Check for updates

Neural Representation of Reward Probability: Evidence from the Illusion of Control

Wouter Kool, Sarah J. Getz, and Matthew M. Botvinick

Downloaded from http://direct.mit.edu/jocn/article-pdf/25/6/852/1937579/jocn_a_00369.pdf by guest on 17 May 2021

Abstract

■ To support reward-based decision-making, the brain must encode potential outcomes both in terms of their incentive value and their probability of occurrence. Recent research has made it clear that the brain bears multiple representations of reward magnitude, meaning that a single choice option may be represented differently—and even inconsistently—in different brain areas. There are some hints that the same may be true for reward probability. Preliminary evidence hints that, even as systematic distortions of probability are expressed in behavior, these may not always be uniformly reflected at the neural level: Some neural representations of probability may be immune from such distortions. This study provides new evidence consistent with this possibility. Participants in a behavioral experiment displayed a classic "illusion of control," providing higher estimates of reward probability for gambles they had chosen than for identical gambles that were imposed on them. However, an fMRI study of the same task revealed that neural prediction error signals, arising when gamble outcomes were revealed, were unaffected by the illusion of control. The resulting behavioral–neural dissociation reinforces the case for multiple, inconsistent internal representations of reward probability, while also prompting a reinterpretation of the illusion of control effect itself. ■

INTRODUCTION

To make principled choices between available lines of action, a decision-maker must evaluate potential outcomes in terms of both value and probability. These two attributes stand as the key ingredients in classical theories of economic choice (Bernoulli, 1954; von Neumann & Morgenstern, 1944), providing the necessary and sufficient materials for computing expected utility. More recent behavioral economic theories have introduced the important idea that subjective representations of outcome value and probability may both be subject to distortions, that is, nonlinear transformations. Even here, however, value and probability maintain their role as the two key pillars supporting decision-making (Hertwig, Barron, Weber, & Erev, 2004; Starmer, 2000; Tversky & Kahneman, 1992; Kahneman, Knetsch, & Thaler, 1990; Kahneman & Tversky, 1979; Allais, 1953).

There has been a major effort over recent years to characterize representations of reward magnitude and probability in the brain, using neurophysiological, neuropsychological, and neuroimaging techniques (Kobayashi, Lauwereyns, Koizumi, Sakagami, & Hikosaka, 2002; Platt & Glimcher, 1999). One important result of such work has been to corroborate the nonlinear transformations in value and probability posited by behavioral economic theories (FitzGerald, Seymour, Bach, & Dolan, 2010; Berns, Capra, Chappelow, Moore, & Noussair, 2008; Paulus & Frank, 2006; Trepel, Fox, & Poldrack, 2005). Another insight, not fully anticipated by economics, is that the brain encodes multiple representations of reward magnitude and probability, which differ in format and functional role. The case for such representational variety is quite strong in the case of reward magnitude, where numerous studies have documented encodings that arise concurrently but differ according to pertinent outcome (Rangel & Hare, 2010; Padoa-Schioppa & Assad, 2006), frame of reference (De Martino, Kumaran, Holt, & Dolan, 2009), current plan of action (Roesch, Singh, Brown, Mullins, & Schoenbaum, 2009), or flexibility in the face of change (Simon & Daw, 2011). Particularly striking are cases where concurrent reward representations are inconsistent with one another (Hutcherson, Plassmann, Gross, & Rangel, 2012; McClure, Ericson, Laibson, Loewenstein, & Cohen, 2007), suggesting that choice options are being encoded in parallel by dissociable decision-making or learning mechanisms (Daw, Niv, & Dayan, 2005).

There is scattered evidence for the possibility that reward probability may also map onto multiple, potentially divergent internal representations. An fMRI study by Tobler, Christopoulos, O'Doherty, Dolan, and Schultz (2008) observed that encodings of reward probability within two regions of pFC were subject to different nonlinear transformations. Intriguingly, the same study found that probability representation in the striatum was essentially linear, leading the authors to propose "a neuronal dissociation between veridical and distorted probability processing in the striatum and prefrontal cortex, respectively"

Princeton University

© 2013 Massachusetts Institute of Technology

Journal of Cognitive Neuroscience 25:6, pp. 852–861
doi:10.1162/jocn_a_00369

(p. 11704). Convergent evidence for such a cortical–subcortical dissociation can be gleaned from other studies. In particular, whereas numerous studies have reported nonlinear representations of outcome probability in cortical areas (FitzGerald et al., 2010; Berns et al., 2008; Paulus & Frank, 2006), Abler and colleagues (Abler, Walter, Erk, Kammerer, & Spitzer, 2006) found that reward prediction error signals in ventral striatum scaled linearly with outcome probability (see also Hsu, Krajbich, Zhao, & Camerer, 2009).

To shed further light on neural encodings of outcome probability, we turned to a behavioral effect known as the "illusion of control" (IOC). First reported by Langer (1975), the IOC is a tendency to overestimate the probability of favorable outcomes in chance situations, when those situations are chosen rather than imposed.[1] In Langer's classic study, participants who chose one from several lottery tickets were later less willing to trade their tickets for others, compared with participants who had simply been given a ticket without choice, although the probability of winning the lottery was the same for both groups. Subsequent research has documented that this behavioral bias reflects, at least in part, a distortion of outcome probability estimates, with chosen gambles being accorded a higher probability of yielding a favorable outcome (for reviews, see Thompson, Armstrong, & Thomas, 1998; Presson & Benassi, 1996).

To our knowledge, the neural correlates of the IOC have not been investigated. However, the "illusion," like other probability distortion effects, offers an opportunity to probe the neural processes that underlie subjective judgments of reward probability. To pursue this, we measured brain activity with fMRI as participants performed an experiment that elicited the IOC.

Our primary focus in analyzing the resulting data was on the moment that reward outcomes were revealed. A large body of research has shown that reward outcomes trigger a reward prediction error (RPE) signal, originating in midbrain dopaminergic nuclei, but detectable in an array of cortical and subcortical structures, most notably the ventral striatum (Schultz, Dayan, & Montague, 1997; Montague, Dayan, & Sejnowski, 1996). The RPE encodes the difference between observed and expected reward (Sutton & Barto, 1998; Rescorla & Wagner, 1972) and thus varies with outcome probability: The less likely a positive outcome is judged to be, the larger the RPE when it occurs (Fiorillo, Tobler, & Schultz, 2003).

Because neural RPE signals reflect prospective estimates of reward probability, it seems plausible that they should be impacted by the IOC. In particular, because the IOC inflates estimates of reward probability, it might be expected to reduce RPE magnitude in the case of positive outcomes, because it effectively makes them less surprising. By the same token, the IOC might be expected to increase the (negative) amplitude of the RPE triggered by nonreward outcomes by making them more surprising.

To test this prediction, we conducted two experiments, one behavioral and one using functional neuroimaging. The behavioral study was aimed at confirming that our chosen behavioral task yields the classic IOC effect, triggering inflated estimates of reward probability. In the neuroimaging experiment, participants performed the same task while undergoing fMRI. Our prediction was that neural RPE signals, for example, within the ventral striatum, would reflect the IOC, being reduced for outcomes in gambles that had been chosen compared with gambles that were imposed.

As detailed in what follows, the data flatly contradicted this prediction. At the level of behavior, participants showed a clear IOC, evincing higher estimates of reward probability following choice. However, striatal RPE signals appeared to be entirely immune to this effect. As we shall argue, this surprising finding adds to the emerging evidence that subcortical probability representations can be impervious to behaviorally expressed distortions, while also inviting a reinterpretation of the IOC effect itself.

EXPERIMENT 1

In this initial experiment, we sought to replicate the behavioral IOC effect in a novel, multitrial task paradigm.

Methods

Participants and Procedures

Fourteen participants from the Princeton University community (aged 18–23 years, 5 women) took part in Experiment 1. All participants received a nominal payment and provided informed consent, following procedures approved by the Princeton University Institutional Review Board.

The task was computer based and programmed using the Psychophysics Toolbox for Matlab (Brainard, 1997; Pelli, 1997). On each of the 96 trials, three spinner dials were presented, depicted as wheels of fortune or spinners: white dials containing a green "win sector" and a black arrow (Figure 1A). They were positioned at the corners of a (hidden) randomly rotated equilateral triangle. The size of the win sector varied from trial to trial but was identical across the three spinners shown together on each individual trial. They were drawn from one of four uniform ranges of width 0.10 around four means (0.2, 0.4, 0.6, and 0.8).

Participants were specifically and truthfully informed that the three spinners in each trial had an equal probability of a win outcome. Despite this, participants were required to select one of the three available options using a mouse. To avoid any bias to a given spinner based on proximity, the mouse cursor was set in the center of the screen at the start of the trial. A blue circle appeared around the selected spinner, indicating the participant's preference.

In the second part of the trial, the approval phase, on an arbitrary one half of trials the participant's choice was

Downloaded from http://direct.mit.edu/jocn/article-pdf/29/5/852/1924357/jocn_a_01090.pdf by guest on 17 May 2021

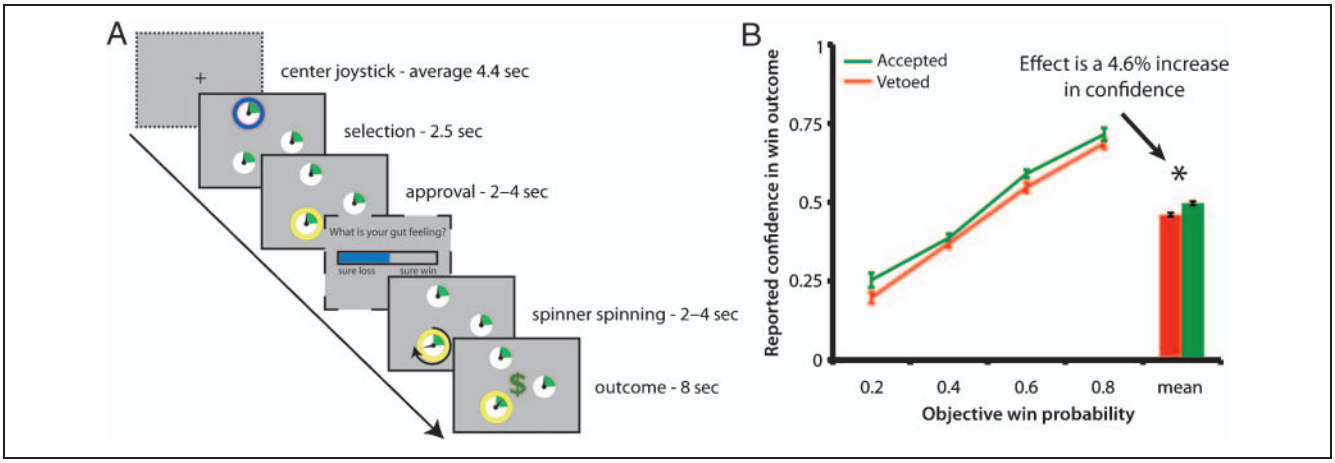

Figure 1. (A) Sequence of events in the experiments. In Experiment 2 (panel with dotted border), each trial started with participants centering the joystick. Participants' selected one of three "spinners" with the mouse (Experiment 1) or the joystick (Experiment 2), and this original decision was marked with a blue circle. When the initial choice was vetoed (as depicted), a yellow circle appeared around another spinner, otherwise a yellow circle would replace the blue circle around the original choice. In the behavioral experiment, participants then reported their confidence using a visual analogue scale (dashed border). Next, the ultimately selected spinner would start revolving with decelerating speed, resulting either in a $1 prize (as depicted) or a 90¢ loss. (B) The reported confidence ratings in a win outcome (n = 14) are plotted against the objective win sector size, separately for trials on which original spinner choice was vetoed and accepted. Participants were more confident in a win outcome when their original spinner choice was accepted. Error bars on the green and red lines indicate *SEM*. Error bars on the columns indicate within-subject *SEM*. *$p < .05$.

"vetoed"; the blue circle was then replaced by a yellow circle around one of the other two spinners. On the other half of the trials, the initial choice was accepted and remained highlighted; the blue circle around the selected spinner then turned yellow.

Participants then indicated their level of confidence that the trial would result in a win outcome. For this purpose, a horizontal rectangle and the question, "What is your gut feeling?" appeared on screen. The rectangle functioned as a visual analogue scale to indicate confidence in the likelihood of a win outcome, ranging from "sure loss" to "sure win." Participants were instructed to click the horizontal point in the rectangle that best reflected their gut feeling.

Finally, all spinners were again presented on screen and the arrow of the selected spinner revolved with decelerating speed. If it ended on the win sector, a green "$" appeared on screen and the running total would increase by $1. Otherwise, a red "X" appeared on screen and the running total decreased by 90¢. A new set of spinners was then presented, beginning a new trial. At the end of the experiment, participants were paid their accumulated final money reward.

Unannounced to participants, the trials comprised 24 instances of each possible combinations of spinner outcome and approval condition (i.e., win/vetoed, lose/vetoed, win/accepted, and lose/accepted). To collect enough data for low probability outcomes, the proportion of wins and losses reflected a compressed probability space of those depicted by the size of the win sectors (see Table 1). The trial sequence randomly intermixed these trial types.[2] Participants were offered a 1-min break after every 16 trials.

Data Analysis

We subjected the participants' confidence ratings to a mixed-effects two-way repeated-measures ANOVA with within-subject factors for choice condition and win sector

Table 1. Frequency of Different Trial Types in Both Experiments

Outcome	*Approval Condition*	*Win Sector Size*	*Frequency*
Win	Vetoed	0.2	3
		0.4	5
		0.6	7
		0.8	9
	Accepted	0.2	3
		0.4	5
		0.6	7
		0.8	9
Lose	Vetoed	0.2	9
		0.4	7
		0.6	5
		0.8	3
	Accepted	0.2	9
		0.4	7
		0.6	5
		0.8	3

Downloaded from http://direct.mit.edu/jocn/article-pdf/25/6/852/1935791/jocn_a_00363.pdf by guest on 17 May 2021

size and the participants as the random factor. Statistical significance was evaluated at $\alpha = .05$. One participant's confidence ratings were heavily nonmonotonic over win sector sizes and were excluded from the analyses. All effects remained significant with this participant included.

Results and Discussion

In this behavioral study, we aimed to confirm that our task gives rise to the standard IOC effect. Following the literature, we predicted that participants would provide higher confidence ratings on trials where the initial spinner choice was accepted than on trials where it was vetoed. Not surprisingly, we found that the confidence ratings were sensitive to win sector size, $F(3, 39) = 163.08$, $p < .001$, $\eta^2 = 0.93$; as shown in Figure 1B, the ratings increased monotonically with win sector size (Bonferonni-corrected pairwise comparisons, all ps $< .001$). More importantly, there was a significant main effect of Approval Phase Outcome. Confidence ratings were significantly higher when participants' original choice was accepted than when it was vetoed, $F(1, 13) = 7.32$, $p < .05$, $\eta^2 = 0.36$, Cohen's $d = 0.72$.

To quantify the magnitude of the IOC effect in subjective probability, we proceeded as follows. For each participant, we ran a linear regression analysis to estimate the slope of the linear relationship between the size of the win sector and the reported confidence in a win outcome for trials in which the player's choice was accepted. Next, we multiplied the multiplicative inverse of the average slope, that is, the increase in win sector size for a one-unit increase in confidence, with the average difference in reported confidence between the choice conditions. The size of this effect was 4.6% ($SE = 1.7$), equivalent to a 16.6° increase in win sector size.[3]

EXPERIMENT 2

Having confirmed that our spinner task yields the IOC, we used fMRI to measure regional brain activity during performance of the task.

Methods

Participants and Procedures

Twenty-nine new participants from the Princeton University community (18–32, 16 women) completed Experiment 2.

The task used in Experiment 2 followed the sequence used in Experiment 1, except that the confidence rating step was omitted and that participants used a nonferromagnetic joystick to indicate initial choice. At the beginning of each trial, participants were required to bring the joystick in its center position to avoid any bias to a given spinner based on proximity.

The timing of task events was changed from Experiment 1 to facilitate estimation of the BOLD signal. The blue circle indicating the initial choice remained on screen 2.5 sec. The yellow circle indicating final choice was presented for 2–4 sec. The arrowhead spun for 2–4 sec; the final outcome remained on screen 2 sec; and there was a 8-sec blank screen between trials.

We used the same number of trials as in Experiment 1, randomly ordered over six scanning runs with equal numbers of trials, which yielded an average of 1139 functional volumes per participant.

Before scanning, participants completed a practice set of 16 trials. At the end of the scanning session, participants completed a questionnaire in which they were asked whether they believed the spinners were equivalent in terms of their likelihood of a win outcome. Two participants indicated that they came to believe there was a difference between the spinners in each display and were excluded from further analysis. All effects reported in the main text remained statistically significant when analyses were repeated with these participants included.

fMRI Acquisition and Preprocessing

Scanning was conducted with a 3T Siemens Allegra scanner at Princeton University. The data were analyzed using AFNI (Cox, 1996) and Matlab. Each session began with a MPRAGE anatomical scan, consisting of 160 1-mm sagittal slices (repetition time = 2.5, echo time = 4.38 msec, flip angle = 8°, field of view = 256 mm). During each functional imaging block, an EPI sequence was used to obtain 34 contiguous 3-mm axial slices aligned to the AC–PC line with repetition time = 2 sec, echo time = 30 msec, flip angle = 90°, matrix = 64 × 64 voxels and field of view = 192 mm, yielding 3-mm isotropic voxels.

Slice acquisition time correction was performed using Fourier interpolation, and images were motion corrected using a six-parameter rigid body transformation to coregister functional scans. Volumes that exhibited a large change in motion parameters or a spike in spatially averaged global signal were excluded from further analysis. The data were spatially smoothed until a 3-D 6-mm FWHM Gaussian kernel approximated the estimated spatial autocorrelation. Finally, we normalized the signal in each voxel to reflect percent change.

Data Analysis

For each participant separately, we analyzed each voxel's time course using a general linear model (GLM), with baseline regressors for zero- through third-order polynomial trends, motion parameters, and the averaged whole-brain signal time course. We used six covariates to model the following task events: (i) fixation point onset, (ii) time until centering of joystick, (iii) spinner presentation,

Downloaded from http://direct.mit.edu/jocn/article-pdf/29/5/852/1923579/jocn_a_01088.pdf by guest on 17 May 2021

(iv) approval phase onset, (v) spinner rotation, and (vi) the moment of spin outcome. The mean-centered size of the win sector was entered into the GLM as parametric regressor at events (iii) and (vi). We modeled the choice condition of each trial with a mean-centered categorical regressor at events (iv) and (vi). Finally, we entered the specific outcome of the trial (win vs. lose) as a mean-centered categorical regressor at event (vi). Each of these covariates was convolved with a canonical hemodynamic response function.

For each participant, the analysis yielded maps of parameter estimates (beta values) for each aforementioned regressor. We spatially normalized these maps by warping the participant's anatomical image to match a template in Talairach space (Talairach & Tournoux, 1988), using a 12-parameter affine and cosine transformation. Next, we applied this transformation to each participant's statistical maps. After spatial normalization, we tested the maps of these parameter estimates in group level t tests. The AFNI tool AlphaSim was used to determine a combination of cluster size and p threshold that controlled whole-brain α to .05.

In one additional GLM, we modeled signed and unsigned RPEs at the moment of spinner outcome, replacing the original win vs. lose and win sector size regressors. The results of this GLM were used to test whether the areas responsive to the approval manipulation were reflective of unsigned RPEs. In a further GLM, we included four regressors to separately model the brain's response in each of the four conditions: win/accepted, win/vetoed, lose/accepted, and lose/vetoed.

Note that the experiment and analyses were designed to avoid collinearity between regressors. All regressors of interest were orthogonal at the outcome event and the results of our ROI analyses should therefore not be affected by artifacts of "double dipping" (Vul & Kanwisher, 2010).

Results and Discussion

The focus of analysis was on neural activity linked to spin outcomes. As explained in the Introduction, we predicted that areas including the ventral striatum would reflect an RPE signal at trial outcome and, critically, that this signal would be modulated by IOC. More specifically, we predicted that areas carrying an RPE signal would also show a main effect of Approval Phase Outcome, displaying lower activity at spin outcome following choice acceptance than following veto events.

To evaluate events at spin outcome, we first contrasted regional brain activity (BOLD signal) for win outcomes against activity for lose outcomes. In keeping with previous studies (Schultz et al., 1997; Montague et al., 1996), we found strong bilateral activation in the BG, with peak activation in ventral striatum (Figure 2A; peaks $x_{left} = 19$, $y_{left} = -5$, $z_{left} = -1$ and $x_{right} = -17$, $y_{right} = -5$, $z_{right} = -1$). For each participant individually, we used this win–lose contrast to construct an ROI in this area (thresholds at $t = 2.5$). Within the resulting ROI, activity at spin outcome varied inversely with win sector size, $t(26) = -6.03$, $p < .001$, as would be expected of an RPE signal (Figure 2B). The ROI also showed phasic activity positively correlating with win sector size at the moment each set of spinners was first presented, $t(26) = 2.07$, $p < .05$, again consistent with an interpretation in terms of RPE signaling (Schultz et al., 1997).

To our surprise, although ventral striatum encoded RPEs at two junctures in our task, signal in the same area at spin outcome displayed no sensitivity to the approval phase outcome, $t(26) = -0.06$, right-sided $p = .52$; VS activity at spin outcome was no greater on trials where

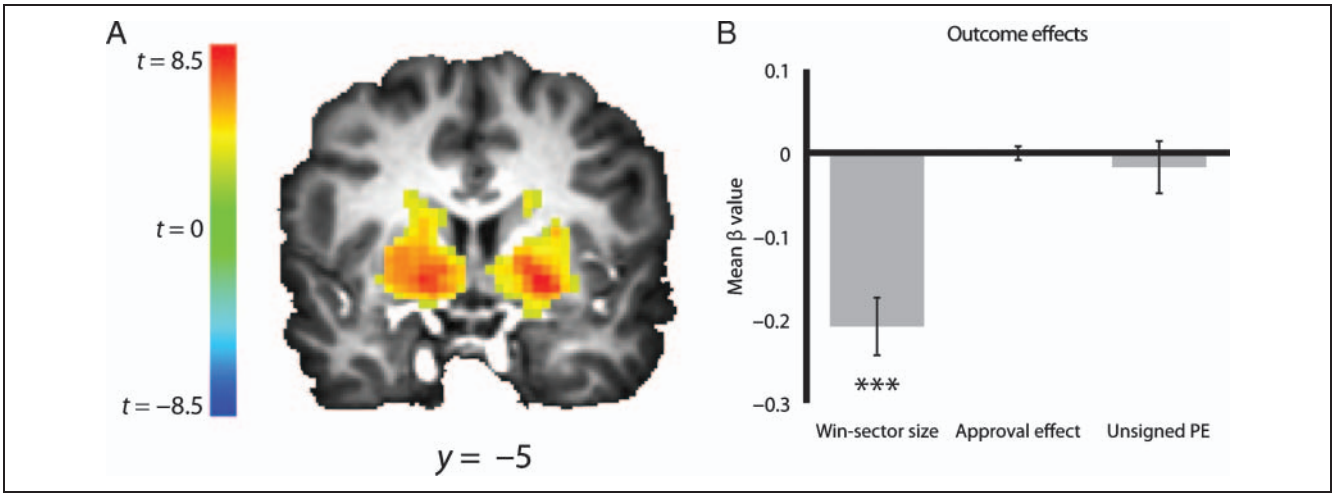

Figure 2. (A) Group analysis results of the Experiment 2. Activity in the BG, with peaks in ventral striatum, showed increased activity in response to a win outcome when compared with a lose outcome. We constructed our subject-wise ventral striatum ROIs based on this contrast. (B) Average parameter of the regressor coding for the effect of the win sector size, the approval condition, and unsigned PE for this ROI in ventral striatum. Our results indicate that striatum only encoded the size of the win sector at spinner outcome.

Downloaded from http://direct.mit.edu/jocn/article-pdf/25/6/852/1935779/jocn_a_00363.pdf by guest on 17 May 2021

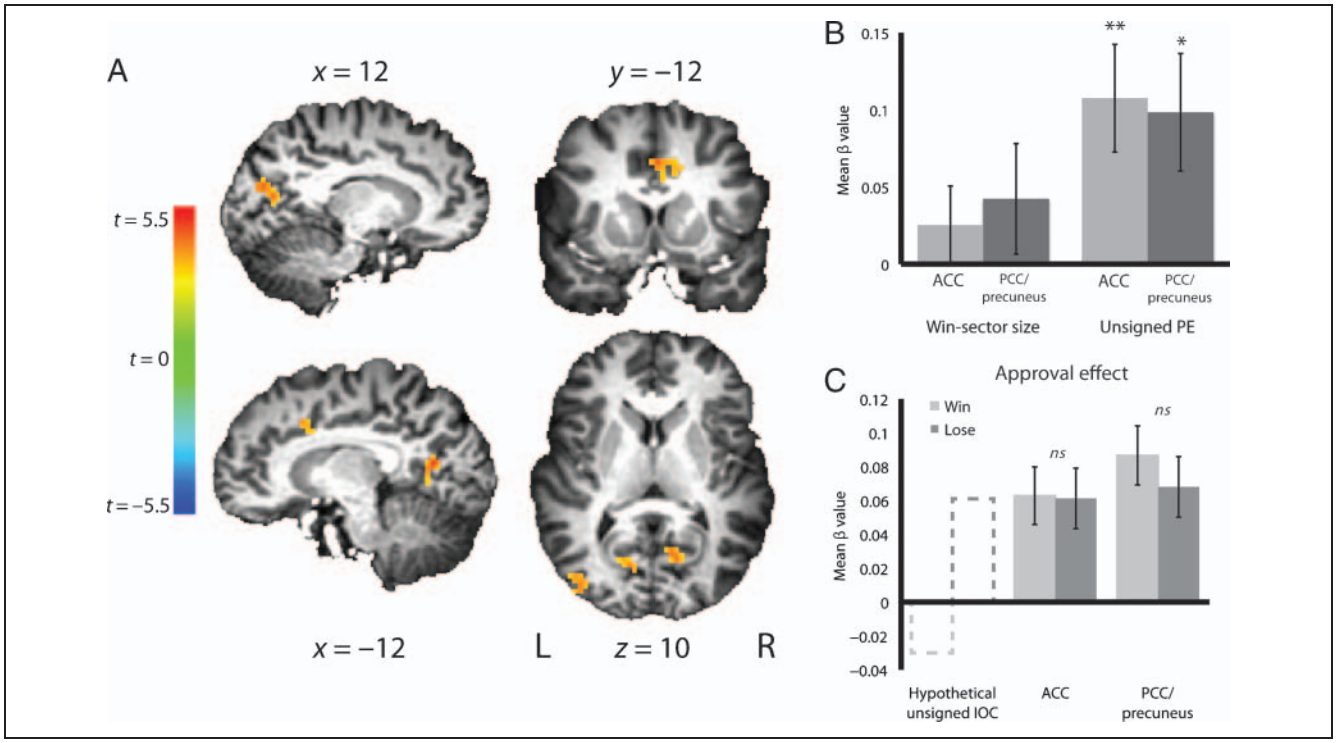

Figure 3. (A) Group analysis results of Experiment 2. Regions responsive to the difference between the outcomes of accepted spinners compared with vetoed spinners included ACC and PCC/precuneus. There were only regions that showed increased activation on accepted trials compared with vetoed trials. Whole-brain threshold at $p < .005$. (B) Average parameter of the regressor coding for unsigned RPEs for ROIs in ACC and PCC/precuneus. Our results indicate that ACC and PCC/precuneus encoded unsigned RPEs or salience signals. (C) Average parameter of the regressor coding for the approval condition (vetoed vs. accepted), separately for trials on which the spinner resulted in a win and a lose outcome for ROIs in ACC and PCC/precuneus. If these regions would have coded an unsigned version of the IOC, this regressor should have differed in sign, that is, in the amount of surprise, between trials involving win and lose outcomes (as depicted). We found, however, that the approval effect was positive for both types of trials and was not significantly different between them ($ps > .40$) in both regions. Error bars indicate *SEM*. $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$.

participants' spinner choice had been vetoed than on trials where it had been accepted. Indeed, there was not even a numerical trend in this direction (average $\beta = -0.0004$, 95% confidence interval $= -0.0173 - 0.0164$; see Figure 2B).

To interpret this null result, it is important to consider whether the experimental design carried sufficient power to detect an approval outcome effect had such an effect been present. As it turns out, the same data provided the opportunity to conduct just such a power analysis. Recall that our behavioral study had indicated that the mean impact of the IOC on outcome prediction was comparable to increasing win sector size by 4.6%. The fact that the ventral striatum ROI displayed a robust effect of win sector size allowed us to estimate the effect of such a win sector expansion on striatal activity.

To do so, we ran new GLMs (100 iterations), which modeled the likelihood of a win outcome separately for two equally sized groups of trials. The trials in these groups were randomly selected with the only constraint that the difference between their average win sector sizes approximated 4.6%, the subjective addition to the win sector size in the case of an accepted spinner as calculated in Experiment 1. For each of these GLMs, we entered an additional regressor, contrasting the two sets, to calculate the effect size of a difference in the size of RPEs induced by a 4.6% difference in win sector size in the striatum ROIs. A post hoc power analysis based on the effect sizes in this set of 100 GLMs indicated a mean Cohen's $d = 0.26$. The aforementioned p value associated with the approval effect revealed that we had considerable power, $1 - \beta = 0.93$,[4] making it unlikely that the absence of an approval phase outcome effect reflected a type II error.

An exploratory whole-brain group analysis, again focused on spin outcome, corroborated this result, revealing no effect of approval phase outcome anywhere in the BG (Figure 3A). The contrast did reveal an effect in several other areas, specifically in left and right posterior cingulate cortex/precuneus (PCC/precuneus; 49 and 29 voxels, peaks $x_{left} = 16$, $y_{left} = 73$, $z_{left} = 20$ and $x_{right} = -14$, $y_{right} = 61$, $z_{right} = 14$), ACC (25 voxels, peak $x = -2$, $y = -11$, $z = 38$), and middle temporal gyrus (34 voxels, peak $x = 40$, $y = 76$, $z = 8$). In all of these areas, activity at spin outcome was greater on trials where the participant's spinner choice had been accepted than when it had been vetoed. Importantly, the same whole-brain analysis and follow-up ROI analyses indicated that none of these

Downloaded from http://direct.mit.edu/jocn/article-pdf/29/5/857/1924357/jocn_a_01088.pdf by guest on 17 May 2021

regions responded monotonically to win sector size, as would be expected from a region carrying an RPE signal (ps $> .25$; Figure 3B). Instead, activity in both ACC and PCC/precuneus (though not temporal gyrus) resembled an "unsigned" RPE (Hayden, Heilbronner, Pearson, & Platt, 2011) or saliency signal (Litt, Plassmann, Shiv, & Rangel, 2011): Activity correlated positively with win sector size on lose outcomes but correlated negatively on win outcomes ($t(26) = 3.08$, $p < .01$ for ACC; $t(26) = 2.58$, $p < .05$ for PCC/precuneus; Figure 3B). Note that this latter result makes it difficult to interpret the approval phase outcome effect in these regions as reflecting a distortion of reward probability. Such an interpretation would require the approval effect, like the win sector effect, to differ in sign between trials involving win and lose outcomes (Figure 3C): Under the IOC, lose outcomes should be more salient than win outcomes. However, the approval effect in each region was positive for both win and lose trials and did not differ significantly between the two (ps $> .40$; see Figure 3C).

Interestingly, although ventral striatum showed no approval effect at spin outcome, it did show such an effect at the moment of choice approval itself. As shown in Figure 4, VS along with ventromedial pFC (vmPFC), thalamus, left caudate, PCC, and middle temporal gyrus displayed greater activity during the approval phase following choice acceptance than after veto (vmPFC: 415 voxels, peak $x = 1$, $y = -38$, $z = -1$; left ventral striatum: 83 voxels, peak $x = 22$, $y = 1$, $z = -4$; bilateral thalamus: 42 voxels, peak $x = 1$, $y = 13$, $z = 8$; left caudate: 32 voxels, $x = 28$, $y = -2$, $z = 20$; PCC: 26 voxels, $x = 10$, $y = 37$, $z = 35$; middle temporal gyrus: 65 voxels, peaks $x_{left} = 49$, $y_{left} = 73$, $z_{left} = 29$ and $x_{right} = -41$, $y_{right} = 64$, $z_{right} = 29$). In addition, we found a broad cluster of activation in occipital cortex and FEFs that displayed lower activity following choice acceptance than after veto (occipital cortex: 4279 voxels, peak $x = -26$, $y = 82$, $z = 23$; left FEF: 70 voxels, peak $x = 37$, $y = 13$, $z = 47$; right FEF: 108 voxels, peak $x = -23$, $y = 7$, $z = 47$).

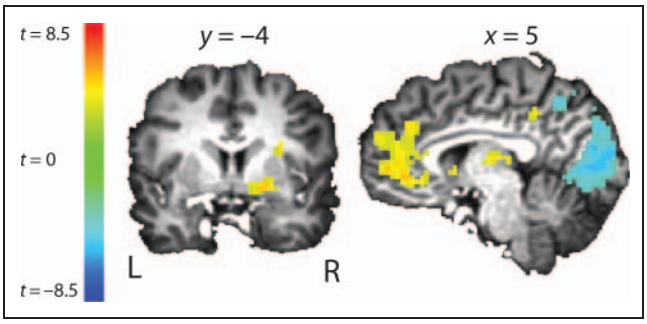

Figure 4. Group analysis results of Experiment 2. The figure depicts the regions that showed differential response between vetoed and accepted trials during the approval phase. We found clusters in occipital cortex, parietal cortex, and FEFs that were significantly increased in response to vetoed spinners. Clusters in vmPFC, right VS, bilateral thalamus, and left caudate showed increased activation to the acceptance of spinner choice compared with a veto event. Whole-brain threshold at $p < .005$.

A final set of analyses tested whether the IOC manipulation might have affected activity in regions previously implicated in risk representation, including anterior insula (Preuschoff, Quartz, & Bossaerts, 2008) and posterior parietal cortex (Huettel, Stowe, Gordon, Warner, & Platt, 2006). Drawing ROIs from relevant studies, we tested for an effect of the choice manipulation both during the approval phase and at spin outcome. No clear effect was observed. Regions in anterior insula described by Preuschoff et al. (2008) were not responsive to the IOC manipulation at either juncture, ps $> .24$. A region in posterior parietal cortex reported by Huettel et al. (2006) showed a marginal IOC effect during the approval phase ($p = .06$), but not during the spin outcome ($p = .26$).

GENERAL DISCUSSION

The question of how the brain represents reward outcome probabilities is central to research on decision-making. Previous studies have revealed reward probability representations in several brain regions, often reflecting nonlinearities or distortions that have been inferred from behavioral studies. A surprising prospect, arising from such work, is that the brain may carry multiple representations of outcome probability that differ in form; probability representations in different brain structures may reflect different distortions or may be immune to the distortions reflected in choice behavior (De Martino et al., 2009).

In this study, we investigated the neural representation of outcome probability in the setting of the IOC (Langer, 1975), a probability distortion effect whose neural correlates have not been studied. In a behavioral experiment, we demonstrated that the classic IOC effect emerged in a novel gambling task: Participants assigned greater probability to positive outcomes for gambles that were chosen than for equivalent gambles that were imposed. Taking a version of the same task to fMRI, we tested a straightforward prediction. Under the IOC, positive outcomes should be less surprising (and negative outcomes more surprising) in chosen gambles. Neural RPEs, as routinely observed for example in the ventral striatum, should therefore presumably be affected by IOC. In our experiment, the specific prediction was for a main effect of choice (approval phase outcome) on RPE magnitude when gamble outcomes were presented.

The neuroimaging results unambiguously contradicted this prediction. Although a strong RPE signal was present in the ventral striatum, it was entirely unaffected by the choice manipulation. Additional analyses indicated that our experiment carried considerable power for detecting an IOC effect, making it unlikely that the absence of an effect reflected a type II error. In short, the results indicate that RPE signals in the ventral striatum are immune to the IOC.

Downloaded from http://direct.mit.edu/jocn/article-pdf/25/6/858/1924357/jocn_a_00363.pdf by guest on 17 May 2021

Our findings add to previous evidence suggesting that veridical neural representations of reward probability can exist alongside distorted representations manifesting either at the neural level or in behavior. As reviewed earlier, Tobler et al. (2008) reported probability representations in ventral striatum that failed to reflect nonlinearities expressed both in choice behavior and in neural representations in other anatomical regions. Similarly, Jessup and O'Doherty (2011) showed that striatal RPEs were not affected by the so-called "gambler's fallacy," another behaviorally expressed distortion of objective probabilities. And Clark, Lawrence, Astley-Jones, and Gray (2009), using a choice manipulation related to our own, found that it magnified the effect of near-misses both on subsequent gambling behavior and on outcome responses in medial frontal cortex, but that it had no such effect on striatal outcome responses.

As in the Jessup and O'Doherty (2011) study, the central finding in the present work involved a behavioral–neural dissociation: We observed behavior directly expressive of outcome probability distortion, in conjunction with neural responses displaying no such distortion. Somewhat surprisingly, especially given previous reports of neural–neural dissociations (Clark et al., 2009; Tobler et al., 2008), our fMRI data revealed no area with activity directly parallelling the distorted probability judgments our participants offered in their behavioral responses. However, the fMRI results did reveal two other effects arising from our choice manipulation, each of which points to a novel explanation for how the IOC may give rise to inflated probability judgments.

The first of these findings involved an effect of the IOC on regional activation during the choice acceptance period: Greater activation was seen in ventral striatum and vmPFC in response to choice acceptance than veto events. This is consistent with recent data on the "value of choice." A number of studies have indicated that the freedom to choose among response alternatives is associated with intrinsic value, and consistent with this, the opportunity to choose has been shown to trigger ventral striatal activation (Leotti & Delgado, 2011; Leotti, Iyengar, & Ochsner, 2010; Bown, Read, & Summers, 2003). A related set of studies has suggested, additionally, that positive affect resulting from free choice (as well as from other sources) can engender optimistic predictions about future events (Isen & Geva, 1987; Isen & Patrick, 1983; Langer & Rodin, 1976). Putting these findings together, the present results are consistent with an interpretation of the IOC according to which the value of choice—reflected in striatal and vmPFC activity—triggers affective changes, which in turn translate into inflated estimates of reward probability.

The second IOC-related effect observed in our data pertained to the spin outcome period, where we observed greater activation in ACC and PCC/precuneus following choice approval than following veto. RPE signals were also detected in both of these regions, making it tempting to interpret the choice approval effect as a modulation of the RPE along the lines we had originally predicted. However, the RPE signals in both ACC and PCC/precuneus (unlike ventral striatum) took an "unsigned" form. That is, signal varied inversely with outcome probability for both gains and losses. This pattern, which has been reported in several other studies of ACC functioning (Browning & Harmer, 2012; Hayden et al., 2011; Litt et al., 2011), makes it difficult to interpret the choice approval effect as reflecting a distortion of reward probability concordant with the IOC, because the latter would predict differential effects on gains and losses. An interpretation that fits better with the data would be in terms of outcome saliency (see also Litt et al., 2011): The effect of the IOC could be interpreted as an enhancement of outcome saliency in the case of chosen gambles. Given that outcomes for chosen gambles may plausibly have been considered more self-relevant, it is also interesting to note that paired ACC–PCC activation has been reported in a number of studies involving self-referential processing (Johnson et al., 2002, 2006; Ochsner et al., 2005; Fossati et al., 2003).

In summary, our findings point to the possibility that the IOC, although it impacts probability judgments, may take root outside the probability domain, arising instead from either affective or self-referential processing. Further research will be necessary to evaluate these possibilities and to flesh out the neural mechanisms by which each might translate into distorted probability judgments. What the present results do show is that, whatever factors underlie the IOC, they do not uniformly affect all neural representations of outcome probability. Consistent with the picture emerging from other recent studies, the IOC appears to give rise to a situation in which distorted probability estimates coexist with separate representations that are resistant to distortion.

Acknowledgments

This work was supported by Collaborative Activity Award from the James S. McDonnell Foundation to M. M. B. We thank Daniel M. Oppenheimer and Andrew R. Conway for useful discussions about the experimental design and statistical analyses.

Reprint requests should be sent to Wouter Kool, Department of Psychology, Green Hall, Princeton University, Princeton, NJ 08540, or via e-mail: wkool@princeton.edu.

Notes

1. The IOC is understood more broadly to cover situations where chance situations carry superficial features that are associated with control over outcomes. Such features include choice, perceived competition, familiarity, the need for control, and mood. Our focus in this study was exclusively on effects of choice.

2. Recent behavioral work on decision-making has highlighted the potentially important distinction between stated and experienced probabilities (Hertwig & Erev, 2009). Given this, it is worth noting that, within our task, outcome probabilities were

Downloaded from http://direct.mit.edu/jocn/article-pdf/25/6/858/1935779/jocn_a_00369.pdf by guest on 17 May 2021

both explicitly communicated (via the win sector size) and directly experienced. The task was thus not designed to distinguish between these two potentially very different sources of probability information.

3. Alternatively, one can compute the subjective magnitude of the IOC effect for each participant separately and then average over these scores. This leads to an IOC effect equivalent to 5.8% (SE = 2.2) or a 21° increase in win sector size. To err on the conservative side, our main analysis focused on the smaller value reported in the main text.

4. A second post hoc power analysis, based on the alternative increase in win sector size, described in Footnote 2, yielded a mean Cohen's $d = 0.32$ and a power of $1 - \beta = 0.96$ to detect a difference in striatal response to the IOC.

REFERENCES

Abler, B., Walter, H., Erk, S., Kammerer, H., & Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *Neuroimage, 31,* 790–795.

Allais, P. M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine. *Econometrica: Journal of the Econometric Society, 21,* 503–546.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica, 22,* 23–36.

Berns, G. S., Capra, C. M., Chappelow, J., Moore, S., & Noussair, C. (2008). Nonlinear neurobiological probability weighting functions for aversive outcomes. *Neuroimage, 39,* 2047–2057.

Bown, N. J., Read, D., & Summers, B. (2003). The lure of choice. *Journal of Behavioral Decision Making, 16,* 297–308.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436.

Browning, M., & Harmer, C. J. (2012). Expectancy and surprise predict neural and behavioral measures of attention to threatening stimuli. *Neuroimage, 59,* 1942–1948.

Clark, L., Lawrence, A. J., Astley-Jones, F., & Gray, N. (2009). Gambling near-misses enhance motivation to gamble and recruit win-related brain circuitry. *Neuron, 61,* 481–490.

Cox, R. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29,* 162–173.

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

De Martino, B., Kumaran, D., Holt, B., & Dolan, R. J. (2009). The neurobiology of reference-dependent value computation. *The Journal of Neuroscience, 29,* 3833–3842.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299,* 1898–1902.

FitzGerald, T. H. B., Seymour, B., Bach, D. R., & Dolan, R. J. (2010). Differentiable neural substrates for learned and described value and risk. *Current Biology, 20,* 1823–1829.

Fossati, P., Hevenor, S. J., Graham, S. J., Grady, C., Keightley, M. L., Craik, F., et al. (2003). In search of the emotional self: An fMRI study using positive and negative emotional words. *American Journal of Psychiatry, 160,* 1938–1945.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: Neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *Journal of Neuroscience, 31,* 4178–4187.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15,* 534–539.

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13,* 517–523.

Hsu, M., Krajbich, I., Zhao, C., & Camerer, C. F. (2009). Neural response to reward anticipation under risk is nonlinear in probabilities. *Journal of Neuroscience, 29,* 2231–2237.

Huettel, S., Stowe, C., Gordon, E., Warner, B., & Platt, M. (2006). Neural signatures of economic preferences for risk and ambiguity. *Neuron, 49,* 765–775.

Hutcherson, C. A., Plassmann, H., Gross, J. J., & Rangel, A. (2012). Cognitive regulation during decision-making shifts behavioral control between ventromedial and dorsolateral prefrontal value systems. *The Journal of Neuroscience, 32,* 13543–13554.

Isen, A. M., & Geva, N. (1987). The influence of positive affect on acceptable level of risk: The person with a large canoe has a large worry. *Organizational Behavior and Human Decision Processes, 39,* 145–154.

Isen, A. M., & Patrick, R. (1983). The effect of positive feelings on risk taking: When the chips are down. *Organizational Behavior and Human Performance, 31,* 194–202.

Jessup, R. K., & O'Doherty, J. P. (2011). Human dorsal striatal activity during choice discriminates reinforcement learning behavior from the gambler's fallacy. *Journal of Neuroscience, 31,* 6296–6304.

Johnson, M. K., Raye, C. L., Mitchell, K. J., Touryan, S. R., Green, E. J., & Nolen-Hoeksema, S. (2006). Dissociating medial frontal and posterior cingulate activity during self-reflection. *Social Cognitive and Affective Neuroscience, 1,* 56–64.

Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2002). Neural correlates of self-reflection. *Brain, 125,* 1808–1814.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *The Journal of Political Economy, 98,* 1325–1348.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society, 47,* 263–292.

Kobayashi, S., Lauwereyns, J., Koizumi, M., Sakagami, M., & Hikosaka, O. (2002). Influence of reward expectation on visuospatial processing in macaque lateral prefrontal cortex. *Journal of Neurophysiology, 87,* 1488–1498.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32,* 311–328.

Langer, E. J., & Rodin, J. (1976). The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting. *Journal of Personality and Social Psychology, 34,* 191–198.

Leotti, L. A., & Delgado, M. (2011). The inherent reward of choice. *Psychological Science, 22,* 1310–1318.

Leotti, L. A., Iyengar, S. S., & Ochsner, K. N. (2010). Born to choose: The origins and value of the need for control. *Trends in Cognitive Sciences, 14,* 457–463.

Litt, A., Plassmann, H., Shiv, B., & Rangel, A. (2011). Dissociating valuation and saliency signals during decision-making. *Cerebral Cortex, 21,* 95–102.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time discounting for primary rewards. *Journal of Neuroscience, 27,* 5796.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience, 16,* 1936–1947.

Ochsner, K. N., Beer, J. S., Robertson, E. R., Cooper, J. C., Gabrieli, J. D. E., Kihlstrom, J. F., et al. (2005). The neural

Downloaded from http://direct.mit.edu/jocn/article-pdf/25/6/852/1935797/jocn_a_00369.pdf by guest on 17 May 2021

correlates of direct and reflected self-knowledge. *Neuroimage, 28,* 797–814.
Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature, 441,* 223–226.
Paulus, M. P., & Frank, L. R. (2006). Anterior cingulate activity modulates nonlinear decision weight function of uncertain prospects. *Neuroimage, 30,* 668–677.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.
Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature, 400,* 233–238.
Presson, P. K., & Benassi, V. A. (1996). Illusion of control: A meta-analytic review. *Journal of Social Behavior and Personality, 11,* 493–510.
Preuschoff, K., Quartz, S., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *The Journal of Neuroscience, 28,* 2745–2752.
Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology, 20,* 262–270.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.
Roesch, M. R., Singh, T., Brown, P. L., Mullins, S. E., & Schoenbaum, G. (2009). Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. *Journal of Neuroscience, 29,* 13365–13376.
Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599.
Simon, D. A., & Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience, 31,* 5526–5539.
Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature, 38,* 332–382.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain*. New York: Theme Medical Publishers.
Thompson, S. Z., Armstrong, W., & Thomas, C. (1998). Illusions of control, underestimations, and accuracy: A control heuristic explanation. *Psychological Bulletin, 123,* 143–161.
Tobler, P. N., Christopoulos, G. I., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2008). Neuronal distortions of reward probability without choice. *Journal of Neuroscience, 28,* 11703–11711.
Trepel, C., Fox, C. R., & Poldrack, R. A. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. *Cognitive Brain Research, 23,* 34–50.
Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 297–323.
von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
Vul, E., & Kanwisher, N. (2010). Begging the question: The non-independence error in fMRI data analysis. In S. J. Hanson & M. Bunzl (Eds.), *Foundational Issues in Human Brain Mapping* (pp. 71–91). Cambridge, MA: MIT Press.

Downloaded from http://direct.mit.edu/jocn/article-pdf/25/6/857/1935797/jocn_a_00369.pdf by guest on 17 May 2021