Behavioral/Systems/Cognitive

The Neural Underpinnings of How Reward Associations Can Both Guide and Misguide Attention

Ruth M. Krebs,[1,2] **Carsten N. Boehler,**[1,2] **Tobias Egner,**[1,3] **and Marty G. Woldorff**[1,3,4]

[1]Center for Cognitive Neuroscience, Duke University, Durham, North Carolina 27708, [2]Department of Experimental Psychology, Ghent University, 9000 Ghent, Belgium, [3]Department of Psychology and Neuroscience, Duke University, Durham, North Carolina 27708, and [4]Department of Psychiatry, Duke University, Durham, North Carolina 27708

It is commonly accepted that reward is an effective motivator of behavior, but little is known about potential costs resulting from reward associations. Here, we used functional magnetic resonance imaging (fMRI) to investigate the neural underpinnings of such reward-related performance-disrupting effects in a reward-modulated Stroop task in humans. While reward associations in the task-relevant dimension (i.e., ink color) facilitated performance, behavioral detriments were found when the task-irrelevant dimension (i.e., word meaning) implicitly referred to reward-predictive ink colors. Neurally, only relevant reward associations invoked a typical reward-anticipation response in the nucleus accumbens (NAcc), which was in turn predictive of behavioral facilitation. In contrast, irrelevant reward associations increased activity in a medial prefrontal motor-control-related region, namely the presupplementary motor area (pre-SMA), which likely reflects the preemption and inhibition of automatic response tendencies that are amplified by irrelevant reward-related words. This view was further supported by a positive relationship between pre-SMA activity and pronounced response slowing in trials containing reward-related as compared with reward-unrelated incongruent words. Importantly, the distinct neural processes related to the beneficial and detrimental behavioral effects of reward associations appeared to arise from preferential-coding mechanisms in visual-processing areas that were shared by the two stimulus dimensions, suggesting a transfer of reward-related saliency to the irrelevant dimension, but with highly differential behavioral and neural ramifications. More generally, the data demonstrate that even entirely irrelevant reward associations can influence stimulus-processing and response-selection pathways relatively automatically, thereby representing an important flipside of reward-driven performance enhancements.

Introduction

Reward is known to be a driving force for adaptive behavior and learning, and numerous studies have demonstrated that animals quickly learn to maximize rewards when interacting with the environment (Schultz, 2002; Wise, 2004). Once established, stimulus–reward associations can influence sensory processing at an early stage, presumably by enhancing the neural coding of salient stimulus features (Lee et al., 2002; Serences, 2008; Pessoa and Engelmann, 2010). Such preferential-coding mechanisms trigger processing cascades that benefit behavior by increasing attention to the evoking stimulus (Small et al., 2005; Engelmann et al., 2009), as well as by interacting with higher order cognitive functions (Locke and Braver, 2008; Kouneiher et al., 2009).

More recently, the view that such reward-driven influences are exclusively beneficial has been challenged by observations that they can concomitantly induce behavioral costs (Pessoa, 2009). In line with this view, we recently demonstrated that reward associations in the task-relevant dimension (i.e., ink color) of a Stroop task facilitated responses, whereas behavioral detriments were found when the task-irrelevant dimension (i.e., word meaning) implicitly referred to reward-predictive colors (Krebs et al., 2010). These opposite effects suggest that reward associations were implicitly transferred to the irrelevant dimension, which in turn led attention astray.

By implementing our behavioral protocol in an fMRI study, we sought to determine the neural underpinnings of such detrimental reward-related effects. Based on the apparent generalization of reward associations across stimulus dimensions, we hypothesized that irrelevant reward-related words, similar to relevant reward-predictive colors, would engage visual object-representation areas within the fusiform gyrus (FG), which are known to be modulated by stimulus saliency and attention (e.g., O'Craven et al., 1999; Vuilleumier, 2005). In the irrelevant case, such a modulation may be triggered by an implicit, albeit salient, association, rather than reflecting a voluntary attentional mechanism.

We further predicted that the beneficial behavioral effects of relevant reward associations, which truly predicted potential reward, would be paralleled by increased activity within key regions of the reward-anticipation circuit, including the ventral striatum, which is particularly sensitive to stimulus–value associations (Knutson et al., 2001a,b; O'Doherty et al., 2004). In contrast, the interference induced by irrelevant reward associations may be

Received Feb. 10, 2011; revised April 11, 2011; accepted May 20, 2011.

Author contributions: R.M.K., C.N.B., and M.G.W. designed research; R.M.K. performed research; R.M.K. analyzed data; R.M.K., C.N.B., T.E., and M.G.W. wrote the paper.

This work was funded by NIH Grants R01-MH060415 and R01-NS051048 to M.G.W. and by DFG Grant BO 3345/1-1 to C.N.B.

Correspondence should be addressed to Dr. Ruth M. Krebs, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium. E-mail: ruthmkrebs@gmail.com.

DOI:10.1523/JNEUROSCI.0732-11.2011

 0270-6474/11/319752-08$15.00/0

due to an unspecific attentional disruption from the main task, similar to the effect of salient distractors (de Fockert et al., 2004). Alternatively, the interference could primarily arise at the response-control level, reflecting prepotent response tendencies to reward-related words. This notion is in line with parallel distributed processing models of Stroop interference proposing that incongruent words automatically coactivate the task-irrelevant stimulus–response pathway (Cohen et al., 1990; MacLeod, 1991; Banich et al., 2001; Botvinick et al., 2001; Zysset et al., 2001). If this were the case, salient, albeit irrelevant, reward-related words would likely engage medial frontal regions implicated in response-conflict monitoring (Ullsperger and von Cramon, 2001; Kerns et al., 2004; Ridderinkhof et al., 2004). In particular, we predicted modulations within the presupplementary motor area (pre-SMA) because of its involvement in situations in which alternative motor plans have to be initiated to overcome prepotent response tendencies (Rushworth et al., 2004; Nachev et al., 2005).

Materials and Methods

Participants and paradigm. Nineteen healthy right-handed subjects participated in the present study (mean age ± SD: 22.6 ± 3.5 years, 10 female). One additional subject had to be excluded from the analysis because of poor general task performance. All participants gave written informed consent in accordance with the Duke Medical Center Institutional Review Board for human subjects and were paid a basic amount of $40 plus an average reward bonus of $15.

Participants performed a version of the classic Stroop task, responding to the ink color of words while ignoring their semantic meaning. Throughout all experimental runs, a small gray fixation square (visual angle 0.3°) was maintained in the center of a black screen. In each trial, a colored capitalized word was presented directly above fixation for 600 ms (Fig. 1*A*), randomly chosen from the following set: RED, YELLOW, BLUE, GREEN, or BROWN (vertical 0.8°, horizontal 2.1–4.6°). The word onsets were pseudorandomly varied with a stimulus onset asynchrony of 1.5–6 s to allow for effective event-related blood oxygen level-dependent (BOLD) response estimation (Hinrichs et al., 2000). Each word was written in one of four ink colors (red, yellow, blue, or green), and participants were instructed to respond as quickly as possible by pressing the button associated with the word's ink color (Color: task-relevant dimension) while ignoring its semantic meaning (Word: task-irrelevant dimension). Responses were given with the index and middle fingers of the left and right hands (color-button assignments and color-reward associations were counterbalanced across subjects). The semantic meaning of a given word could be congruent (Wc; e.g., GREEN in green ink), fully incongruent (Wi; e.g., RED in green ink), or incongruent-ineligible (Wii; e.g., BROWN in green ink) with respect to the ink color. Notably, although the latter trials (Wii) are incongruent at the perceptual level, similar to fully incongruent trials (Wi), they do not invoke any competition at the response level, which is why they are often labeled "incongruent response-ineligible" (Milham et al., 2001; Chen et al., 2006). Traditionally, these trials are used as a baseline to quantify the amount of behavioral benefits for congruent trials and detriments for fully incongruent ones as they introduce an intermediate level of processing (MacLeod, 1991; van Veen et al., 2001). Although we included incongruent-ineligible trials to match the previous behavioral protocol, they were not considered in the imaging analysis because of their exceptional position.

Two of the four possible ink colors represented typical Stroop trials (no reward, termed Color0), and responses to the remaining two colors were associated with monetary incentives (potential reward, termed Color$) (Fig. 1*B*). Fast and correct responses to the latter resulted in a 10-cent gain, whereas incorrect or slow responses resulted in a 10-cent penalty. To keep all participants at a similar reward ratio of 70% throughout the experiment, the response time (RT) window was adjusted dynamically based on individual performance, leading to a mean monetary gain of $3 per run. Specifically, after each trial, the hit rate was routinely updated in the background, and the response timeout for the next trial was shortened or extended by 10 ms if this rate was above or below 70%, respectively. Importantly, however, all analyses regarding RT and accuracy were based on the actual responses within a window of 150–1200 ms after word onset, whereas the adjustments only affected the interim visual feedback. Following a short training session, subjects completed five 8 min runs inside the fMRI scanner, yielding a total of 500 Color$ and 500 Color0 trials. During four breaks within each run, as well as at the end of each run, the updated dollar amount was displayed, serving as intermediate performance feedback.

Importantly, because of the relevant color–reward associations, the irrelevant semantic meaning of incongruent words could implicitly refer to a rewarded color (termed Wi$) or not (termed Wi0), which was equally distributed for Color$ and Color0 trials (Fig. 1*B*). Despite the words' implicit relation to the different ink–color subsets, word meanings were always task-irrelevant and never predictive of any reward. To simplify the nomenclature, however, we refer to incongruent words with implicit reward-related meaning as "irrelevant reward" trials with respect to the data analysis. This manipulation allowed us to investigate the direct effects of reward associations in the relevant dimension (Color$ vs Color0), as well as their indirect effects that were entirely irrelevant to the task (Wi$ vs Wi0).

The averaged RTs and error rates within the potential-reward and no-reward word categories were submitted to repeated-measures ANOVAs to verify the overall main effects of the relevant (Color: Color$, Color0) and irrelevant (Word: Wc, Wii, Wi$/Wi0) dimensions. To investigate differential effects of reward-related irrelevant information, we conducted additional 2 × 2 repeated-measures ANOVAs focusing on the two types of incongruent trials (Color$ vs Color0 × Wi$ vs Wi0).

fMRI data acquisition. fMRI images were acquired using a 3 tesla GE MR750 scanner with an eight-channel head-coil array. Each functional run consisted of 324 images acquired in an axial slice orientation (36 slices, 3 mm) using an interleaved scanning order (inward spiral sequence with SENSE acceleration factor of 2, TR = 1.5 s, TE = 25 ms, FOV = 192 mm, matrix size = 64 × 64, in-plane resolution = 3 × 3 mm). The first five volumes of each run were discarded to allow a steady magnetization to be reached.

For each participant, a T1-weighted high-resolution whole-brain anatomical scan (3D FSPGR sequence, FOV = 256 mm, yielding a voxel size of 1 × 1 × 1 mm) was acquired to enable coregistration and normalization. Participants were required to keep their eyes fixated on a centered square throughout the task, and fixation performance was monitored using an MR-compatible eye-tracking system (Viewpoint, Arrington Research).

fMRI data analysis. Images were preprocessed and analyzed using the Statistical Parametric Mapping software package (SPM5). Anatomical images were coregistered to the SPM template and spatially normalized using the gray- and white-matter segmentation routine implemented in SPM5. Functional images were corrected for acquisition delay, spatially realigned, and coregistered to the original T1-weighted image. After resampling to a final voxel size of 2 × 2 × 2 mm, functional images were smoothed with an isotropic 6 mm full-width half-maximum Gaussian kernel. Before model estimation, a high-pass temporal filter of 128 s was applied.

A two-stage model was used for statistical analysis (Friston et al., 1995). BOLD responses were modeled by delta functions at the stimulus onsets for each event type, which were then convolved with a canonical hemodynamic response function (HRF) plus temporal and dispersion derivatives. The resulting regressors were entered into a general linear model together with the six realignment parameters for each run. The analysis focused on correct trials only, and all erroneous trials were therefore modeled as regressors of no interest. Individual participants' contrast images were calculated using the amplitude HRF parameter and entered into a random-effects analysis using one-sample *t* tests for voxelwise comparisons (corrected for multiple comparisons on the cluster level $p < 0.05$; auxiliary *p*-value threshold $p < 0.001$; extent threshold $k = 15$) (Figs. 2, 3). Brain activity for relevant reward was derived by contrasting potential-reward and no-reward trials in the absence of irrelevant

reward associations (RELEVANT REWARD contrast: Color$ > Color0, excluding Wi$ trials); activity related to irrelevant reward was derived by contrasting incongruent reward-related words to incongruent reward-unrelated words in the absence of relevant reward (IRRELEVANT REWARD contrast: Color0_Wi$ > Color0_Wi0).

To further verify the observed differential effects within NAcc and the pre-SMA, we performed region of interest (ROI) analyses using the MarsBar analysis toolbox (Brett et al., 2002). Spheres were centered around the local activity maxima (4 mm radius) derived from the voxelwise relevant reward (MNI coordinates x, y, z: −10, 10, 2) (Fig. 3*A*) as well as the irrelevant reward contrasts (x, y, z: 2, 8, 56) (Fig. 3*B*). The BOLD signal change was extracted for each condition of interest and analyzed with respect to the two types of reward associations, analogous to the voxelwise comparisons (Color$ − Color0; Color0_Wi$ − Color0_Wi0). It should be emphasized that, based on the ROI selection, one of the respective comparisons (as indicated by asterisks in Fig. 3) naturally reiterates the results of the voxelwise analyses (Kriegeskorte et al., 2009). Therefore, these ROI results should only be considered illustrative rather than inferential. The main purpose of this ROI-based signal extraction was, however, to assess whether the identified regions are sensitive to reward information in the respective opposite stimulus dimension. This approach allowed us to rule out that potential modulations by the respective opposite reward information in these regions remained undetected in the voxelwise comparisons. In addition, the extracted signal-change values were used to perform correlational analyses with RTs.

The main aim of the present study was to examine effects of relevant and irrelevant reward in the total absence of the respective opposite factor, that is, the RELEVANT REWARD contrast included only trials free from reward information in the irrelevant dimension and vice versa. These comparisons guaranteed that the observed effects are not directly confounded by the opposite factor, but they did not allow us to investigate a potential interaction between the two dimensions. To investigate such interactions across the whole brain, we performed a 2 × 2 voxelwise repeated-measures ANOVA focusing on incongruent trials only, analogous to the behavioral data analysis (significance threshold $p < 0.001$ uncorrected; extent threshold $k = 15$) (see Table 4).

Results

Behavioral results

Participants responded to the word's ink color (Color: task-relevant), which could be associated with obtaining reward or not (Color$ vs Color0), while ignoring the semantic meaning (Word: task-irrelevant), which could be congruent, incongruent, or incongruent-ineligible with respect to the ink color. A full representation of the task performance is provided in Table 1.

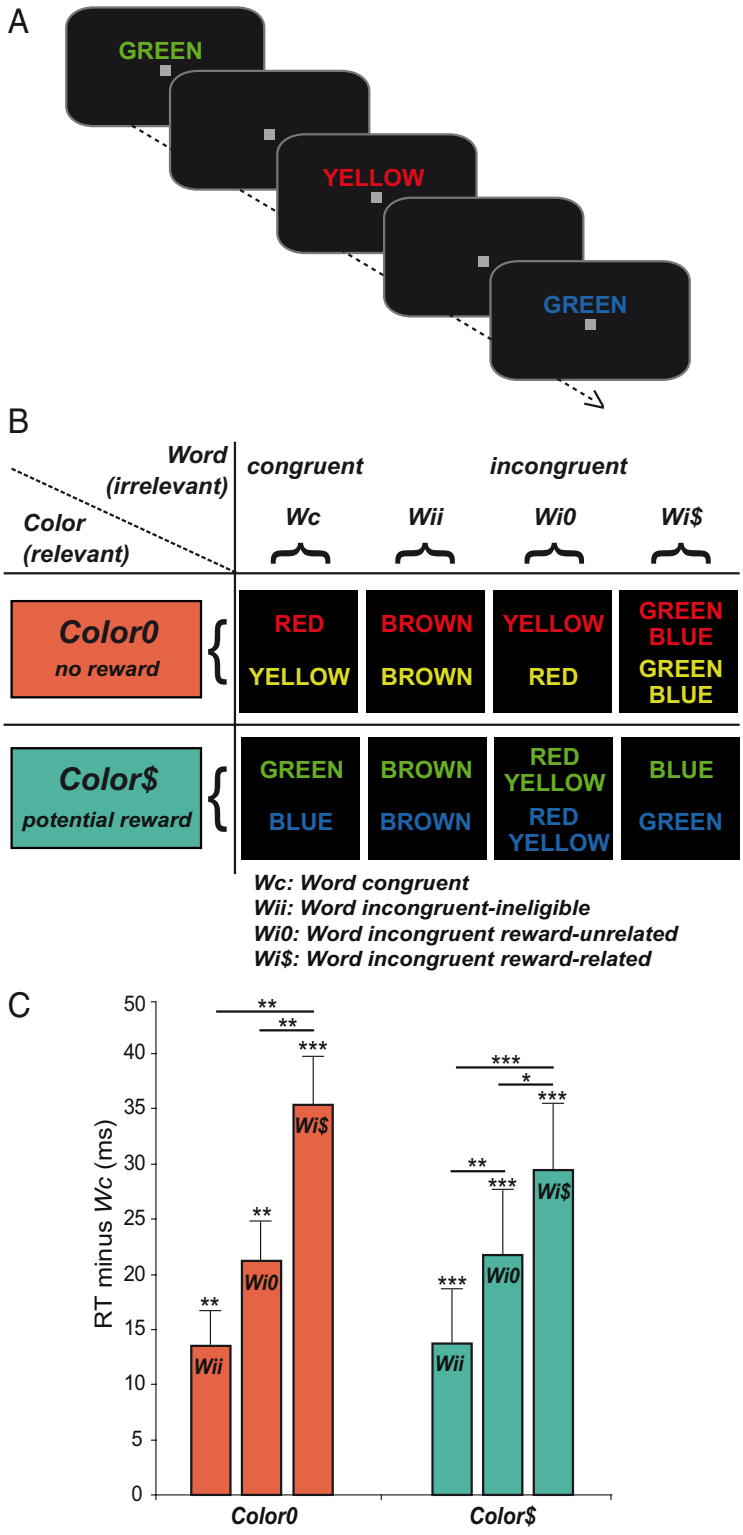

Figure 1. Experimental paradigm and behavioral results. ***A***, Colored words were presented for 600 ms each on a black background, separated by a stimulus onset asynchrony of 1.5–6 s. Participants responded to the word's ink color (task-relevant dimension) while ignoring its semantic meaning (task-irrelevant dimension). ***B***, A subset of ink colors and the respective responses were associated with potential-reward (termed Color$; e.g., green/blue), and the remaining colors were not (termed Color0; e.g., red/yellow). Accordingly, task-irrelevant word meanings could not only be congruent, incongruent-ineligible, or fully incongruent to the ink color, but the latter word category could moreover implicitly refer to the potential-reward (Wi$) or the no-reward (Wi0) ink-color subset. ***C***, Relative RT differences are depicted for all incongruent word categories compared with the congruent ones within both ink-color subsets (Color$: turquoise bars; Color0: orange bars). Error bars represent the SEM for the difference values; asterisks indicate significant paired t tests (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; above single bar: difference relative to congruent words; between bars: difference between two word categories).

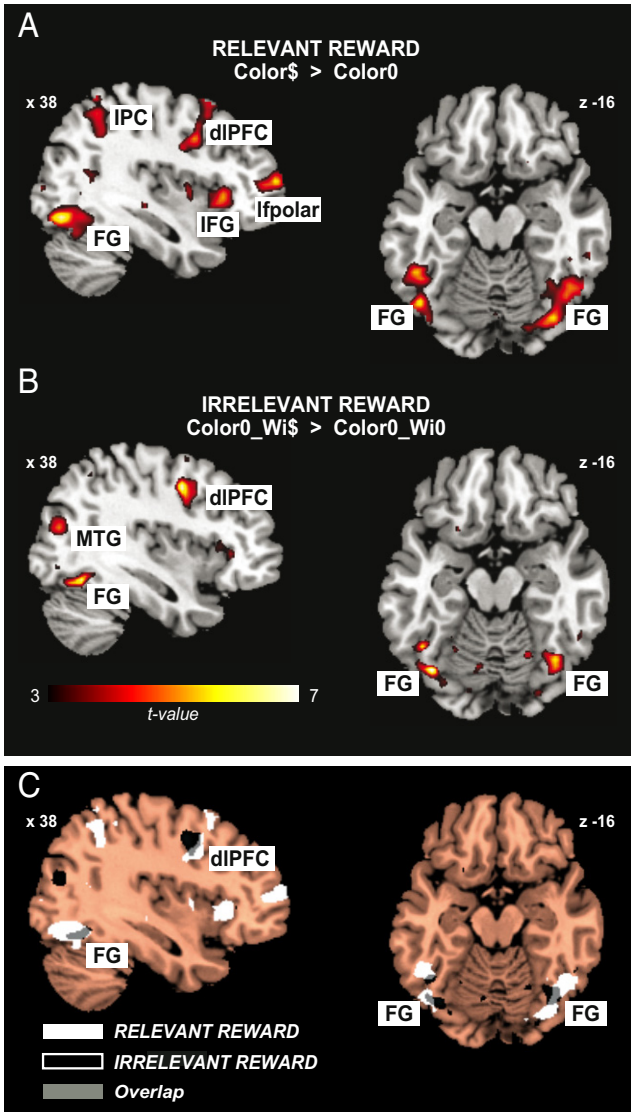

Figure 2. Cortical activity associated with relevant and irrelevant reward. ***A***, The actual prospect of reward, defined by comparing relevant potential-reward with no-reward trials, engaged various prefrontal regions, including right dlPFC and inferior frontal junction, right IFG, and lateral frontopolar cortex (lf polar), as well as bilateral FG and inferior parietal cortex (IPC). ***B***, Irrelevant reward information, defined by comparing incongruent reward-related with incongruent reward-unrelated words in no-reward trials, was associated with increased activity in dlPFC and FG, as well as in the right middle temporal gyrus (MTG). ***C***, Neural activity associated with relevant reward (white) and irrelevant reward (black) is overlaid to illustrate the activity overlap (gray) within right dlPFC and bilateral FG (display cutoff $p < 0.001$).

Table 1. Performance in the color-naming Stroop task

	Word (irrelevant)			
	Congruent	Incongruent		
Color (relevant)	Wc	Wii	Wi0	Wi$
Color0				
RT ms (SD)	592 (67.0)	606 (69.5)	614 (76.3)	628 (78.6)
Errors % (SD)	9.2 (5.7)	10.0 (6.5)	10.9 (6.7)	11.4 (7.3)
Color$				
RT ms (SD)	521 (75.3)	535 (79.8)	543 (85.0)	551 (87.9)
Errors % (SD)	6.2 (4.4)	6.6 (4.9)	8.3 (6.1)	8.9 (6.0)

Overall, responses were faster in potential-reward trials (Color$, mean RT = 538 ms) as compared with no-reward trials (Color0, mean RT = 610 ms), reflected statistically in a main effect of relevant reward in the ANOVA ($F_{(1,18)} = 41.27$, $p < 0.001$). Furthermore, in keeping with typical Stroop-interference effects, RTs were significantly modulated by the word meaning in a given trial, with fastest responses in congruent trials, intermediate in incongruent-ineligible trials, and slowest in incongruent ones ($F_{(1,17)} = 33.0$, $p < 0.001$). No interaction between potential reward and conflict was observed ($p > 0.7$). Participants also committed fewer errors in potential-reward as compared with no-reward trials (Color$ < Color0: $F_{(1,18)} = 12.37$, $p = 0.002$), as well as fewer errors in trials with congruent and incongruent-ineligible as compared with fully incongruent words ($F_{(1,17)} = 6.35$, $p = 0.004$).

To test the interference effects specifically induced by reward-related (Wi$) incongruent words, we conducted an additional 2 × 2 repeated-measures ANOVA focusing on incongruent trials of the potential-reward and no-reward trial types (Color$ vs Color0 × Wi$ vs Wi0). We again observed faster responses in potential-reward as compared with no-reward trials (Color$ < Color0: $F_{(1,18)} = 35.71$, $p < 0.001$), but interference effects were significantly greater for incongruent words that were semantically related to reward (Wi$) across potential-reward and no-reward trials ($F_{(1,18)} = 21.45$, $p < 0.001$) (Fig. 1*C*). No interaction of these factors was observed ($p > 0.2$).

Neural activity associated with relevant reward

To identify brain regions that are specifically modulated by the actual prospect of reward as indicated by the ink color, we compared potential-reward to no-reward trials (RELEVANT REWARD contrast: Color$ > Color0) (Figs. 2*A*, 3*A*, Table 2). Importantly, we excluded all trials containing irrelevant reward associations (Wi$) to avoid interaction effects.

The voxelwise comparison yielded enhanced activity for potential-reward trials in the bilateral NAcc (Fig. 3*A*), a region that has been strongly associated with the processing of reward, along with regions related to cognitive control: the right dorsolateral prefrontal cortex (dlPFC), right inferior frontal gyrus (IFG) extending into the anterior insula, and lateral frontopolar cortex bilaterally (Fig. 2*A*). The cluster within the dlPFC comprises parts of the posterior middle frontal gyrus and extends into the inferior frontal junction. Furthermore, we found bilateral clusters of increased activity in the fusiform gyri (FG).

Neural activity associated with irrelevant reward

To isolate neural responses reflecting the processing of irrelevant reward information, we directly compared incongruent reward-related and incongruent reward-unrelated words of the no-reward trial category (IRRELEVANT REWARD: Color0_Wi$ > Color0_Wi0) (Figs. 2*B*, 3*B*, Table 3). These trial types differed exclusively regarding the reward-related meaning of the word, while both entailed response conflict and were nonpredictive of reward, thereby avoiding interactions between relevant and irrelevant reward processing in the same trial.

The voxelwise analysis revealed enhanced activity in medial and lateral frontal cortex (Figs. 2*B*, 3*B*). The cluster within the medial frontal wall corresponds closely to the pre-SMA region as it is located just rostral to the vertical commissure anterior (Vorobiev et al., 1998). The activity within the dlPFC mainly comprises the posterior middle frontal gyrus. Furthermore, irrelevant reward signals activated the bilateral FG, extending into the inferior temporal gyrus (Fig. 2*B*).

Differential activity patterns and relation to task performance

Despite the considerable overlap in lateral prefrontal and posterior cortical regions (Fig. 2*C*), the voxelwise comparisons revealed two regions that were highly selective for relevant and irrelevant reward information, respectively, namely the NAcc and the pre-SMA (Fig. 3). To further verify this dissociation, we extracted the BOLD response for all conditions of interest based on the ROIs derived from the relevant reward and irrelevant reward contrasts separately (for further details on ROI selection, see fMRI data analysis). These ROI analyses naturally reflected the results of the voxelwise comparisons (as indicated by asterisks), but revealed moreover that the NAcc (Fig. 3*A*) was indeed insensitive to irrelevant reward, whereas the pre-SMA (Fig. 3*B*) was insensitive to relevant reward (as indicated by ns).

To relate these differential activity patterns to task performance, we submitted the ROI-based difference values delineated in Figure 3 to an across-subjects correlational analysis with RT measures. Response speeding related to the processing of relevant reward, defined as the RT difference between Color$ and Color0 trials, was negatively correlated with the corresponding neural responses within left NAcc (Fig. 3*C*), indicating a facilitating influence of activity in this region on performance. In contrast, response slowing associated with the occurrence of irrelevant reward-related words, defined as the RT difference between Color0_Wi$ and Color0_Wi0 trials, was associated with activity increases in the pre-SMA (Fig. 3*D*), presumably reflecting this region's role in the inhibition of inappropriate responses associated with the task-irrelevant stimulus dimension and in the selection of the appropriate response (Isoda and Hikosaka, 2007). To assess whether the observed differential behavioral effects were independent, we tested whether the NAcc-associated speeding and the pre-SMA-associated slowing were statistically related across subjects. We found that neither the relative RT speeding (relevant) and RT slowing (irrelevant; $r = -0.05$, $p = 0.819$) nor the BOLD response within NAcc and pre-SMA ($r = 0.22$, $p = 0.365$) were significantly correlated across participants, indicating that both processes are mediated relatively independently by distinct regions.

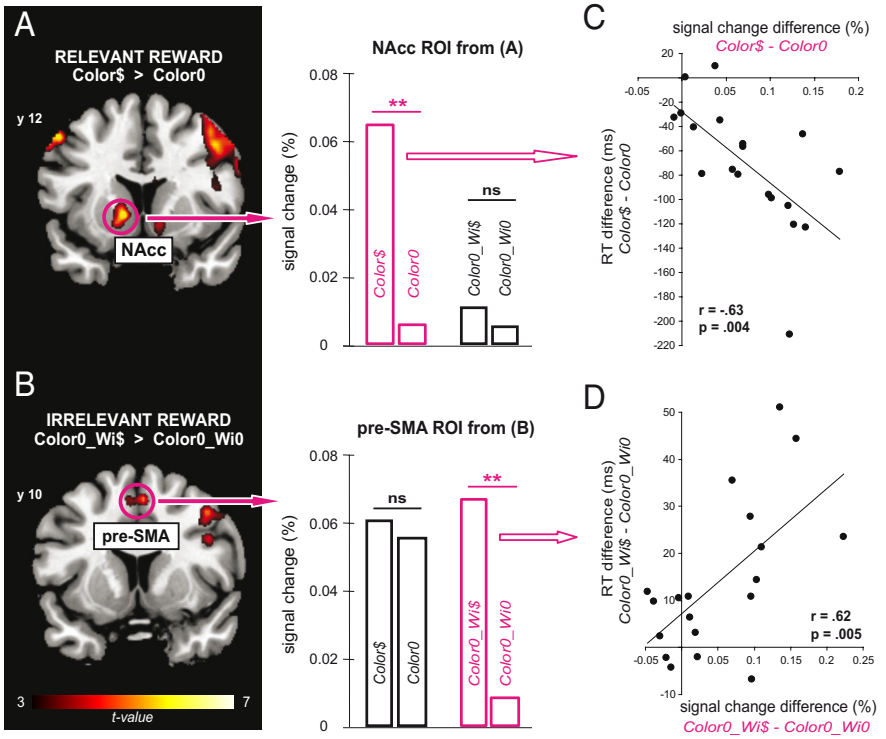

Figure 3. Differential activity patterns related to relevant and irrelevant reward associations. The NAcc was activated by relevant reward (***A***), whereas the pre-SMA was activated by irrelevant reward (***B***). ROI analyses of the extracted signal-change values confirmed that these regions were insensitive to the respective opposite type of reward information as indicated by ns (nonsignificant comparison). Asterisks indicate the reiterations of the voxelwise results (significant at $p < 0.01$). ***C***, Furthermore, the NAcc activity increase that was selective for relevant reward (depicted as difference between Color$ and Color0) was correlated with response acceleration (depicted as RT difference between Color$ and Color0). ***D***, In contrast, increased pre-SMA activity related to reward-related words (depicted as differences between Color0_Wi$ and Color0_Wi0) was associated with response slowing (depicted as RT difference between Color0_Wi$ and Color0_Wi0).

Voxelwise interactions between relevant and irrelevant reward

To investigate whether any of the regions isolated by the independent contrasts exhibited an interaction between the two reward dimensions, we performed a voxelwise 2 × 2 repeated-measures ANOVA focusing on incongruent trials only. Three cortical regions—right dlPFC, right IPC, and right FG—exhibited a significant interaction between relevant and irrelevant reward (Table 4). When extracting the signal change from these activated clusters, we found that the interaction in all three regions was due to a relatively larger activity increase for incongruent reward-related words in no-reward as compared with reward trials. In contrast, no voxelwise interactions were found in the other key areas of interest, namely the NAcc and the pre-SMA.

Discussion

Consistent with our previous behavioral study, we found that relevant reward associations (i.e., ink colors indicating potential-reward trials) facilitated behavioral responses, whereas irrelevant reward associations (i.e., word meaning related to reward-predicting ink colors) impaired performance. These results suggest that the established reward associations in the relevant dimension generalized to a different dimension of the stimulus on the basis of a shared abstract color representation. Based on this "transfer" of reward associations, the entirely irrelevant word meaning seemed to also acquire saliency, thereby impairing performance. The present study focused on the neural underpinnings of these striking behavioral effects.

Preferential coding of reward-associated stimuli

We hypothesized that reward associations, in both the relevant and irrelevant dimensions, would be reflected in increased neural activity within visual object-representation areas. Such a shared preferential-coding mechanism was indeed supported by the highly overlapping enhanced activity observed in the FG in response to reward associations in both dimensions. Considering the involvement of the FG in the processing of visual objects, along with its sensitivity to stimulus saliency and attentional modulations (e.g., O'Craven et al., 1999; Vuilleumier, 2005), the observed activity

Table 2. Activity clusters associated with relevant reward (Color$ > Color0)

Region	L/R	k	MNI coordinates x	y	z	t[a]
Dorsolateral prefrontal cortex (BA 8)	R	613	54	6	46	6.36
Inferior frontal junction			34	8	38	5.79
Dorsolateral prefrontal cortex (BA 9)			46	16	42	5.16
Inferior occipital gyrus (BA 19)	R	342	36	−76	−12	6.03
Fusiform gyrus (BA 37)			46	−58	−16	4.51
Nucleus accumbens	L	85	−10	10	−2	6.00
Nucleus accumbens			−15	14	−6	4.58
Superior temporal gyrus (BA 39)	R	183	42	−58	16	5.98
Superior temporal gyrus (BA 39)			50	−54	14	4.82
Lateral frontopolar region (BA 10)	L	110	−46	48	22	5.79
Lateral frontopolar region (BA 10)			−38	58	18	4.99
Lateral frontopolar region (BA 10)	R	116	40	56	8	5.76
Lateral frontopolar region (BA 10)			46	52	0	4.39
Inferior frontal gyrus (BA 47)	R	129	34	26	−4	5.53
Inferior parietal cortex (BA 40)	L	264	−50	−50	56	5.43
Inferior parietal cortex (BA 40)			−56	−42	54	4.91
Fusiform gyrus (BA 37)	L	389	−46	−74	−6	5.11
Fusiform gyrus (BA 37)			−46	−66	−16	5.04
Fusiform gyrus (BA 37)			−48	−48	−18	4.85
Nucleus accumbens	R	45	8	20	−6	4.48
Nucleus accumbens			14	16	−8	4.15
Inferior parietal cortex (BA 40)	R	228	42	−42	52	4.26
Inferior parietal cortex (BA 40)			38	−54	52	4.19

[a]*t* value: corrected cluster-size threshold $p = 0.05$; auxiliary *p*-value threshold $p < 0.001$; extent threshold $k = 15$.

Table 3. Activity clusters associated with irrelevant reward (Color0_Wi$ > Color0_Wi0)

Region	L/R	k	MNI coordinates x	y	z	t[a]
Inferior parietal cortex (BA 7)	R	65	20	−54	52	6.83
Inferior temporal gyrus (BA 37)	L	65	−42	−50	−22	6.56
Inferior temporal gyrus (BA 37)			−44	−56	−14	5.66
Fusiform gyrus (BA 37)	L	49	−38	−70	−16	6.38
Dorsolateral prefrontal cortex (BA 8)	R	190	40	0	44	5.89
Fusiform gyrus (BA 37)	R	91	38	−62	−14	5.69
Presupplementary motor area (BA 6)	L/R	28	2	8	56	4.66
Middle temporal gyrus (BA 19)	R	188	34	−76	16	4.77
Fusiform gyrus (BA 37)			36	−54	−10	4.07

[a]*t* value: corrected cluster-size threshold $p = 0.05$; auxiliary *p*-value threshold $p < 0.001$; extent threshold $k = 15$.

Table 4. Voxelwise interaction between relevant and irrelevant reward (Wi$ and Wi0 trials only)

Region	L/R	k	MNI coordinates x	y	z	F[a]
Dorsolateral prefrontal cortex (BA 8)	R	100	46	30	52	22.00
Dorsolateral prefrontal cortex (BA 8)			36	2	40	13.55
Inferior parietal cortex (BA 7)	R	230	20	−54	46	18.64
Inferior parietal cortex (BA 7)			24	−64	44	16.21
Superior parietal cortex (BA 7)			30	−66	38	14.66
Fusiform gyrus (BA 37)	R	47	34	−62	−14	18.04

[a]*F* value: uncorrected cluster-size threshold $p = 0.001$; extent threshold $k = 15$.

in this region likely reflects increased attention to the Stroop stimuli that hold salient information, regardless of the underlying dimension. Interestingly, the spatial spread of the FG activity appeared to be relatively more posterior for the color dimension and relatively more inferior for the word-meaning dimension. However, although it might appear tempting to propose a dimension-specific attentional enhancement in color-specific (McKeefry and Zeki, 1997) versus word-specific (McCandliss et al., 2003) areas within FG, the present data do not allow us to establish such a dissociation.

Despite the considerable activity overlap in the FG, the opposing behavioral effects related to the source dimension of the reward information suggest differential neural cascades following such preferential coding.

Facilitation related to relevant reward associations

The observed behavioral facilitation in potential-reward trials is consistent with repeatedly observed beneficial effects of reward on performance in various task domains (Ramnani and Miall, 2003; Small et al., 2005; Locke and Braver, 2008; Engelmann et al., 2009; Kiss et al., 2009). Because relevant reward associations in the present paradigm are in line with the higher order task goal, preferential coding of these stimuli likely promotes cognitive-control functions to accomplish the task. Our imaging data support this view, in that we observed greater activity on potential-reward trials in lateral prefrontal regions implicated in the representation and maintenance of task goals, including dlPFC and IFG (Miller and Cohen, 2001; Braver et al., 2007). Moreover, these regions are sensitive to the behavioral relevance of a task (Taylor et al., 2004; Locke and Braver, 2008; Kouneiher et al., 2009). In the present study, not only should color naming per se be represented in the current task set, but also the specific color-reward mapping. In line with this view, we found increased activity in a lateral frontopolar region that has been related to maintaining a higher order goal while performing an ongoing task (Koechlin et al., 1999; Pochon et al., 2002).

Importantly, the anticipation of actual reward reliably activated the NAcc, a key region of the dopaminergic circuits that have been proposed to assign incentive values to behaviorally relevant stimuli (Knutson et al., 2001a; Schultz, 2002; O'Doherty et al., 2004; Wise, 2004; Knutson and Cooper, 2005). The highly selective NAcc response to relevant reward associations in the present paradigm presumably arises through interactions with prefrontal regions that promote the higher order goal of the task (Wise, 2004; Goto and Grace, 2005), an inference further supported by the inverse relationship between reward-related activity increase within this region and response speed across subjects.

It is important to consider whether the observed facilitation in reward trials is primarily related to the anticipation of reward or a specific combination of reward and punishment contingencies. First, the bonuses participants could potentially win were an addition to the hourly payment, which makes it unlikely that participants actually "feared" to lose money. Second, there is evidence that feedback related to both winning and avoiding-to-lose similarly activates the ventral striatum (e.g., Kim et al., 2006). Most importantly, the direct comparison between versions of the task that exclusively used reward and those entailing both reward and punishment (Krebs et al., 2010) did not reveal any differences regarding the behavioral facilitation.

Interference related to irrelevant reward associations

The inducement of increased visual processing of reward-predictive stimuli in the relevant dimension (i.e., color naming) benefited performance, but reward associations in the irrelevant dimension (i.e., word meaning) appeared to be disruptive. This detrimental influence could be related to different mechanisms. One possibility is that preferential coding of a salient, albeit irrelevant, stimulus leads to a relatively unspecific attentional distraction from the color-naming task (de Fockert et al., 2004), thereby delaying responses. Alternatively, the acquired saliency of irrelevant reward-related words could act at the level of stimulus–response mappings by enhancing prepotent response tendencies

for the irrelevant word meaning (MacLeod, 1991; Botvinick et al., 2001).

The present neural data support the latter notion, in that irrelevant but salient reward-related words specifically engaged the pre-SMA, which has been closely associated with the detection as well as the resolution of conflict at the response level (Ullsperger and von Cramon, 2001; Rushworth et al., 2002, 2004; Garavan et al., 2003; Nachev et al., 2005). In particular, the increased pre-SMA activity in response to irrelevant reward-related words might reflect the controlled selection of the correct response to overcome automatic response tendencies (Nachev et al., 2005; Isoda and Hikosaka, 2007) that have been amplified by implicitly transferred reward associations. According to this view, the correct motor output comes at the cost of delaying the responses (van Gaal et al., 2010, 2011), which is consistent with the observed positive relationship between pre-SMA activity and response slowing associated with reward-related incongruent words in the present data. The concomitant engagement of the right dlPFC in these trials presumably reflects reactive processes to help enforce the relevant task representation against the irrelevant one (Miller and Cohen, 2001; Egner and Hirsch, 2005).

Despite their prominent effects in cortical areas, there was no evidence that irrelevant reward associations were related to actual reward anticipation. The NAcc was virtually silent in these trials, which is in line with proposals that its activation strongly depends on the actual behavioral relevance of reward-related stimuli (Zink et al., 2004; Bjork and Hommer, 2007). The lack of robust neural reward-anticipation responses may thus indicate that reward-related words in the present paradigm acquire some kind of low-level saliency that is independent of the original reward context.

Considering that word reading is a highly overlearned process in itself, it appears possible that this low-level saliency may trigger the associated response-selection pathways in a rather automatic fashion, thereby bypassing not only reward-processing regions but also conscious awareness of response conflict. In line with this notion, it has been demonstrated that the motor-control functions of the pre-SMA, and in particular the inhibition and replacement of automatic responses, are even evident under subliminal conflict situations (Sumner et al., 2007; Sumner and Husain, 2008; van Gaal et al., 2010, 2011).

Of note, our analysis focused on the effects of irrelevant reward associations on trials in which no actual reward was at stake, thereby avoiding interaction effects between the processing of relevant and irrelevant reward associations in the same trial. Although beyond our main focus, it is likely that the processing of irrelevant reward signals is different in trials in which cognitive-control functions, and consequently performance, are concomitantly promoted by relevant reward (Pochon et al., 2002; Locke and Braver, 2008). Hence, responses to the relevant, and therefore truly reward-predicting, stimulus may be less vulnerable to any interference from the irrelevant dimension. This notion is supported by the observed voxelwise interaction in several areas that were part of the overlapping networks in the independent comparisons—the dlPFC, the FG, and the IPC. In all three of these regions, the interaction was driven by a relatively larger activity increase for incongruent reward-related words in no-reward as compared with reward trials.

Conclusions

The present study investigated the neural underpinnings of the recent behavioral observations that reward associations, which are meant to enhance performance for certain facets of a task, can concomitantly disrupt behavior for others (Pessoa, 2009; Padmala and Pessoa, 2010; Krebs et al., 2010; Rutherford et al., 2010). The present neural data indicate that reward associations influence low-level visual-processing and subsequent response-selection pathways relatively automatically, even when they are only implicitly transferred and entirely irrelevant to the task. Together, these results uncover a fundamental flipside of reward associations on behavior and neural processing. Accordingly, they may also provide valuable insights into the mechanisms at play in a derailed reward system, such as in addiction (van Ree et al., 1999), where overrepresented reward signals (Bradley et al., 2003; Heinz et al., 2004) can substantially disrupt behavior and have major clinical ramifications.

References

Banich MT, Milham MP, Jacobson BL, Webb A, Wszalek T, Cohen NJ, Kramer AF (2001) Attentional selection and the processing of task-irrelevant information: insights from fMRI examinations of the Stroop task. Prog Brain Res 134:459–470.

Bjork JM, Hommer DW (2007) Anticipating instrumentally obtained and passively received rewards: a factorial fMRI investigation. Behav Brain Res 177:165–170.

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD (2001) Conflict monitoring and cognitive control. Psychol Rev 108:624–652.

Bradley BP, Mogg K, Wright T, Field M (2003) Attentional bias in drug dependence: vigilance for cigarette-related cues in smokers. Psychol Addict Behav 17:66–72.

Braver TS, Gray JR, Burgess GC (2007) Explaining the many varieties of working memory variation: dual mechanisms of cognitive control. New York: Oxford UP.

Brett M, Anton J-L, Valabregue R, Poline J-P (2002) Region of interest analysis using an SPM toolbox (abstract). Available on CD-Rom. Neuroimage 16.

Chen Q, Wei P, Zhou X (2006) Distinct neural correlates for resolving Stroop conflict at inhibited and noninhibited locations in inhibition of return. J Cogn Neurosci 18:1937–1946.

Cohen JD, Dunbar K, McClelland JL (1990) On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psychol Rev 97:332–361.

de Fockert J, Rees G, Frith C, Lavie N (2004) Neural correlates of attentional capture in visual search. J Cogn Neurosci 16:751–759.

Egner T, Hirsch J (2005) Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. Nat Neurosci 8:1784–1790.

Engelmann JB, Damaraju E, Padmala S, Pessoa L (2009) Combined effects of attention and motivation on visual task performance: transient and sustained motivational effects. Front Hum Neurosci 3:4.

Friston K, Holmes AP, Worsley KJ, Poline J-B, Frith CD, Frackowiak RS (1995) Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2:189–210.

Garavan H, Ross TJ, Kaufman J, Stein EA (2003) A midline dissociation between error-processing and response-conflict monitoring. Neuroimage 20:1132–1139.

Goto Y, Grace AA (2005) Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. Nat Neurosci 8:805–812.

Heinz A, Siessmeier T, Wrase J, Hermann D, Klein S, Grüsser SM, Flor H, Braus DF, Buchholz HG, Gründer G, Schreckenberger M, Smolka MN, Rösch F, Mann K, Bartenstein P (2004) Correlation between dopamine D2 receptors in the ventral striatum and central processing of alcohol cues and craving. Am J Psychiatry 161:1783–1789.

Hinrichs H, Scholz M, Tempelmann C, Woldorff MG, Dale AM, Heinze HJ (2000) Deconvolution of event-related fMRI responses in fast-rate experimental designs: tracking amplitude variations. J Cogn Neurosci 12:76–89.

Isoda M, Hikosaka O (2007) Switching from automatic to controlled action by monkey medial frontal cortex. Nat Neurosci 10:240–248.

Kerns JG, Cohen JD, MacDonald AW 3rd, Cho RY, Stenger VA, Carter CS (2004) Anterior cingulate conflict monitoring and adjustments in control. Science 303:1023–1026.

Kim H, Shimojo S, O'Doherty JP (2006) Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. PLoS Biology 4:e233.
Kiss M, Driver J, Eimer M (2009) Reward priority of visual target singletons modulates event-related potential signatures of attentional selection. Psychol Sci 20:245–251.
Knutson B, Cooper JC (2005) Functional magnetic resonance imaging of reward prediction. Curr Opin Neurol 18:411–417.
Knutson B, Adams CM, Fong GW, Hommer D (2001a) Anticipation of increasing monetary reward selectively recruits nucleus accumbens. J Neurosci 21:RC159.
Knutson B, Fong GW, Adams CM, Varner JL, Hommer D (2001b) Dissociation of reward anticipation and outcome with event-related fMRI. Neuroreport 12:3683–3687.
Koechlin E, Basso G, Pietrini P, Panzer S, Grafman J (1999) The role of the anterior prefrontal cortex in human cognition. Nature 399:148–151.
Kouneiher F, Charron S, Koechlin E (2009) Motivation and cognitive control in the human prefrontal cortex. Nat Neurosci 12:939–945.
Krebs RM, Boehler CN, Woldorff MG (2010) The influence of reward associations on conflict processing in the Stroop task. Cognition 117:341–347.
Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540.
Lee TS, Yang CF, Romero RD, Mumford D (2002) Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency. Nat Neurosci 5:589–597.
Locke HS, Braver TS (2008) Motivational influences on cognitive control: behavior, brain activation, and individual differences. Cogn Affect Behav Neurosci 8:99–112.
MacLeod CM (1991) Half a century of research on the Stroop effect: an integrative review. Psychol Bull 109:163–203.
McCandliss BD, Cohen L, Dehaene S (2003) The visual word form area: expertise for reading in the fusiform gyrus. Trends Cogn Sci 7:293–299.
McKeefry DJ, Zeki S (1997) The position and topography of the human colour centre as revealed by functional magnetic resonance imaging. Brain 120:2229–2242.
Milham MP, Banich MT, Webb A, Barad V, Cohen NJ, Wszalek T, Kramer AF (2001) The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. Brain Res Cogn Brain Res 12:467–473.
Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.
Nachev P, Rees G, Parton A, Kennard C, Husain M (2005) Volition and conflict in human medial frontal cortex. Curr Biol 15:122–128.
O'Craven KM, Downing PE, Kanwisher N (1999) fMRI evidence for objects as the units of attentional selection. Nature 401:584–587.
O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454.
Padmala S, Pessoa L (2010) Interactions between cognition and motivation during response inhibition. Neuropsychologia 48:558–565.
Pessoa L (2009) How do emotion and motivation direct executive control? Trends Cogn Sci 13:160–166.
Pessoa L, Engelmann JB (2010) Embedding reward signals into perception and cognition. Front Neurosci 4:17.
Pochon JB, Levy R, Fossati P, Lehericy S, Poline JB, Pillon B, Le Bihan D, Dubois B (2002) The neural system that bridges reward and cognition in humans: an fMRI study. Proc Natl Acad Sci U S A 99:5669–5674.
Ramnani N, Miall RC (2003) Instructed delay activity in the human prefrontal cortex is modulated by monetary reward expectation. Cereb Cortex 13:318–327.
Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S (2004) The role of the medial frontal cortex in cognitive control. Science 306:443–447.
Rushworth MF, Hadland KA, Paus T, Sipila PK (2002) Role of the human medial frontal cortex in task switching: a combined fMRI and TMS study. J Neurophysiol 87:2577–2592.
Rushworth MF, Walton ME, Kennerley SW, Bannerman DM (2004) Action sets and decisions in the medial frontal cortex. Trends Cogn Sci 8:410–417.
Rutherford HJ, O'Brien JL, Raymond JE (2010) Value associations of irrelevant stimuli modify rapid visual orienting. Psychon Bull Rev 17:536–542.
Schultz W (2002) Getting formal with dopamine and reward. Neuron 36:241–263.
Serences JT (2008) Value-based modulations in human visual cortex. Neuron 60:1169–1181.
Small DM, Gitelman D, Simmons K, Bloise SM, Parrish T, Mesulam MM (2005) Monetary incentives enhance processing in brain regions mediating top-down control of attention. Cereb Cortex 15:1855–1865.
Sumner P, Husain M (2008) At the edge of consciousness: automatic motor activation and voluntary control. Neuroscientist 14:474–486.
Sumner P, Nachev P, Morris P, Peters AM, Jackson SR, Kennard C, Husain M (2007) Human medial frontal cortex mediates unconscious inhibition of voluntary action. Neuron 54:697–711.
Taylor SF, Welsh RC, Wager TD, Phan KL, Fitzgerald KD, Gehring WJ (2004) A functional neuroimaging study of motivation and executive function. Neuroimage 21:1045–1054.
Ullsperger M, von Cramon DY (2001) Subprocesses of performance monitoring: a dissociation of error processing and response competition revealed by event-related fMRI and ERPs. Neuroimage 14:1387–1401.
van Gaal S, Ridderinkhof KR, Scholte HS, Lamme VA (2010) Unconscious activation of the prefrontal no-go network. J Neurosci 30:4143–4150.
van Gaal S, Scholte HS, Lamme VA, Fahrenfort JJ, Ridderinkhof KR (2011) Pre-SMA gray-matter density predicts individual differences in action selection in the face of conscious and unconscious response conflict. J Cogn Neurosci 23:382–390.
van Ree JM, Gerrits MA, Vanderschuren LJ (1999) Opioids, reward and addiction: an encounter of biology, psychology, and medicine. Pharmacol Rev 51:341–396.
van Veen V, Cohen JD, Botvinick MM, Stenger VA, Carter CS (2001) Anterior cingulate cortex, conflict monitoring, and levels of processing. Neuroimage 14:1302–1308.
Vorobiev V, Govoni P, Rizzolatti G, Matelli M, Luppino G (1998) Parcellation of human mesial area 6: cytoarchitectonic evidence for three separate areas. Eur J Neurosci 10:2199–2203.
Vuilleumier P (2005) How brains beware: neural mechanisms of emotional attention. Trends Cogn Sci 9:585–594.
Wise RA (2004) Dopamine, learning and motivation. Nat Rev Neurosci 5:483–494.
Zink CF, Pagnoni G, Martin-Skurski ME, Chappelow JC, Berns GS (2004) Human striatal responses to monetary reward depend on saliency. Neuron 42:509–517.
Zysset S, Müller K, Lohmann G, von Cramon DY (2001) Color–word matching Stroop task: separating interference and response conflict. Neuroimage 13:29–36.