Published in final edited form as:
Neuron. 2013 May 8; 78(3): 563–573. doi:10.1016/j.neuron.2013.03.023.

Social Manipulation of Preference in the Human Brain

Keise Izuma[1,*] and **Ralph Adolphs**[1]

[1]Division of Humanities and Social Sciences, California Institute of Technology, 1200 E. California Blvd, Pasadena, CA, 91125, USA

SUMMARY

Our preferences are influenced by what other people like, but depend critically on how we feel about those people, a classical psychological effect called “cognitive balance”. Here we manipulated preferences for goods by telling participants the preferences of strongly liked or disliked groups of other people. Participants’ preferences converged to those of the liked group, but diverged from the disliked group. Activation of dorsomedial prefrontal cortex (dmPFC) tracked the discrepancy between one’s own preference and its social ideal, and was associated with subsequent preference change (towards the liked and away from the disliked group), even several months later. A follow-up study found overlapping activation in this same region of dmPFC with negative monetary outcomes, but no overlap with nearby activations induced by response conflict. A single social encounter can thus result in long-lasting preference change, a mechanism that recruits dmPFC and that may reflect the aversive nature of cognitive imbalance.

INTRODUCTION

Our preferences for goods are influenced by what other people prefer (Cialdini and Goldstein, 2004). As the internet has become pervasive, we are often exposed to information about what is popular among a certain group of people (e.g., one’s friends, one’s school, people in other countries, and so forth). Having preferences that are different from those of people we like, and having preferences that are similar to those of people that we dislike, are both undesirable. One would therefore expect that we not only change our preferences to be more similar to those of people we like, but also to be more dissimilar to those of people we dislike, an effect with a long history in social psychology dubbed “cognitive balance” (Heider, 1946, 1958). Since the theory was proposed more than five decades ago, it has stimulated a huge number of studies (see Abelson et al., 1968; Gawronski and Strack, 2012; Insko, 1984; Zajonc, 1968), and because of the theory’s wide applications, it has remained strongly influential across all of social psychology (see Greenwald et al., 2002; Walther & Weil, 2012). Although there have been several recent investigations of the neural mechanisms underlying how our preferences are influenced by the opinions of others in general (i.e., social conformity) (Berns et al., 2010; Campbell-Meiklejohn et al., 2010; Campbell-Meiklejohn et al., 2012a; Campbell-Meiklejohn et al., 2012b; Klucharev et al., 2009; Klucharev et al., 2011; Zaki et al. 2011), it remains unknown how the brain incorporates our attitude towards other people into such influence, and whether such effects are transient or more permanent. Elucidating a specific neural mechanism to explain the ubiquitous effect of other people’s opinions on our own preferences is of high relevance also

*Correspondence to: izuma@caltech.edu.

for marketing and advertising, and would importantly inform classical social psychology theories that are based on our need to maximize cognitive consistency and congruency (Abelson et al., 1968; Festinger, 1957; Gawronski and Strack, 2012; Greenwald et al., 2002; Heider, 1946, 1958; Osgood and Tannenbaum, 1955).

To investigate how people's preferences are influenced by those of others, we used the framework of balance theory (Heider, 1946, 1958), which states that our preferences for objects change so as to become similar to the preferences of people we like, and dissimilar to the preferences of people we dislike (Figure 1). In balance theory, balance depends on a triadic relationship between self, another person(s) and an object being evaluated (Figure 1a). A triadic relationship includes three attitudes: 1) one's attitude toward the other person, 2) one's attitude toward the object, and 3) the other person's attitude toward the same object. A balanced triad occurs when all the attitudes are positive, or two are negative and one is positive (for example, two individuals have a negative attitude towards an object, but they like each other). If a state is imbalanced, an individual is motivated to change one of their attitudes (either about the other person, or about the object) in order to restore balance (Heider, 1946, 1958).

In the present study, regional brain activation was measured with BOLD-fMRI while participants (N=18) rated their preferences for t-shirts (N=174; Figure 1b & c), saw other people's preferences for the same t-shirts (Figure 1d), and then re-evaluated the t-shirts a second time. Critically, we employed two groups of other people, validated to be strongly liked or disliked (see Supplemental Text for details): in some trials, participants were given feedback about how their fellow students rated the t-shirt ("students"; liked group) whereas in other trials, they found out how sex offenders rated it ("offenders"; disliked group). A further question of interest was whether the preference changes induced by cognitive imbalance in our study might persist. Four months after the fMRI study, we therefore asked participants (15/18) to rate the same t-shirts a third time.

We hypothesized that posterior medial frontal cortex (pMFC), which includes dorsal anterior cingulate cortex (dACC), dorsomedial prefrontal cortex (dmPFC) and pre-supplementary motor area (pre-SMA), plays a key role in representing imbalanced states and subsequent preference change. The pMFC is activated by a variety of aversive outcomes (Hikosaka and Isoda, 2010; Ridderinkhof et al., 2004; Shackman et al., 2011) including monetary loss (O'Doherty et al., 2003), reduced reward (Bush et al., 2002; Shima and Tanji, 1998) and negative feedback (Jocham et al., 2009; Miltner et al., 1997; Ullsperger and von Cramon, 2003), error detection (Falkenstein et al., 1991; Holroyd and Coles, 2002), physical and social pain (Eisenberger and Lieberman, 2004), as well as by demands for cognitive control (e.g., response conflict) (Botvinick et al., 2001). Several theories of pMFC function proposed to account for these diverse findings (Alexander and Brown, 2011; Botvinick, 2007; Botvinick et al., 2001; Holroyd and Coles, 2002; Rushworth et al., 2004; Yeung et al., 2004) all agree on an important role in detecting changes in the environment (especially negative ones including unfavorable outcomes and response conflict) that entail subsequent behavioral adjustment (Botvinick et al., 2001; Eisenberger and Lieberman, 2004; Holroyd and Coles, 2002; Ridderinkhof et al., 2004; Shackman et al., 2011; Ullsperger et al., 2004). It was also recently proposed that pMFC may be a key area in inducing palliative responses following various types of inconsistency described in social psychological theories (Proulx et al., 2012). While this idea of pMFC function generally fits well with balance theory, we nonetheless used a whole-brain approach in our initial analyses.

To provide initial insight into the neural processes corresponding to the hypothesized pMFC activation underlying balance theory, we also had our subjects complete two independent localizer tasks known from prior studies to activate sectors of this region. It could be argued

(Harmon-Jones, 2004) that cognitive imbalance constitutes a type of cognitive conflict, recruiting the pMFC for the same reason that this region is activated in simple speeded-response tasks such as Stroop or flanker tasks: the simultaneous activation of incompatible response representations (response conflict) (Botvinick et al., 2001). The other possibility, as originally stated in balance theory (Heider, 1958), is that cognitive imbalance might simply be aversive, recruiting the pMFC as a form of negative feedback or aversive outcome (Miltner et al., 1997; Ullsperger and von Cramon, 2003). These two possibilities may reflect functional dissociations mapping to anatomically distinct sectors of the pMFC (Hikosaka and Isoda, 2010; Ridderinkhof et al., 2004; Rushworth et al., 2004). To provide further insight, we used two independent localizer tasks to identify, within pMFC, areas previously implicated in processing response conflict (using the Multi-Source Interference Task or MSIT) (Bush and Shin, 2006) and areas involved in processing aversive outcomes (using the Monetary Incentive Delay Task or MIDT) (Knutson et al., 2000).

RESULTS

Behavioral results: Impression ratings

From the multiple attitude ratings that participants gave about the two social groups, all endorsed a positive attitude toward their fellow Caltech students (mean rating = 9.23) but a strongly negative attitude toward sex offenders (mean = 1.85, on a scale from 1 to 14). The difference in the impression ratings between two groups was highly significant ($t_{(17)} = 12.9$, $p < 0.001$).

Behavioral results: the effect of cognitive imbalance on preference

Participants' preferences for the t-shirts changed as predicted by balance theory (Heider, 1946, 1958): towards the preferences of the student group and away from those of the offender group (Figure 2a). Specifically, subjects' preferences for t-shirts they initially disliked increased after they learned that other students liked or offenders disliked the same items, whereas their preferences for t-shirts they initially liked decreased after they learned that students disliked or offenders liked the same items. This was borne out by a 2 (Group) x 2 (Self preference) x 2 (Other preference) repeated-measures ANOVA showing a significant group-by-other preference interaction ($F_{(1,17)} = 9.41$, $p = 0.007$).

As preferences were rated using a 14-point scale (see Methods for details), the degree of cognitive imbalance in each trial could be quantified parametrically as the difference between a participant's ratings and the (binary) preference of the other group (Cognitive Imbalance Index, CII, ranging from 0 to 13; see Methods and Table S1). Thus, the greater the difference between self ratings and student preferences, or the smaller the difference between self ratings and offender preferences, the greater the CII. To test whether larger CIIs motivate greater preference change as hypothesized by balance theory (Heider, 1946, 1958), we conducted linear regressions, separately for each group (students, offenders), with preference change (2nd ratings – 1st ratings) as the dependent variable and the signed CII as the predictor variable (see Methods for details). CII significantly predicted subsequent preference change for both groups (ps < 0.001), and there was no significant difference between the effect of the two groups ($p = 0.41$, n.s.; Figure S1a). All effects remained significant when including the raw first rating as a separate regressor, to account for any possible effects of regression-to-the-mean (see Supplemental Text and Figure S2 for details).

Remarkably, subjects' preferences remained socially influenced even after several months had passed. We observed the same significant group-by-other preference interaction ($F_{(1,14)} = 9.15$, $p = 0.009$), and the CII significantly predicted subjects' preferences for both groups

after four months (Figure S1b & c). This persistent preference change was all the more striking as participants did not in fact remember how the t-shirts had been rated by students or offenders four months ago, performing at chance on a forced-choice memory task (see Supplemental Text).

fMRI results: tracking cognitive imbalance

To identify brain regions within which activation correlated with the CII on a trial-by-trial basis for both liked and disliked groups, we carried out a conjunction analysis (see Methods for details). This showed activation in pMFC, especially in dorsomedial prefrontal cortex (dmPFC) (Figure 2b), as well as left insula, left inferior frontal gyrus (IFG) and posterior cingulate cortex (PCC) (see Table 1 and Figure S3). A further quantification of this finding was provided by an ROI analysis that found that the dmPFC region showed particularly increased activation in conditions featuring cognitive imbalance, and a significant 3-way interaction (group X self preference X other preference; $F_{(1,17)} = 19.7$, $p < 0.001$) (Figure 2c) that paralleled the behavioral findings described above. These results confirm that dmPFC activation in our task depends on a triadic relationship between self, other and an object of shared evaluation, providing a direct neural correlate of cognitive imbalance according to balance theory (Heider, 1946, 1958). Left insula and left IFG also showed the same 3-way interaction (ps < 0.01), whereas PCC did not ($p = 0.83$, n.s.) (see Figure S3). As a confirmatory analysis, when we directly explored the 3-way interaction contrast (i.e., imbalance vs. balance contrast), significant activations were found in dmPFC and left IFG (see Figure S4a).

There was no region significantly negatively correlated with the CII. However, when the four balanced conditions were compared with the four imbalanced conditions, we found significant activation in ventral striatum, consistent with prior reports (Campbell-Meiklejohn et al., 2010; Klucharev et al., 2009). Furthermore, this striatal area overlapped with the region sensitive to reward that was identified by our MIDT (Figure S4b & c), consistent with the hypothesis that not only agreeing with liked others (Campbell-Meiklejohn et al., 2010; Klucharev et al., 2009) but also disagreeing with disliked others may both be rewarding. Right insula and right IFG were also activated in this contrast (see Figure S4b and Supplemental Text for details).

We also tested whether any region showed an association with CII that might differ between the two groups (students, offenders) and found only one: right middle temporal gyrus (MTG; x = 40, y = −68, z = 18; 256 voxels) was more strongly associated with the CII for students than for offenders (no area was found in the reverse contrast). Overall, the findings suggest that a qualitatively similar mechanism operates to represent cognitive imbalance evoked by both liked and disliked groups of other people, notably recruiting the dmPFC.

Association between brain activation and preference change

Activation of pMFC has been associated with behavioral adjustment following negative or unexpected outcomes (Bush et al., 2002; Jocham et al., 2009; O'Doherty et al., 2003; Ridderinkhof et al., 2004; Shima and Tanji, 1998; Ullsperger et al., 2004). As the CII in our study is significantly associated both with subsequent preference change as well as activation within specific brain regions (Table 1), it naturally follows that activations in these regions should also be associated with preference change. To show this we pooled the within-subject correlation between brain activation and preference change across the eight experimental conditions of interest (see Figure 1d). Among all brain regions identified above, the dmPFC showed the strongest association between CII-evoked activation and subsequent behavioral preference change (Figure 3a and Table 2). Our analysis thus confirmed that the higher the dmPFC activation when viewing others' preference for the

same item, the larger the subsequent preference change in the direction predicted by balance theory (Heider, 1946, 1958).

The association between dmPFC activation during the initial fMRI experiment and later preference change also remained significant even after four months (Figure 3b), and the dmPFC was the only region that survived a correction for multiple comparisons across both time points (initial change and four months later; see Table 2). As subjects no longer remembered the original ratings at four months later, a single episode of feedback about the opinions of other people thus results in dmPFC activation that correlates with long-lasting preference change likely to be implicit.

Two localizer tasks

The two commonly used localizer tasks we employed in our study (Bush and Shin, 2006; Knutson et al., 2000) successfully identified areas within pMFC especially sensitive to response conflict (pre-SMA) or negative outcome (posterior part of dmPFC) (Figure 4a; see also Supplemental Text and Figure S5). The region related to response conflict was posterior to that related to negative outcome, consistent with prior findings (Hikosaka and Isoda, 2010; Ridderinkhof et al., 2004; Rushworth et al., 2004). The functional role of these anatomical regions in each of the two localizer tasks was qualitatively distinct (Henson, 2006): the beta values for each condition of each localizer task from each peak within pre-SMA and posterior dmPFC showed a significant 3-way interaction (anatomical region X localizer task X task condition; $F_{(1,17)} = 9.03$, $p = 0.008$) (Figure 4b). Importantly, the cognitive imbalance-related dmPFC region we described earlier largely overlapped with the area involved in negative outcome, whereas there was no overlap at all with the pre-SMA involved in response conflict (Figure 4a). There was a significant 4-way interaction (region X group X self preference X other preference; $F_{(1,17)} = 11.8$, $p = 0.003$) (Figure 4c & d): whereas activation within the posterior dmPFC was modulated by a triadic relationship among self, others and objects as postulated by balance theory (significant 3-way interaction; $F_{(1,17)} = 12.7$, $p = 0.002$) (Figure 4d), activation within the pre-SMA was insensitive to this ($p = 0.87$, n.s.) (Figure 4c). Thus, the present results argue against the idea that cognitive imbalance and response conflict share neural mechanisms, and instead suggest that cognitive imbalance activates the dmPFC because it represents an aversive outcome requiring subsequent adjustment.

A final point of clarification concerned whether the dmPFC activation we found might have arisen from expectancy violation (Somerville et al., 2006) or surprise signal (unsigned reward prediction error) (Hayden et al., 2011): participants might simply expect other students to give preference ratings similar to their own, and sex offenders to have different preferences, confounding the CII with the degree of expectancy violation. To address this issue, we asked participants at the end of the experiment to guess the other group's preference for those t-shirts that had been presented in the control condition (about which no feedback had been given). Expectations of the other group's preferences were not related to self preferences for either group (mean correlation coefficients were not significantly different from zero; $ps > 0.80$) (Figure S6a). Furthermore, there was no change in the strength of the association between CII and dmPFC activation across the sequence of three runs of our study (Figure S6b), thus showing no indication of a change in expectancy with learning over sessions. Simple expectation violation is thus unlikely to be the mechanism whereby CII drives activation within the dmPFC (see Supplemental Text for further details).

DISCUSSION

Our attitudes and preferences are potently influenced by other people, effects formalized in social psychology theories (Abelson et al., 1968; Heider, 1946, 1958; Osgood and

Tannenbaum, 1955). The present study showed that an individual's preference for goods is influenced by the difference between one's own preference and the preferences of other people; however, it also depends on one's attitude toward those other people. The present results show that cognitive imbalance, a key concept in classical social psychology, is associated with activation in the dmPFC. Consistent with the idea of balance theory (Heider, 1946, 1958) that cognitive balance depends on a triadic relationship among self, others and objects, we found that the dmPFC activation depended on these three factors, and its activity tracked the degree of cognitive imbalance on a trial-by-trial basis. While the dmPFC showed greater activation the further a subject's preference was from that of a liked group of other people, this relationship was completely reversed for a disliked group of people. The pattern of activations we observed clearly indicates that our brain does not simply encode the difference between our own and another's preference, but that it also depends on how we feel about the other person. In other words, the dmPFC encodes the difference between a person's current preference state and a cognitively balanced state.

Our study further demonstrated that the effect of cognitive imbalance on preference change was remarkably long lasting. People's preference change remained influenced by the opinion of others even after four months, despite the fact that they no longer remembered how each t-shirt had been rated by others. We also confirmed that the dmPFC activation is correlated with the degree of preference change. Although this significant correlation itself is not surprising as the dmPFC region is identified by using the CII which is related to preference change (Figure S1a & c), the fact that among all regions related to cognitive balance, the dmPFC activation was the most strongly associated with preference change both immediately after social feedback as well as four months later suggests that the dmPFC plays a pivotal role in adjusting one's preference so as to restore cognitive balance.

The present findings are consistent with two previous neuroimaging studies (Izuma et al., 2010; van Veen et al., 2009) showing that the pMFC (including dACC as well as dmPFC) is activated by "cognitive dissonance," another form of cognitive inconsistency which is induced by the discrepancy between what we believe and what we do (Festinger, 1957). Since pMFC activation is sensitive to the degree of cognitive dissonance (Izuma et al., 2010) and cognitive imbalance on a trial-by-trial basis, neural activation within this region may be a direct physiological measure of cognitive inconsistency in general. While cognitive consistency theories once dominated social psychology especially during 1950s–70s, their influence has declined in the past few decades. One reason for the decline may be the historically heavy reliance on self-report measures (Greenwald et al., 2002). Early research on balance theory using self-report produced findings that were at times unreliable and contradictory, arising especially from the complexity of cases where the attitude of self toward the other person is negative (Insko, 1984). The present results may provide a more valid measure of cognitive inconsistency that is conceptually aligned with theory, as we observed a similar degree of dmPFC activation and preference change in all imbalanced conditions, whether the other social group was one whose opinions were to be avoided or to be emulated. The symmetry of our results may indicate that pMFC activation is a conceptually more valid measure of cognitive inconsistency than is verbal report by itself. Given that the framework of balance theory has been applied to many central topics in social psychology, such as stereotyping, self-esteem, and self-conceptualization (Greenwald et al., 2002; see also Gawronski, 2012), it will be an important future project to confirm that the brain regions we identify in the present study also play a key role in other social processes.

The dmPFC activation found in the present study may reflect a number of different processes, none of them mutually exclusive. One possibility is that cognitive imbalance in our task engages processing analogous to emotional reappraisal (Etkin et al., 2011; Ochsner and Gross, 2008). dmPFC and dACC activations are often associated with reduction in

negative emotion through reappraisal (Etkin et al., 2011; Ochsner and Gross, 2008). Social psychologists have often conceptualized cognitively inconsistent states (cognitive imbalance and cognitive dissonance) as emotionally aversive, providing the driving force for subsequent attitude change (Abelson et al., 1968; Festinger, 1957; Gawronski and Strack, 2012; Heider, 1946, 1958), similar to the change in evaluation that drives emotion reappraisal. As dmPFC activation is associated with a decrease (not increase) of negatively valenced affect through reappraisal, it may be that the dmPFC activity found in the present study does not simply reflect the level of negative affect *per se*, but rather the motivation for cognitive reappraisal or adjustment which leads to the desired consequence of cognitively balanced states.

Two other possibilities for the processes corresponding to our observed dmPFC activation are the registration of negative (aversive) outcomes, or response conflict. As stated above, previous fMRI studies on cognitive dissonance consistently found that pMFC is involved in cognitive dissonance (Izuma et al., 2010; van Veen et al., 2009). Based solely on the general finding that pMFC is activated by cognitive dissonance (i.e., using reverse inference), these studies interpreted the pMFC activation as meaning that a common neural mechanism is responsible for response-level conflict and cognitive-level conflict (Izuma et al., 2010; van Veen et al., 2009). However, because pMFC is known to be activated by a variety of cognitive processes other than response conflict, reverse inference based on pMFC activation is especially problematic (Poldrack, 2011). To circumvent this problem, the present study used two independent localizer tasks in the same participants: we found that the dmPFC region tracking cognitive imbalance overlapped with a region sensitive to negative outcome, but not at all with activation related to response conflict. While these findings will need to be followed up with other methods, they suggest that response conflict is unlikely to come into play during cognitive imbalance, whereas aversive outcomes may be a candidate for what is represented by the dmPFC in our task. Similarly, it has been suggested that pMFC generates a negative prediction error-like signal which induces social learning (conformity) (Klucharev et al., 2009; see also Proulx et al., 2012). However, it should be noted that the present findings suggest that dmPFC activation cannot be explained merely by expectation violation. Rather, as we noted above, the dmPFC appears to encode the discrepancy between actual outcomes and outcomes that would be the most cognitively consistent (i.e., what participants would hope to get, rather than what they expected to get).

When proposed more than five decades ago, theories of cognitive consistency revolutionized the way psychologists thought about how human behavior is motivated, replacing traditional Behaviorist views of reward and punishment with cognitive versions (Greenwald et al., 2002). The present findings in a sense come full circle by arguing that cognitive imbalance works through basic mechanisms for adjusting behavior following aversive outcomes. The pMFC is known to respond to a variety of aversive outcomes (Botvinick et al., 2001; Bush et al., 2002; Eisenberger and Lieberman, 2004; Falkenstein et al., 1991; Hikosaka and Isoda, 2010; Holroyd and Coles, 2002; Miltner et al., 1997; O'Doherty et al., 2003; Ridderinkhof et al., 2004; Shackman et al., 2011; Shima and Tanji, 1998; Ullsperger and von Cramon, 2003), play a key role in detecting stimuli which signal the deviation from a desired state, and induce subsequent behavioral adjustment leading to desired consequences (Botvinick et al., 2001; Eisenberger and Lieberman, 2004; Hikosaka and Isoda, 2010; Holroyd and Coles, 2002; Ridderinkhof et al., 2004; Shackman et al., 2011; Ullsperger et al., 2004). Our present results suggest that cognitive imbalance might motivate preference change in much the same way as aversive outcomes motivate behavioral change. It is important to note that activation of the same brain region does not guarantee the involvement of the same neural processes (Henson, 2006), rendering our process interpretation preliminary at this stage. While it has been argued (Harmon-Jones, 2004) that cognitive inconsistency might recruit the same region involved in simple response conflict, this idea was never formally tested. Our present

finding that the area tracking cognitive inconsistency was dissociated from the area related to response conflict is inconsistent with this idea and indicates that neural processes underlying cognitive conflict such as cognitive imbalance (and cognitive dissonance) and response conflict are likely distinct.

A final possibility to consider in interpreting the dmPFC activation we observed is suggested by the accumulating data that dmPFC activation is found also in tasks that involve theory-of-mind (Ochsner et al., 2004), the ability to reason about other people's mental states such as their beliefs. It has been argued that dmPFC plays a key role in the uniquely human representation of triadic relations between two minds and an object (Saxe, 2006), a schema quite similar to the basic idea of balance theory. It might well be that our dmPFC activation, at least to some degree, reflects this representation of triadic relations as its activation was higher in our eight conditions of interest (where the preference of others was presented) compared to the two control conditions (where no feedback about others' preferences was presented). However, the mentalizing hypothesis cannot account for the significant 3-way interaction pattern found in dmPFC and its significant association with preference change, suggesting that this is unlikely to be the whole explanation.

It should be noted that although we framed Caltech students as a liked group and sex offenders as a disliked group in the present study, it is likely that these two groups differ also on dimensions other than likeability (or what balance theory called a "sentiment" relationship). For example, subjects might perceive Caltech students as similar and sex offenders as dissimilar to themselves, evoking ingroup and outgroup biases that could influence cognitive consistency irrespective of overt likeability. It will be important in future studies to decorrelate these and other attributes in order to determined precisely which are primarily driving the effect we observed. On the other hand, it seems reasonable to suppose that several attributes, including valenced likeability, similarity to self, and so forth, all contribute. Indeed, balance theory postulates that not only a "sentiment relationship" but also a "unit relationship" (i.e., similarity, proximity, causality, membership, possession, or belonging) influence social balance (Heider, 1946, 1958); to what extent these involve overlapping or distinct neural substrates will be important topics to be explored in the future.

The dmPFC signal that we found appears to reflect unfavorable outcomes that motivate preference change. This interpretation of dmPFC function in cognitive balance should be situated within a network of other brain regions involved in reward processing, several of which we also found activated in some contrasts in our study. Thus, imbalance versus balance differentially activated the IFG, a region implicated in cognitive control and change in opinion (Sharot et al., 2011); the insula, possibly related to evoked emotional responses; and the ventral striatum, involved in processing reward prediction-errors and activated when others agree with one's own opinion (Campbell-Meiklejohn et al., 2010; Klucharev et al., 2009). It will be important to determine the causal roles of these regions, and whether they are truly identical to the ones for general reward learning or contain specific neuronal populations that code cognitive balance separately from general reward. As the phenomenon of cognitive balance is unlikely to occur in nonhuman animals, such future investigations will require lesion studies and rare intracranial recordings in human patients for which the present study provides focused hypotheses.

EXPERIMENTAL PROCEDURES

Subjects

Twenty Caltech undergraduate or graduate students who had never rated any of our t-shirt designs previously completed two fMRI experiments. Data from 2 participants were removed due to excessive head motion in the scanner and all reported analyses are based on

the remaining 18 (6 female, mean age = 22.5 ± 2.6 years, all right-handed with no history of neurological or psychiatric illness). Fifteen of the 18 subjects also participated in the third preference rating task (4 female). All subjects gave written informed consent for participation, and the study was approved by the Institutional Review Board of the California Institute of Technology.

Experimental paradigm

The experiment consisted of three phases performed on three separate days: 1) first and second preference rating tasks as described above (fMRI experiment; Figure 1), 2) two localizer tasks (Bush and Shin, 2006; Knutson et al., 2000) (fMRI experiment), and 3) third preference rating task (behavioral experiment; in 15/18 participants). The order of the two fMRI experiments was counterbalanced across subjects, and these two experiments were separated by a mean of 12.7 (±6.2) days. The mean interval between first and third preference rating tasks was 124.9 (±20.2) days.

First and second preference rating tasks

On the first day, inside the fMRI scanner, subjects were presented with a front image of a t-shirt and the scale on the monitor, and asked to rate how much they liked each t-shirt using a 14-point scale (Figure 1c). Discrete ratings were given using a button box with three buttons: subjects used the right index finger to shift a cursor one point leftward in the scale, the middle finger to shift it one point rightward, and then decided by pressing a button with the ring finger. Subjects had no time limit to indicate their preference but were instructed not to think too much about it. The mean response time was 4.16 (±1.28) sec. Immediately after giving their ratings, subjects were presented with their preference depicted by a thumbs-up (for ratings ≥ 8) or thumbs-down icon (for ratings ≤ 7) for 0.5 sec, and then presented with how others rated the same item in terms of thumbs-up (others liked the item) or thumbs-down (others disliked the item) for 2 sec (Figure 1d).

Subjects rated a total of 174 t-shirt designs, which had been selected from an internet t-shirt store, randomly pre-assigned to one of three conditions (Caltech, Offenders or control). In 72 trials, subjects were presented with ratings they were told were given by other Caltech students. In another 72 trials, they were presented with ratings they were told were given by sex offenders who are currently in jail. These two groups were chosen based on extensive piloting to maximize their influence on the preferences of the population from which our subject sample was drawn (Caltech students; see Supplemental Text for further details). In the remaining trials (30 trials), no information about other people's preference was presented (control condition). Before the fMRI experiment, all subjects were led to believe that we had collected preference data for the same 174 t-shirts from other Caltech students and from sex offenders a few months previously. Participants were further told that only the worst sex offenders were included in this study, and all subjects read brief descriptions about the sex offenders before the fMRI experiment began. In reality, others' ratings were experimentally determined such that Caltech students' (or sex offenders') likes or dislikes were approximately equally distributed between t-shirts liked or disliked by the participants themselves. Post-study questions verified that all participants believed they received feedback from Caltech students and sex offender groups as we intended.

About five minutes after the first preference rating task, subjects were asked to rate the same 174 t-shirts again using a 14-point scale inside the scanner. No information about others' preference was presented this time. After these two rating tasks, subjects came out of the fMRI scanner and were asked to guess others' (Caltech students' or sex offenders') preferences for the 30 t-shirts presented in the control condition, using the same 14-point scale. In this task, subjects were instructed to give a best guess about the average preference

of Caltech students or sex offenders for each t-shirt (these latter data were used to test for expectancy violation as a possible mechanism in our findings; cf. Main Text and Supplemental Text); the order of ratings (guessing Caltech students' preference first or sex offenders' preference first) was counterbalanced across subjects.

Finally, all subjects were asked to rate their impression toward other Caltech students or sex offenders (see below for details). Six items each were used for the impression ratings about the two social groups (How much do you identify yourself as a member of Caltech students?; How likable do you find Caltech students in general?; How similar do you think you are to other Caltech students?; How much do you think you and other Caltech students have in common?; To what extent would you use the term 'we' to describe yourself and other Caltech students?; How much do you think you might like to interact with other Caltech students at some future time?; the phrase "other Caltech students" was replaced with "sex offenders" when rating sex offenders.). Impressions were rated on a 14-point scale (1 = not at all, 14 = very much). These six items showed good internal consistency for the rating of both groups (Caltech students Cronbach's $\alpha = 0.92$; sex offenders Cronbach's $\alpha = 0.75$).

Two localizer tasks

All subjects also completed two localizer tasks: 1) the Multi-Source Interference Task (Bush and Shin, 2006) (MSIT), and 2) a Monetary Incentive Delay Task (Knutson et al., 2000) (MIDT). The MSIT is well-known to robustly and reliably activate response conflict-related pMFC regions (Bush and Shin, 2006; Bush et al., 2003). The MIDT is a simple button-response task involving monetary reward (Knutson et al., 2000). As it allows subjects to get performance-contingent positive (hit) or negative (miss) feedback multiple times within a short period of time it reliably activates brain areas responsive to negative feedback compared to positive feedback (although it may not be entirely specific in this regard, also encompassing processes such as reward prediction error).

During the MSIT, subjects were presented with a set of three numbers (1, 2, 3 or 0) on the center of the screen, and one number was always different from the other two numbers. The task was to report the identity of the number that is different from the other two numbers by using a button box with three buttons each of which represents one, two and three from left to right. It was emphasized that subjects needed to report what the target number was regardless of its position. In the control blocks, the target number (1, 2 or 3) always matched its spatial position and non-target numbers were always zero (e.g., 100, 020, 003). On the other hand, in the interference blocks, the target number never matched its spatial position, and the distracters (non-target numbers) were themselves potential targets (e.g., 212, 233, 311, etc). Within each block (42 sec), subjects performed 24 control or 24 interference trials. Each set of numbers was presented for 500 ms with an inter-stimulus interval (ISI) of 1250 ms. Within an fMRI run, subjects completed three control and three interference blocks in alternating order. 30-sec fixation periods were inserted at the beginning and the end of each fMRI run. All subjects completed two fMRI runs (each ~7 min).

During the MIDT, subjects were first presented with one of three cues indicating \$0, \$0.2 or \$2 for 0.5 sec. After a random delay (2 to 2.5 sec), a white square was just briefly presented at the center of the screen, and the task was to press a button as soon as this white square was detected. If a subject pressed the button while the square was still on the screen, they could obtain the amount of money specified during the cue period. The duration of the square presentation was dynamically adjusted between 160 ms to 300 ms on a trial-by-trial basis for each subject during the task so that subjects could hit the target about 66% of trials. If they successfully hit the target, a hit feedback appeared on the screen (e.g., "You won \$2.00") for 0.8 sec. Similarly, if they missed the target, a miss feedback was presented (e.g., "Miss!"). Inter-trial interval (ITI) was 1.5 sec. Subjects performed a total of 45 trials in

randomized order in each fMRI run (i.e., 15 trials each for $0, $0.2 and $2 trials), and they completed two fMRI runs (each ~ 6 min).

Before the fMRI experiment, subjects performed a brief practice session using a laptop computer for both MSIT and MIDT outside the scanner. At the end of the two localizer experiments, subjects were given all the money they earned during the MIDT.

Third preference rating task

After approximately four months from the first and second preference rating tasks, all 18 subjects were asked to take part in a follow-up behavioral experiment, and 15 out of 18 subjects (4 female) participated. In this follow-up experiment, subjects were asked to rate the same 174 t-shirts again using the same 14-point scale, followed by a memory task. The memory task presented each subject with the 144 t-shirts which had been previously associated with either Caltech students' or sex offenders' ratings (i.e., t-shirts presented in the control conditions were not presented during the memory test). Below each t-shirt, there were four options (Caltech students Liked, Caltech students Disliked, Sex offenders Liked or Sex offenders Disliked), and they were asked to pick one.

fMRI data acquisition

All fMRI data were acquired using a Siemens 3.0 Tesla Trio scanner with a 32-channel phased array headcoil. For functional imaging, interleaved T2*-weighted gradient-echo echo-planar imaging (EPI) sequences were used to produce 44 contiguous 3-mm-thick trans-axial slices covering nearly the entire cerebrum (repetition time [TR] = 2,500 ms; echo time [TE] = 30 ms; flip angle [FA] = 80°; field of view [FOV] = 192 mm; 64 × 64 matrix; voxel dimensions = 3.0 × 3.0 × 3.0 mm). A high-resolution anatomical T1-weighted image (1mm isotropic resolution) was also acquired for each subject.

fMRI data pre-processing

The data were analyzed using SPM8 (Wellcome Department of Imaging Neuroscience) implemented in Matlab 7.8 (Mathworks). Before data processing and statistical analysis, we discarded the first four volumes to allow for equilibration. After correcting for differences in slice timing within each image volume, head motion was corrected. Following realignment, the volumes were normalized to MNI space using a transformation matrix obtained from the normalization of the first EPI image of each individual subject to the EPI template. The normalized fMRI data were spatially smoothed with an isotropic Gaussian kernel of 8 mm (full-width at half-maximum).

fMRI Data analysis

We used two general linear models (GLM) to analyze the fMRI data for the first preference rating task; one GLM was intended to identify brain regions correlated with the degree of cognitive imbalance on a trial-by-trial basis, and the other GLM was for extracting brain activation for each of 10 conditions (Figure 1d) for detailed analysis of activation patterns. Since the present study focuses on brain activation while subjects perceive balanced or imbalanced situations (feedback period during the first preference rating task), the fMRI data from the second preference rating task was not analyzed.

For the first GLM, we first quantified the degree of imbalance in each trial as the discrepancy/similarity between subject's own and others' rating (i.e., Cognitive Imbalance Index; CII). For Caltech students' rating, the greater the difference between a participant's and Caltech students' rating, the larger the CII. In contrast, for sex offenders' rating, the more similar a participant's rating is to sex offenders' ratings, the larger the CII. Therefore, the CII can be quantified as follows:

CII = 14 – subject's first rating (if Caltech students like or sex offenders dislike items)

CII = subject's first rating – 1 (if Caltech students dislike or sex offenders like items)

The CII for items in the two control conditions are always 0. Thus, the CII has the possible range of 0 to 13 (see Table S1).

Therefore, the first model included: 1) each t-shirt presentation (duration = subject's response time), 2) t-shirt presentation modulated by subject's preference for each t-shirt, 3) t-shirt presentation modulated by the number of button pressed until subjects decide their rating, 4) Caltech students' feedback presentation (2 sec), 5) Caltech students feedback presentation modulated by the CII, 6) sex offenders' feedback presentation (2 sec), 7) sex offenders feedback presentation modulated by the CII. Items in the control condition were randomly divided into halves and allocated to either Caltech students or sex offender conditions.

The second GLM modelled the feedback period (2 sec) separately for each of 10 conditions (see Figure 1d), and thus included following regressors: 1) each t-shirt presentation (duration = response time), 2) t-shirt presentation modulated by subject's preference for each t-shirt, 3) t-shirt presentation modulated by the number of button pressed until subjects decide their rating, 4) feedback presentation for the Self-Like and Caltech students-Like condition, 5) feedback presentation for Self-Like and Caltech students-Dislike condition, 6) feedback presentation for the Self-Dislike and Caltech students-Like condition, 7) feedback presentation for Self-Dislike and Caltech students-Dislike condition, 8) feedback presentation for the Self-Like and sex offenders-Like condition, 9) feedback presentation for Self-Like and sex offenders-Dislike condition, 10) feedback presentation for the Self-Dislike and sex offenders-Like condition, 11) feedback presentation for Self-Dislike and sex offenders-Dislike condition, 12) feedback presentation for the Self-Like-Control condition, and 13) feedback presentation for the Self-Dislike-Control condition.

We first carried out a conjunction analysis with a masking procedure to define common networks of CII-related activations for both student and offender groups using the first GLM. We employed an inclusive masking procedure so that areas identified by the conjunction analysis were significantly associated with CIIs for both the student as well as the sex offender conditions ($p < 0.001$ uncorrected, and cluster $p < 0.05$ FWE-corrected), an approach logically analogous to a conjunction analysis with a conjunction null hypothesis (Friston et al., 2005; Nichols et al., 2005). Subsequently, ROI analyses were carried out using a leave-one-subject-out (LOSO) cross-validation procedure (Esterman et al., 2010) in order to eliminate non-independence bias for plots of parameter estimates. We re-ran the second level analysis (first GLM) 18 times with a different single subject left out in each, and each second level analysis was used to determine the ROIs for each left-out subject. Beta values were extracted from relevant local maxima for each subject for 10 conditions (Figure 1d) by using the second GLM, and these values were averaged to plot overall effect sizes (Figure 2c & Figure S3).

For fMRI data analysis on the MSIT, the model included two regressors for each of the control and interference blocks (42 sec). To account for phasic activations that might be accompanied by error processing, error trials regardless of conditions were also modelled as a separate regressor of no interest.

For fMRI data analysis on the MIDT, the model included the following regressors: 1) cue onset, 2) cue onset modulated by reward level ($0, $0.2 or $2), 3) feedback onset for $0-Hit trials, 4) feedback onset for $0-Miss trials, 5) feedback onset for $0.2-Hit trials, 6) feedback onset for $0.2-Miss trials, 7) feedback onset for $2-Hit trials, 8) feedback onset for $2-Miss

trials, 9) feedback onset for error trials (i.e., trials with no response or too early response). For both localizer tasks, average effect sizes were plotted using the above-mentioned LOSO cross-validation procedure (Figure 4b).

For all GLMs, the regressors were calculated using a box-car function convolved with a hemodynamic-response function. Other regressors that were of no interest, such as six motion parameters, the session effect, and high-pass filtering (128 sec) were also included.

For all fMRI results reported (preference rating task and two localizer tasks), a whole-brain statistical threshold was set at $p < 0.001$ voxelwise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank John P. O'Doherty and Dan Kennedy for helpful comments on the manuscript, and the staff of the Caltech Brain Imaging Center for their assistance in conducting the fMRI. This work was funded by the Japan Society for the Promotion of Science Fellows (K.I.), as well as the National Science Foundation, a Conte Center grant from the National Institute of Mental Health, the Gordon and Betty Moore Foundation, and the Tamagawa University Global Centers of Excellence grant from the Ministry of Education, Culture, Sports, Science and Technology, Japan (R.A.).

References

Abelson, RP.; Aronson, E.; McGuire, WJ.; Newcomb, TM.; Rosenberg, MJ.; Tannenbaum, PH. Theories of cognitive consistency: A sourcebook. Chicago: Rand McNally; 1968.

Alexander WH, Brown JW. Medial prefrontal cortex as an action-outcome predictor. Nature Neurosci. 2011; 14:1338–1163. [PubMed: 21926982]

Berns GS, Capra CM, Moore S, Noussair C. Neural mechanisms of the influence of popularity on adolescent ratings of music. Neuroimage. 2010; 49:2687–2696. [PubMed: 19879365]

Botvinick MM. Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. Cogn Affect Behav Neurosci. 2007; 7:356–366. [PubMed: 18189009]

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. Conflict monitoring and cognitive control. Psychol Rev. 2001; 108:624–652. [PubMed: 11488380]

Bush G, Shin LM. The Multi-Source Interference Task: an fMRI task that reliably activates the cingulo-frontal-parietal cognitive/attention network. Nature Protoc. 2006; 1:308–313. [PubMed: 17406250]

Bush G, Shin LM, Holmes J, Rosen BR, Vogt BA. The Multi-Source Interference Task: validation study with fMRI in individual subjects. Mol Psychiatr. 2003; 8:60–70.

Bush G, Vogt BA, Holmes J, Dale AM, Greve D, Jenike MA, Rosen BR. Dorsal anterior cingulate cortex: a role in reward-based decision making. Proc Natl Acad Sci USA. 2002; 99:523–528. [PubMed: 11756669]

Campbell-Meiklejohn DK, Bach DR, Roepstorff A, Dolan RJ, Frith CD. How the opinion of others affects our valuation of objects. Curr Biol. 2010; 20:1165–1170. [PubMed: 20619815]

Campbell-Meiklejohn DK, Kanai R, Back DR, Dolan RJ, Frith CD. Structure of orbitofrontal cortex predicts social influence. Curr Biol. 2012a; 22:R123. [PubMed: 22361146]

Campbell-Meiklejohn DK, Simonsen A, Jensen M, Wohlert V, Gjerloff T, Scheel-Kruger J, Moller A, Frith CD, Roepstorff A. Modulation of social influence by methylphenidate. Neuropsychopharmacol. 2012b; 37:1517–1525.

Cialdini RB, Goldstein NJ. Social influence: Compliance and conformity. Annu Rev Psychol. 2004; 55:591–621. [PubMed: 14744228]

Eisenberger NI, Lieberman MD. Why rejection hurts: a common neural alarm system for physical and social pain. Trends Cogn Sci. 2004; 8:294–300. [PubMed: 15242688]

Esterman M, Tamber-Rosenau RJ, Chiu Y, Yantis S. Avoiding non-independence in fMRI data analysis: Leave one subject out. Neuroimage. 2010; 50:572–576. [PubMed: 20006712]

Etkin A, Egner T, Kalisch R. Emotional processing in anterior cingulate and medial prefrontal cortex. Trends Cogn Sci. 2011; 15:85–93. [PubMed: 21167765]

Falkenstein M, Hohnsbein J, Hoormann J, Blanke L. Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. Electroencephalogr Clin Neurophysiol. 1991; 78:447–455. [PubMed: 1712280]

Festinger, L. A Theory of Cognitive Dissonance. Stanford: Stanford University Press; 1957.

Friston KJ, Penny WD, Glaser DE. Conjunction revisited. Neuroimage. 2005; 25:661–667. [PubMed: 15808967]

Gawronski B. Back to the future of dissonance theory: Cognitive consistency as a core motive. Soc Cognition. 2012; 30:652–668.

Gawronski, B.; Strack, F., editors. Cognitive Consistency: A Fundamental Principle in Social Cognition. 1. New York: Guilford Press; 2012.

Greenwald AG, Banaji MR, Rudman LA, Farnham SD, Nosek BA, Mellott DS. A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. Psychol Rev. 2002; 109:3–25. [PubMed: 11863040]

Harmon-Jones E. Contributions from research on anger and cognitive dissonance to understanding the motivational functions of asymmetrical frontal brain activity. Biol Psychol. 2004; 67:51–76. [PubMed: 15130525]

Hayden BY, Heilbronner SR, Pearson JM, Platt ML. Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. J Neurosci. 2011; 31:4178–4187. [PubMed: 21411658]

Heider F. Attitudes and cognitive organization. J Psychol. 1946; 21:107–112. [PubMed: 21010780]

Heider, F. The psychology of interpersonal relations. New York: Wiley; 1958.

Henson R. Forward inference using functional neuroimaging: dissociations versus associations. Trends Cogn Sci. 2006; 10:64–69. [PubMed: 16406759]

Hikosaka O, Isoda M. Switching from automatic to controlled behavior: cortico-basal ganglia mechanisms. Trends Cogn Sci. 2010; 14:154–161. [PubMed: 20181509]

Holroyd CB, Coles MGH. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. Psychol Rev. 2002; 109:679–709. [PubMed: 12374324]

Insko, CA. Balance theory, the Jordan paradigm, and the Wiest tetrahedron. In: Berkowitz, L., editor. Advances in experimental social psychology. San Diego, CA: Academic Press; 1984. p. 89-140.

Izuma K, Matsumoto M, Murayama K, Samejima K, Sadato N, Matsumoto K. Neural correlates of cognitive dissonance and choice-induced preference change. Proc Natl Acad Sci USA. 2010; 107:22014–22019. [PubMed: 21135218]

Jocham G, Neumann J, Klein TA, Danielmeier C, Ullsperger M. Adaptive coding of action values in the human rostral cingulate zone. J Neurosci. 2009; 29:7489–7496. [PubMed: 19515916]

Klucharev V, Hytonen K, Rijpkema M, Smidts A, Fernandez G. Reinforcement learning signal predicts social conformity. Neuron. 2009; 61:140–151. [PubMed: 19146819]

Klucharev V, Munneke MAM, Smidts A, Fernandez G. Downregulation of the posterior medial frontal cortex prevents social conformity. J Neurosci. 2011; 31:11934–11940. [PubMed: 21849554]

Knutson B, Westdorp A, Kaiser E, Hommer D. FMRI visualization of brain activity during a monetary incentive delay task. Neuroimage. 2000; 12:20–27. [PubMed: 10875899]

Miltner WHR, Braun CH, Coles MGH. Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a “generic” neural system for error detection. J Cogn Neurosci. 1997; 9:788–798.

Nichols T, Brett M, Andersson J, Wager T, Poline JB. Valid conjunction inference with the minimum statistic. Neuroimage. 2005; 25:653–660. [PubMed: 15808966]

O'Doherty J, Critchley H, Deichmann R, Dolan RJ. Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. J Neurosci. 2003; 23:7931–7939. [PubMed: 12944524]

Ochsner KN, Gross JJ. Cognitive emotion regulation: Insights from social cognitive and affective neuroscience. Curr Dir Psychol Sci. 2008; 17:153–158.

Ochsner KN, Knierim K, Ludlow DH, Hanelin J, Ramachandran T, Glover G, Mackey SC. Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. J Cogn Neurosci. 2004; 16:1746–1772. [PubMed: 15701226]

Osgood CE, Tannenbaum PH. The principle of congruity in the prediction of attitude change. Psychol Rev. 1955; 62:42–55. [PubMed: 14357526]

Poldrack A. Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. Neuron. 2011; 72:692–697. [PubMed: 22153367]

Proulx T, Inzlicht M, Harmone-Jone E. Understanding all inconsistency compensation as a palliative response to violated expectations. Trends Cogn Sci. 2012; 16:285–291. [PubMed: 22516239]

Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuiss S. The role of the medial frontal cortex in cognitive control. Science. 2004; 306:443–447. [PubMed: 15486290]

Rushworth MFS, Walton ME, Kennerley SW, Bannerman DM. Action sets and decisions in the medial frontal cortex. Trends Cogn Sci. 2004; 8:410–417. [PubMed: 15350242]

Saxe R. Uniquely human social cognition. Curr Opin Neurobiol. 2006; 16:235–239. [PubMed: 16546372]

Shackman AJ, Salomons TV, Slagter HA, Fox AS, Winter JJ, Davidson RJ. The integration of negative affect, pain and cognitive control in the cingulate cortex. Nature Rev Neurosci. 2011; 12:154–167. [PubMed: 21331082]

Sharot T, Korn CW, Dolan RJ. How unrealistic optimism is maintained in the face of reality. Nature Neurosci. 2011; 14:1475–U1156. [PubMed: 21983684]

Shima K, Tanji J. Role for cingulate motor area cells in voluntary movement selection based on reward. Science. 1998; 282:1335–1338. [PubMed: 9812901]

Somerville LH, Heatherton TF, Kelley WM. Anterior cingulate cortex responds differentially to expectancy violation and social rejection. Nature Neurosci. 2006; 9:1007–1008. [PubMed: 16819523]

Ullsperger M, Volz KG, von Cramon DY. A common neural system signaling the need for behavioral changes. Trends Cogn Sci. 2004; 8:445–446. [PubMed: 15450505]

Ullsperger M, von Cramon DY. Error monitoring using external feedback: specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging. J Neurosci. 2003; 23:4308–4314. [PubMed: 12764119]

van Veen V, Krug MK, Schooler JW, Carter CS. Neural activity predicts attitude change in cognitive dissonance. Nature Neurosci. 2009; 12:1469–1474. [PubMed: 19759538]

Walther, E.; Weil, R. Balance principles in attitude formation and change: The desire to maintain consistent cognitions about people. In: Gawronski, B.; Strack, F., editors. Cognitive Consistency: A Fundamental Principle in Social Cognition. New York: Guilford Press; 2012. p. 351-368.

Yeung N, Botvinick MM, Cohen JD. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol Rev. 2004; 111:931–959. [PubMed: 15482068]

Zajonc, RB. Cognitive theories in social psychology. In: Lindzey, G.; Aronson, E., editors. The Handbook of Social Psychology. Reading, MA: Addison-Wesley; 1968. p. 320-411.

Zaki J, Schirmer J, Mitchell JP. Social influence modulates the neural computation of value. Psychol Sci. 2011; 22:894–900. [PubMed: 21653908]

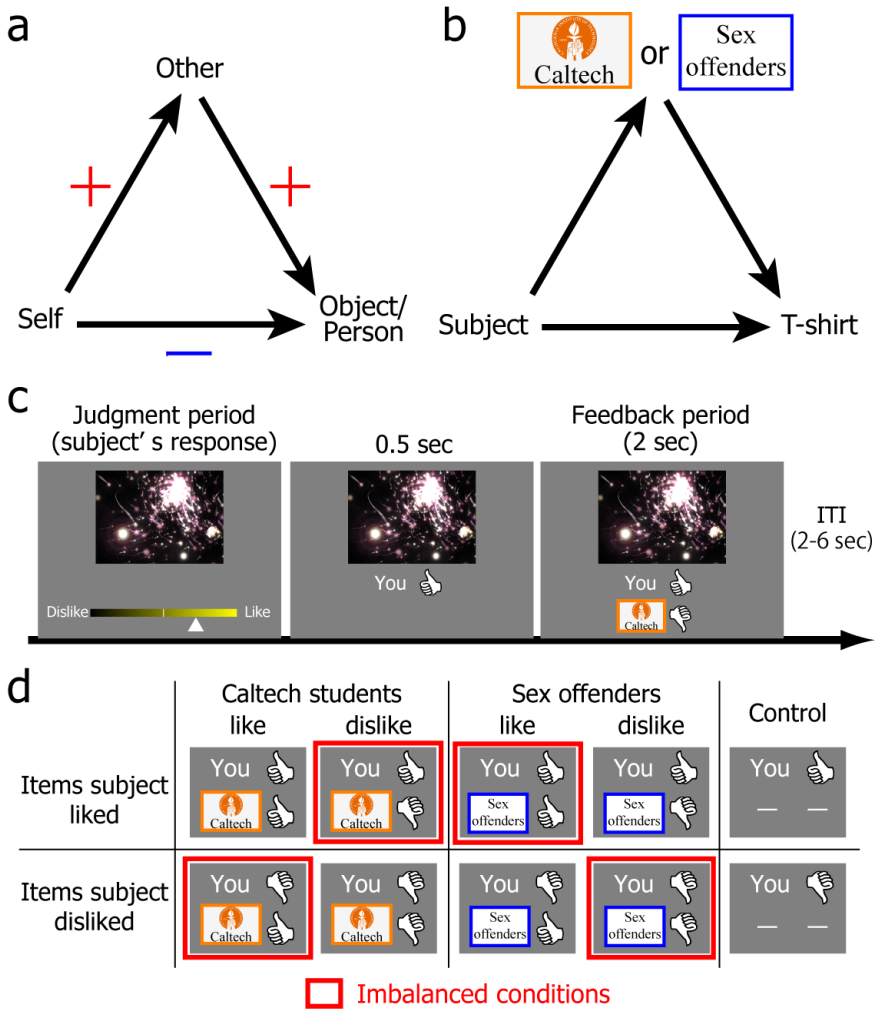

Figure 1.

Balance theory and experimental protocol. (**a**) Arrows represent the direction of evaluation together with indicated valence (+, like; −, dislike). Any imbalanced state has an odd number of negative (−) attitudes. The figure shows an example of an imbalanced state (e.g., one negative attitude) that would motivate a change in one's evaluation of the object (towards increased preference in this example). (**b**) Present experiment. Participants were students at the California Institute of Technology (Caltech). Their attitude toward others was manipulated by using a validated liked group (fellow Caltech students) and disliked group (sex offenders). Participants rated their preferences for t-shirts and were subsequently given feedback about the other group's preferences for the same t-shirts. (**c**) During the first preference rating task, subjects rated 174 t-shirt designs using a 14-point scale. Immediately after rating a t-shirt, subjects viewed their own preference and the preference of one of the two groups (either Caltech students or sex offenders), dichotomized as liked or disliked. Thumbs-up corresponded to ratings ≥ 8 and thumbs-down to ratings ≤ 7 on the 14-point scale. (**d**) Eight possible combinations of subjects' preference and others' preference (plus two control conditions). Four of them represent imbalanced states (highlighted by red squares) according to balance theory.

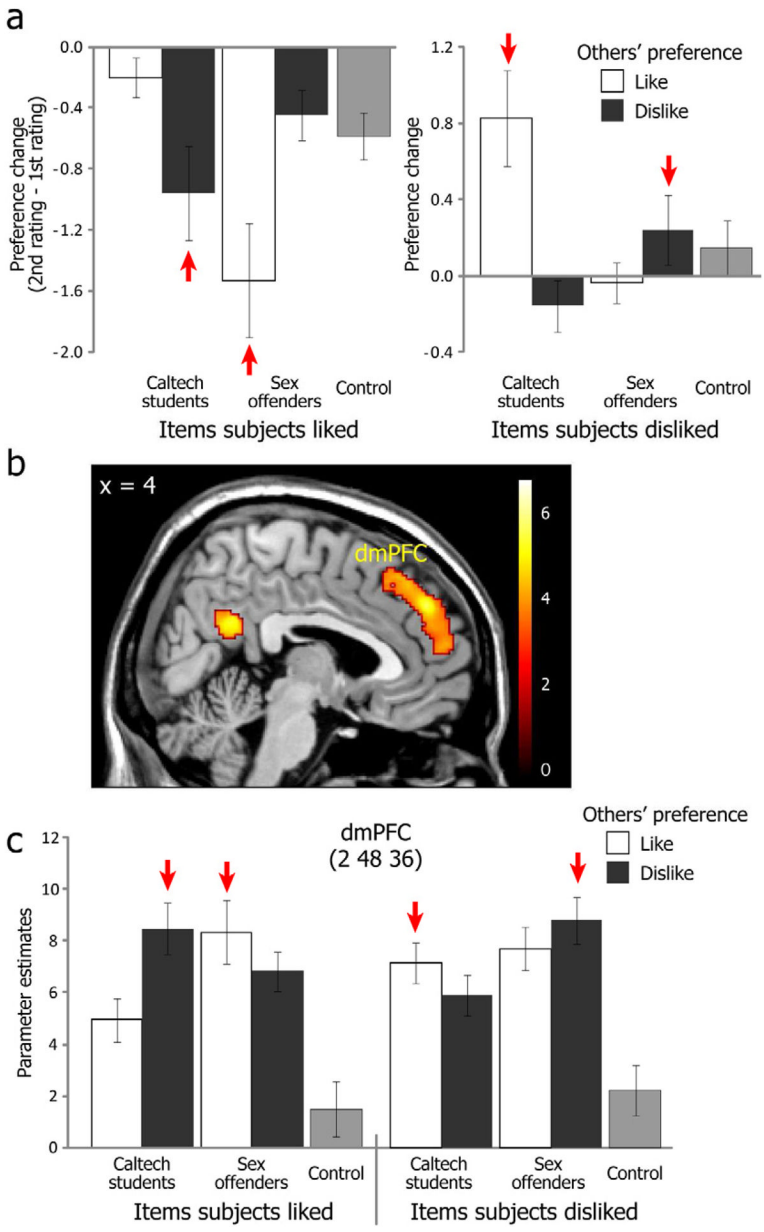

Figure 2.

Preference change and dmPFC activation induced by cognitive imbalance. (**a**) Self-reported preference change between second and first ratings. Red arrows indicate imbalanced conditions. (**b**) dmPFC regions significantly correlated with the degree of cognitive imbalance (Cognitive Imbalance Index; CII) in each trial. (**c**) Breakdown of activation patterns in dmPFC during the feedback period of the first preference rating tasks. Beta values were extracted using a leave-one-subject-out (LOSO) cross-validation procedure from the nearest local maximum from the peak activation identified by the conjunction analysis (see Methods for more details). Especially high activations were observed in imbalanced conditions (red arrows). Means and S.E.M. shown.

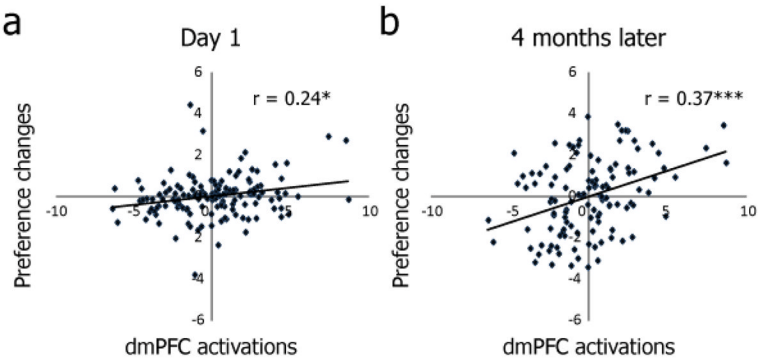

Figure 3.

Pooled within-subject correlations between preference change and dmPFC activation. dmPFC activations significantly predicted subsequent preference change, (**A**) several minutes after viewing others' preferences (18 subjects X 8 conditions = 144 data points) and, (**B**) even after four months (15 subjects X 8 conditions = 120 data points). Y-axis indicates preference change for each condition in the predicted direction (i.e., higher value indicates preference increase in Caltech students-like or sex offenders-dislike conditions, and preference decrease in Caltech students-dislike or sex offenders-like conditions). For both preference changes and brain activations, subject-mean centering was performed to remove between-subjects variance before computing correlations. * $p < 0.05$, *** $p < 0.001$ (after Bonferroni correction for 14 tests, see Table 2).

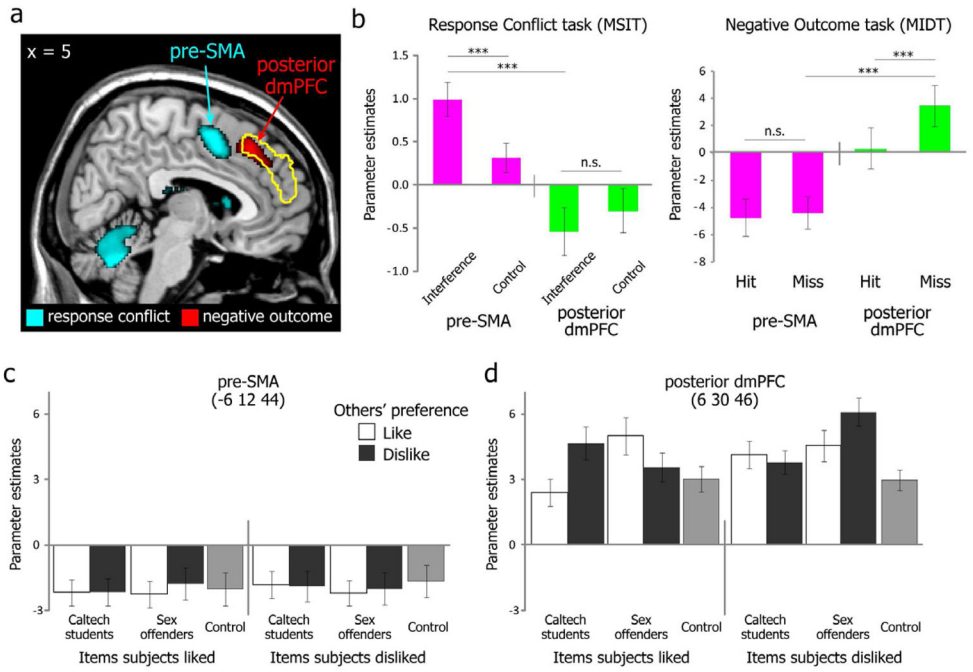

Figure 4.

pMFC areas sensitive to response conflict and negative outcome. (**a**) Response conflict-related areas were localized by the contrast of Interference vs. Control conditions in the MSIT task (Bush and Shin, 2006), which activated the pre-SMA (x = −6, y = 12, z = 44). Negative outcome-related areas were localized by the contrast of Miss vs. Hit feedback in the MIDT task (Knutson et al., 2000) which activated the posterior part of dmPFC (x = 6, y = 30, z = 46). The yellow outline indicates the dmPFC areas significantly correlated with CII in our balance task (cf. Figure 2b). (**b**) Activation patterns in pre-SMA and posterior dmPFC during the two localizer tasks. Beta values were extracted using a leave-one-subject-out (LOSO) cross-validation procedure from the local maxima from the peak activation identified by the Interference vs. Control contrast (pre-SMA) and the Miss vs. Hit contrast (posterior dmPFC) (see Methods for more details). Activation patterns in (**c**) pre-SMA and (**d**) posterior dmPFC during the feedback period of the t-shirt rating task. Means and S.E.M. shown.

Table 1

CII-related brain regions for both liked and disliked groups.

Location	BA	MNI coordinate			Z	Cluster size
		x	y	z		
dmPFC	8/9/10/32	2	48	36	4.81	948
Left IFG	44/45	−44	20	10	4.82	638
Left insula	13	−30	18	−26	5.40	
PCC	31	4	−54	26	4.54	288

Areas are identified by a conjunction analysis between CIIs for Caltech students and CIIs for sex offenders. BA, Brodmann Area. A statistical threshold was set at $p < 0.001$ voxelwise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons).

Table 2

Correlations between regional activation and preference changes.

	Day 1		4 months later	
Brain area	**$r_{(144)}$**	**p value**	**$r_{(120)}$**	**p value**
dmPFC	0.24	0.027 *	0.37	0.001 ***
L insula	0.17	0.29	0.30	0.007 **
L IFG	0.05	1	0.15	0.76
PCC	0.03	1	0.10	1
R ventral striatum	−0.19	0.14	−0.18	0.31
R insula	−0.12	1	−0.28	0.015 *
R IFG	−0.13	0.91	−0.24	0.062

dmPFC, left insula, left IFG and PCC were significantly positively correlated with the CII (see Figure 2 and Figure S3). Right ventral striatum, right insula and right IFG were significantly activated by the contrast of all balanced conditions vs. all imbalanced conditions (see Figure S4b). Correlations for Day 1 are based on 18 subjects so that there are 144 data points (18 subjects X 8 conditions). Correlations for four months later are based on 15 subjects so that there are 120 data points (15 subjects X 8 conditions). For both preference changes and brain activities, subject-mean centering was performed to remove between-subjects variance before computing correlations. Reported p values (one-tailed) are Bonferroni corrected for multiple comparisons (a total of 14 correlations).

* $p < 0.05$,

** $p < 0.01$,

*** $p < 0.001$.