Research report

Outcome representations, counterfactual comparisons and the human orbitofrontal cortex: Implications for neuroimaging studies of decision-making

Stefan Ursu*, Cameron S. Carter

Department of Psychiatry, University of California, Davis, CA 95616, USA

Accepted 4 January 2005
Available online 16 February 2005

Abstract

Recent research suggests that the primate orbitofrontal cortex (OFC) is critical for representations of outcomes of actions and their subsequent impact on the control of behavior. In parallel, a recent theory of decision-making called decision affect theory (Mellers, Schwartz, and Ritov, *Psychological Science*, 1997) emphasizes the role of anticipated affective impact of outcomes in guiding choices, and the effects of comparisons with alternative outcomes (i.e., counterfactual effects). In the context of decision affect theory, we present results from two event-related functional MRI experiments consistent with two hypotheses regarding the role of the human OFC in guiding behavior through outcome representation: (1) counterfactual effects are manifested in the human OFC during expectation of outcomes, such that the anticipated affective impact of outcomes is modulated by the nature of the various possible alternative outcomes; (2) a regional specialization exists in the human prefrontal cortex, such that affective impact of potential negative outcomes of actions is represented mainly by the lateral areas of the OFC, while areas situated progressively more medial and dorsal on the ventral and medial PFC are specifically involved in representing the impact of positively valenced outcomes. We also discuss some of the implications that these hypotheses have for neuroimaging studies of reward processing and decision-making, and for studies of neuropsychiatric disorders in which these processes are thought to be disturbed.

© 2005 Elsevier B.V. All rights reserved.

Theme: Neural basis of behavior
Topic: Motivation and emotion

Keywords: Orbitofrontal cortex; Decision-making; Reward; fMRI; Motivation

1. Introduction

The adaptive control of behavior is the result of multiple cognitive processes, including representation of goals, maintenance and manipulation of task rules associated with any particular goal, selection of responses from an array of possibilities, and execution of the selected response. The term “executive control” can be used to refer to any or all of these types of processes. While the degree of separation between cognitive and emotional/motivational processes is a matter of considerable debate, there is strong support for the claim that motivational and emotional factors interact with cognition [25]. This interaction occurs in the implementation of executive control functions, presumably by modulating the strength or quality of goal representation and by influencing the decision-making and task execution processes. In terms of the neural underpinnings of these processes, all the aforementioned aspects of behavioral regulation have been at some point linked to the prefrontal cortex (PFC).

While it is widely accepted that the PFC plays a critical role in executive control of behavior, the precise mechanisms through which it implements control are unclear.
Consequently, the functional organization within the PFC continues to be the focus of extensive research. Of most interest to us are the differences in specialization of the dorsolateral vs. ventral areas of the PFC. A long standing view has been that the dorsal and lateral PFC are involved in mnemonic and attentional processes, while the orbital areas of the PFC are necessary for inhibitory function [19]. A more recent line of research suggests that inhibitory functions are a general property of the PFC, but that the type of information being inhibited differs between the dorsolateral PFC (DLPFC) and the orbitofrontal cortex (OFC) [38,39]. Theoretical work using computational modeling proposes that the prefrontal cortex is involved in the control of behavior by actively maintaining patterns of activity that represent task goals and the intermediate steps necessary to achieve them [30]. The “executive” aspect of the activity in these prefrontal neural networks may be exerted through a biasing effect on the activity of other subsystems involved in task execution at various levels, for example, sensory processing, mnemonic operations (recognition, retrieval, etc), response selection. Thus, the PFC implements executive control of cognitive processes by biasing the “flow of neural activity” through pathways which ensure appropriate mapping of inputs onto outputs, with the ultimate goal of reaching optimal internal states [30]. In other words, the PFC is thought to ensure that patterns of activation in networks directly involved in task execution are appropriate relative to the specific context provided by current goals and task rules.

As mentioned before, the involvement of specific areas of the PFC in aspects of executive functions is still a matter of active investigation. However, in the last decade or so, results generated by several lines of investigation have suggested that the orbital part of the PFC is consistently involved in tasks in which performance is modulated by emotional/motivational factors [1,41,50]. Such reinforcer-induced changes in OFC activity have been documented in electrophysiological experiments on awake behaving monkeys [21,40,42,44,46,50,51] and in human neuroimaging studies [7,18,33,35,49]. The proposed involvement of the OFC in reward-based behavioral control has also been supported by patterns of deficits in humans with ventral prefrontal lesions. These patients show perseveration in high-reward, high-risk choices despite overall losses incurred as a result of those choices [1,2,4,22,43]. Several specific mechanisms through which the OFC is involved in control of behavior based on rewards and punishments have been proposed, including representations of future consequences of actions [1], reversals of stimulus–reinforcer associations [16,41,44], and inhibition of irrelevant stimulus–reward associations [37]. These findings, as well as our own results from a series of fMRI experiments [52,53], are consistent with the idea that the human OFC is involved in executive control of behavior by representing the expected rewarding or punishing effects of potential outcomes of actions. Such a function accounts well for findings that humans with lesions of the OFC are impaired in tasks in which optimal performance depends on processing of reinforcers, but are unimpaired in tasks which rely mainly on maintenance and manipulation of other task-relevant information [3,5]. Furthermore, human imaging studies have demonstrated that anticipation and experience of pleasant and unpleasant stimuli like tastes or foods, odors, touch, and faces, modulate activity in the OFC [11,32,34,45,47]. Thus, the hypothesis that the OFC is involved in representing potential outcomes can account for a wide range of existing results from studies of motivational processes.

In this article, we will argue that tests of this hypothesis, especially when functional neuroimaging methods are used to measure indices of brain activation, should also take into account the effects of counterfactual outcomes on anticipated or experienced emotions. Counterfactual effects, studied systematically in a series of behavioral economics experiments [26,28,29], can be readily exemplified by referring to a situation frequently encountered in real life: a neutral outcome (e.g., no win, no loss) elicits significantly different emotional reactions depending whether the alternative outcome under a different state of the world would have been a gain or a loss. Similarly, when subjects are informed with regards to the nature of the alternative outcome of a given decision, the affective reaction to gains or losses resulting from that decision can change significantly, depending on the nature of the alternative outcome. For instance, when subjects make a risky decision which results in a loss of $5, they report feeling worse when the alternative outcome would have been losing only $3 compared to when the alternative would have been losing $12. Conversely, a $5 gain feels better if the unobtained outcome was winning only $3 compared to when it was winning $12 [26]. In extreme cases, counterfactual comparisons may render the affective impact of outcomes paradoxical, such as when small actual losses (incurred in the context of an alternative possible large loss) feel slightly better than small actual gains (when the alternative is a much larger gain that was not obtained). The emotional impact of “what could have been” has been formally represented in a recent theory of choice called decision affect theory [28]. Decision affect theory has successfully described the pleasure and pain of anticipated and experienced outcomes of decisions in a wide range of experimental and observational studies [23,27,29].

Only a few studies to date have investigated the neural substrates of counterfactual comparisons. Nonetheless, several experiments in humans and non-human primates have generated results consistent with the idea that activity in structures of the limbic and paralimbic circuits parallels the counterfactual comparisons of outcomes (i.e., the activity changes not only depending on the potential outcome of the current trial but also depending on the nature of possible alternative outcomes). In one study,
conducted by Breiter et al. [7], subjects were imaged while performing a “wheel of fortune” type task. Changes in blood oxygenation level dependent (BOLD) signal in the nucleus accumbens and the sublenticular extended amygdala in response to a neutral monetary outcome (i.e., no gain, no loss) were relatively stronger if the counterfactual outcome was a loss than if the counterfactual outcome was a gain. The fact that Breiter et al. [7] did not find evidence for counterfactual effects of prospects or outcomes on the activity of the OFC is not necessarily surprising given that such effects are most likely relatively subtle and imaging the OFC using fMRI poses additional challenges compared to most other areas of the brain.

A few other studies have also generated results consistent with a link between the OFC and the effects of counterfactual outcomes on the representation of various reinforcers. For instance, in a study in non-human primates, Tremblay and Schultz [50] recorded the activity of OFC neurons in response to stimuli predicting future availability of rewards. On each trial, monkeys were presented with pairs of instruction cues (i.e., pictures), and were required to choose one of the two pictures in order to receive the reward associated with it. Three different rewards were used in the experiment (e.g., A, B and C), and the monkeys’ choices of cues indicated a clear preference order over rewards (for example A more preferred than B, and B more preferred than C). In each block of trials, monkeys had to discriminate and choose pictures associated with only two of the three rewards (A and B, B and C, or A and C). In this study, Tremblay and Schultz [50] identified a population of OFC neurons which responded selectively to cues predicting the preferred reward in a given block (for example, they fired more to the instructional cue which predicted the availability of reward A than to the instruction which predicted reward B). Interestingly, in blocks in which the monkey had to respond to cues which predicted reward B or C, the same neurons which fired more to reward A cues than to reward B cues, now fired more to presentation of the reward B cues, since reward B was the preferred one relative to the alternative (reward C). Therefore, the changes in activity in these OFC neurons not only reflected the monkey’s behavioral preference for any given reward, but this preference was developed relative to the range of the choices immediately available (i.e., depending on what the alternative reward was in a given choice).

A third piece of evidence in support of the involvement of the OFC in counterfactual thinking comes from a recent study comparing, in a gambling task, the emotional reaction of healthy subjects and of patients with lesions of the OFC [8]. The task required subjects to choose, in each trial, between playing one of two gambles with varying possible outcomes (200, 50, −50, and −200) and varying probabilities (0.2, 0.8, and 0.5). In some of the trials, subjects found out only the outcome of the chosen gamble (partial feedback), whereas in others they could observe the outcomes of both the chosen and the rejected gamble. When full feedback was available, the ratings by healthy subjects of the emotional effect of obtained outcomes (as well as skin conductance responses to outcome presentation) were modulated by the outcome of an unchosen gamble, whereas the ratings of patients with ventromedial PFC lesions were not affected by the nature of the outcome of the rejected gamble. Authors interpreted these results as evidence that the OFC contributes critically to decision-making by supporting the counterfactual processes leading to the experience of regret, which in turn exert top-down biases on the processes leading to choices.

Taken together, the studies mentioned above are consistent with the hypothesis that the OFC carries out complex computations in which affective and cognitive factors interact to modulate the emotional effects of decision outcomes, which in turn influence the nature of decisions. If correct, this hypothesis implies that in studying the neural substrates of decision-making, and in general of reward- or punishment-related processes, one needs to carefully consider the possible counterfactual effects of available reinforcers when examining the changes in brain activity. These effects would become especially important if their presence could lead to changes in activity (relative to a neutral baseline) that are qualitatively different from those expected in the absence of counterfactual comparisons. In the following sections, we will illustrate this idea with our own results from a recent series of fMRI studies. We will provide evidence that the activity in the human OFC is consistent with a role in representing prospectively the affective impact of outcomes of actions. Since a detailed discussion of possible alternative accounts of OFC function (such as involvement in various aspects of cognitive control) is presented elsewhere [55], this paper will focus on data suggesting that representations of anticipated outcomes are modulated by the full range of alternative outcomes that may be expected in the context of a given task.

2. Methods

We have recently conducted two event-related fMRI experiments examining the neural substrates of the modulation of cognitive control processes by motivational factors in 36 young healthy subjects ($N = 19$ in the first experiment, $N = 17$ in the second) [52,53,55]. In both studies, subjects were presented with a cue stimulus, which had to be retained over a 12-s interval in order to respond correctly to a subsequent target stimulus (Fig. 1). Accurate performance at the time of probe presentation depended on the correct identification of the shape of the cue. However, the task requirements of responding to the target were completely orthogonal to the motivational manipulation and therefore is not detailed here. Independent of stimulus shape, the color of the stimulus informed subjects of the monetary incentives associated with a correct and fast response to the probe. In Experiment 1, one color informed subjects that a $0.5
reward would be obtained for a correct and fast response
(Reward), while a second color indicated that no reward was
available (No-reward). In the second experiment, a third
color was added, that is, one color signaled that $1 would be
obtained for correct and fast responding to the trial probe
(Reward), a second color meant that $1 would be lost upon
incorrect or late responding (Penalty), and the third color
indicated that no money would be won or lost in that trial
(No-reward). Additionally, in Experiment 1, the deadline for
a fast response was always 1 s; in Experiment 2, the
pressure for speeded responses was implemented through
the use of an individualized dynamic response deadline
which, unbeknownst to subjects, decreased when they made
repeated correct and fast responses to incentive trials and
increased after errors or late correct responses [53,55]. In
both experiments, subjects were informed that the response deadline was slightly shorter than the target
presentation time and that late responses counted as errors (see Methods). Immediately after the offset of the target, feedback was presented for 0.5-s, followed
by 10.5-s of inter-trial interval. Functional scans were acquired every 1.5-s, synchronized with the stimulus onset. The analyses presented in the paper are
focused on the scans covering the interval between cue and target stimuli (i.e., S01–S08).

Fig. 1. Both studies used a slow event-related design. Each trial consisted of 1-s presentation of a cue stimulus (a single star), an 11-s cue-target interval, and a
1-s presentation of the target stimulus (two stars presented side by side, one of which was always identical to the sample stimulus). Subjects were instructed to
remember the cue and, upon target presentation, respond by pressing one of two buttons (on the same side or on the opposite side of the stimulus that matched
the cue; this specific response rule was given at the beginning of each block of trials). Randomized from trial to trial, the color of the cue indicated the monetary
consequences of the response to the target. In Experiment 1, two colors were used: one meant that $0.50 would be won for a correct response (Reward trials),
and the other color that no monetary reward will be available (No-reward trials). Subjects were also told that the response deadline was 1-s and late responses
would count as errors, and therefore, they should be as fast and as accurate as possible. In Experiment 2, three colors were used. Two colors coded for the same
type of trials as in the first experiment (a $1 Reward or No-reward), while the third indicated Penalty (i.e., incorrect or slow responses would result in a $1 loss
from their winnings, and no reward would be won for correct responses). Subjects were informed that the response deadline was slightly shorter than the target
presentation time and that late responses counted as errors (see Methods). Immediately after the offset of the target, feedback was presented for 0.5-s, followed
by 10.5-s of inter-trial interval. Functional scans were acquired every 1.5-s, synchronized with the stimulus onset. The analyses presented in the paper are
focused on the scans covering the interval between cue and target stimuli (i.e., S01–S08).

Functional imaging parameters are detailed elsewhere
[55], but it is important to note that a reverse spiral
acquisition protocol was used in order to reduce magnetic
susceptibility effects in the OFC [20]. Data were analyzed
using an ANOVA model using subject as random factor,
MR signal as dependent variable, and incentive and scan as
repeated measures [52,54]. Voxels were defined as differentially active under changing incentive conditions if the Incentive by Scan interaction (for the 8 scans covering the anticipatory cue-target interval) was significant. A cluster-size threshold was imposed on the activations in order to correct for multiple comparisons at the 0.05 level. The directionality of effects were verified by computing the timeseries of the regions of interest and conducting planned t tests of the signal change (relative to the first scan of the trial). These tests were conducted for the middle scan (scans 2 through 7) with the largest difference between the mean signal change of the conditions of interest, and only if at least one of the conditions showed a positive mean of the signal change. In one case, when the two condition means reversed order in the later part of the interval, a test was conducted for each part of the interval using the same criteria as above. All P values reported are two tailed unless otherwise specified.

3. Results and discussion

During the preparatory cue-target, Experiment 1 identified areas of the lateral OFC (BA 11/47) in which cues signaling absence of monetary incentives elicited more activity relative to incentive cues \[t(18) = 2.47, P = 0.02, \text{ see Fig. 2a}\]. This was the first piece of evidence consistent with a counterfactual effect of a neutral outcome since, in this experiment, the high probability of obtaining the reward in incentive trials may turn the lack of opportunity for monetary gains into a relatively punishing event. Our results suggested that the specificity of the lateral OFC for potential punishment is manifested early on, during the expectancy phase of the task, when negative outcomes are represented prospectively. This hypothesis was confirmed by results of the second experiment, when the same area of the lateral OFC (BA 11/47) was maximally engaged after presentation of cues signaling potential penalties \[\text{Penalty} > \text{Reward}, \text{ see Fig. 2b}\].

![Fig. 2. Timecourse of the BOLD signal averaged across all voxels in the left lateral OFC regions of interest. Asterisks mark the scans for which pairwise tests of effects were conducted (see Methods). (a) Experiment 1: activity in the left lateral OFC (BA11/47, MNI coordinates of the peak voxel: \(-24, 38, -16\)) increased in response to cues signaling the absence of rewards for correct and fast responses. (b) Experiment 2: the same area of the lateral OFC (MNI coordinates of the peak: \(-31, 30, -10\)) was maximally activated by Penalty cues. In this experiment, however, Reward cues also induced positive signal changes, consistent with anticipation of counterfactual negative affect in response to potential missed rewards.]

According to decision affect theory [27–29], the lateral OFC reacts specifically to the potential for negative outcomes, examination of the pattern of activation during the reward trials in the two studies provided additional evidence for the presence of counterfactual effects. Overall, the magnitude of available rewards (0.5 and 0.7 in Experiment 1 and 2, respectively), and the frequency of their occurrence (as a function of frequency of rewarded responses, approximately 0.90 and 0.66, respectively), the average expected gain per reward trial increased from $0.45 per trial in Experiment 1 to $0.66 in Experiment 2. Therefore, if the lateral OFC activity simply reflected the expected impact of negative outcomes based only on the current trial, it should be inversely related to the expected gain of each trial, and consequently, the signal changes elicited by reward cues should decrease in the second experiment relative to the first one. In fact, we observed exactly the opposite, as can be seen from comparing the time course of activation of reward trials in Fig. 2a with the time course in Fig. 2b. In other words, since this brain area activated maximally during the trials with the worst outcomes, the activity during Reward trials of Experiment 2 is hard to reconcile with classical utility theories, which posit the status quo (i.e., current assets) as the only reference point in computing the expected utility of an outcome. In contrast, this pattern of activation fits well with predictions generated by decision affect theory [27–29].

According to this theory, in Experiment 2, subjects might expect to feel worse during the preparatory phase of reward trials (relative to reward trials of Experiment 1) if they consider the other incentive conditions present in the task. Indeed, missing a reward in Experiment 2 had a greater impact on the total winnings due to the presence of the Penalty trials (which tended to diminish overall gains).

Obviously, these comparisons with the predictions of decision affect theory should be regarded as preliminary, since detailed debriefing of subjects was not available in one of the experiments, and because of differences in task and assessment methods relative to previous studies of decision-making [26,28,29]. Nevertheless, it is remarkable that, aside from this aspect of the data which relies on across-group comparisons, within-subject effects were also consistent with the idea that multiple conditions (i.e., incentive levels) and their likely outcomes influenced the affective impact of the current trial/outcome. Most notably, both the ratings of negative affect and the relative ordering of cue-related lateral OFC activity in Experiment 2 was inconsistent with a simple inverse relationship with the expected utility of the current trial (which would predict a monotonic increase in activity from Reward to No-reward to Penalty). Specifically, 14 out of 17 subjects rated neutral outcomes in Reward trials (i.e., missed rewards) as having a negative emotional impact (3 subjects rated them as neutral). Furthermore, only 7 out of 17 subjects rated actual losses as more negative than missed rewards, while 7 subjects rated them equally negative and 3 subjects rated missed rewards as more upsetting than actual losses. Given that error feedback in No-reward trials was rated neutral by 13 of the 17 subjects (and only mildly negative by the other four), the lateral OFC activity seemed to follow closely these affective ratings.

The aforementioned results suggest that the lateral orbitofrontal cortical areas are involved in representing the anticipated feelings of disappointment associated with the outcomes of actions. Such a functional specialization would in turn argue for at least partial segregation of the brain circuits subserving avoidance behaviors from those underlying approach. Some other aspects of signal changes observed in our experiments support such a segregation. Since previous studies have suggested that the medial aspects of the OFC are specifically involved in processing of rewards [24,33,47], we also examined the changes in BOLD responses in prefrontal regions of activation situated progressively more medially and dorsally. Consistent with a bias toward processing of positively valenced outcomes, in both experiments, the BOLD signals in the ventro-medial PFC tended to peaked higher during the expectation phase of reward trials (Fig. 3). In Experiment 1, a region of interest in BA 32/10 showed a trend for higher activity in Reward trials relative to No-reward ($t(18) = 1.95, P = 0.09$). In Experiment 2, we examined two regions of activation on the ventro-medial PFC. In BA 32/10, paired contrasts revealed trends for greater signal changes in Reward trials: Reward > No-reward ($t(16) = 1.81 P = 0.09$), Reward > Penalty [$t(16) = 0.81 P = 0.43$], Penalty > No-reward [$t(16) = 0.66 P = 0.52$]. In a more dorsal and caudal medial prefrontal area (BA 32), Reward > No-reward [$t(16) = 2.81 P = 0.01$], Reward > Penalty [$t(16) = 2.42 P = 0.03$], Penalty > No-reward [$t(16) = 1.70 P = 0.11$]. Even though some of these differences did not reach statistical significance, a qualitative trend can be noted in the pattern of activations in regions located progressively more medial and more dorsal on the medial wall. As can be seen in Fig. 3, the increase in activity during Reward trials relative to either Neutral or Penalty trials seemed to occur in the ventro-medial foci of activation which included subgenual cingulate cortex and ventro-medial Brodmann’s area 10 (Figs. 3a and b, top graph), and reached statistical significance in post hoc tests in more dorsal areas of the medial wall (see Fig. 3b, bottom graph).

The clear trend of the activity in these areas to be driven by the prospect of positive outcomes in both experiments is the basis for the argument that counterfactual effects of alternative outcomes were manifested in these areas as well. Analogous to the case made for anticipated negative
emotionality and the lateral OFC activation, the neural correlates of anticipated positive feelings might be expected to be activated, counterfactually, by the prospect of avoiding a penalty. This parallel between the relative ordering of activity in the three incentive levels, and ratings of positive affect, seemed to also be tentatively
supported by our data. Indeed, 8 out of 17 subjects rated avoiding a loss in a Penalty trial (which happened with the same frequency as gains in Reward trials) as equally positive or more positive than gains, consistent with the counterfactual effects described by decision affect theory. In parallel, the BOLD signal changes in the ventro-medial prefrontal cortex followed the same pattern, with anticipatory activity during the Penalty trials being lower than Reward but higher than No-reward activity (i.e., in the more dorsal BA32), or statistically equivalent to Reward and No-reward activity (such as in BA32/10). Even in the case of the late increase in activity during No-reward trials in the ventromedial BA32/10 (Fig. 3b), one compelling (albeit post hoc) account would be consistent with the presence of counterfactual effects of alternative outcomes. Namely, if activity in this area reflects positive affect which results from interactions of expected and alternative outcomes, it is possible that, as the time of responding to targets approached, the prospect of alternative outcomes becomes more salient. This would in turn make the prospect of negative affect more likely in Reward and Penalty trials (i.e., after missed rewards or losses), but not during No-reward trials, thus explaining the relatively increased activity seen in the No-reward condition.

In summary, evidence accumulates suggesting that effects of counterfactual comparisons can be detected not only in behavior but also in its neural substrates. It is important that such effects should be carefully considered in future studies investigating the neural underpinnings of reward-related processes and, in particular, of decision-making. Indeed, it is possible that the influence of alternative possible outcomes is present in results of existing studies, and accounts for some of the apparent discrepancies between various reports. For instance, probabilistic reward delivery has frequently been used [13,18,36], while other studies used as incentives, in the same trial, the prospect of a monetary reward for correct performance and of a punishment for incorrect responses [48]. It is therefore possible, in such designs, that some of the brain activity observed in response to high-reward trials to reflect the anticipated affective impact of the alternative outcome (missing a reward or facing a punishment, respectively). Similarly, the effects of reinforcing feedback should be evaluated in the context of the overall background set by the outcomes with the greatest subjective affective impact, which may be a function of their magnitude, probability, or both [6]. Furthermore, one can even speculate that the link between the necessarily two-faceted nature of reinforcement-related phenomena (e.g., that avoiding an aversive stimulus has rewarding properties) interacts closely with counterfactual comparisons.

Consequently, the issue of counterfactual comparisons becomes particularly germane to any area of research for which the degree of overlap between neural basis of approach and avoidance behaviors is relevant. For instance, it should be carefully considered when interpreting results of studies using complex gambling tasks [1,2], in which the impact of both rewards and punishments is critical for optimal performance. Moreover, numerous models have proposed a central role for disturbed emotional and/or motivational processes in pathogenesis of important neuropsychiatric disorders such as mood disorders [10,14,15,17], anxiety disorders [31], addiction [12,56], and impulse control [9]. However, it appears that the nature of the alternative possible outcomes may lead to non-intuitive changes in indices of brain activity to any given reinforcer, and therefore the design of experiments should include control conditions to ensure that appropriate baselines are used as reference in computing activity changes. Alternatively (or, ideally, in addition to this), careful debriefing of subjects should also help constrain the interpretation of results from imaging experiments. Such considerations become particularly important since behavioral deficits which characterize the aforementioned psychopathologies may reflect not only dysfunctional representations and/or impact of the obtained outcomes but also abnormal counterfactual comparisons (i.e., increased or decreased focus on alternative outcomes).

Obviously, the characteristics and precise neural mechanisms underlying the counterfactual effects of alternative outcomes of actions will need to be studied in more detail. In doing so, future studies investigating these issues can benefit from the explanatory and predictive power of recent theories of decision making such as decision affect theory [28], which can inform and constrain hypotheses regarding the brain circuitry involved in reward processing and its disturbance in psychopathology [7].

Acknowledgment

Authors thank Barbara A. Mellers for valuable comments and suggestions on an earlier version of this manuscript.

References


