimal whenever the relative value of exploiting is positive; when this value is negative, the exploratory action is optimal

The *Ideal Actor*, with knowledge of the task structure, integrates its past beliefs with present observations to produce a belief about the currently higher-paying option, using simple Bayesian inference. Figure 1b depicts this evolving belief, inferred from the example participant's sequence choices in Fig. 1a. These optimal beliefs are then transformed into action-values by using existing techniques in RL (see the Appendix), resulting in action-values associated with exploitative and exploratory actions. Figure 1c depicts the Ideal Actor's calculated relative value of the exploitative action [expressed as Q(exploit) − Q(explore)] as a function of its belief. Intuitively, the Ideal Actor prescribes an exploitative action whenever this quantity is greater than zero and an exploratory action otherwise.

The Ideal Actor's directed form of exploration contrasts with the purely random exploration exhibited by the Naïve RL model—in which action-values are updated in a reflexive fashion to match directly observed payoffs. Unlike the reflective Ideal Actor, this reflexive model does not fully utilize environment structure. Beyond the qualitatively different patterns of behavior ascribed to the reflexive and reflective accounts of choice, we also intuited that the two modes of choice impose different requirements on cognitive resources. On this view, usage of reflective choice should be constrained by available central executive resources, given its comparatively greater computational expenses—a prediction we also test in this report.

In Experiment 1, we examine anticipatory SCRs while participants negotiate the exploration–exploitation trade-off in the leapfrog task. Since the option payoffs continually change in the task, the identities of the actions (exploration versus exploitation) are uncoupled from the choice stimuli (options A and B),

ruling out the possibility that anticipatory arousal is tied to specific stimulus items, as in the Iowa gambling task, discussed above. Our modeling approach uniquely positions us to examine the level of sophistication in arousal signals that accompany choice under uncertainty and the choices themselves. To foreshadow, we find that SCRs in exploratory choice are in accord with the Ideal Actor, suggesting that SCRs index reflective and nuanced assessments of the expected benefit of the chosen action. That is, anticipatory arousal appears to register computations deeper than static associations between emotional arousal and particular actions or stimuli (as evidenced by Bechara et al., 1996), evincing a more sophisticated trial-by-trial calculation of the relative benefit of the two actions.

Having demonstrated that anticipatory SCRs and concomitant choice behavior manifest reflective, rather than reflexive, signatures, we then provide two additional confirmations of the computational framework's predictions. First, by manipulating the volatility level in the leapfrog paradigm, we reveal that anticipatory SCRs appear to be rationally modulated by the current level of environment volatility, in line with the Ideal Actor's predictions. And second, to highlight the computational expense of the reflective choice processes presumed to drive behavior and SCRs in Experiment 1, we demonstrate that concurrent cognitive demand reverts decision-makers to more reflexive choice behavior. In doing so, we corroborate—indirectly—the role of reflective calculations in the anticipatory SCRs seen in Experiment 1.

Experiment 1

Method

Forty-three undergraduates at the University of Texas completed 200 trials of the leapfrog task (Fig. 1a). In the choice task, the payoffs for the two options continually alternated in superiority over the course of the task. Payoffs for options A and B started at 10 and 20, respectively, and alternated in superiority by increasing (i.e., "jumping") by 20 points with probability $P(\text{jump}) = .075$ after each choice. In order to facilitate a full understanding of the task structure, participants were provided with the following instructions:

> Option A and B will both keep getting more valuable over the course of the experiment. Option A and B will take turns being the better option. The only way to know which option is currently better is by sampling the options. The better option will always give you 10 more points than the worse option. When the worse option becomes the better option, it will jump in value by 20 points.

To avoid complications associated with participants exhibiting diminishing sensitivity to payoff differences as the payoff magnitudes rise (Tversky & Kahneman, 1992), participants were paid 5 cents per "correct" choice, defined by whether they had chosen the option with the superior payoffs at the time of choice. At the outset of the experiment, participants were informed which option would give the higher initial payoff (20 points) and which option had the lower initial payoff (10 points).

Participants were first presented with a prechoice (anticipatory) period, during which they were instructed to "THINK ABOUT YOUR CHOICE," followed by a prompt to make their choice. Participants had 1.5 s to choose by keyboard. When a participant failed to respond, he or she was presented with a screen that read "TOO SLOW, TRY AGAIN," and the trial was repeated. After their choice, the outcome was displayed for 1 s. An intertrial interval (ITI) occurred after each trial, with a duration ranging from 2 to 6 s (Poisson, mean = 3 s).

SCR was measured via Ag-AgCl electrodes attached to the crease between the distal and middle phalanges of the first and second digits of the left hand and were recorded with a BIOPAC unit at 200 Hz. We employed a deconvolution technique, based on a physiological model of the general SCR shape, that allows for separation and quantification of the fast-varying (phasic) and slow-varying (tonic) components of the skin conductance signal (Benedek & Kaernbach, 2010). We calculated anticipatory sympathetic arousal by integrating (i.e., summing over time) the phasic driver signal during the 7.5-s anticipatory period starting at the onset of the "THINK ABOUT YOUR CHOICE" prompt and ending at feedback onset. Average optimized time constants τ_1 and τ_2 were 0.93 and 2.79. To avoid the influence of task novelty on SCRs, the first 10 trials were excluded from analysis. Finally, SCR magnitudes were log-transformed to remove skew and z-transformed within participants.

Results and discussion

Choice behavior

We first assessed sequential dependence in choice behavior (for which the reflective and reflexive models make divergent qualitative predictions), using what we termed the *hazard rate* of exploration. This hazard rate is calculated as the probability of making an exploratory choice as a function of the number of consecutive exploitative choices made prior to choice. The two candidate models of choice behavior make divergent predictions about the hazard rate of exploratory choice (Fig. 2a). The Ideal Actor prescribes that the hazard rate for exploration monotonically increases with the streak length of exploitative choices of the observed superior option, because the model's uncertainty about which option currently has the higher payoff increases in the absence of exploration. In contrast, the reflexive Naïve RL model acts only on direct observations of payoffs and does not perform an inference about the goodness of the observed inferior option. Since its sole source of exploration is trial-independent choice randomness, it prescribes a flat hazard rate of exploration.

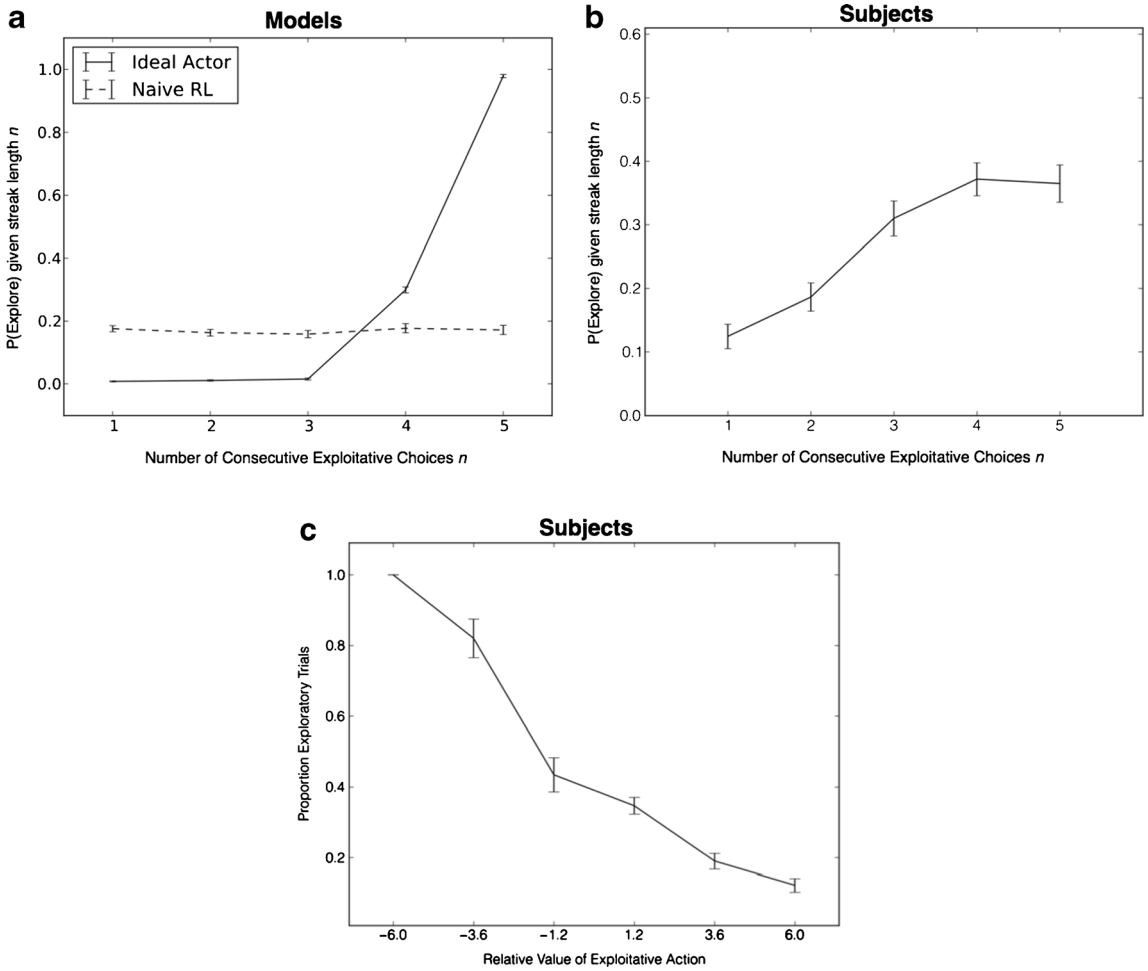

Fig. 2 Choice behavior in Experiment 1. **a** Predicted hazard rates of exploration, calculated as the probability of making an exploratory choice given an exploitative choice streak of length *n*, according to the belief-directed Ideal Actor (blue line) and the Naïve RL model reinforcement learning (red line) accounts. **b** Observed hazard rate for participants in Experiment 1, which roughly exhibits the monotonically increasing signature of the Ideal Actor. **c** Proportion of participants' exploratory choices as a function of the Ideal Actor calculated relative value of choosing the exploitative option, revealing a monotonically decreasing relationship. Error bars represent standard error of the mean

Qualitatively, participants' hazard rates (Fig. 2b) manifest contributions of both reflective and reflexive strategies. Consistent with the reflective model of choice, participants by and large made more exploratory choices as the number of stable exploitative choices increased, $F(4, 38) = 40.85$, $p < .0001$. Still, this group hazard rate appears less sloped than that prescribed by the reflective model, suggesting that behavior also bore the influence of a reflexive strategy. Quantitative assessment of model fits (via maximum likelihood) reveals that 88 % of participants were better described by the Ideal Actor model[1] (binomial test, $p < .0001$; see Table 1 for parameter values and goodness-of-fit measures). The psychometric curve plotted in Fig. 2c further illustrates the reflective character of participants' choices.

Psychophysiological results

We began examining autonomic responses by analyzing the phasic (fast-varying) component of participants' anticipatory SCRs as a function of simple trial type. We found that SCRs accompanying exploratory actions were significantly larger than those accompanying exploitative actions (Fig. 3a), $F(1,$

[1] We also examined the goodness of fit of a baseline model (Yechiam & Busemeyer, 2005), which assumes that choice probabilities for each option are constant and statistically independent across trials, to identify "nonlearners" who were not responsive to the options' changing payoffs. We found that no participants in this experiment were best fit by the baseline model (using BIC), as compared with the two other models. This analysis supports an interpretation that the model fits elucidate the type of learning strategy taken and not, say, whether a participant demonstrated learning versus nonlearning in response to choice outcomes.

Table 1 Summary of model fit in Experiment 1 [*P*(jump) = .075]

Model	% Best Fit	Total BIC	*P*(jump) (*SD*)	γ (*SD*)
Naïve	12	11,402.24	-	0.07 (0.05)
Ideal Actor	88	9,513.41	.03 (.03)	0.34 (0.13)

38) = 7.33, $p < .01$. One might expect intuitively that choices to the observed inferior option would evoke a larger emotional response than would choices to the observed superior option; indeed, this mirrors previous work revealing heightened arousal in anticipation of future monetary (Bechara et al., 1996) or cognitive (Botvinick & Rosen, 2009) costs, as well as a subjective experience of regret (Camille et al., 2004). This coarse analysis lends credibility to a reflexive account of arousal. But could these signals provide evidence for a reflective process, following participants' behavior, that makes inferences on the basis of unobserved payoff changes?

To perform such an assessment, we leveraged the Ideal Actor's trial-by-trial action prescriptions to understand how participants' beliefs about the currently optimal action might drive SCRs at choice. Specifically, the model affords classification of choices to the observed superior option (exploitation) as optimal versus suboptimal because, for example, there are situations where the model prescribes that the observed inferior option has become superior and exploratory action is actually optimal. Likewise, choices of the observed inferior option (exploration) are classified by the Ideal Actor as optimal versus suboptimal on the basis of how recently the observed inferior option has been explored. Accordingly, we reexamined the choice SCRs in Fig. 3a, classifying them as "explore optimal" or "exploit optimal" as defined by the model's optimal prescription at the time of choice. A reflective arousal signature critically entails that SCRs differentiate according to the agreement between the chosen action and the reflectively calculated benefit of that action—that is, an interaction, in contrast to the main effect of chosen action, as predicted by a reflexive account. Figure 3b reveals that, upon

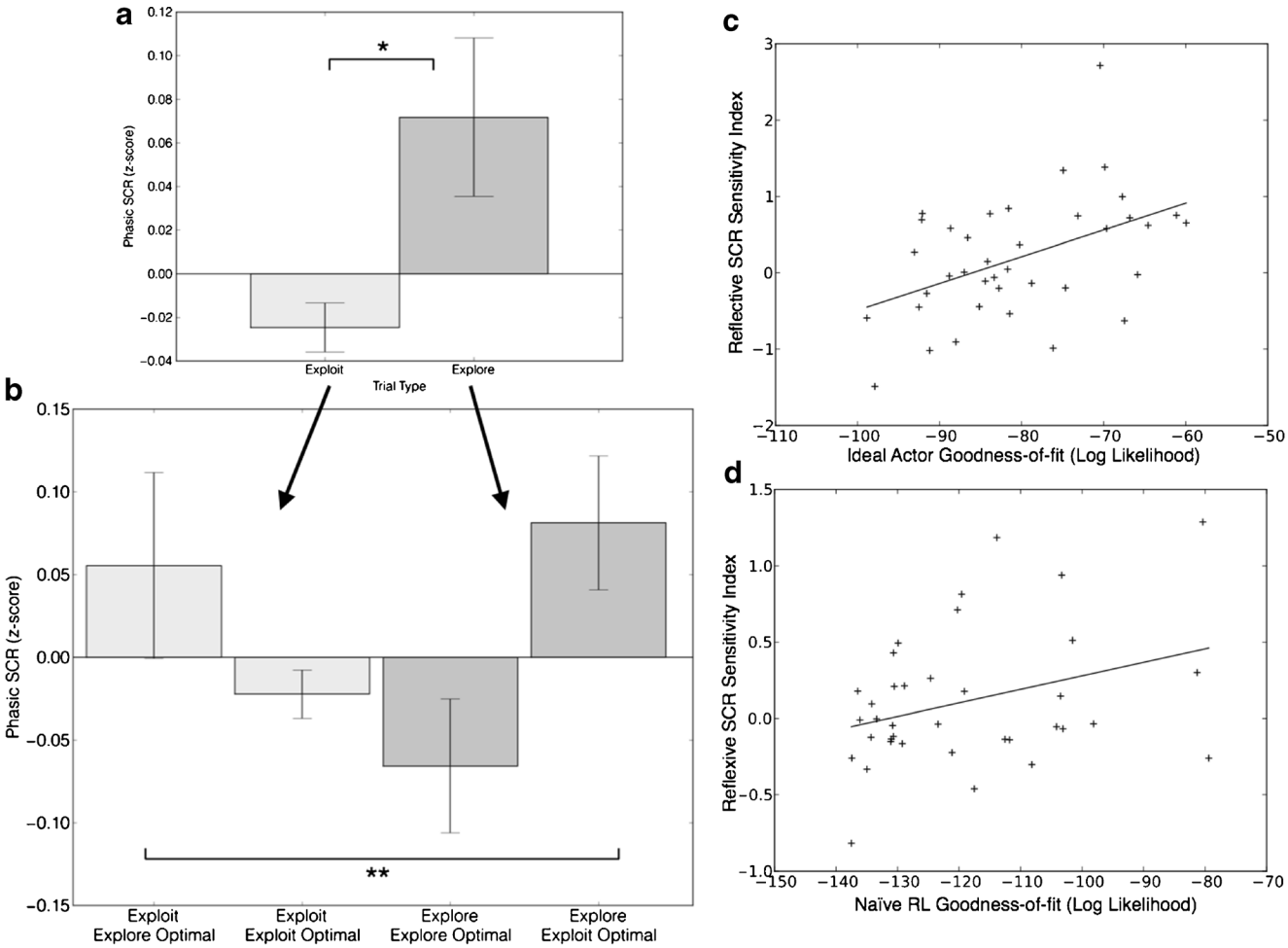

Fig. 3 Anticipatory skin conductance responses (SCRs) in Experiment 1. **a** Log-transformed phasic SCRs accompanying exploitative versus exploratory choices. **b** Exploratory and exploitative choices decomposed according to the Ideal Actor's optimal prescription at time of choice. The outer two bars depict SCRs accompanying choices where decision-makers acted against the prescription of the model. Conversely, the inner two bars depict SCRs in choices in which decision-makers acted in accordance with the model. Here, participants appear to differentiate physiologically between optimal and suboptimal choices as calculated by a reflective model of choice. **c** Behavioral agreement with the reflective Ideal Actor (quantified using model log-likelihood) predicts the Reflective SCR Sensitivity Index (the extent to which an individual physiologically differentiated between suboptimal and optimal actions). **d** Behavioral agreement with the reflexive, Naïve RL model reinforcement learning (RL) model predicts the Reflexive SCR Sensitivity Index (the extent to which an individual physiologically differentiated between exploratory and exploitative choices). All SCRs are reported as *z*-scores computed from the log-transformed integrated phasic SCR. Error bars represent standard error of the mean. $*p < .05$, $**p < .01$

closer inspection, anticipatory SCRs indicate greater arousal accompanying suboptimal choices (outer bars) and comparatively less arousal on optimal choices (inner bars). The interaction was statistically reliable, $F(1, 38) = 7.02$, $p < .01$.

We also directly compared the explanatory power of the ideal-actor-based factorial analysis (leveraging *both* chosen action and model prescriptions; Fig. 3b) with that of the more coarse analysis (examining only actions; Fig. 3a). Indeed, from a baseline linear model predicting SCR as a function of the action type (exploratory or exploitative), adding the Ideal Actor's prescription and its interaction with the chosen action resulted in a significant improvement in the amount of variance explained, $\chi(9) = 18.81$, $p < .05$. This more formal assessment suggests, compellingly, that anticipatory arousal signals some form of reflective assessment of the chosen action's benefit.

Notably, these findings hold at the level of the individual: Decision-makers who manifested more reflective SCRs behaved more consistently with the Ideal Actor. For each participant, we calculated a Reflective SCR Sensitivity Index, which quantifies the extent to which their SCRs differentiated between suboptimal versus optimal choices:

$$[SCR(\text{Exploit}|\text{Explore Optimal}) - SCR(\text{Explore}|\text{Exploit Optimal})] - [SCR(\text{Exploit}|\text{Exploit Optimal}) - SCR(\text{Explore}|\text{Explore Optimal})].$$

The Reflective SCR Sensitivity Index is akin to interaction size in Fig. 3b. We found that the more pronounced a participant's physiological responses to suboptimal choice was, the more reflective his or her choice behavior appeared (Fig. 3c), $r(41) = .46$, $p < .01$. Intuitively, because reflective exploration is the optimal strategy in this task, we found that reflective choice behavior (quantified by Ideal Actor log-likelihood) significantly predicted total obtained payoff, $r(41) = .54$, $p < .01$.

We also examined the converse: Did participants who behaved more reflexively (i.e., made choices more consistent with the Naïve RL model) exhibit more reflexive SCRs—akin to the coarse effect in Fig. 3a? We calculated a Reflexive SCR Sensitivity Index as *SCR*(Explore) − *SCR*(Exploit). Indeed, participants who exhibited more reflexive behavior displayed more reflexive SCRs (Fig. 3d), $r(41) = .33$, $p < .05$. Critically, reflective and reflexive behavioral indices did not correlate with each other, $r(41) = .12$, $p = .46$, and moreover, a permutation test supported the pairing of the two models with their respective SCR indices, $p < .025$. In other words, physiological differentiation dovetails with the choice strategy employed, suggesting against the possibility that choice-related arousal merely reflects a general form of task engagement. Full correlations between behavioral and physiological metrics are reported in Table 2.

Physiological arousal, a marker of the involvement of emotion in decision making, is well documented to accompany choices made under uncertainty, but past work has been unable to characterize the intelligent nature of these signals: Namely, does arousal stem from a "cognitive" assessment of value that utilizes the full environment structure, or does it merely signal a coarse, reflexive assessment of the possible consequences of choices? In this experiment, we found compelling evidence for a reflective source of anticipatory arousal accompanying choice as participants negotiate an exploration–exploitation task, mirroring the observed behavior. Our subsequent experiments bolster this account, by demonstrating how changing the task environment (a factor external to the decision-maker) affects arousal and choice behavior in a manner predicted by our reflective model (Experiment 2) and how manipulating available processing resources (a factor internal to the decision-maker) attenuates the reflective signature observed in the behavior here (Experiment 3).

Table 2 Correlations between skin conductance response (SCR) sensitivity indices and choice model goodness of fit

	Ideal Actor Log-Likelihood	Naïve RL Log-likelihood
Reflective SCR sensitivity	$r = .46$, $p = .003$	$r = .04$, $p = .79$
Reflexive SCR sensitivity	$r = -.13$, $p = .44$	$r = .33$, $p = .04$

Experiment 2

Having found suggestive evidence in Experiment 1 that decision-makers' anticipatory arousal patterns manifest a reflective assessment of the chosen action's value, we sought confirmation of two corollary predictions made by the Ideal Actor. By manipulating the task volatility, we demonstrate that participants' choices and arousal patterns change in a manner predicted by the reflective account. First, more volatile environments necessitate higher rates of exploratory choice (Humphries, Khamassi, & Gurney, 2012), because the Ideal Actor's beliefs about which action is optimal change more rapidly. Second, as environment volatility increases, the model's certainty about the currently higher-payoff action decreases, and in turn, the value differential between the exploratory and exploitative actions decreases.

Accordingly, we manipulated the environment volatility in a counterbalanced within-subjects fashion, such that *P*(jump) varied between .025 (low volatility) and .125 (high volatility), and obtained model predictions of exploratory choice rates for the two volatility rates (Fig. 4a; see the Appendix for model simulation details). We predicted, intuitively, that participants should explore more during high-volatility blocks. We reasoned further that the value indifference brought about by increased volatility (expressed as average *Q*-values differential; Fig. 4b) should attenuate the physiological differentiation

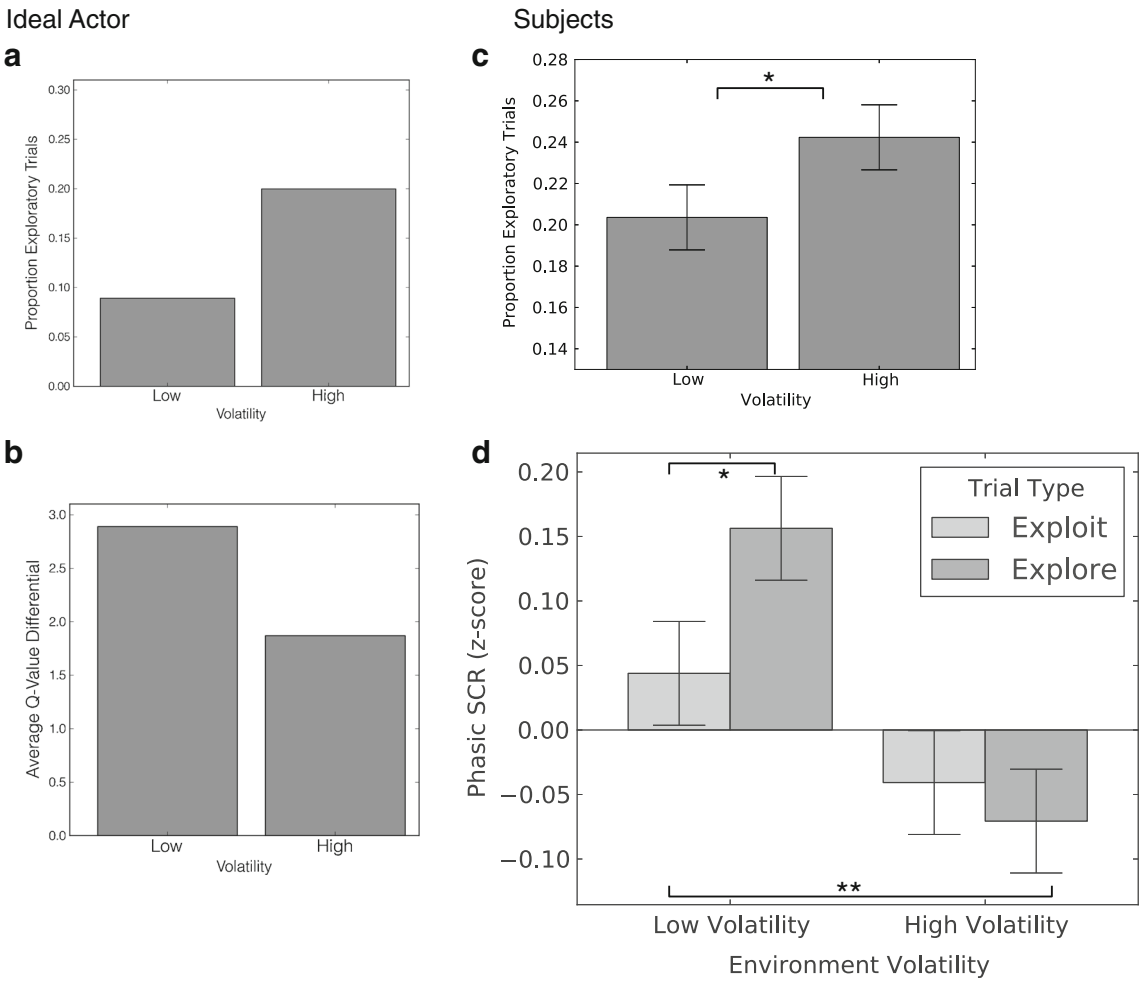

Fig. 4 Model predictions and behavioral and psychophysiological results in Experiment 2. Simulations reveal that the Ideal Actor prescribes **a** a greater rate of exploration in high- versus low-volatility environments and **b** greater indifference with respect to action-values expressed in terms of average *Q*-value differences over all trials and choice types. **c** Participants exhibit the same ordinal relationship between rate of exploratory choice and environment volatility. **d** The model's prediction of increased indifference in high-volatility environments is manifested in participants' skin conductance response differentiation between exploratory versus exploitative actions in low- versus high-volatility environments. Error bars represent standard errors of the means. *$p < .05$, **$p < .01$

between exploratory and exploitative choices in our participants (Fig. 3a). In other words, the arousal accompanying exploration (as compared with exploitation) should decrease in volatile environments, as compared with stable environments. Critically, the Naïve RL model, since its action-value estimates do not depend on environment volatility, makes no such prediction.

Method

Thirty-two undergraduates at the University of Texas made choices in the same 200-trial choice task as in Experiment 1, with one exception: The volatility rate began at .025 (low) for one 100-trial block and changed to .125 (high) for the other 100-trial block. Critically, the order of low- and high-volatility blocks was counterbalanced across participants. Participants were provided with the same instructions as in Experiment 1. SCR was measured and analyzed in the same manner as in Experiment 1. To avoid the influence of task novelty on SCRs and to allow for participants to adapt their behavior to the change in volatility, the first 10 trials of each block were excluded from analysis. The SCR deconvolution procedure (Benedek & Kaernbach, 2010) yielded average optimized fast (τ_1) and slow (τ_2) time constants of 0.79 and 2.68, respectively.

Results and discussion

Behavioral results

Figure 4c reveals that participants made more frequent exploratory choices in high-volatility blocks than in low-volatility blocks, $F(1, 30) = 6.64$, $p < .05$, in accordance with the qualitative predictions of the Ideal Actor model (Fig. 4a).

Moreover, the increasing hazard rate of exploration—characteristic of reflective choice (Fig. 2a)—was evident across both low [effect of streak length: $F(4, 27) = 5.68, p < .01$] and high, [$F(4, 27) = 9.28, p < .001$] volatility levels, mirroring the behavior seen in Experiment 1.

Psychophysiological results

Following Experiment 1, we examined anticipatory phasic SCRs for both exploratory and exploitative trials across high- and low-volatility blocks. These results, shown in Fig. 4d, suggest that environment volatility attenuated decision-makers' physiological response to exploratory choice. An ANOVA conducted on SCRs revealed a significant interaction between volatility (high/low) and trial type (explore/exploit), $F(1, 31) = 8.64$, $p < .01$. During low-volatility periods, SCRs significantly differentiate between exploratory and exploitative choices, $F(1, 31) = 4.84$, $p < .05$, but in high-volatility periods—where the Ideal Actor's value differential between the exploratory and exploitative actions is markedly smaller— this differentiation was absent, $F(1, 31) = 0.82$, $p = .36$.

Although the order of blocks was counterbalanced across participants, we also examined whether order effects (whether a participant experienced a high-volatility or a low-volatility block first) could play a role in the observed volatility effect. Adding block order as a factor to the above ANOVA, we found no significant three-way interaction, $F(1, 31) = 0.77$, $p = .38$, suggesting against the possibility that the attenuated physiological response in high-volatility blocks was the result of an order effect. Furthermore, tonic SCR—the slow-varying component of SCR—did not differ significantly across volatility blocks, $t(41) = 0.21$, $p = .83$, ruling out the possibility that heightened environment volatility increased the tonic SCR signal and reduced our capability to detect phasic SCR peaks.

Taken together, these results confirm that both choice-related autonomic arousal and exploratory choice, in accordance with the predictions of the reflective account of choice, are sensitive to environment volatility. Following our model predictions (Fig. 4a, b), participants increased their exploration rates during more volatile periods. Furthermore, decision-makers show less physiological differentiation between exploratory and exploitative choices during more volatile periods, as compared with less volatile periods, providing suggestive evidence, beyond that of Experiment 1, that the arousal accompanying choice is the result of a reflective computation of action-value.

Experiment 3

In a third experiment, we examine reflective choice more deeply, demonstrating through its cognitive costs that the patterns of choice and SCRs observed here manifest a reflective and sophisticated computation of value. Both here and in past work (Blanco et al., 2013; Knox et al., 2012), we have shown that under normal circumstances, people exhibit signatures of both reflective and reflexive strategies in their negotiation of the exploration–exploitation trade-off. Here, we reveal how, with concurrent cognitive demands (via working memory [WM] load), decision-makers revert to more reflexive behavior, highlighting the computational sophistication of the processes presumed to underpin behavior and anticipatory SCRs in Experiment 1.

In our framework, the reflective and reflexive modes of choice are differentiated, in part, by their computational expense: Reflective choice requires maintaining a belief about the environment's state and prospectively planning, while reflexive choice involves stochastic, unprincipled exploration. This is echoed in contemporary two-system theories of RL: A reflective account would be regarded as "model-based" because it utilizes a model of the environment structure to prospectively evaluate the values of actions, while the reflexive account would be considered "model-free" because it is informed only by directly experience payoffs and eschews the full environment structure (Daw, Niv, & Dayan, 2005). Indeed, the two hypothesized choice systems characteristically impose different computational costs (Daw et al., 2005; Keramati, Dezfouli, & Piray, 2011).

Here, we sought to disentangle the two sources of exploration with the intuition that reflective exploratory choice imposes greater requirements on decision-makers' cognitive resources than does reflexive exploratory choice. We placed participants under WM load during the leapfrog task to examine whether, with concurrent cognitive demands, decision-makers would revert to a reflexive exploration strategy. In past work, WM load manipulations have been shown to foster reliance on implicit classification strategies (Foerde, Knowlton, & Poldrack, 2006; Zeithamova & Maddox, 2006) and cognitively inexpensive model-free choice strategies during sequential decision making (Gershman, Markman, & Otto, 2014; Otto, Gershman, Markman, & Daw, 2013). Such a demonstration here would further affirm that the choice behavior seen in these experiments arises from a computationally sophisticated action-selection process like that of the reflective Ideal Actor.

Method

Sixty-eight undergraduates at the University of Texas were randomly assigned to two groups: the single-task condition and the dual-task condition. Both groups completed 300 trials of the leapfrog choice task using the same volatility rate as in Experiment 1 [$P(\text{jump}) = .075$]. The dual-task condition followed the general tone-counting procedure of Foerde et al. (2006), which we modified to ensure that the concurrent task persisted over all stages of the decision task (Otto, Taylor, & Markman, 2011). We used a deadline procedure to ensure

that, between conditions, a fixed amount of time elapsed on each trial. On each trial, participants saw the word "CHOOSE" and had 1.5 s to make a response, after which the resultant payoff was displayed for 1 s, followed by a variable length ITI (2–6 s). To ensure that WM load did not interfere with learning of the volatility rate, participants completed a passive viewing of a random instantiation of the task for 500 trials before the choice task.

In the dual-task condition, two types of tones—high pitched (1000 Hz) and low pitched (500 Hz)—were played during each trial. The choice period of each trial was divided into 10 intervals of 250 ms, with tones occurring in intervals 3–7 (500–1,750 ms after trial onset). The number of tones presented during each trial varied uniformly between one and four. The base rate of high tones was determined every 50 trials and was sampled from a uniform distribution between 0.3 and 0.7. Participants were instructed to maintain a running count of the number of high tones, while ignoring the low-pitched tones. At the end of each 50-trial block, participants reported their counts and were instructed to restart their count at zero.

Results and discussion

To ensure that dual-task participants did not trade off performance on the concurrent task in order to complete the choice task, we excluded the data of 7 dual-task participants who exhibited a root mean squared error of 40 or greater on the concurrent task. Sixty-one participants (31 single task, 30 dual task) remain in the analyses that follow.

Exploratory choice rates

Critically, we found no significant effect of WM load upon overall rate of exploratory choice [condition: $F(1) = 0.22$, $p = .64$], ruling out the possibilities that WM load impeded exploration altogether or rendered participants insensitive to the options' payoffs (Fig. 5a). Furthermore, these exploration rates reached a steady state early on across both groups [25-trial block × condition: $F(1, 11) = 0.74$, $p = .48$], since pretraining both groups of participants on the environmental volatility rate presumably stabilized exploration rates early on.

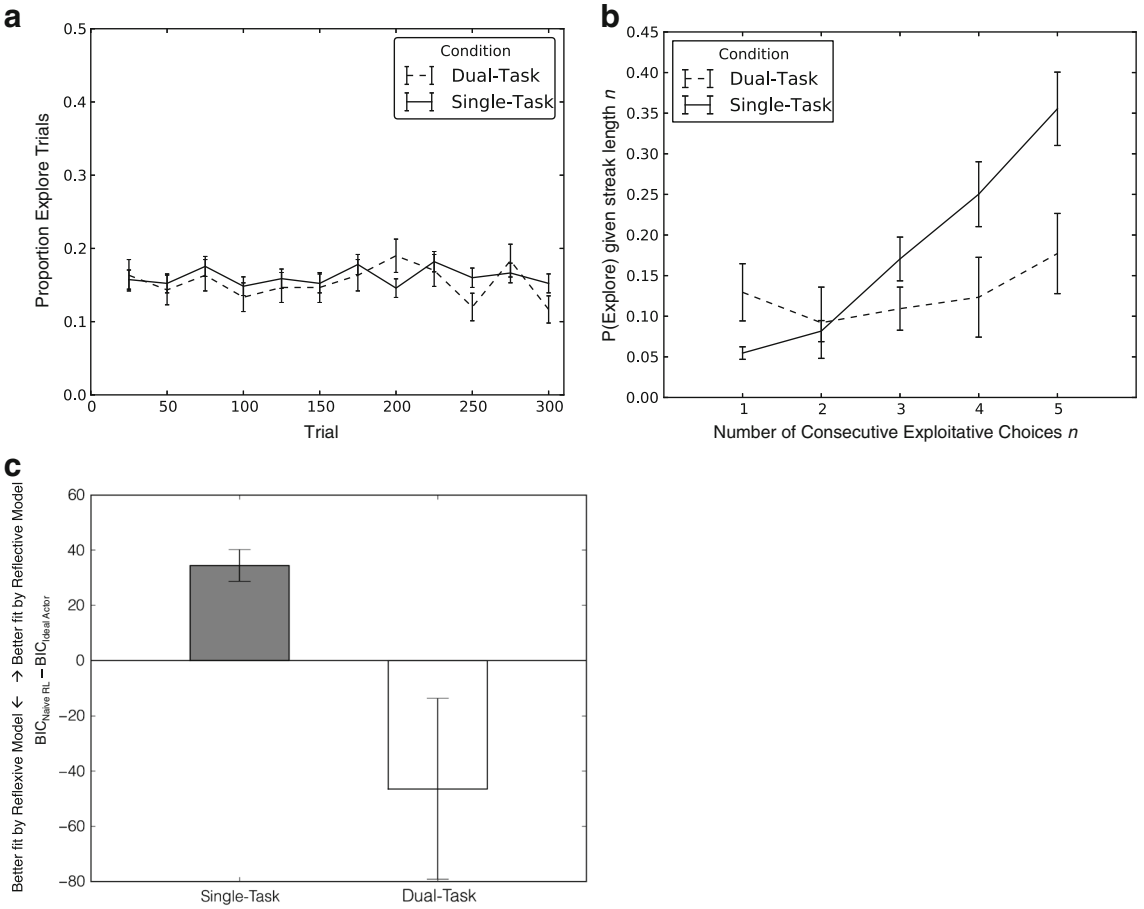

Fig. 5 **a** Proportion of exploratory choice across single-task (solid line) and dual-task (dashed line) participants, as a function of 25-trial block. **b** Hazard rate of exploratory choice averaged across single-task and dual-task participants. **c** Average relative goodness of fit between the Ideal Actor and the Naïve reinforcement learning (RL) model (calculated as a difference in BIC scores) between single-task and dual-task participants. Positive values indicate that a decision-maker's choices were better described by the Ideal Actor, and negative values indicate that a decision-maker's choices were better described by the Naïve RL model. Error bars represent standard errors of the means

Hazard rates of exploration

Qualitatively, we predicted that the hazard rates (calculated the same way as in the previous experiments) of dual-task participants would be less sloped than those of single-task participants, since their reliance on a strategy resembling the Naïve RL model should yield a flatter hazard rate. Figure 5b reveals that these predictions were borne out in the hazard rates of exploration between conditions, since dual-task participants exhibited markedly less sloped hazard rates. Critically a two-way ANOVA revealed a significant interaction between group (single task vs. dual task) and exploitative choice streak length n, $F(1, 4) = 5.49$, $p < .05$, indicating that the slopes of the hazard rates were different across single-task and dual-task groups. Additionally, there was a significant main effect of streak length n, $F(4) = 32.31$, $p < .0001$. Corroborating the previous overall exploration rate analysis, there was no significant main effect of condition (single task vs. dual task), $F = 0.19$, $p = .66$.

Individual models

To quantitatively illustrate single-task and dual-task groups' differential reliance upon exploration strategies, we performed a model comparison, examining the relative goodness of fit of the Ideal Actor model and the Naïve RL model across the two conditions. We fit both models to participants' choices, using the procedure detailed in Experiment 1, and calculated a relative BIC score, $BIC_{Naive} - BIC_{Actor}$, which quantifies how much better a participant's behavior is described by the Ideal Actor model than by the Naïve RL model. Intuitively, positive scores indicate a better fit by the Ideal Actor model, while negative scores indicate a better fit by the Naïve RL model. Figure 5c depicts relative BIC scores of single-task and dual-task participants, revealing that single-task participants were significantly better characterized by the Ideal Actor model than by the Naïve RL model, whereas dual-task participants were significantly better characterized by the Naïve RL model (paired samples $t = 2.79$, $p < .01$). Model goodness-of-fit scores and parameter estimates, by condition, are reported in Table 3.

Using this metric, we also examined the possibility that dual-task participants could be trading off performance in the concurrent task in order to make choices in the leapfrog task. If this were true, we might find that larger relative BIC scores (indicating better description of behavior by the Ideal Actor) accompany more erroneous tone-counting performance. We found no significant relationship between the two quantities, $r = .22$, $p = .48$, suggesting against the possibility of an internal trade-off occurring.

Hybrid model

We also considered a hybrid model that assumes that choices at each trial are the result of a weighted combination of the Naïve RL and the Ideal Actor models. This approach follows past computational accounts indicating that individuals exhibit a mixture of model-based and model-free contributions (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Gläscher, Daw, Dayan, & O'Doherty, 2010; Otto, Raio, Chiang, Phelps, & Daw, 2013). We hypothesized that, in the present experiment, the weighting of the Ideal Actor's contribution to choice should decrease in the dual-task condition, in favor of an increased weighting of the Naïve RL model. Accordingly, we fit a constrained hybrid model whose weighting parameter quantifies the contribution of the reflective Ideal Actor relative to the reflexive Naïve RL model (see the Appendix for model details). We found that the best-fitting weighting parameter was larger for the single-task participants ($M = 0.62$, $SD = 0.12$) than for the dual-task participants ($M = 0.49$, $SD = 0.21$) and that this difference was significant, $t = 2.80$, $p < .01$, indicating that concurrent cognitive demand decreased participants' behavioral expression of reflective choice strategies.

Both of these modeling approaches corroborate our empirically assessed hazard rates of exploration, elucidating how concurrent cognitive demands attenuated the influence of a reflective and belief-based exploratory choice strategy and fostered increased reliance on a Naïve, stochastic strategy. Moreover, revealing the cognitively demanding nature of reflective choice provides an indirect hint about the computational sophistication underlying SCRs observed in Experiment 1.

General discussion

Using a novel task and modeling approach, we demonstrated that the arousal signals accompanying choice, widely interpreted as evidence for interplay between emotion and

Table 3 Summary of model fits across single-task and dual-task conditions in Experiment 3 [P(jump) = .075]

Condition	Model	% Best Fit	Total BIC	P(Jump) (SD)	γ (SD)
Single task	Naïve RL	10	8,815.46	.03 (.03)	0.14 (0.02)
	Ideal Actor	90	7,747.28	–	0.5 (0.18)
Dual task	Naïve RL	58	8,487.27	–	0.14 (0.06)
	Ideal Actor	42	8,256.24	.02 (.02)	0.38 (0.18)

decision making (Critchley, 2005), bear a reflective and planning-oriented signature. The apparent sophistication of these signals, particularly pronounced in participants whose choices were most reflective, challenges accounts purporting that there is a sharp divergence between emotional responses and cognitive evaluations of action outcomes (Loewenstein et al., 2001; Rolls, 1999). That the level of cognitive sophistication manifested in behavior predicts the character of physiological differentiation underscores the close relationship between reflective choice behavior and these concomitant sophisticated physiological signals. Furthermore, we demonstrated that this sort of reflective behavior is computationally expensive, suggesting that—with the assumption that similar processes drive behavior and anticipatory arousal—SCRs can also originate, at least in part, from a cognitively sophisticated evaluation of an action's benefit.

The pronounced arousal accompanying suboptimal choices (Fig. 3b) raises an interesting question: Why do people make actions that they appear to "know," at some level (insofar as registering autonomic responses), are disadvantageous? The observed hazard rates of exploratory choice (Fig. 2b) may provide a hint: Individuals exhibit a mixture of reflexive and reflective exploration strategies, suggesting that they occasionally make reflexively-guided (i.e., purely stochastic) exploratory actions that violate the prescriptions made by a reflective choice strategy. Accordingly, the SCRs seen here may index a form of discord between the chosen action—however it came about—and a reflective assessment of the action's advantageousness. Indeed, the observation that these putative "errors" register physiologically mirrors the SCRs observed to accompany response errors in more basic control tasks (Critchley, Tang, Glaser, Butterworth, & Dolan, 2005; Hajcak, McDonald, & Simons, 2003).

We note that in Experiment 1, it is conceivable that SCRs could reflect uncertainty at the time of choice, such that larger SCRs are simply associated with greater uncertainty[2] (Critchley, Mathias, & Dolan, 2001). However, a regression examining trial-by-trial SCRs as a function of belief state uncertainty (i.e., Shannon entropy) revealed no significant or positive effect of uncertainty. The factorial analysis depicted in Fig. 3b may yield insight about why uncertainty does not play a critical role in SCRs here. Because uncertainty-driven exploration distinguishes the Ideal Actor, we can deduce that the model is most uncertain in situations where it prescribes exploration ("Exploit–Explore Optimal" and "Explore–Explore Optimal" in Fig. 3b). However, SCRs are lower on those trials. Furthermore, the far right cell ("Explore–Exploit Optimal") corresponds to choices where uncertainty would be low (i.e., the model is most certain that the highest-observed option is still the highest observed), but SCRs are elevated in this situation. The impression obtained, then, is that uncertainty alone appears unable to explain the pattern of SCRs; the key element driving SCR here is the chosen action itself and one's belief about the currently optimal action.

In Experiment 2, we found that heightened environment volatility both evoked higher rates of exploration and modulated the physiological response to exploratory choice. This finding raises a question about the source of the behavioral and autonomic regulation. Although both changes are straightforward predictions of the Ideal Actor model—that is, more rapidly evolving beliefs yield increased action-value indifference—an emotion regulation process (Martin & Delgado, 2011) could explain the effect of volatility on choice and arousal patterns accompanying these choices. Previous work (Sokol-Hessner et al., 2009) suggested that explicitly coaching decision-makers to intentionally reinterpret their actions decreased loss aversion in risky choice and decreased arousal in response to losses. Here, a marked increase in volatility (which warrants more frequent exploration) could have brought about a regulation strategy, decreasing decision-makers' aversion to the potential regret associated with exploratory choice in that it entails the risk of obtaining an inferior payoff. Such a form of cognitive regulation, then, could conceivably underpin the observed choice strategy change and modulation of physiological responses observed here.

It is worth noting that modern theoretical treatments of emotion posit two orthogonal dimensions to emotional experience: arousal and valence (Lang, Greenwald, Bradley, & Hamm, 1993). However, SCR and, more broadly, any index of sympathetic nervous system activity captures only the arousal component of emotional response. Thus, the present work does not address the valence dimension—the pleasant/unpleasant dimension of affective experience— associated with choice-related emotional responding. While it is conceivable that the arousal signals accompanying suboptimal choices here indicate negative emotions (e.g., anxiety surrounding undesirable outcomes), it is also possible that positive emotions may also be involved in decision making under uncertainty (Schonberg, Fox, & Poldrack, 2011), since past work suggests that some individuals actively seek out stimulation associated with risky actions (Figner et al., 2009).

The nuanced source of observed SCRs is (at least conceptually) predicted by lesion and functional imaging work. Notably, vmPFC lesion patients fail to register anticipatory SCRs when making disadvantageous choices (Bechara et al., 1999; Bechara et al., 1996), and at the same time, their actions appear largely insensitive to negative consequences of these actions (Fellows & Farah, 2005). Not surprisingly, functional neuroimaging work implicates these same prefrontal structures in the generation of SCRs (Critchley, Elliott, Mathias, & Dolan, 2000; Mitchell, 2011; Nagai, Critchley, Featherstone, Trimble, & Dolan, 2004) and, moreover, highlights a role for the vmPFC in flexible and intricate calculations of action-value (Fellows, 2007; Hampton, Bossaerts, &

[2] We thank two anonymous reviewers for pointing out this possibility.

O'Doherty, 2006; Hare, Camerer, & Rangel, 2009; Nicolle et al., 2012; Rushworth, Noonan, Boorman, Walton, & Behrens, 2011). Taken together, this work supports our contention that anticipatory SCRs could indeed arise from an insightful, reflective valuation process.

Here, we have demonstrated how physiological arousal accompanying choice—a marker of the involvement of emotion in decision making—exhibits sophisticated and reflective contributions, enriching previous characterizations of these signals. Along these lines, pupil diameter changes (a putative index of locus coeruleus–noradrenergic activity) have been shown to reflect assessments of environment state (Nassar et al., 2012), uncertainty (Preuschoff, Hart, & Einhäuser, 2011), and cognitive control state (Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010). Correspondingly, recent imaging work examining decision making emphasizes model-based (i.e., planning-oriented) contributions to value signals in the brain (Daw et al., 2011; Wunderlich, Dayan, & Dolan, 2012). The pattern of arousal signals observed here evinces a similar form of value computation, suggesting that the emotions accompanying choices, much like the observed behavior, exhibit hallmarks of a forward-planning and reflective process.

Although we have provided a preliminary demonstration of how computational modeling can elucidate the level of cognitive sophistication present in autonomic arousal, future work is needed in order to understand more precisely, beyond the reflective/reflexive distinction, the information sources that play into these computations and, moreover, the nature of the computations themselves. Experiment 1, in particular, reveals that individuals who choose in a more reflective manner (which presumably requires a rich representation of the decision environment) manifest a more sophisticated pattern of arousal than do individuals who choose more reflexively (which presumably requires only a crude understanding of the task environment and a simpler choice strategy). That is, arousal responses accompanying choices appear to mimic, in part, the cognitive process guiding choice. Still, these physiological and behavioral signatures could be the result of qualitatively different understandings of the task structure—that is, a decision-maker who appears reflective because he or she conceptualizes the task as less structured than it actually is—rather than as reflecting differential engagement of a more "intelligent" system. Such a level of characterization may afford critical insight into the causal nature of the interaction between behavior and arousal in decision making under uncertainty, a question historically beset with methodological and theoretical issues (Dunn et al., 2006).

Acknowledgements The experiments reported here were part of A.R.O.'s doctoral dissertation at the University of Texas at Austin. During this period A.R.O. was supported by a Mike Hogg Endowment Fellowship from the University of Texas at Austin. The authors thank Todd Gureckis, Russ Poldrack, Alex Huk, Nathaniel Daw, Tom Schönberg, Yael Niv and Tyler Davis for helpful conversations.

Appendix

The appendix describes the two models used and the model-fitting and comparison procedures we employed.

Naïve reinforcement learning model

The Naïve RL model is a single-parameter reflexive model that assumes that action-values are updated reactively to directly experienced rewards. The model reflexively maintains beliefs about payoffs based only on what it has seen. In other words, it believes that the point payoffs for each action are those most recently observed. Accordingly, the Naïve RL model assumes that Action H (that with highest observed payoff; the "exploit" action) and $\neg H$ (the alternative action with inferior observed payoffs; the "explore" action) give rewards of 1 and 0, respectively, corresponding to the higher and lower payoffs, respectively. Its expectation of each action's reward, $Q(H)$, is input into a softmax action selector (Sutton & Barto, 1998), giving it a constant probability of exploring or exploiting:

$$P(H) = \exp(\gamma \cdot Q(H)) / [\exp(\gamma \cdot Q(H)) + \exp(\gamma \cdot Q(\neg H))],$$

where γ is an inverse temperature parameter. As γ increases, the probability that H (the highest-observed action) will be chosen increases. This model is algorithmically equivalent to the softmax models used by Otto, Markman, Gureckis, and Love (2010) and Worthy, Maddox, and Markman (2007).

Ideal Actor model

The Ideal Actor is a two-parameter reflective model that maintains an optimal belief about the probability that each action has higher immediate payoffs. These beliefs are then used to compute optimal Q-values. A full specification of this belief update and Q-value computation is provided below. The Ideal Actor prescribes the exploitative action when the Q-value for the exploitative action is greater than zero and, conversely, prescribes the exploratory action when the Q-value for the exploitative action is less than zero. This model has two free parameters: P(jump), the environment volatility it uses to perform belief updates, and γ, the inverse temperature parameter used in the softmax rule.

In the main text, we described that the current highest-payoff option switches over time as the payoffs "jump" and that these jumps are not necessarily observed. The Ideal Actor model maintains a probabilistic distribution over possible underlying payoff states, represented as a belief B, which is the probability that the exploitative action—that is, the action

with the currently highest observed payoffs—yields the larger immediate payoff.

In this formulation, the underlying payoff states can also be thought of as the number of unobserved (i.e., true) jumps at a given time point. Accordingly, if there are zero or two unobserved jumps at the time of choice, the exploitative action yields the higher immediate payoff. Conversely, if there is one unobserved jump at the time of choice, then the exploratory option (i.e., not the option with the highest observed payoff) yields the higher immediate payoff. In the model, beliefs are optimally updated after each choice and observation of resultant payoff. The update to calculate the belief B_{t+1}—the probability distribution over the number of unobserved jumps (zero, one, or two) before taking the action at trial t+1—depends on the previous belief state B_t, the action taken at trial t (exploratory or exploitative), the observed number of payoff jumps o seen as a result of that action at t (which can take on the values of zero and one jumps in the case of exploratory choices and zero and two jumps in the case of exploitative choices), and the assumed volatility rate of the environment [P(jump), a free parameter].

The state transition matrix on exploitative choices is defined by

Unobserved jumps at time t+1

Unobserved jumps at time t	0	1	2
0	$\begin{cases} B_t \times (1 - P(jump)) & \text{if o=1 or o=2} \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} B_t \times P(jump) & \text{if o=0} \\ 0 & \text{otherwise} \end{cases}$	0
1	0	$\begin{cases} (1 - B_t) \times (1 - P(jump)) & \text{if o=0} \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} (1 - B_t) \times P(jump) & \text{if o=2} \\ 0 & \text{otherwise} \end{cases}$

The state transition matrix on exploratory choices is defined by

Unobserved jumps at time t+1

Unobserved jumps at time t	0	1	2
0	$\begin{cases} B_t \times (1 - P(jump)) & \text{if o=1} \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} B_t \times P(jump) & \text{if o=1} \\ 0 & \text{otherwise} \end{cases}$	0
1	0	$\begin{cases} (1 - B_t) \times (1 - P(jump)) & \text{if o=1} \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} (1 - B_t) \times P(jump) & \text{if o=1} \\ 0 & \text{otherwise} \end{cases}$

These individual state transition probabilities are then combined and normalized to form a posterior belief:

$$B_{t+1} = \frac{P(s_{0,t+1}, s_{0,t}) + P(s_{2,t+1}, s_{1,t})}{P(s_{0,t+1}, s_{0,t}) + P(s_{2,t+1}, s_{1,t}) + P(s_{1,t+1}, s_{1,t}) + P(s_{1,t+1}, s_{0,t})}.$$

Note that, above, the state $s_{i,t+1}$ refers to the number of unobserved jumps after the choice and payoff observation were made, while $s_{i,t+1}$ refers to the number of unobserved jumps before the choice.

If the choice was exploratory and a jump is observed, it is intuitive that the subsequent identities of the exploratory and exploitative actions should swap. The change of reference point necessitated by this situation is accomplished by inverting the belief:

$$B_{t+1} = \begin{cases} (1-B_{t+1}) \, if \, a_t = \text{explore and} \, o = 1 \\ B_{t+1} \, \text{otherwise} \end{cases}.$$

Using these optimally maintained beliefs, the Ideal Actor employs another step that optimally converts beliefs into action-values ("*Q*-values"). To do this, we make use of methods for solving partially observable Markov decision processes (POMDPs: Kaebling, Littman, & Cassandra, 1998), whereby each option's action-value expresses the statistical expectation of the sum of future reward given that the option is chosen and subsequent actions are chosen optimally. Importantly, in this leapfrog task, an action's relative value—that is, the difference in its value and the other actions'—is dependent only on the probability of the action yielding an immediate payoff and the informational value of the action's observation, which affects the actor's ability to accurately assess the value of future actions. To calculate these optimal action-values, we employed Cassandra, Littman, and Zhang's (1997) incremental pruning algorithm, an exact inference method that calculates *Q*-values for each possible belief state at each time horizon (i.e., number of choices remaining). These routines are implemented in Cassandra et al.'s (1997) POMDP-Solve library.

The true Ideal Actor deterministically (i.e., greedily) chooses $\text{argmax}_a \, Q_t(a, t)$. However, for the purpose of fitting this model to participants' choices, we utilized a softmax choice rule (identical to that used by the Naïve RL model) to generate response probabilities from these *Q*-values:

$$P(a_i, t) = \frac{\exp[\gamma \cdot Q(a_i, t)]}{\sum_{j=1}^{2} \exp[\gamma \cdot Q(a_j, t)]}$$

where γ is an inverse temperature parameter governing the choice rule's sensitivity to value differences. As in the Naïve RL model, as γ approaches infinity, the Ideal Actor becomes deterministic (greedy) in its choices; as γ approaches zero, the Ideal Actor moves toward choosing actions with uniform randomness.

Hybrid model

The hybrid model in Experiment 3 assumes that the final response probabilities are governed by a weighted combination of the response probabilities of two models: (1) a deterministic (i.e., $\gamma = \infty$) version of the Ideal Actor with *P*(jump) set to the optimal, ground-truth volatility rate of .075 and (2) a Naïve RL model with the inverse temperature parameter γ set to the optimal value of 0.143, as determined by simulations of a pure Naïve RL model. Maximally constraining the parameters of the constituent models in a theory-neutral manner reduces undesirable trade-offs in parameter values estimation, which reduce identifiabilility and interpretability of parameters. In the hybrid model, the weighted mixture is governed by

$$P(Explore) = w \cdot P(\text{Actor Explore}) + (1-w) \cdot P(\text{Naive RL Explore}),$$

where *w* is a mixture parameter indicating the influence of the Ideal Actor, relative to the Naïve RL model, in determining final response probabilities.

Model fitting and comparison

Our model-fitting procedure sought parameter values that maximized the log-likelihood of participants' choices given their previous rewards and choices. We conducted an exhaustive grid search to optimize parameter values for each participant. To compare goodness of fit across different models, we utilized the Bayesian information criterion (Schwarz, 1978), defined as $\text{BIC} = -2 \times LL + k \times \log(n)$, where *k* is the number of free parameters in the model, *LL* is the log likelihood of the model given the participants' data, and *n* is the number of choices fit. Lower BIC values indicate better fit. Summary statistics for best-fitting parameter values and goodness-of-fit measures are provided in Table 2.

Model simulation details

To generate the model-predicted hazard rates of exploration in Experiment 1 (Fig. 2a), each of the two models was yoked to a participant's particular instantiation of the leapfrog payoff structure and, consequently, the environment volatility rate. To simulate model choice behavior, we used the average of participants' best-fitting parameter values for each model and averaged across each simulated participant to calculate hazard rates. To derive predicted exploration rates and *Q*-value differentials in Experiment 2 (Fig. 4a, b), we simulated the Ideal

Actor using *P*(jump) parameter values of .025 for the low-volatility condition and .125 for the high-volatility condition. The *Q*-value differentials across 100 such simulations for each condition were then averaged.

References

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron, 73*(3), 595–607.

Bechara, A., Damasio, H., Damasio, A. R., & Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience, 19*(13), 5473–5481.

Bechara, A., Tranel, D., Damasio, H., & Damasio, A. R. (1996). Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. *Cerebral Cortex, 6*(2), 215–225.

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods, 190*(1), 80–91.

Blanco, N. J., Otto, A. R., Maddox, W. T., Beevers, C. G., & Love, B. C. (2013). The influence of depression symptoms on exploratory decision-making. *Cognition, 129*(3), 563–568.

Botvinick, M. M., & Rosen, Z. B. (2009). Anticipation of cognitive demand during decision-making. *Psychological Research Psychologische Forschung, 73*(6), 835–842.

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment, 14*(3), 253–262.

Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J.-R., & Sirigu, A. (2004). The involvement of the orbitofrontal cortex in the experience of regret. *Science, 304*(5674), 1167–1170.

Cassandra, A., Littman, M. L., & Zhang, N. L. (1997). Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 54–61.

Cohen, J. D., McClure, S. M., & Yu, A. 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*(1481), 933–942.

Critchley, H. D. (2005). Neural mechanisms of autonomic, affective, and cognitive integration. *The Journal of Comparative Neurology, 493*(1), 154–166.

Critchley, H. D., Elliott, R., Mathias, C. J., & Dolan, R. J. (2000). Neural activity relating to generation and representation of galvanic skin conductance responses: A functional magnetic resonance imaging study. *Journal of Neuroscience, 20*(8), 3033–3040.

Critchley, H. D., Mathias, C. J., & Dolan, R. J. (2001). Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron, 29*(2), 537–545.

Critchley, H. D., Tang, J., Glaser, D., Butterworth, B., & Dolan, R. J. (2005). Anterior cingulate activity during error and autonomic response. *NeuroImage, 27*(4), 885–895.

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876–879.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science, 298*(5596), 1191–1194.

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience & Biobehavioral Reviews, 30*(2), 239–271.

Fellows, L. K. (2007). Advances in understanding ventromedial prefrontal function the accountant joins the executive. *Neurology, 68*(13), 991–995.

Fellows, L. K., & Farah, M. J. (2005). Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans. *Cerebral Cortex, 15*(1), 58–63.

Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: Age differences in risk taking in the Columbia Card Task. *Journal of Experimental Psychology Learning, Memory, and Cognition, 35*(3), 709–730.

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences, 103*(31), 11778–11783.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General, 143*(1), 182–194.

Gilzenrat, M., Nieuwenhuis, S., Jepma, M., & Cohen, J. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 252–269.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*(4), 585–595.

Hajcak, G., McDonald, N., & Simons, R. F. (2003). To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behavior. *Psychophysiology, 40*(6), 895–903.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience, 26*(32), 8360–8367.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science, 324*(5927), 646–648.

Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Decision Neuroscience, 6,* 9.

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration–exploitation trade-off: Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience, 23*(7), 1587–1596.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence, 101*(1–2), 99–134.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology, 7*(5), e1002055.

Knox, W. B., Otto, A. R., Stone, P. H., & Love, B. C. (2012). The nature of belief-directed exploratory choice by human decision-makers. *Frontiers in Psychology, 2,* 398.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology, 30*(3), 261–273.

Ledoux, J. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York: Simon and Schuster.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127*(2), 267–286.

Lovibond, P. F. (2003). Causal beliefs and conditioned responses: Retrospective revaluation induced by experience and by instruction.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(1), 97–106.

Martin, L. N., & Delgado, M. R. (2011). The influence of emotion regulation on decision-making under risk. *Journal of Cognitive Neuroscience, 23*(9), 2569–2581.

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science, 8*(6), 423–429.

Mitchell, D. G. V. (2011). The nexus between decision making and emotion regulation: A review of convergent neurocognitive substrates. *Behavioural Brain Research, 217*(1), 215–231.

Nagai, Y., Critchley, H., Featherstone, E., Trimble, M., & Dolan, R. (2004). Activity in ventromedial prefrontal cortex covaries with sympathetic skin conductance level: A physiological account of a "default mode" of brain function. *NeuroImage, 22*(1), 243–251.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience, 15*(7), 1040–1046.

Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. J. (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron, 75*(6), 1114–1121.

Öhman, A., & Soares, J. J. F. (1994). "Unconscious anxiety": Phobic responses to masked stimuli. *Journal of Abnormal Psychology, 103*(2), 231–240.

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after pavlovian, observational, and instructed fear. *Psychological Science, 15*(12), 822–828.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013a). The curse of planning dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science, 24*(5), 751–761.

Otto, A. R., Markman, A. B., Gureckis, T. M., & Love, B. C. (2010). Regulatory fit and systematic exploration in a dynamic decision-making environment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 797–804.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013b). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences, 110*(52), 20941–20946.

Otto, A. R., Taylor, E. G., & Markman, A. B. (2011). There are at least two kinds of probability matching: Evidence from a secondary task. *Cognition, 118*(2), 274–279.

Preuschoff, K., Hart, B., & Einhäuser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Decision Neuroscience, 5,* 115.

Rolls, E. T. (1999). *The brain and emotion*. Oxford: Oxford University Press.

Rushworth, M. F. S., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron, 70*(6), 1054–1069.

Schonberg, T., Fox, C. R., & Poldrack, R. A. (2011). Mind the gap: Bridging economic and naturalistic risk-taking with cognitive neuroscience. *Trends in Cognitive Sciences, 15*(1), 11–19.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F., & Phelps, E. A. (2009). Thinking like a trader selectively reduces individuals' loss aversion. *Proceedings of the National Academy of Sciences, 106*(13), 5035–5040.

Studer, B., & Clark, L. (2011). Place your bets: Psychophysiological correlates of decision-making under risk. *Cognitive, Affective, & Behavioral Neuroscience, 11*(2), 144–158.

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

Suzuki, A., Hirota, A., Takasawa, N., & Shigemasu, K. (2003). Application of the somatic marker hypothesis to individual differences in decision making. *Biological Psychology, 65*(1), 81–88.

Tomb, I., Hauser, M., Deldin, P., & Caramazza, A. (2002). Do somatic markers mediate decisions on the gambling task? *Nature Neuroscience, 5*(11), 1103–1104.

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

Whitney, P., Hinson, J. M., Wirick, A., & Holben, H. (2007). Somatic responses in behavioral inhibition. *Cognitive, Affective, & Behavioral Neuroscience, 7*(1), 37–43.

Worthy, D. A., Maddox, W. T., & Markman, A. B. (2007). Regulatory fit effects in a choice task. *Psychonomic Bulletin & Review, 14*(6), 1125–1132.

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013). Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models. *Psychonomic Bulletin & Review, 20*(2), 364–371.

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22406551

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review, 12*(3), 387–402.

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist, 39*(2), 117–123.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition, 34,* 387–398.