frontiers in
BEHAVIORAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 08 February 2010
doi: 10.3389/neuro.08.057.2009

Modeling the value of strategic actions in the superior colliculus

Dhushan Thevarajah[1†], Ryan Webb[2†], Christopher Ferrall[2] and Michael C. Dorris[1]*

[1] Department of Physiology, Centre for Neuroscience Studies and Canadian Institutes of Health Research Group in Sensory-Motor Systems, Queen's University, Kingston, ON, Canada
[2] Department of Economics, Queen's University, Kingston, ON, Canada

Edited by:
Daeyeol Lee, Yale University School of Medicine, USA

Reviewed by:
Tim Behrens, University of Oxford, UK
Hiroshi Abe, Yale University School of Medicine, USA

***Correspondence:**
Michael C. Dorris, Department of Physiology, Room 440 Botterell Hall, Queen's University, Kingston, ON, Canada K7L3N6.
e-mail: dorrism@biomed.queensu.ca

[†]Dhushan Thevarajah and Ryan Webb contributed equally to this work.

In learning models of strategic game play, an agent constructs a valuation (action value) over possible future choices as a function of past actions and rewards. Choices are then stochastic functions of these action values. Our goal is to uncover a neural signal that correlates with the action value posited by behavioral learning models. We measured activity from neurons in the superior colliculus (SC), a midbrain region involved in planning saccadic eye movements, while monkeys performed two saccade tasks. In the strategic task, monkeys competed against a computer in a saccade version of the mixed-strategy game "matching-pennies". In the instructed task, saccades were elicited through explicit instruction rather than free choices. In both tasks neuronal activity and behavior were shaped by past actions and rewards with more recent events exerting a larger influence. Further, SC activity predicted upcoming choices during the strategic task and upcoming reaction times during the instructed task. Finally, we found that neuronal activity in both tasks correlated with an established learning model, the Experience Weighted Attraction model of action valuation (Camerer and Ho, 1999). Collectively, our results provide evidence that action values hypothesized by learning models are represented in the motor planning regions of the brain in a manner that could be used to select strategic actions.

Keywords: decision, macaque, mixed strategy, motor intention, saccade, reinforcement, game theory, EWA

INTRODUCTION

In reinforcement learning models, an individual's choice is a probabilistic function of the current values of possible actions, which in turn are functions of past choices and past rewards (Sutton and Barto, 1998). These learning models are based on the concept of choice reinforcement, traced back to the Law of Effect (Thorndike, 1898; Erev and Roth, 1998).

Empirical studies have supported such learning models in a variety of strategic environments with mixed strategy equilibria (Mookherjee and Sopher, 1994, 1997; Erev and Roth, 1998; Camerer and Ho, 1999; Ho et al., 2007, 2008). However, because learning models predict serial dependence in sequential choices, they conflict with independent (uncorrelated) choice predicted by repetition of the stage game Nash Equilibrium in a repeated game. For example, while laboratory studies of the matching pennies game in humans confirm the equilibrium prediction of a 50/50 ratio of choices, sequential dependencies in individual choices remain (Mookherjee and Sopher, 1994; Ochs, 1995). Similar results have been observed against a computer opponent in studies of both humans (Spiliopoulos, 2008) and monkeys (Lee et al., 2004; Thevarajah et al., 2009). Studies of a broader class of mixed strategy games also exhibit similar choice dependencies though not all the authors address learning models directly (O'Neill, 1987; Brown and Rosenthal, 1990; Rapoport and Boebel, 1992; Rapoport and Budescu, 1992; McCabe et al., 2000).

The goal of this study is to look for evidence of neuronal signals that correlate with the action values predicted by the Experience Weighted Attraction (EWA) learning model (Camerer and Ho, 1999). We use EWA because it is both empirically established and a general formulation. It incorporates simple reinforcement learning (Win/Stay-Lose/Shift), both cumulative reinforcement learning and average reinforcement learning (or Q-Learning) (Watkins, 1989; Erev and Roth, 1998), and belief-based models (Fudenberg and Levine, 1998), as special cases of its parameterization. In fact, it is entirely reasonable for behaviour to lie in some middle ground of the above model restrictions of EWA, and empirical evidence suggests it does (Camerer and Ho, 1999; Ho et al., 2008).

Evidence that learning models are instantiated by the brain has been found from measuring neural signals while humans and animals decide. Evaluative signals are encoded, in part, via dopaminergic structures which represent the difference between realized and expected reward following an action (Schultz, 2004; Caplin et al., 2010). In addition, neural signals have been found that encode the combination of actions and their associated outcomes during adaptive decision-making (Barraclough et al., 2004; Lau and Glimcher, 2007; Seo et al., 2007; Luk and Wallis, 2009). Finally, some neural signals reflect the value of potential actions. Thus they may play an important role in driving the choice process (Platt and Glimcher, 1999; Dorris and Glimcher, 2004; Rushworth et al., 2004; Sugrue et al., 2004; Samejima et al., 2005; Padoa-Schioppa and Assad, 2006; Kennerley et al., 2006; Lau and Glimcher, 2008; Jocham et al., 2009).

We build on this previous work by looking for action value signals within a brain region quite close to the motor output, the intermediate layers of the superior colliculus (SCi). The SCi has a number features that suggest it may encode action value. The

SCi is topographically organized as a map of potential saccadic eye movements (Robinson, 1972; Schiller and Stryker, 1972) and determines when and where a saccade will be directed (Glimcher and Sparks, 1992; Dorris et al., 1997). The SCi receives input signals from upstream brain regions involved in choosing saccades in both strategic environments (Barraclough et al., 2004; Dorris and Glimcher, 2004; Seo et al., 2007; Seo and Lee, 2008) and non-strategic environments (Schultz, 1998; Sugrue et al., 2004; Samejima et al., 2005; Lau and Glimcher, 2007, 2008). The topographic organization of the SCi ensures that any value-related signals we observe are closely associated with specific actions. Moreover, strong lateral inhibition between distant SCi locations could play an important role in selecting between action values associated with competing saccades (Munoz and Istvan, 1998; Dorris et al., 2007). Finally, the SCi sends commands to premotor neurons in the brainstem (Moschovakis and Highstein, 1994), as well as providing feedback to dopaminergic neurons in the ventral tegmental area and substantia nigra (Comoli et al., 2003; Dommett et al., 2005).

We measured preparatory activity in the SCi while a monkey played a simultaneous move game of matching pennies against a computer algorithm designed to exploit serial dependence in the monkey's choices. To control for any serial dependence outside of strategic competition, we also measured activity during a sequential move game with random payoffs. First, we hypothesize that SCi activity displays serial dependence based on both previous saccades and their outcomes, and that more recent events will exert a stronger influence. Second, we hypothesize that SCi activity predicts upcoming strategic choices. Finally, we hypothesize that activity in the SCi provides a signal that is correlated with the current value of actions in the EWA learning model. Collectively, our results support the conclusion that action value signals are represented in the motor planning regions of the brain in a manner suitable for selecting strategic actions.

MATERIALS AND METHODS

Electrophysiological experiments were conducted on two male rhesus monkeys (Macaca mulatta), weighing between 9–13.5 kg each, while they performed saccadic eye movement tasks. All procedures were approved by the Queen's University Animal Care Committee and complied with the guidelines of the Canadian Council on Animal Care. Animals were under the close supervision of the university veterinarian. Physiological recording techniques as well as the surgical procedures have been described previously (Munoz and Istvan, 1998; Thevarajah et al., 2009).

GENERAL METHODOLOGY

Behavioral paradigms, visual displays, delivery of liquid reward, and storage of both neuronal discharge and eye position data were under the control of a PC computer running a real-time data acquisition system (Gramalkn, Ryklin Software). Red visual stimuli (11 cd/m^2) were produced with a digital projector (Duocom InFocus SP4805, refresh rate 100 Hz) and back-projected onto a translucent screen that spanned 50° horizontal and 40° vertical of the visual space. Right eye position was recorded at 500 Hz with resolution of 0.1° using an infra-red eye tracker system (Eyelink II, SR Research). Trials were aborted online if eye position was not maintained within ±3° of the appropriate spatial location or if saccades were initiated outside the 70–300 ms temporal window following target presentation. We have further discussion of aborted trials in Section "Results".

The activity of single neurons was recorded with tungsten microelectrodes (Frederick Haer, 1–2 MΩ at 1 kHz) and sampled at 1 kHz. Data analysis was performed offline using Matlab, version 7.6.0 (Mathworks Inc.) on an Intel Core 2 Duo processor. To quantify neuronal activity, each spike train was convolved with a post-synaptic activation function with a rise time of 1 ms and a decay time of 20 ms (Thompson et al., 1996).

NEURONAL CLASSIFICATION

We recorded the activity from saccade-related neurons located between 1.0 and 3.0 mm below the surface of the SC. The center of each neuron's response field was defined as the location, relative to central fixation, associated with the most vigorous activity during target-directed saccades. One target was always placed at this location (referred to hereafter as *in*) and the other at the mirror-image location in the opposite hemi-field (*out*) except ten experiments where two neurons located in opposite colliculi were recorded simultaneously. For these dual neuron experiments, the two targets were located in opposite hemifields corresponding to the response fields of the two neurons under study. To be included in our analysis, neurons had to meet two requirements: (1) *motor burst*, a transient burst of activity that was time-locked to onset of the saccade into the response field that surpassed 100 spikes/s and (2) *preparatory activity*, neural activity during the 50 ms that followed presentation of the mixed-strategy targets that exceeded 30 spikes/s and was significantly greater than the mean activity 100 ms before fixation point offset (paired t-test, $p < 0.01$). Note that in the modelling Section "Value, SCi Activity and Actions", this preparatory activity will be designated SC^s_{it}.

BEHAVIORAL TASKS

Monkeys performed two behavioral tasks. In the strategic task, monkeys were free to choose between two saccade targets while they competed against an adaptive computer opponent playing the matching pennies game. In the instructed task, monkeys were instructed which saccade to make with the presentation of a single saccade target on each trial. The purpose of the instructed task is to characterize how SCi activity is shaped by previous choices and outcomes. The strategic task is used to emphasize this relationship between SCi activity and the history of the game, and determine whether SCi activity is predictive of choice in a strategic decision making environment.

Strategic task

Monkeys competed in a saccadic version of the repeated mixed-strategy game matching-pennies against an adaptive computer opponent (**Figure 1**). Each trial, both the subject and computer reveal a strategy *in* or *out*. The monkey, pre-designated the "matcher", wins if their strategies match, and the computer, pre-designated the "non-matcher", wins if their strategies differ. The unique Minimax/Nash Equilibrium in mixed strategies is for each player to play *in* and *out* with equal probability (von Neumann and Morgenstern, 1947; Nash, 1951), though our analysis does not require that equilibrium play is achieved. Because our experimental setup limits the ability for the monkey to suffer a loss we replaced a loss with a withholding of reward, though the equilibrium remains unchanged. The payoff

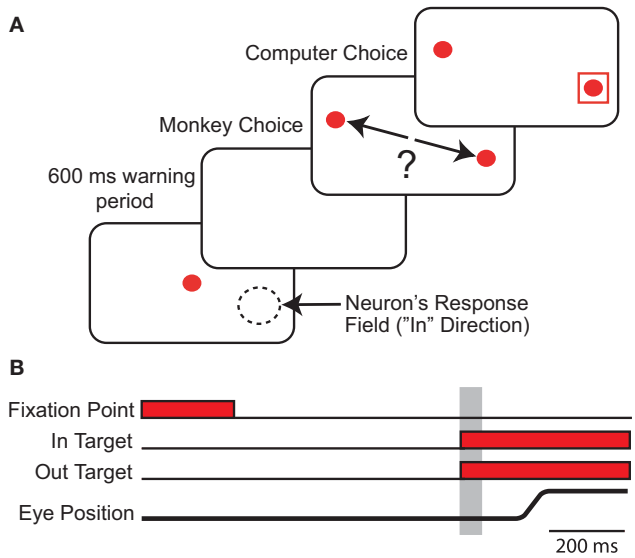

FIGURE 1 | Strategic Task. (A) Each panel represents successive visual displays presented to the monkey. Red circles represent the central fixation point and choice targets respectively. In the third panel, arrows indicate the monkeys possible saccadic choices. One of the saccade targets was always placed in the center of the neuron's response field (i.e., *in* target) as indicted with the dashed circle. The *out* target was placed in the opposite hemifield. The red square indicates the choice of the computer opponent. **(B)** Time-line of strategic task. Grey shaded region indicates the 50 ms epoch during which SCi preparetory activity was sampled for neuronal analyses. The stimuli setup and time-line were identical for the instructed task (not shown) except that only one target was presented per trial and the red square surrounded the target only on rewarded trials.

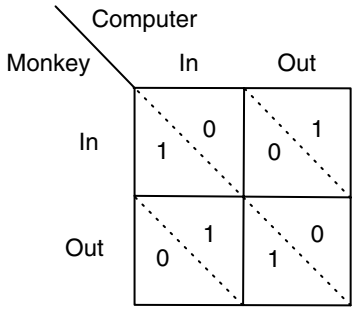

FIGURE 2 | Payoff matrix for strategic task.

matrix is given in **Figure 2** and has been previously studied experimentally in humans (Mookherjee and Sopher, 1994) and monkeys (Lee et al., 2004; Thevarajah et al., 2009).

Subjects were required to maintain central gaze fixation throughout the 800 ms presentation of the fixation point, and after its removal during a fixed 600 ms warning period. Subjects were free to saccade towards either of two simultaneously presented targets, i.e. *in* and *out* of the response field. The fixed warning period and known target locations facilitated advanced selection and preparation of saccades (Thevarajah et al., 2009). After fixating on the target stimulus for 300 ms, a red square, which indicated the computer opponent's choice, appeared around one of the targets for 500 ms. The monkey received a 0.3 mL liquid reward if both players chose the same target and nothing otherwise. The computer opponent performed statistical analyses on the subject's history of previous choices and payoffs and exploited systematic biases in their choice strategy (see algorithm 2 from Lee et al., 2004 for specific details).

Instructed task

The instructed task was identical to the strategic task with two exceptions. First, only a single saccade target was presented on each trial. This target was equally likely to be presented *in* or *out*. Second, reward was equally likely to be received or withheld for successful completion of each trial. Therefore, the expected value of the instructed task is equal to the equilibrium payoff of the strategic task, but saccadic choice was under sensory instruction in the former and under voluntary control in the latter.

DEPENDENCE ON PREVIOUS CHOICES AND REWARDS

To examine any biases exerted by previous saccades and rewards, we segregated SCi activity and saccadic responses on the current trial t based on past ($t-n$, where $1 \leq n \leq 7$) and future ($t+n$, where $1 \leq n \leq 3$) events (Maljkovic and Nakayama, 1994). Future events were examined for control purposes as these should not exert any influence on the current trial. This sequential analysis is illustrated in **Figures 5 and 6** which shows neuronal activity on the current trial segregated into four categories based on four possible events that occurred on the previous trial. (1) a rewarded saccade into the response field (*in*/*R*), (2) an unrewarded saccade into the response field (*in*/*U*), (3) a rewarded saccade out of the response field (*out*/*R*), and (4) an unrewarded saccade out of the response field (*out*/*U*).

We estimated preparatory activity from the postsynaptic spike activation function during the 50 ms following target presentation (**Figure 5**, grey bar). This represented the neuronal firing rate just before saccadic responses were made yet still uncontaminated by visual inputs related to target presentation (Dorris et al., 2000).

The same sequential analysis was performed on choice selection during the strategic task. Response biases were quantified by determining the probability of the monkey selecting the *in* target on the current trial based on past or future events.

Comparatively, for the instructed task, sequential analysis was performed on SRTs rather than saccade choice since saccade location was instructed. SRTs were defined as the time to initiate a saccade following target presentation. Computer software determined the beginning and end of each saccade using a velocity and acceleration threshold. These events were verified by an experimenter to ensure accuracy. Response biases were quantified by examining the influence of an event n trials in the past or future on trials only where saccades were instructed to *in*.

Sequences of trials were constructed from the raw data based on the following criteria. First only sequences of 5 or more consecutive non-aborted trials in length were analyzed. Second, single aborted trials were removed and the sequence was treated as continuous. Third, sequences were started anew if two or more aborted trials occurred in succession. We felt these criteria struck

a balance between providing sufficient sequential data for the analysis in this section while removing those sequences with poor continuity.

EWA LEARNING

The behavior of the subject in trial t of experiment i is coded as

$$s_{i,t} = \begin{cases} in & \text{saccade into response field in trial } t \text{ of experiment } i. \\ out & \text{otherwise.} \end{cases} \quad (1)$$

Let $s'_{i,t}$ denote the computer opponent's choice. Whether reward is received depends on both choices and the experiment being conducted. Let $\pi_{i,t} = 1$ indicate that a reward was received in trial t of experiment i and 0 otherwise.

$$\pi_{i,t} = \begin{cases} R & \text{if } s_{i,t} = s'_{i,t} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

In both tasks, a reward is only received when the choices match, $s_{i,t} = s'_{i,t}$. During the strategic task the computer opponent makes its choice simultaneously, and if the choices match the subject is rewarded with $R = 1$. During the instructed task, $s'_{i,t}$ is chosen before $s_{i,t}$, but even if the choices agree the monkey is only rewarded half the time:

$$R = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Therefore the expected payoff during the instructed task equals the equilibrium expected payoff in the strategic task.

An EWA learning model posits an action value $A^s_{i,t}$ for each strategy s in trial t in experiment i, and includes free parameters which control how action value evolves. On a given trial, it yields a continuous propensity to choose each action, $s_{i,t+1}$, as a monotonic function of current action values $A_{i,t} = [A^{in}_{i,t}, A^{out}_{i,t}]$.

At the start of the experiment $A^{in}_{i0} = A^{out}_{i0}$ for each strategy so that values are equal in the first trial. In general, after trial t the current value of strategy s is updated according to a formula that depends on whether s was chosen or not. If strategy s was chosen then its updated value can be written as a combination of past value (with weight ϕ) and current reward:

$$\text{if } s = s_{i,t}, \text{then } A^s_{i,t} = \frac{\phi N_{t-1} A^s_{i,t-1} + \pi(s_{i,t}, s'_{i,t})}{N_t}. \quad (4)$$

Alternatively, if strategy s was not chosen then its updated value depends on past value (with weight ϕ) and *foregone* payoffs:

$$\text{if } s \neq s_{i,t}, \quad A^s_{i,t} = \frac{\phi N_{t-1} A^s_{i,t-1} + \delta\pi(s, s'_{i,t})}{N_t}. \quad (5)$$

The weight δ is the foregone payoff the subject would have received had it counterfactually chosen s. In both equations, N_t is a trial weight which evolves according to

$$N_t = \rho N_{t-1} + 1. \quad (6)$$

On a given trial, the probability of choosing $s_{i,t} = in$ is defined as

$$P(s_{i,t} = in) = \frac{e^{\lambda A^{in}_{t-1}}}{e^{\lambda A^{in}_{t-1}} + e^{\lambda A^{out}_{t-1}}}, \quad (7)$$

and the parameters $\lambda, \phi, \delta, \rho$, and N_0 are estimated via maximum likelihood. The estimated parameters (except λ) are then used to generate a sequence of fitted action values which we use in our analysis. Importantly, $A_{i,t}$ is computed using only choices and rewards (both actual and fictitious) through trial t, which implies that it can directly enter a model of choice for the next trial, $t+1$. For a complete definition of the EWA model and estimation procedure, see the APPENDIX.

WIN-STAY/LOSE-SWITCH LEARNING

Since EWA is based on a reinforcement premise, it includes a Win-Stay, Lose-Switch choice dependency as a special case. Relative to trial $t + 1$, a Win-Stay outcome for strategy s is coded with an indicator for $s = s_{i,t}$ and $\pi_{i,t} = 1$. A Lose-Switch outcome is $s \neq s_{i,t}$ and $\pi_{i,t} = 0$. This behavior can be captured by a different value, $\text{WSLS}^s_{i,t}$ with its own updating formula,

$$\text{WSLS}^s_{i,t} = \begin{cases} \pi_{i,t} & s = s_{it} \\ 1 - \pi_{i,t} & \text{otherwise.} \end{cases} \quad (8)$$

As in the EWA model, current reward affects the evolution of action value (here represented by $\text{WSLS}^s_{i,t}$). Similarly, the strength of the connection between $\text{WSLS}_{i,t}$ and $s_{i,t+1}$ can be modulated with additional parameters (see Eq. 7). But unlike Eqs. 4 and 5, the WSLS model of value in Eq. 8 does not account for past events before period t nor does it account for a fictitious assessment of actions not chosen (foregone payoffs).

Both Win-Stay/Lose-Switch and the more general EWA models of value predict dependence in the sequence of actions $s_{i,t}$ across adjacent trials. One method for exploring this dependence is to use the updating equations to generate predictors for actions in the following trial. First, we can rewrite Eq. 8 as the sum of two terms,

$$\text{WSLS}^{in}_{i,t} = \text{WS}^{in}_{i,t} + \text{LS}^{in}_{i,t} = 1_{[s_{i,t}=in]}\pi_{i,t} + 1_{[s_{i,t}=out]}\left(1 - \pi_{i,t}\right). \quad (9)$$

This formulation motivates a probit model for choice of the form:

$$P\left(s_{i,t} = in\right) = \Phi\left(\nu_i + \alpha_1 1_{[s_{i,t-1}=in]} + \alpha_2 \text{WS}^{in}_{i,t-1} + \alpha_3 \text{LS}^{in}_{i,t-1}\right), \quad (10)$$

for $t = 1, \ldots, T_i$ and $\Phi()$ denotes the standard normal distribution function (see Wooldridge, 2001 for a discussion of the probit and tobit model introduced below). The term ν_i is a fixed effect for experiment i. $1_{[s_{i,t}=in]}$ is the indicator function which yields 1 if $s_{i,t} = in$ and 0 otherwise. A simple Win-Stay/Lose-Switch hypothesis would predict $\alpha_1 = 0$ since the WS and LS variables would capture all the dependence in the sequence of decisions. Further, it would predict that $\alpha_2 = \alpha_3$, since the effects of winning and losing are symmetric.

VALUE, SCi ACTIVITY AND ACTIONS

To address how value is encoded in neural signals, we introduce our measurement of $SC^s_{i,t}$, defined as the SCi activity associated with saccade target s in trial t of experiment i. In 10 experiments we observe $SC^s_{i,t}$ for both choices; for the other 58 we

observe it only for one choice, $s = in$. To test whether SCi activity encodes the value of actions, in the form of a choice, we estimate the probit

$$P(s_{i,t} = in) = \Phi(\gamma_1 + \gamma_2 SC^{in}_{i,t} + \nu_i), \quad (11)$$

for $t = 1,\ldots,T_i$. Associating $SC^{in}_{i,t}$ with the value of $s = in$ is the hypothesis that $\gamma_2 > 0$. Rejecting the hypothesis $\gamma_2 = 0$ in favour of $\gamma_2 > 0$ is a necessary condition for SC_{i,t} to encode value, but is not sufficient proof that it does.

For the 10 experiments in which we measure SCi activity associated with both choices, we can also estimate a probit of the form

$$Prob(s_{i,t} = in) = \Phi(\mu_1 + \mu_2 \Delta\overline{SC}_{i,t} + \nu_i), \quad (12)$$

where $\Delta\overline{SC}_{i,t} = (SC^{in}_{i,t} - \overline{SC}^{in}_i) - (SC^{out}_{i,t} - \overline{SC}^{out}_i)$ is the difference in SCi activity across actions relative to their within-experiment means, $\overline{SC}^s_i$. A positive value for $\Delta\overline{SC}_{i,t}$ indicates the de-meaned activity associated with the *in* target was larger than for *out*. If choice depends on the comparative value of actions, and value is encoded in SCi activity, then choice probabilities should depend on differences in SCi activity. Thus we hypothesize that $\mu_2 > 0$.

Our final hypothesis is that SCi activity reflects the action valuation in the EWA model. To test it, we consider a random-effects regression of the form:

$$SC^{in}_{i,t} = \beta_1 + \beta_2 D_i + \beta_3 A^{in}_{i,t-1} + \beta_4 D_i A^{in}_{i,t-1} + \beta_5 A^{out}_{i,t-1} + \beta_6 D_i A^{out}_{i,t-1} + \nu_i + \varepsilon_{i,t}. \quad (13)$$

For experiments involving the strategic task, we define $D_i = 1$, with $D_i = 0$ for the instructed task. The constant term, β_1, records the conditional mean activity for the sample of neurons examined, while β_2 measures the effect of the strategic task on this baseline activity. The coefficient β_3 captures the relationship between SCi activity (*in* the response field) and the EWA action value of choosing *in*. The strength of association between SCi activity and action value in the strategic task is determined by the value of the interaction parameter β_4. To capture any relationship between SCi activity (for *in*) and the valuations of alternative actions we include $A^{out}_{i,t-1}$ as a regressor with parameter β_5. Again, this relationship in the strategic task is reflected by the interaction parameter β_6. Our hypothesis is that only EWA action value for *in* positively influences SC activity: $\beta_3 > 0$, $\beta_3 + \beta_4 > 0$, $\beta_5 \leq 0$, $\beta_5 + \beta_6 \leq 0$.

Since SCi activity varies continuously, we can estimate equation 13 as a regression. However, on some trials there is no SCi activity measured during our 50 ms preparatory epoch, thus there is left-censoring at zero of the endogenous variable $SC^s_{i,t}$ for a small but sizeable portion of trials. We account for this censoring by estimating equation 13 as a tobit model.

We should emphasize the timing of our regression equations 12 and 13, presented graphically in **Figure 3**. The EWA valuation A^s_{t-1} is a function of all observed choices and rewards through trial $t-1$ (see Appendix). SCi activity in trial t, SC^s_t, is a function of A^s_{t-1}, therefore is a function of all choices and rewards through trial $t-1$. Finally, the chosen action s_t is a function of the SCi activity in trial t. Importantly, A^s_{t-1} does not include any information from the trial t choice. Thus the maintained hypothesis is that *past* action predicts *current* SCi activity which predicts upcoming choice in the *current* trial.

ALGORITHM FOR COMPUTER OPPONENT

The computer algorithm which the monkey competes against is primarily designed to elicit equilibrium behavior from the monkey, that is, a 50/50 randomization of choices. In doing so, the algorithm does not play the Nash strategy itself. This somewhat paradoxical setup is a result of the unstable nature of mixed strategies highlighted by

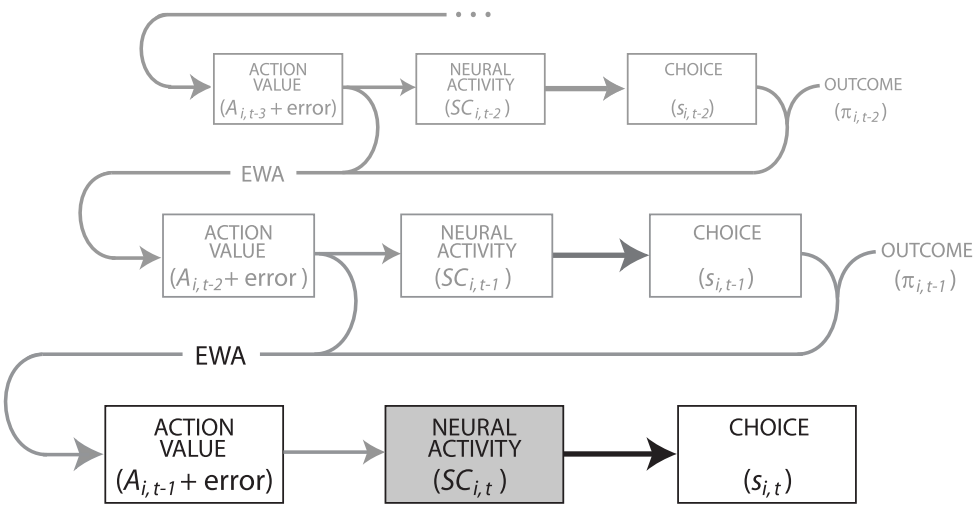

FIGURE 3 | Schematic illustrating the recursive EWA computation and its hypothesized influence on SCi activity. In our experiments, we record and analyze current neural activity ($SC_{i,t}$, gray box). $SC_{i,t}$ lies at the nexus between action value, calculated from past events, and choice on a trial t. Action value ($A_{i,t-1}$) is recursively updated based on an EWA learning model whose main inputs are past choices ($s_{i,t-1}$) and their realized or forgone outcomes ($\pi_{i,t-1}$). Action value also includes an error term to highlight that the calculated $A_{i,t-1}$ only approximates and does not fully capture underlying valuation, in addition to any error arising from noisy neural signals. As such, the relationship between hypothesized $A_{i,t-1}$ and neural activity is likely weaker (thin arrows) than between neural activity and the observable $s_{i,t}$ (thick arrows).

Harsanyi (1973). When the computer is not adaptive, but simply randomizes its choices, the monkey is indifferent between his strategies (any strategy the monkey chooses will be rewarded on half of the trials) and the monkey's choices become strongly biased in one direction (Lee et al., 2004). For this reason, the algorithm was designed to exploit the monkey's choice biases, perhaps more in line with what constitutes (approximate) equilibrium in such games. Refer to algorithm 2 from Lee et al. (2004) for additional details on the computer opponent.

RESULTS

We begin by characterizing the effects of current and previous trials on both behavior and SCi activity in Sections "Analysis of Current Trial", "Dependence of Choice on Previous Trial" and "Sequential Dependence of Choice". In section "Dependence of Choice on Previous Trial", we formally test for a Win-Stay/Lose-Switch strategy. The ability of SCi neurons to predict choice is examined in section "Neuronal Choice Prediction". In section "Behavioural EWA Estimates", we fit the EWA model to choice data and generate sequences of action values for each monkey. Finally, having established that choice is dependent on previous trials, and SCi activity predicts choice on a given trial, in Section "Encoding EWA Action Value" we test our hypothesis that SCi neurons represent the action-specific valuations posited by EWA.

We have data from 68 experiments where neurons satisfied our criteria for inclusion (See Section "Materials and Methods"). In 10 of these experiments, we were able to measure SCi activity associated with both saccades simultaneously, 20 neurons total. In the remaining 58 experiments, we were able to measure SCi activity associated with only one of the potential saccades.

The data consists of a choice, preparatory SCi activity, and a saccadic response time (SRT) for a set of $i = 1 \ldots 78$ neurons respectively with T_i ordered trials. In 38 of these experiments, data were collected for both the strategic task and the instructed task control. This sub-sample of 38 neurons is used in Sections "Analysis of Current Trial", and "Dependence of choice on previous trial", and "Sequential dependence of choice". In this sub-sample, a mean of 246 ± 11 SEM trials per neuron were analyzed during the strategic task and a mean of 146 ± 8 SEM trials per neuron were analyzed during the instructed task. The full sample is used in Section "Neuronal Choice Prediction", while Sections "Behavioural EWA Estimates" and "Encoding EWA Action Value" drop experiments in which greater than 30% of the trials were aborted. These experiments were dropped since many aborted trials within an experiment may interrupt the sequence of valuation posited by EWA learning. The cut-off 30% was set to balance choice sequence consistency and sample size.

ANALYSIS OF CURRENT TRIAL

We will briefly characterize saccade behaviors and SCi preparatory activity on the current trial before examining the effects of events on previous trials. A more detailed current trial analysis can be found in Thevarajah et al. (2009). All reported statistics are (mean ± se).

The allocation of saccade choices did not differ between the two targets during the strategic task [$p(in) = 49.8 \pm 0.6\%$; paired t-test $p > 0.05$]. Moreover, SRTs did not differ between the two targets during the instructed task (*in*: 192.9 ± 4.2 ms, *out*: 186.1 ± 3.7 ms, $p > 0.05$). These behavioral measures suggest that, on average, saccade preparation processes were not biased towards any one particular target location during either task. However, in the strategic task the monkey was rewarded on only 42.2% of the trials, whereas in the instructed task the monkey was rewarded half the time (**Table 1**).

In both tasks, neuronal activity steadily increased during the warning period in advance of choosing either target (**Figure 4**). Overall preparatory activity was greater regardless of saccade direction during the strategic task compared to the instructed task (*in*: $p < 0.05$, *out*: $p < 0.05$). Moreover, in the strategic task activity was segregated for saccades *in* (99.9 ± 8.8 spikes/s) and *out* (80.2 ± 7.2 spikes/s, paired t-test, $p < 0.001$), whereas activity was not segregated between *in* (63.5 ± 6.5 spikes/s) and *out* (64.5 ± 6.5 spikes/s) saccades during the instructed task (paired t-test, $p > 0.05$). This greater overall activation and neuronal selectivity during the strategic task may occur because saccades are under voluntary control and can be planned in advance. In the instructed task the monkey must wait for the presentation of the target.

DEPENDENCE OF CHOICE ON PREVIOUS TRIAL

We examine sequential choice dependencies by segregating behavior and neuronal activity on the events of the previous trial (i.e., previous choice and its reward outcome). Particularly, we test for the prevalence of a WS/LS strategy.

The influence of previous trials on subsequent saccadic responses

We begin by summarizing the frequencies of WS/LS choice patterns in the strategic task over all experiments (**Table 1**). Choices were repeated in a WS/LS pattern in 55.5% of the trials. A WS was observed in 62.1% of post-win trials vs. LS observed in 50.6% of post-loss trials, which suggests a WS/LS strategy is solely due to a Win-Stay rather than Lose-Shift bias. The larger percentage of losing trials suggests the computer opponent was able to exploit this tendency in choice patterns.

To further assess the influence of previous trial events, we estimate Eq. 10 which models choice as a function of lagged choice and the Win-Stay and Lose-Switch variables. Estimates of the fixed-effects probit are presented in **Table 2**. The explicit prediction of the simple WS/LS strategy is rejected because the estimated coefficients α_2 and α_3 are significantly different from each other: the tendency to repeat rewarded actions is greater than the tendency to switch from unrewarded actions. We can also note that the tendency to repeat choices is largely due to the Win-Stay bias since α_1 is not significantly different from zero.

To measure any biases in the instructed task, saccadic reaction times (SRTs) were examined in **Table 3**. Considering that target location and outcome were stochastic, therefore unpredictable, these previous events had a surprisingly large influence on SRTs. Repeating an action resulted in faster SRTs than switching actions (Stay vs. Switch, binomial test $p < 10^{-11}$). SRTs were particularly biased if a saccade direction was previously rewarded (Win-Stay vs. Lose-Stay, binomial test $p < 10^{-5}$; Win-Switch vs. Lose-Switch, binomial test $p < 10^{-3}$). This suggests preparation biases were a function of both previous choices and their outcomes.

The influence of previous trials on SCi preparatory activity

Figure 5 illustrates how SCi activity was also influenced by the previous trial. The black dashed line shows mean activity over all experiments. Each of the coloured lines depicts how current trial activity was influenced by choices and outcomes on the previous

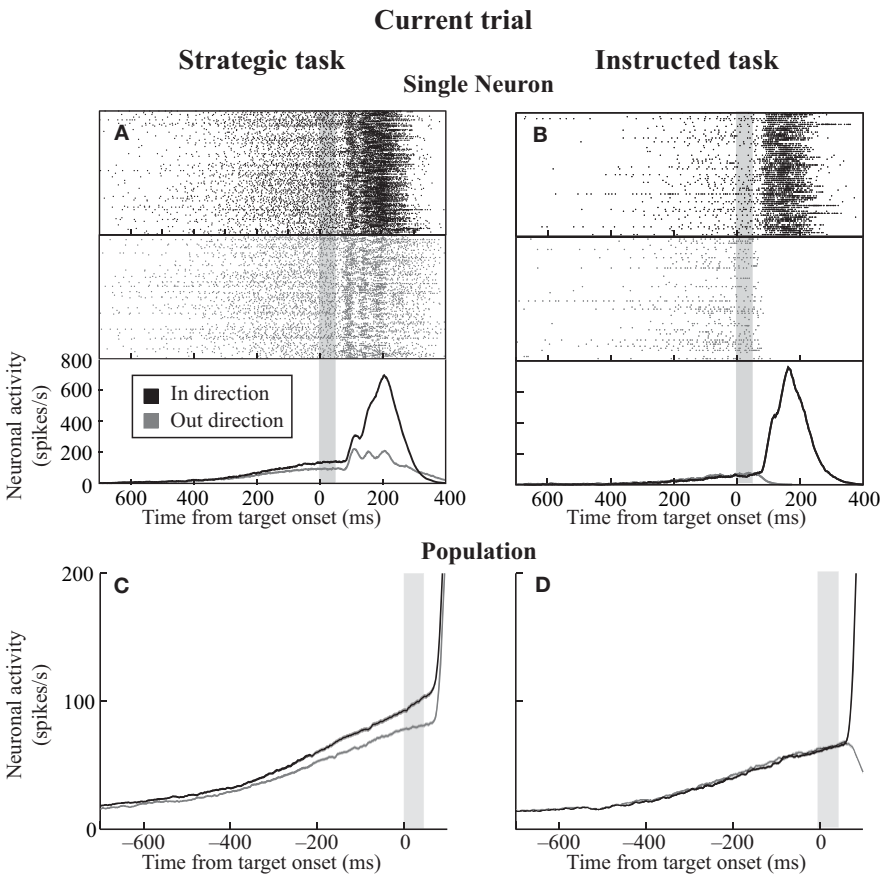

FIGURE 4 | SCi activity during the current trial. (A, B) Activity of a representative SCi neuron during the strategic **(A)** and instructed **(B)** tasks. Rasters (top panels) and post-synaptic activation functions (bottom panel) are sorted based on saccades directed *in* (black) and *out* (gray) of the neuron's response field. The shaded gray bar denotes the epoch during which preparatory activity was analyzed. **(C, D)** Mean activity of neuron sample in which both strategic **(C)** and instructed **(D)** tasks were recorded (38 neurons).

Table 1 | Frequencies of choice dependencies in strategic game.

Previous trial dependency	Proportion (%)
Win	42.2
Loss	57.8
Win-Stay/Lose-Switch	55.5
Win-Stay	62.1
Lose-Switch	50.6

Table 2 | Probit estimates of $s_{i,t+1}$ on lagged choice and Win-Stay/Lose-Switch outcomes.

Variable	Coefficient	Estimate	Standard error	*p*-Value
$I(s^{in}_{i,t-1})$	α_1	0.1415	0.1095	0.20
$WS^{in}_{i,t}$	α_2	0.5478	0.1118	0.00
$LS^{in}_{i,t}$	α_3	0.252	0.0986	0.01

Estimates of Eq. 10 on sample of 33 experiments. (5 redundant paired experiments dropped). Sample size = 8809. Estimates of the 33 experimental effects ν_i *not reported. Reported standard errors are clustered within experiments.*

Table 3 | Reaction time dependencies in instructed task.

Previous trial dependency	Reaction time (ms)	Standard error
Win-Stay	170.8	4.0
Stay	177.2	3.2
Lose-Stay	183.6	5.0
Lose-Switch	194.5	5.0
Switch	200.1	4.0
Win-Switch	205.7	6.4

trial. This influence is most prevalent at the end of the warning period (gray-shaded area). Therefore we will use SCi activity in this epoch for the sequential analysis that follows.

SEQUENTIAL DEPENDENCE OF CHOICE

Having observed a dependency in choices and outcomes in the previous trial, we will now characterize this dependency over multiple trials. Two sequential patterns were evident in both tasks (**Figure 6**). First, more recent events had the greatest influence. Second, actions that were rewarded generally had a more pronounced effect, both in terms of magnitude and duration, than

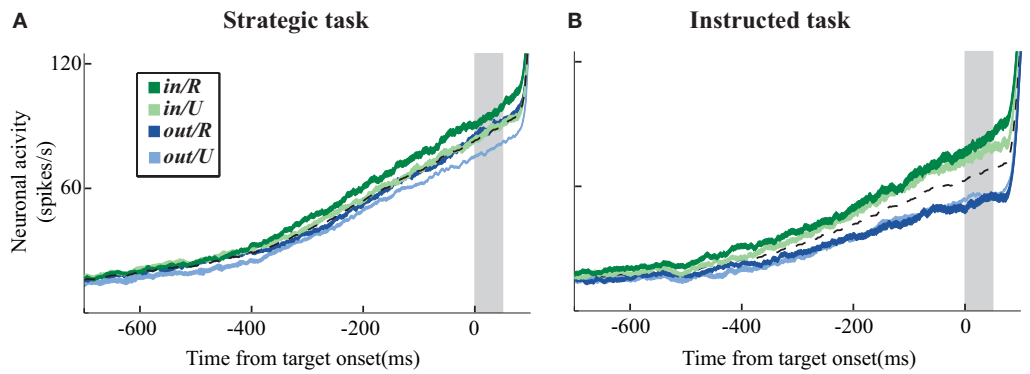

FIGURE 5 | Mean activity of neuronal sample (38 neurons) segregated on previous trials during the strategic (A) and instructed (B) tasks respectively. The black dashed line represents the mean activity for all trials. The colored lines also represent current trial activity, but this activity is segregated into four categories based on the combination of target choice (*in* or *out*) and their outcomes (*R* or *U*). Note that the mean activity is a sum of each of the four colored lines weighted by the proportion of trials in each category. Line widths represent the SEM. The shaded gray bar denotes the epoch during which preparatory activity was sampled during subsequent analyses.

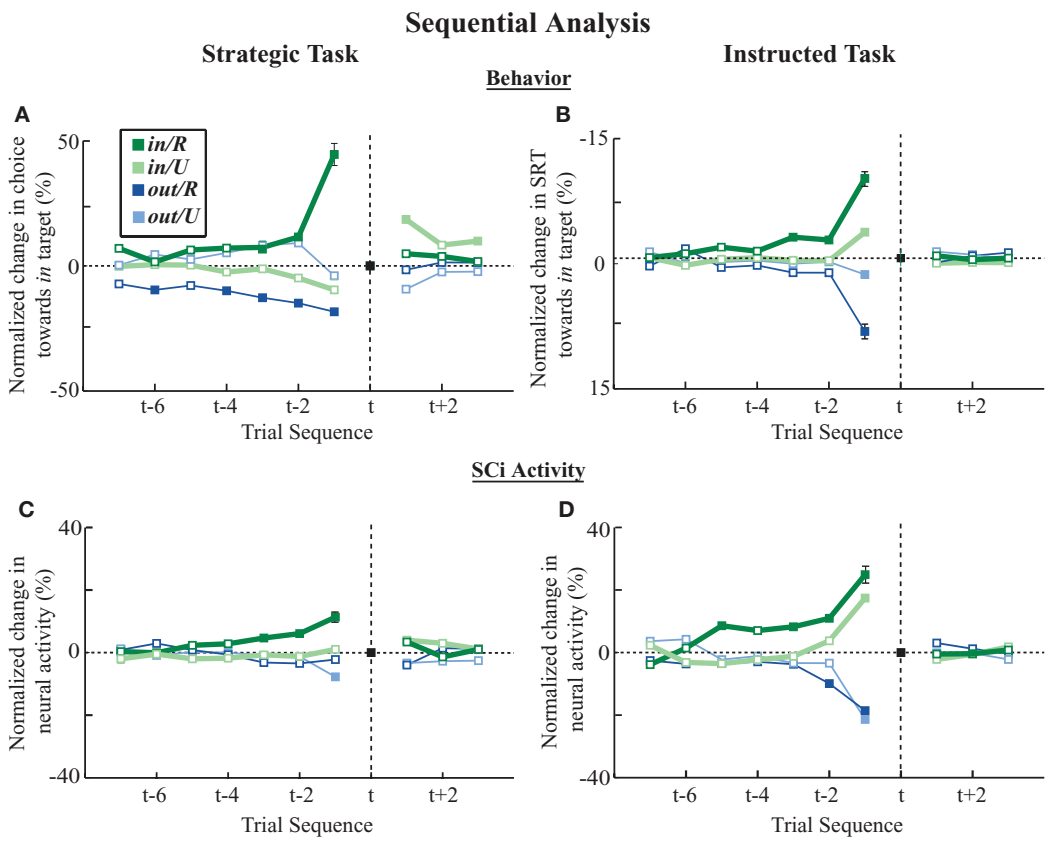

FIGURE 6 | Saccade behavior and SCi activity segregated on previous and future trials. The data for each trial sequence is presented as the percentage change from the mean data for all trials at trial *t* (black dots). Note that each data point in the trial sequence represents the influence of an earlier or later trial on the current trial. Therefore, the four colored data points at each time sequence always sum to the mean data point at time *t* when weighted by the proportion of trials in each category. **(A)** Changes in *in* target choices during the strategic task. **(B)** Changes in *in* target SRTs during the instructed task. Note that the the ordinate axis has been flipped because SRTs are negatively correlated with SCi activity. **(C)** Changes in SCi activity during the strategic task. **(D)** Changes in SCi activity during the instructed task. Filled squares indicate significant differences from the mean activity. Representative standard errors are shown for *in*/*R* data points.

unrewarded actions (**Figure 6**, dark colored lines vs. light colored lines). Whether a previous trial was rewarded or not, did not, by itself, affect SC activity or saccade behaviors. Instead, the effects of reward influenced a particular saccade location rather than providing a general motivating or alerting effect for both actions.

The strategic and instructed tasks also differed in two ways during this sequential analysis. First, future events were correlated with choice selection in the strategic but not the instructed task (**Figure 6A**). This seemingly paradoxical finding is a consequence of the computer exploiting the monkey's Win-Stay bias. That is, monkeys were more likely to lose following a rewarded trial as they tended to repeat actions. This phenomena is evident in the Lose-Stay bias observed in future choices in **Figure 6A**. Second, modulation of SCi activity by past events was greater for the instructed task than for the strategic task. For example, the change in activity imposed by the previous trial was approximately three times as large during the instructed task compared to the strategic task (compare the spread in data along the vertical axis in **Figures 6C vs. D**).

NEURONAL CHOICE PREDICTION

Having characterized serial dependency in choices, the second step in determining whether neurons in the SCi encode action value is to determine if activity predicts choice. The ten experiments where we measured two neurons simultaneously, one for each target, allows us to specify how opposing SCi activity is compared in Eq. 12. Results for the fixed-effects probit estimation are given in **Table 4**.

The parameter μ_2 measures the impact of SCi activity on the probability of an *in* saccade and is both positive and highly significant. To interpret the magnitude of the coefficient μ_2, we take the predicted probabilities from the regression and compare them to the observed choices by two methods. The first rounds the probabilities to the nearest integer and compares them to the choices, resulting in a prediction rate of 65%. The second simulates choices from the binomial distribution using the predicted probabilities, and compares the simulated choices to the actual choices, resulting in a prediction rate of 56% for 1000 simulations. Comparatively, 1000 independent draws from a 50/50 binomial distribution would predict 56% of the trials (560 matches of the monkey's choice) with probability 6.3×10^{-5}. Results did not change significantly when we estimated on half the sample and predicted out of sample.

For the entire 78 neuron sample, we can also assess how well single neurons predict choice from Eq. 11. Results are reported in **Table 5**. Again, we observe that the estimate of SCi activity, γ_2, is both positive and highly significant. As before, our assessment of the magnitude of the parameter γ_2 relies on in-sample prediction. Rounding the fitted probabilities results in a 60% prediction rate for the 78 individual neurons, while simulating the choices results in a 53% prediction rate. As expected, the single neuron is a worse predictor compared to the the paired neuron analysis, presumably because choice is based on a comparison of valuation between the two targets. Again, 1000 independent 50/50 draws would still only predict 53% with probability 0.03.

Table 4 | Probit estimates of $s_{i,t}$ based on difference in activity from neuronal pairs.

Variable	Coefficient	Estimate	Standard error	*p*-Value
Constant	μ_1	0.0435	0.0015	0.00
$\Delta \overline{SC}_{i,t}$	μ_2	0.0054	0.0005	0.00

Estimates of Eq. 12 using ten experiments with paired neuronal measures. Fixed effect estimates are not reported. Standard errors were clustered at the experiment level.

BEHAVIOURAL EWA ESTIMATES

To generate a sequence of action values which reflect each monkey's valuation on a given trial, we estimated the EWA model on choice data (see Section "EWA Learning" and APPENDIX). Estimates are reported in **Tables 6 and 7.** We observe significant heterogeneity in the fitted EWA parameters, similar to Ho et al. (2008). Estimates suggest Monkey H (54/78 experiments) is a cumulative reinforcement learner ($\delta = 0$, $\rho = 0$), while Monkey B (24/78 experiments) has a fictive learning component and averages rewards as in Q-Learning ($\delta > 0$, $\rho = \phi$). For each monkey, the estimates for ϕ, δ, ρ, and N_0 are used to generate the sequence $A^s_{i,t}$ which we use in section "Encoding EWA Action Value".

Table 5 | Probit estimates of $s_{i,t}$ based on activity from individual neurons.

Variable	Coefficient	Estimate	Standard error	*p*-Value
Constant	γ_1	0.0045	0.0008	0.00
$SC_{i,t} - \overline{SC}_i$	γ_2	0.0053	0.0005	0.00

Estimates of Eq. 11 using 78 experiments. Fixed effect estimates are not reported. Standard errors were clustered at the experiment level.

Table 6 | EWA Estimates for Monkey B.

Parameter	Estimate	Standard Error
λ	3.68	3.54
ϕ	0.78	0.08
δ	0.12	0.07
ρ	0.91	0.08
N_0	3.73	8.46

Sample of 19 experiments for monkey B (4/24 experiments dropped due to >30% aborted trials; 1/54 redundant paired experiments additionally dropped).

Table 7 | EWA Estimates for Monkey H.

Parameter	Estimate	Standard Error
λ	0.45	0.29
ϕ	0.52	0.04
δ	0.00	0.05
ρ	0.00	0.64
N_0	1	0

Sample of 27 experiments for monkey H (20/54 experiments dropped due to >30% aborted trials; 7/54 redundant paired experiments additionally dropped). The restriction $N_0 = 1$ was imposed to ensure identification of ρ (see Appendix).

ENCODING EWA ACTION VALUE

The EWA action value is a function of the observed choices and reward structure of the game. Our final hypothesis is that SCi activity reflects the fitted action values from Section "Behavioural EWA Estimates". To test this hypothesis, we estimate Eq. 13 separately for each monkey and its appropriate action value $A^{s}_{i,t}$. Results are reported in **Tables 8 and 9**.

For monkey H (**Table 8**), the instructed task relationship between EWA action value and SCi activity for target *in* is positive, significant and large in magnitude $(\beta_3 = 24.46)$. Over the observed range of the EWA action value $(0.00 < A_{i,t} < 1.96)$, this represents an 81% change in SCi activity relative to baseline activity of 59.86 spikes/s. Notably, this relationship is partially offset by the *out* EWA action value (H_0: $\beta_3 + \beta_5 = 0$, $p = 0.36$). If the action values of the two targets were equal ($A^{in}_{i,t} = A^{out}_{i,t}$), the estimates predict there would still be an increase in SCi activity for the *in* target. This suggests that a given SC neuron encodes the action value for the target it is associated with on the topographic map, but other neurons (valuable targets) can partially inhibit this valuation.

Table 8 | Estimates of $SC^{in}_{i,t}$ on EWA action values and task type for monkey H.

Variable	Coefficient	Estimate	Standard error	*p*-Value
Constant	β_1	59.86	12.94	0.00
D_i	β_2	24.69	9.46	0.01
$A^{in}_{i,t}$	β_3	24.46	3.72	0.00
$D_i A^{in}_{i,t}$	β_4	–9.07	4.58	0.05
$A^{out}_{i,t}$	β_5	–17.95	5.74	0.01
$D_i A^{out}_{i,t}$	β_6	17.25	5.91	0.00
Var(ν_i)	σ^2_ν	59.69	9.86	0.00
Var($\varepsilon_{i,t}$)	σ^2_ε	62.99	5.43	0.00

Random-effects tobit estimates of Eq. 13 on 36 neurons for monkey H (18/54 experiments were dropped due to > 30% aborted trials). Sample size=10704, 998 observations censored at 0. σ^2_ν is the variance of the random effect ν_i; σ^2_ε is the variance of $\varepsilon_{i,t}$. $A^{s}_{i,t}$ is generated using behavioural EWA estimates for monkey H (see Section "Behavioural EWA Estimates"). Standard errors are calculated by means of clustered bootstrap with 1000 bootstrap samples, re-sampling within experiment i.

Table 9 | Estimates of $SC_{i,t}$ on EWA action values and task type for monkey B.

Variable	Coefficient	Estimate	Standard Error	*p*-value
Constant	β_1	48.72	10.39	0.00
D_i	β_2	20.51	9.30	0.03
$A^{in}_{i,t}$	β_3	28.06	21.10	0.18
$D_i A^{in}_{i,t}$	β_4	–5.45	31.28	0.86
$A^{out}_{i,t}$	β_5	8.71	12.62	0.49
$D_i A^{out}_{i,t}$	β_6	–10.93	19.14	0.57
Var(ν_i)	σ^2_ν	35.29	7.28	0.00
Var($\varepsilon_{i,t}$)	σ^2_ε	39.78	4.59	0.00

Random-effects tobit estimates of Eq. 13 on 19 neurons for monkey B (5/24 experiments were dropped due to >30% aborted trials). Sample size = 5907, 258 observations censored at 0. σ^2_ε is the variance of the random effect ν_i; σ^2_ε is the variance of $\varepsilon_{i,t}$. $A^{s}_{i,t}$ is generated using behavioural EWA estimates for monkey B (see Section "Behavioural EWA Estimates"). Standard errors are calculated by means of clustered bootstrap with 1000 bootstrap samples, re-sampling within experiment i.

As expected from our sequential analysis, the relationship between SC activity and action value is attenuated in the strategic task ($\beta_4 < 0$) though it is still positive and significant (H_0: $\beta_3 + \beta_4 = 0$, $p = 0.00$). The estimates yield a 36% increase in SC activity relative to baseline ($\beta_1 + \beta_2 = 84.5$ spikes/s) over the range of $A^{in}_{i,t}$. However the *out* EWA action value now has no impact (H_0: $\beta_5 + \beta_6 = 0, p = 0.82$) suggesting no inhibition from *out* target neurons during this measurement epoch of the strategic task.

Estimation results for monkey B have considerably more variance (**Table 9**). In the instructed task, we still observe a positive coefficient for $A^{in}_{i,t}$ ($\beta_3 = 28.06$) but with a larger p-value ($p = 0.18$) and a smaller magnitude relative to baseline (33%) over the observed range of action values ($0.00 < A_{i,t} < 0.58$). While the estimate for attenuation in the strategic sample is of the correct sign ($\beta_4 < 0$), it is not significantly different from zero ($p = 0.86$). The estimates for the *out* action value are also highly variable and not significantly different from zero in either task. We should note that the sub-sample for monkey B contains half as many observations and neurons as the sub-sample for monkey H, though this efficiency loss likely does not account for all of the increased variability of the estimates.

DISCUSSION

SUMMARY OF FINDINGS

This study examined whether a valuation of future actions, constructed as a function of previous choices and rewards, is represented by the superior colliculus in a strategic environment. Our results show that SCi preparatory activity was shaped by both previous saccades and their outcomes, particularly a Win-Stay bias, and more recent events had a more pronounced effect. These sequential biases were reflected in upcoming choices during the strategic task and upcoming saccadic reaction times during the instructed task.

SCi activity was also predictive of upcoming strategic saccades on a trial-by-trial basis (**Tables 4 and 5**); at a rate of 60% for single neurons and 65% for opposing neuron pairs. Although our pool of neuron pairs was small (10 pairs), this improvement in prediction suggests that it is not the absolute level of activity, but the relative level of activity between potential actions, that is best correlated to choice.

The fact that SCi activity was both shaped by previous choices and rewards and predicted future choices suggest it as a candidate neural correlated of action values posited by behavioural learning model. Our analysis demonstrated that SCi activity was correlated on a trial-by-trial basis with the EWA learning valuation. Specifically, SCi activity was positively correlated with the action value for its response field, with some evidence that it is negatively correlated with the action value of the alternative target. Collectively, our empirical and modelling results suggest that hypothesized action value signals are represented in the motor planning regions of the brain in a manner that could be used to select strategic actions.

EFFECTS OF PREVIOUS ACTIONS AND REWARDS

Serial dependence of choices has previously been observed in strategic and non-strategic environments. Consistent with previous studies, more recent events had a greater influence on both

choices (Juttner and Wolf, 1992; Maljkovic and Nakayama, 1994; Dorris et al., 2000; Barraclough et al., 2004; Lee et al., 2004; Lau and Glimcher, 2005) and neuronal activity (Dorris et al., 2000; Bayer and Glimcher, 2005; Seo and Lee, 2007), and these influences decayed with time (**Figure 6**). Unlike the computer opponent which weighed all past events equally, monkeys gave more weight to recent events when selecting actions. This policy may be an efficient solution for using past events to predict future rewarded actions given organisms have a limited memory store (Anderson et al., 1996; Callicott et al., 1999), and it allows organisms to more readily adapt to a changing environment.

Sequential effects have been characterized previously in the SCi during a task similar to our instructed task (Dorris et al., 2000). Although target location was unpredictable in this previous study, all saccades were rewarded; therefore the contribution from repeating a motor action, or repeating a rewarded location, remained unclear. By allocating rewards unpredictably, we were able to isolate the contribution of these factors. Previously unrewarded actions had a biasing effect, but to a lesser extent than previously rewarded actions. We found no effect of previously rewarded trials when analyzed independently of actions, which suggested that reward, at least our task, did not have a generalized alerting or motivating effect. Instead, SCi activity was found to be influenced by a combination of both previous actions and rewards. These biases, in turn, were reflected in saccade behaviors (**Figure 6**).

Finally, we observed differences in how SC activity was influenced by previous events during the two tasks. First, the overall level of SC activity was greater preceding strategic than instructed saccades (i.e., compare black dashed lines in **Figures 5A vs. B**). Strategic saccades may have been more fully prepared because the locations of the two targets were known in advance whereas the location of the single target had to be identified before the saccade preparation processes could be completed in the instructed task. Second, previous events exerted less influence on SCi activity during the strategic task (i.e., compare **Figures 6C vs. D**). This was observed in the magnitude of the sequential dependencies and the number of previous trials which exerted an influence. Although having sequential biases was seemingly unnecessary in the instructed task, as the monkey could neither control nor predict saccade direction or reward, having such biases were relatively inconsequential. In the strategic task however, sequential biases led to exploitation by the computer opponent as evidenced by a reduced reward rate (**Table 1** and Barraclough et al., 2004). Our results suggest the influence of previous events, borne out in sequential dependencies, can be attenuated in strategic situations.

WIN-STAY BIAS

Though the analysis in Sections "Dependence of Choice on Previous Trial" and "Sequential Dependence of Choice" revealed notable choice tendencies in the strategic sample, many of which are incorporated in the EWA learning model, there is one in particular we wish to highlight. Although both effects were significant, subjects repeated winning choices more often than switching from losing choices controlling for repeated choices ($\alpha_2 > \alpha_3$), or a Win-Stay bias. This observation is a rejection of a strict Win-Stay/Lose-Switch model of choice in repeated games.

However, a stronger Win-Stay bias is compatible with our candidate model of action value (EWA). If unchosen winning actions are updated by a fraction $\delta < 1$ relative to chosen winning actions, the difference in the action value after a rewarded trial is larger than after an unrewarded trial:

$$\Delta A_t \equiv A_t^{s_t} - A_t^{-s_t} = \begin{cases} \frac{\phi N_{t-1}\left(A_{t-1}^{s_t} - A_{t-1}^{-s_t}\right)+1}{N_t} & \text{if } \pi_t = 1 \\ \frac{\phi N_{t-1}\left(A_{t-1}^{s_t} - A_{t-1}^{-s_t}\right)-\delta}{N_t} & \text{if } \pi_t = 0 \end{cases} \quad (14)$$

Therefore

$$\left(\Delta A_t\right)_{\pi_t=1} - \left(\Delta A_t\right)_{\pi_t=0} > \frac{1-\delta}{N_t} > 0. \quad (15)$$

This result holds generally for all models nested by EWA, as long as $\delta < 1$. A Win-Stay bias may be exacerbated in our experiment because our payoff matrix is not zero-sum (**Figure 2**); not matching the opponent constituted a withholding of reward rather than a loss of reward. This asymmetry in payoffs may bias the subject's responses in favour of rewarded trials.

PREDICTING CHOICE

Our results indicate that the activity of individual SCi neurons can predict upcoming choices with 60% reliability. Although significantly better than chance, the SCi may not appear to be a particularly impressive predictor. However, a number of issues must be taken under consideration to make this judgment.

The predictive capability of SCi neurons depends on the number of neurons in the population, the correlation in their firing patterns, and the manner in which downstream structures read-out these predictive signals. Although we only had a sample of 10 neuronal pairs, our results demonstrate that simply comparing the relative firing of two opposing neurons increases prediction from 60% to 65%. Moreover, while the predictive capability of any one (or two) neuron(s) may be weak, this is a very consistent prediction across the neuronal population (see **Figure 5D** from Thevarajah et al., 2009). Therefore, these small individual biases can be amplified to provide a strong signal for selecting strategic actions.

Although the SCi is required for generating saccades (Hanes and Wurtz, 2001) and manipulating SCi activity alters saccadic choices (Carello and Krauzlis, 2004; McPeek and Keller, 2004; Dorris et al., 2007; Thevarajah et al., 2009), the robust activity for *out* direction saccades (**Figure 4**) demonstrates that the reverse is not true; executing a saccade is not a pre-requisite for preparatory SCi activity. This evidence strongly suggests that a causal arrow passes from SCi to choice uni-directionally (**Figure 3**). Similarly, if action value is indeed a function of past choices, then it must be action value that influences SCi activity. If these arrows were not uni-directional then current activity or choices would paradoxically cause past choices.

NEURONAL CORRELATES OF EXPERIENCE WEIGHTED ATTRACTION

Our preliminary analysis has shown that both behaviour and SCi activity are correlated with previous choices and rewards, particularly through a reinforcement of rewarded choices (Win-Stay). To formalize this result, we found a neural correlate of a general learning model based upon this reinforcement premise. This model calculates an action value on each trial as a function of the history of observed

choices and payoff structure of the game. Therefore, our results in Section "Encoding EWA Action Value" are consistent with the hypothesis that neurons in the SCi encode the history of the two tasks in the form of learned action values for each potential action. A given neuron in the SCi is correlated with the action value of its target in both tasks, though the magnitude of this relationship is attenuated in the strategic task. Further, SCi activity is negatively correlated with the action values of competing targets in the instructed task, but not in the strategic task during the period we measure. This suggests that both the attenuation of the value/SCi relationship, and the lack of inhibition from competing neurons *within* the preparatory period we measure, may serve a strategic purpose.

The EWA model we use in this study (Camerer and Ho, 1999) is a general learning model that has proven successful in predicting play both in and out of sample in a wide variety of games. The role EWA plays in our analysis is akin to an objective valuation. It is a function of past choices and rewards which reflects a component of the relative value of each strategy. As such, there remain unaddressed components of value. Learning models do not assess the forward-looking value of an action. That is, there is no consideration of repeated game strategies such as "leading" an opponent in order to exploit him in later periods (though we should emphasize the only unique repeated game equilibrium in matching pennies is the stage game equilibrium). Our analysis also does not address satiation in the experiment nor learning between experiments. However, the relative success of EWA in predicting choice in a strategic environment suggests that its historical, objective component is important in the ultimate valuation of an action.

As a theoretical construct of valuation, both the simplifying assumptions mentioned above and additional neural and/or behavioural factors will combine to limit the explanatory power of EWA (referred to in **Figure 3**). But even if the SC is not coding action value as specified by EWA, the fact that EWA action value significantly predicts SC activity suggests that the correct model will share many features of the EWA formulation. Whether a complete model actually nests EWA as a special case remains an open question that is beyond the scope of this paper.

There has been some progress in identifying the neural correlates of the functional elements of EWA. It has been previously observed that the striatum encodes the difference between realized and expected reward, suggesting the striatum may form part of a learning system in the brain (Schultz, 1998; Caplin et al., 2010). Rewriting Eq. 18 for only the chosen strategy s_{it} highlights the role the striatum may play in a general EWA formulation:

$$A_{it}^{s_{it}} = \frac{(1+\phi)N_{i,t-1}A_{i,t-1}^{s_{it}} + \Delta_{it}}{N_t}, \tag{16}$$

where

$$\Delta_{it} = \left[\pi\left(s_{it}, s'_{it}\right) - A_{i,t-1}^{s_{it}}\right], \tag{17}$$

and Δ_{it} is the dopaminergenic response system analyzed in Caplin et al. (2010). Left unspecified here is the means by which all action values for unchosen actions, $s \neq s_{it}$, are updated (see Lohrenz et al., 2007).

Other important components associated with reinforcement learning models are also encoded in a network of cortical structures that send projections to the SCi. In contrast to the SCi, the signals carried by these cortical structures are much more heterogeneous across individual neurons. A proportion of neurons in the dorsolateral prefrontal cortex (Barraclough et al., 2004), dorsal anterior cingulated cortex (Seo and Lee, 2008) and lateral intraparietal cortex (Platt and Glimcher, 1999; Dorris and Glimcher, 2004; Seo et al., 2009) encode relevant information necessary to construct action value such as past choices, opponent's choices, the animal's reward history, as well as functions of action value. Like the SCi, some cortical signals display serial dependencies over trials (Seo and Lee, 2007).

ROLE OF THE SCi WITHIN THE SACCADE DECISION CIRCUIT

We propose that the SCi is involved in three important aspects of selecting strategic saccades:

1. *integrating* value related inputs and tagging action values to particular saccade vectors;
2. *selecting* a saccade in a process where action value representations are compared;
3. *providing feedback* of choices to dopaminergic centres.

First, as outlined in Section "Neuronal Correlates of Experience Weighted Attraction", the SCi receives inputs from regions that encode functional elements of action value learning models. Because the SCi integrates many inputs, and outputs to pre-motor neurons, its representations of action value may be particularly suited for choosing final actions. Moreover, the topographic organization of the SCi allows value representations to be tagged to particular saccade vectors.

Second, the SCi provides a platform where multiple action value representations can compete and ultimately be resolved to choose a particular action. The topographic map within the SCi is organized based on the principle of local excitation and distant inhibition (Munoz and Istvan, 1998; Trappenberg et al., 2001; Dorris et al., 2007). Once activity reaches a certain threshold level on this map, a saccade command is sent to pre-motor neurons in the brainstem (see Moschovakis and Highstein, 1994 for review). Therefore, the SCi is perhaps the last site within the visuosaccadic circuit where action value can be represented to influence saccade selection without directly triggering (or necessarily resulting in) saccades.

Third, the SCi sends direct mono-synaptic projections to dopaminergic neurons in the substantia nigra and ventral tegmental area (Comoli et al., 2003; Dommett et al., 2005). Therefore, the SCi may provide feedback on selected actions, thus providing a critical component for the reinforcement learning circuitry of the striatum.

CONCLUSION

Our results suggest that the evolutionarily old SCi does not simply execute sensory-driven reflexive saccades but also encodes action value signals that can be used to select voluntary, strategic saccades. As would be expected from a brain region involved in the decision process, SCi activity simultaneously reflects past choices and their outcomes, and predicts future choice. Similarly, learning models, such as EWA, recursively compute action values from past events to probabilistically choose future actions. We demonstrate that these small trial-to-trial fluctuations in SCi activity are not entirely random but have serial dependencies which can be captured, in part, by the EWA learning model.

APPENDIX

The goal of EWA learning is to construct a model that predicts play across a wide variety of games yet retains a framework that is psychologically sound. In an EWA learning model, each strategy has an attraction (which we re-labelled action value) that is updated based on observed choices and the payoff structure.

We introduce EWA in the context of a player who faces a single opponent. Each period the player chooses s from one of two alternatives, $s \in \{in, out\}$. For each trial t, the subject makes a choice s_t, the opponent chooses $s_t' \in \{in, out\}$, and the subject receives a payoff $\pi_t(s_t, s_t')$ as defined in Section "EWA Learning". We drop the experiment subscript i here for illustration.

Once a choice is made and payoff received in trial t, the attraction of strategy s in trial t is defined as a recursive function of past attractions, choices, and rewards by means of

$$A_t^s = \frac{\phi_t N_{t-1} A_{t-1}^s + [\delta + (1-\delta) 1_{[s=s_t]}]\,\pi(s, s_t')}{N_t} \tag{18}$$

where $A_t^s \in A_t$ and $A_1^s = 0.5$ at the beginning of each experiment.

The first component of Eq. 18, $\phi N_t A_{t-1}^s$ is a depreciation of the previous period's action value. The second component, $[\delta + (1-\delta) 1_{[s=s_t]}]\pi(s, s_t')$, is determined by the choice and reward of the current trial. The experience weight N_t is given by

$$N_t = \rho N_{t-1} + 1 \tag{19}$$

with $N_0 = 0$. Assuming $\rho < 1$, $\lim_{t \to \infty} N_t = \frac{1}{1-\rho}$.

Finally, after a choice is made and a reward is determined in trial t, A_t^s is updated to reflect the valuation of every candidate choice in trial t. On a given trial, the probability of choosing $s_{i,t} = in$ is defined as

$$P(s_{i,t} = in) = \frac{e^{\lambda A_{t-1}^{in}}}{e^{\lambda A_{t-1}^{in}} + e^{\lambda A_{t-1}^{out}}}, \tag{20}$$

which yields a likelihood function for our observed choices

$$L = \prod_{i=1}^{I} \prod_{t=1}^{T_i} P(s_{i,t} = in) \tag{21}$$

which is estimated via maximum likelihood using the log-likelihood function

$$LL = \sum_{i=1}^{I} \sum_{t=1}^{T_i} 1_{[s_t = in]} P(s_{i,t} = in) + 1_{[s_t = out]} (1 - P(s_{i,t} = in)). \tag{22}$$

In addition to the identification restrictions detailed in Ho et al. (2008), we had to make an additional identification assumption for monkey H. We found that the restriction $N_0 = \frac{1}{1-\rho}$ was always binding, so we restricted $N_0 = 1$ for this monkey to ensure identification of ρ, although the estimates are robust to $N_0 \leq 1$.

ACKNOWLEDGMENTS

This work was supported by a Career Development Award from the Human Frontier Science Program (HFSP), a Discovery Grant from the National Science and Engineering Research Council (NSERC) of Canada and a group grant from the Canadian Institutes of Health Research (CIHR) awarded to MCD. We thank J. Green, S. Hickman, M. Lewis, F. Paquin and R. Pengelly for technical assistance. J. Turner provided programming expertise and E. Ryklin customized the data acquisition program. D. Byrne and D. Standage provided constructive feedback regarding the manuscript.

REFERENCES

Anderson, J. R., Reder, L. M., and Lebiere, C. (1996). Working memory: activation limitations on retrieval. *Cogn. Psychol.* 30, 221–256.

Barraclough, D. J., Conroy, M. L., and Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nat. Neurosci.* 7, 404–410.

Bayer, H. M., and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47, 129–141.

Brown, J., and Rosenthal, R. (1990). Testing the minimax hypothesis: a re-examination of O'Neill's game experiment. *Econometrica* 58, 1065–1081.

Callicott, J. H., Mattay, V. S., Bertolino, A., Finn, K., Coppola, R., Frank, J. A., Goldberg, T. E., and Weinberger, D. R. (1999). Physiological characteristics of capacity constraints in working memory as revealed by functional MRI. *Cereb. Cortex* 9, 20–26.

Camerer, C., and Ho, T. (1999). Experienced-weighted attraction learning in normal form games. *Econometrica* 67, 827–874.

Caplin, A., Dean, M., Glimcher, P. W., and Rutledge, R. (2010). Measuring beliefs and rewards: a neuroeconomic approach. *Q. J. Econ.* (forthcoming).

Carello, C. D., and Krauzlis, R. J. (2004). Manipulating intent: evidence for a causal role of the superior colliculus in target selection. *Neuron* 43, 575–583.

Comoli, E., Coizet, V., Boyes, J., Bolam, J. P., Canteras, N. S., Quirk, R. H., Overton, P. G., and Redgrave, P. (2003). A direct projection from superior colliculus to substantia nigra for detecting salient visual events. *Nat. Neurosci.* 6, 974–980.

Dommett, E., Coizet, V., Blaha, C. D., Martindale, J., Lefebvre, V., Walton, N., Mayhew, J. E., Overton, P. G., and Redgrave, P. (2005). How visual stimuli activate dopaminergic neurons at short latency. *Science* 307, 1476–1479.

Dorris, M. C., and Glimcher, P. W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* 44, 365–378.

Dorris, M. C., Olivier, E., and Munoz, D. P. (2007). Competetive integration of visual and preparatory signals in the superior colliculus during saccadic programming. *J. Neurosci.* 27, 5053–5062.

Dorris, M. C., Pare, M., and Munoz, D. P. (1997). Neuronal activity in monkey superior colliculus related to the initiation of saccadic eye movements. *J. Neurosci.* 17, 8566–8579.

Dorris, M. C., Pare, M., and Munoz, D. P. (2000). Immediate neural plasticity shapes motor performance. *J. Neurosci.* 20, 1–5.

Erev, I., and Roth, A. (1998). Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* 88, 848–881.

Fudenberg, D., and Levine, D. (1998). The Theory of Learning in Games. Cambridge, MIT Press.

Glimcher, P. W., and Sparks, D. L. (1992). Movement selection in advance of action in the superior colliculus. *Nature* 355, 542–545.

Hanes, D. P., and Wurtz, R. H. (2001). Interaction of the frontal eye field and superior colliculus for saccade generation. *J. Neurophysiol.* 85, 804–815.

Harsanyi, J. (1973). Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points. *Int. J. Game Theory,* 2, 1–23.

Ho, T., Camerer, C., and Chong, J. (2007). Self-tuning experience weighted attraction learning in games. *J. Econ. Theory* 133, 177–198.

Ho, T., Wang, X., and Camerer, C. (2008). Individual differences in EWA learning with partial payoff information. *Econ. J.* 118, 37–59.

Jocham, G., Neumann, J., Klein, T. A., Danielmeier, C., and Ullsperger, M. (2009). Adaptive coding of action values in the human rostral cingulate zone. *J. Neurosci.* 29, 7489–7496.

Juttner, M., and Wolf, W. (1992). Occurrence of human express saccades depends on stimulus uncertainty and stimulus sequence. *Exp. Brain Res.* 89, 678–681.

Kennerley, S. W., Walton, M. E., Behrens, T. E. J., Buckley, M. J., and Rushworth M. F. S. (2006). Optimal decision making and the anterior cingulate cortex. *Nat. Neurosci.* 9, 940–947.

Lau, B., and Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* 84, 555–579.

Lau, B., and Glimcher, P. W. (2007). Action and outcome encoding in the

primate caudate nucleus. *J. Neurosci.* 27, 14502–14514.
Lau, B., and Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron* 58, 451–463.
Lee, D., Conroy, M. L., McGreevy, B. P., and Barraclough, D. J. (2004). Reinforcement learning and decision making in monkeys during a competitive game. *Cogn. Brain Res.* 22, 45–58.
Lohrenz, T., McCabe, K., Camerer, C. F., and Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proc. Natl. Acad. Sci. U.S.A.* 104, 9493–9498.
Luk, C. H., and Wallis, J. D. (2009). Dynamic encoding of responses and outcomes by neurons in medial prefrontal cortex. *J. Neurosci.* 29, 7526–7539.
Maljkovic, V., and Nakayama, K. (1994). Priming of pop-out.1. Role of features. *Mem. Cognit.* 22, 657–672.
McCabe, K., Mukherji, A., and Runkle, D. (2000). An experimental study of information and mixed-strategy play in the three-person matching-pennies .... *Econ. Theory*, 15, 421–462.
McPeek, R. M., and Keller, E. L. (2004). Deficits in saccade target selection after inactivation of superior colliculus. *Nat. Neurosci.* 7, 757–763.
Mookherjee, D., and Sopher, B. (1994). Learning behavior in an experimental matching pennies game. *Games Econ. Behav.* 7, 62–91.
Mookherjee, D., and Sopher, B. (1997). Learning and decision costs in experimental constant sum games. *Games Econ. Behav*. 19, 97–132.
Moschovakis, A. K., and Highstein, S. M. (1994). The anatomy and physiology of primate neurons that control rapid eye movements. *Annu. Rev. Neurosci.* 17, 465–488.
Munoz, D. P., and Istvan, P. J. (1998). Lateral inhibitory interactions in the intermediate layers of the monkey superior colliculus. *J. Neurophysiol.* 79, 1193–1209.
Nash, J. (1951). Non-cooperative games. *Ann. Math. Second Series* 54, 286–295.
Ochs, J. (1995). Games with unique, mixed strategy equilibria: an experimental study. *Games Econ. Behav.* 10, 202–217.
O'Neill, B. (1987). Nonmetric test of the minimax theory of two-person zerosum games. *Proc. Natl. Acad. Sci. U.S.A.* 84, 2106–2109.
Padoa-Schioppa, C., and Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature* 441, 223–226.
Platt, M. L., and Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature* 400, 233–238.
Rapoport, A., and Boebel, R. (1992). Mixed strategies in strictly competitive games: a further test of the minimax hypothesis. *Games Econ. Behav.* 4, 261–283.
Rapoport, A., and Budescu, D. (1992). Generation of random series in two-person strictly competitive games. *J. Exp. Psychol. Gen.* 121, 352–363.
Robinson, D. A. (1972). Eye-movements evoked by collicular stimulation in alert monkey. *Vision Res.* 12, 1795–1808.
Rushworth, M. F., Walton, M. E., Kennerley, S. W., and Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends Cogn. Sci.* 8, 410–417.
Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science* 310, 1337–1340.
Schiller, P. H., and Stryker, M. (1972). Single-unit recording and stimulation in superior colliculus of alert rhesus-monkey. *J. Neurophysiol.* 35, 915–924.
Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27.
Schultz, W. (2004). Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. *Curr. Opin. Neurobiol.* 14, 139–147.
Seo, H., Barraclough, D. J., and Lee, D. (2007). Dynamic signals related to choices and outcomes in the dorslateral prefrontal cortex. *Cereb. Cortex* 17, I110–I117.
Seo, H., Barraclough, D. J., and Lee, D. (2009). Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. *J. Neurosci.* 29, 7278–7289.
Seo, H., and Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *J. Neurosci.* 27, 8366–8377.
Seo, H., and Lee, D. (2008). Cortical mechanisms for reinforcement learning in competitive games. *Philos. Trans. R. Soc.* 363, 3845–3857.
Spiliopoulos, L. (2008). Humans versus computer algorithms in repeated mixed strategy games. *MPRA Paper No.* 6672, 68.
Sugrue, L. P., Corrado, G. S., and Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science* 304, 1782–1787.
Sutton, R., and Barto, A. G. (1998). Reinforcment Learning: An Introduction. Cambridge, MIT Press.
Thevarajah, D., Mikulic, A., and Dorris, M. (2009). Role of the superior colliculus in choosing mixed-strategy saccades. *J. Neurosci.* 89, 1998–2008.
Thompson, K. G., Hanes, D. P., Bichot, N. P., and Schall, J. D. (1996). Perceptual and motor processing stages identified in the activity of macaque frontal eye field neurons during visual search. *J. Neurophysiol.* 76, 4040–4055.
Thorndike, E. L. (1898). Animal intelligence: An Experimental Study of the Associative Processes in Animals. New York, Macmillan.
Trappenberg, T., Dorris, M., and Munoz, D. (2001). A model of saccade initiation based on the competitive integration of exogenous and .... *J. Cogn. Neurosci.* 13, 256–271.
von Neumann, J., and Morgenstern, O. (1947). Theory of Games and Economic Behaviour. Princeton University Press, Princiton.
Watkins, C. (1989). Learning from Delayed Rewards. Ph.D. thesis, Cambridge University, Cambridge.
Wooldridge, J. M. (2001). Econometric Analysis of Cross Section and Panel Data. Cambridge, MIT Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 July 2009; paper pending published: 04 August 2009; accepted: 01 December 2009; published online: 08 February 2010.
*Citation: Thevarajah D, Webb R, Ferrall C and Dorris MC (2010) Modeling the value of strategic actions in the superior colliculus. Front. Behav. Neurosci. **3**:57. doi: 10.3389/neuro.08.057.2009*