doi:10.1093/scan/nsq054 SCAN (2011) 6, 460–467

The neural basis of rationalization: cognitive dissonance reduction during decision-making

Johanna M. Jarcho,[1,2] Elliot T. Berkman,[3] and Matthew D. Lieberman[1,3]

[1]Department of Psychiatry & Biobehavioral Sciences, [2]Center for Neurobiology of Stress, University of California, Los Angeles, CA 90025, and [3]Department of Psychology, University of Oregon, Eugene, OR 97403, USA

People rationalize the choices they make when confronted with difficult decisions by claiming they never wanted the option they did not choose. Behavioral studies on cognitive dissonance provide evidence for decision-induced attitude change, but these studies cannot fully uncover the mechanisms driving the attitude change because only pre- and post-decision attitudes are measured, rather than the process of change itself. In the first fMRI study to examine the decision phase in a decision-based cognitive dissonance paradigm, we observed that increased activity in right-inferior frontal gyrus, medial fronto-parietal regions and ventral striatum, and decreased activity in anterior insula were associated with subsequent decision-related attitude change. These findings suggest the characteristic rationalization processes that are associated with decision-making may be engaged very quickly at the moment of the decision, without extended deliberation and may involve reappraisal-like emotion regulation processes.

Keywords: cognitive dissonance; rationalization; attitudes; decisions; neuroimaging; right inferior frontal gyrus; insula

INTRODUCTION

Decision-making is a ubiquitous part of daily life and people often make difficult choices between equally attractive alternatives. Yet, there are unexpected consequences for making such decisions. After a choice is made between initially matched options, people no longer find the alternatives similarly desirable (Brehm, 1956; Harmon-Jones and Harmon-Jones, 2002). Rather, people adjust their attitudes to support their decision by increasing their preference for the selected option, decreasing their preference for the rejected option or both. This rationalization is thought to be motivated by the drive to reduce 'cognitive dissonance', an aversive psychological state aroused when there is a discrepancy between actions and attitudes (Festinger, 1957; Zanna and Cooper, 1974; Elliot and Devine, 1994). In situations when decisions cannot be reversed, or when doing so requires great effort, this discrepancy is often reduced by adjusting attitudes to be in line with decisions.

Despite decades of research characterizing decision-related attitude change, relatively little is known about the psychological mechanisms supporting it. Though self-report measures can provide a detailed account of magnitude and direction of attitude change, they shed less light on the cognitive and neural processes engaged in producing this change (Elliot and Devine, 1994). Moreover, attitude change associated with difficult decisions is known as 'post-decisional' attitude change even though most theories of cognitive dissonance are agnostic regarding the temporal course of attitude change. This name appears to reflect when attitudes are measured in the experimental process, rather than an empirically based reference to when processes driving this change are implemented. Nevertheless, the term has propagated the belief that attitude change is driven by relatively slow, reflective cognitive processes, engaged well after decisions have been made, during post-decision attitude assessment, which typically occurs many minutes after decision making has taken place (for examples see Lieberman *et al.*, 2001).

In contrast to traditional assumptions about decision-related attitude change, more recent models of cognitive dissonance suggest that the psychological distress associated with cognitive dissonance can begin to be resolved rapidly, with attitude change processes being engaged as an unintentional byproduct of decision making itself (Shultz and Lepper, 1996; Lieberman *et al.*, 2001; Simon *et al.*, 2004; Egan *et al.*, 2007). These models are supported by evidence from functional magnetic resonance imaging (fMRI) studies that demonstrate motor and cognitive conflict, as well as affective distress can be resolved within seconds, often as a function of activity in right inferior frontal gyrus (IFG) (Goel and Dolan, 2003; Aron *et al.*, 2004; Ochsner and Gross, 2005). Given that deciding between equally attractive options by definition provokes conflict, and attitude change

Received 18 March 2010; Accepted 19 May 2010
Advance Access publication 12 July 2010

For their generous support, the authors also wish to thank the Brain Mapping Medical Research Organization, Brain Mapping Support Foundation, Pierson-Lovelace Foundation, The Ahmanson Foundation, William M. and Linda R. Dietel Philanthropic Fund at the Northern Piedmont Community Foundation, Tamkin Foundation, Jennifer Jones-Simon Foundation, Capital Group Companies Charitable Foundation, Robson Family and Northstar Fund. This work was also supported by a National Research Service (Award #F31 DA021951) from the National Institute on Drug Abuse and Research Training in Psychobiological Sciences (#T32 MH017140) from the National Institute of Mental Health (to J.M.J.); Neuroimaging Training (Grant #T90 DA022768) from the National Institute on Drug Abuse (to E.T.B.); a grant from the National Institute of Mental Health (#MH 071521 to M.D.L.).

Correspondence should be addressed to Matthew D. Lieberman, Department of Psychiatry and Behavioral Sciences, University of California, Los Angeles, CA 90025, USA or Department of Psychology, University of Oregon, Eugene, OR 97403, USA. E-mail: lieber@ucla.edu

resolves that conflict, decision-related attitude change might involve reappraisal processes, which are often associated with rapid increases in right IFG, and decreases in limbic activity (Ochsner *et al.*, 2004; Kalisch *et al.*, 2005; Lieberman, 2007a; Tabibnia *et al.*, 2008). As such, activity in brain regions associated with conflict resolution during decision making, such as right IFG, may be associated with decision-related attitude change.

In order to investigate brain activity during decision making and determine whether conflict resolution processes occurring in that moment are associated with attitude change, we conducted an experiment with fMRI using a novel, scanner-compatible paradigm for inducing decision-related cognitive dissonance. Classic studies suggest 27–59% of subjects experience decision-related attitude change (Brehm, 1956). Since the goal of the current study was to investigate the neural mechanisms specific to decision-related attitude change, an a priori decision was made to limit analyses to individuals who exhibited the phenomenon. Just as neuro-imaging studies of placebo response typically exclude non-responders from analyses (Mayberg *et al.*, 2002; Sarinopoulos *et al.*, 2006) as a means of isolating specific, homogeneous mechanisms underlying those effects, individuals who did not demonstrate significant levels of attitude change ('non-responders') were not included in analyses.

METHODS

Subjects

Twenty-one subjects participated in a protocol approved by UCLA's Institutional Review Board, and were paid $30 for their time. Given the novelty of the current paradigm, sample size was based on an estimate derived from a pilot study (see Supplementary Data), which suggests 60–70% of subjects demonstrate reliable decision-related attitude change across trials. The majority of subjects (60%; $N = 12$; three male; mean age 22 ± 3.42 years) demonstrated reliable decision-related attitude change across trials, thus analyses were limited to this group of interest. One other subject was excluded for technical difficulties.

Behavioral procedures

Paradigms employed in classic behavioral studies of decision-related attitude change do not conform to the constraints of event-related fMRI. In behavioral studies, subjects make a small number of decisions between similarly rated items, with statistical power generated by sample size. In contrast, fMRI samples are typically smaller, with each subject being exposed to numerous trials. We designed a novel paradigm with many more decisions than classic studies, and confirmed its efficacy in a behavioral pilot study (reported in the Supplementary Data).

Prior to entering the scanner, subjects rated their liking for 140 names and 140 paintings on a 1 (strongly dislike) to 100 (strongly like) scale (Supplementary Figure S1). Once a subject completed their ratings, one experimenter positioned them in the scanner, while another identified a subset of 80 similarly rated pairs (40 names/40 paintings), defined as ≤ 10 points apart on the 100-point scale, for presentation during decision making.

In the scanner, subjects were told they would be presented with pairs of names and paintings, and they were to choose which item in each pair they preferred. Because decision-related attitude change is more likely to occur when decisions are meaningful (Festinger, 1957), subjects were asked to select names based on naming their future child, and paintings based on which they would rather hang in their home, with the belief they would receive posters of two selected paintings.

Stimuli were presented with MRI-compatible goggles over two counterbalanced 5-min runs (one names, one paintings). Each run consisted of 40 5-s trials during which subjects indicated by button press which item in each pair they preferred. Following the choice, the screen became blank for the duration of the trial and remained blank during inter-trial intervals, which were jittered randomly with a γ-distribution ($M = 2.5$ s) (Wager and Nichols, 2003).

After exiting the scanner, subjects again rated all 280 items, then learned they would not receive posters, but were compensated an additional $10.

Quantification of attitude change scores for each item

Each item from the 80 pairs of stimuli presented during decision making was classified *post hoc* as selected or rejected. The 120 items that were rated twice, but excluded from decision making, were classified as 'no choice' items. Attitude change was computed by subtracting initial from final ratings for each item; a positive score indicates an increase in liking during the study. Attitude change for each type of item (selected, rejected, no choice) was assessed with repeated measures *t*-tests. Magnitude of attitude change did not differ based on order of presentation or stimulus type. Data were therefore pooled across runs and stimulus type for analyses.

Quantification of attitude change scores for each trial

An attitude change score was calculated for each decision-making trial by computing the difference in attitude change for selected and rejected items as follows:

$$(\text{selected item} : \text{final} - \text{initial rating})$$
$$-(\text{rejected item} : \text{final} - \text{initial rating})$$

Higher scores indicate that selected items changed in the positive direction more than rejected items. Within-subjects *t*-tests were performed on attitude change scores across all trials for each subject to identify subjects exhibiting significant decision-related attitude change. These subjects were included in reported analyses.

fMRI data acquisition

Imaging was performed using a 3T Siemens Allegra scanner at UCLA's Ahmanson-Lovelace Brainmapping Center. We acquired 306 functional T2*-weighted echo-planar images (EPI) [slice thickness, 2 mm; 36 axial slices; repetition time (TR), 2 s; echo time (TE), 25 ms; flip angle, 90°; matrix, 64×64; field of view (FOV), 20 cm]. Two volumes were discarded at the beginning of each run to allow for T1 equilibrium effects. A T2-weighted matched-bandwidth high-resolution anatomical scan (same slice prescription as EPI) and magnetization-prepared rapid-acquisition gradient echo were acquired for each subject for registration purposes (TR, 5 s; TE, 33 ms; FOV, 20 cm; matrix, 128×128; sagittal plane; slice thickness, 3 mm; 36 slices).

fMRI data processing

fMRI data were analyzed using SPM5 (http://www.fil.ion.ucl.ac.uk/spm/software/spm5). Images were realigned within-subject to correct for head motion, slice-time corrected to adjust for timing within each TR, normalized into Montreal Neurological Institute standard stereotactic space and smoothed with an 8-mm Gaussian kernel, full width at half maximum. Voxel size was $2 \times 2 \times 2$ mm. A priori regions of interest (ROIs) were defined for right- and left-pars triangularus, which encompasses IFG, using the WFU Pickatlas (Maldjian *et al.*, 2003) and the AAL atlas (Tzourio-Mazoyer *et al.*, 2002). Although we specifically hypothesized a relationship would occur between attitude change and activity in right, but not left IFG, both regions were assessed to determine if laterality effects were present. ROI analyses were conducted using small volume correction (SVC) with significance level of $P < 0.05$ for magnitude of activation and extent threshold of 10 voxels for each of the above specified regions. We verified the false detection rate within our a priori ROIs was >0.05 using Monte Carlo simulations as implemented in the AlphaSim routine (part of the AFNI package). Whole-brain analyses were conducted using significance level of $P < 0.005$ for magnitude of activation and extent threshold of 10 voxels, which provides a reasonable balance with respect to Types I and II error concerns consistent with the false discovery rate in typical behavioral science papers (Lieberman and Cunningham, 2009).

Functional brain activity associated with attitude change

Data were modeled within subjects with an event-related regression analysis to determine whether brain activity during decision making was associated with attitude change. For each subject, each of the 80 trials was modeled as an event with a 5-s duration. The corresponding attitude change score for each of the trials was entered as a continuous regressor. Using this analysis technique results in each trial being convolved with the hemodynamic response function, with the attitude change score for the trial as a scaling factor for the hemodynamic response function. This analysis produced a single activation map for each subject reflecting regions significantly associated with attitude change. Results were entered into a random-effects analysis at the group level in the standard manner (i.e. a one-sample *t*-test at each voxel across subjects). This single-step, group-level analysis produced a map reflecting activity that reliably correlated with attitude change in the brain regions reported below.

Functional connectivity of brain regions associated with attitude change

Analyses were conducted to examine functional connectivity among attitude change-related brain regions. Psychophysiological interaction (PPI) analyses were used to quantify the extent to which the relationship between the time courses in a pair of brain regions changes as a function of the trial type (Friston *et al.*, 1997).

New design matrices were computed for each subject, with trials categorized as having large (>10%) or small amounts of decision-related attitude change (≤10%). The differential relationship between a source region and all other voxels in the brain during large and small attitude change trials were then contrasted. Whole-brain regression analyses searched for voxels whose time course correlated with the source region more negatively during large than small attitude change trials, controlling for task design itself and the zero-order correlation between the time courses of the two regions. Results were brought to the group level for a random effects analysis, producing an activation map of regions that were functionally connected with the source region.

RESULTS

Behavioral results

Behavioral results revealed the expected pattern of decision-related attitude change (Figure 1), replicating a prior pilot study of these methods (Supplementary Data). Change in attitudes toward selected items ($M = 5.08$, s.d. = 3.29) was significantly different than change toward rejected items [$M = -5.95$. s.d. = 4.07; $t(11) = 9.14$, $P < 0.001$]. Attitudes toward selected items became significantly more positive [$t(11) = 5.35$, $P < 0.001$], while attitudes toward rejected options became significantly more negative [$t(11) = -5.07$, $P < 0.001$], and attitudes about items rated twice without an intervening decision remained unchanged ($M = 4.35$, s.d. = 14.10; $t = 1.07$, $P =$ ns).

Neuroimaging results

Among the ROIs investigated, attitude change was positively associated with activity in right but not left pars triangularis. Whole-brain analyses further specified that within the pars triangularis, attitude change was positively associated with activity in right IFG (BA45), along with medial prefrontal cortex (BA10), precuneus (BA7), ventral striatum and parahippocampal gyrus. Bilateral anterior insula and lateral parietal cortex were the only regions throughout the brain

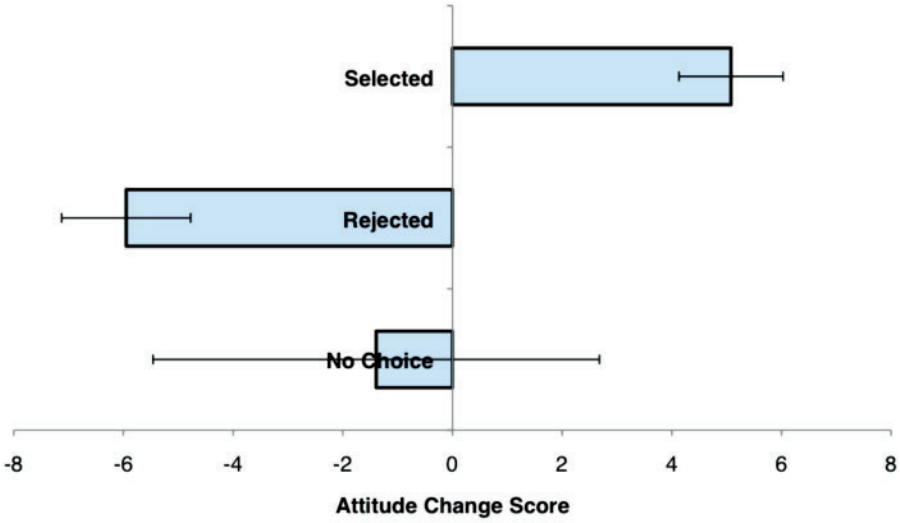

Fig. 1 Average change in attitudes from pre-choice to post-choice ratings. Items were categorized retrospectively based on whether the item was 'selected' or 'rejected' during decision making, or was rated twice but excluded from the decision-making phase of the study ('no choice').

whose activity was negatively correlated with attitude change (Figure 2 and Table 1).

These results support recent models of cognitive dissonance that suggest decision-related attitude change can occur during decision making as an immediate byproduct of conflict resolution processes. The positive correlation between IFG activity and attitude change, coupled with the negative correlation between anterior insula activity and attitude change is consistent with the suggestion that dissonance reduction may partially be a consequence of IFG downregulation of distress or arousal responses in the anterior insula, via selection of more decision-consistent interpretations of the stimuli. To examine this possibility further, functional connectivity with anterior insula (5-mm sphere centered at the peak voxel of activation identified in initial regression analysis) was assessed in a whole-brain PPI analysis. Consistent with a regulation account, the time course of a cluster in right IFG [54, 30, 20; $k = 67$; $t(11) = 4.32$, $P < 0.001$], overlapping with the right-IFG activation reported above, was negatively correlated with the time course of activity in left-anterior insula to a greater extent during trials with large compared with small amounts of attitude change (Figure 3).

DISCUSSION

Faced with a difficult decision between equally attractive alternatives, people often adjust their attitudes to support their choice. We demonstrated that processes associated with decision-related attitude change are engaged even when making numerous decisions in quick succession over a relatively short period of time. Using fMRI, we examined brain activity while difficult decisions were made, and observed that greater shifts in attitude were associated with increased activity in right IFG, medial fronto-parietal regions and ventral striatum, and decreased activity in anterior insula. Further analyses revealed that activity in right IFG was more negatively correlated with activity in left-anterior insula during trials with large compared to small amounts of attitude change.

These findings are consistent with newer models of cognitive dissonance, which suggest that cognitive mechanisms supporting attitude change can be engaged rapidly, without extended deliberation, as a by-product of the decision-making process itself (Shultz and Lepper, 1996; Lieberman *et al.*, 2001; Simon *et al.*, 2004; Egan *et al.*, 2007). A plausible mechanism underlying this change in attitudes is suggested by studies of emotion regulation, which demonstrate that selecting more desirable interpretations of threat is associated with relief from psychological distress, increased activity in IFG, and downregulation of limbic responses (Berkman and Lieberman, 2009). Extending this logic to the current study, conflict or distress produced early in the decision-making process (Supplementary Data) may be relieved by increased activity in right IFG, which can both facilitate a shift toward decision-consistent attitudes, and modulate activity in anterior insula, which often accompanies experiences of arousal, affective distress or discomfort (Sanfey *et al.*, 2003; reviewed by Ochsner and Gross, 2005). In a recent study, a positive relationship was found between activity in anterior insula and attitude change in the context of performing counter-attitudinal behavior, a situation also thought to elicit cognitive dissonance (van Veen *et al.*, 2009). This further supports the theory that arousal, conflict or distress set attitude change in motion, however, unlike the current study, van Veen and colleagues did not identify potential neural mechanisms activated during counter-attitudinal behavior associated with the resolution of distress that presumably facilitate attitude change. This suggests that

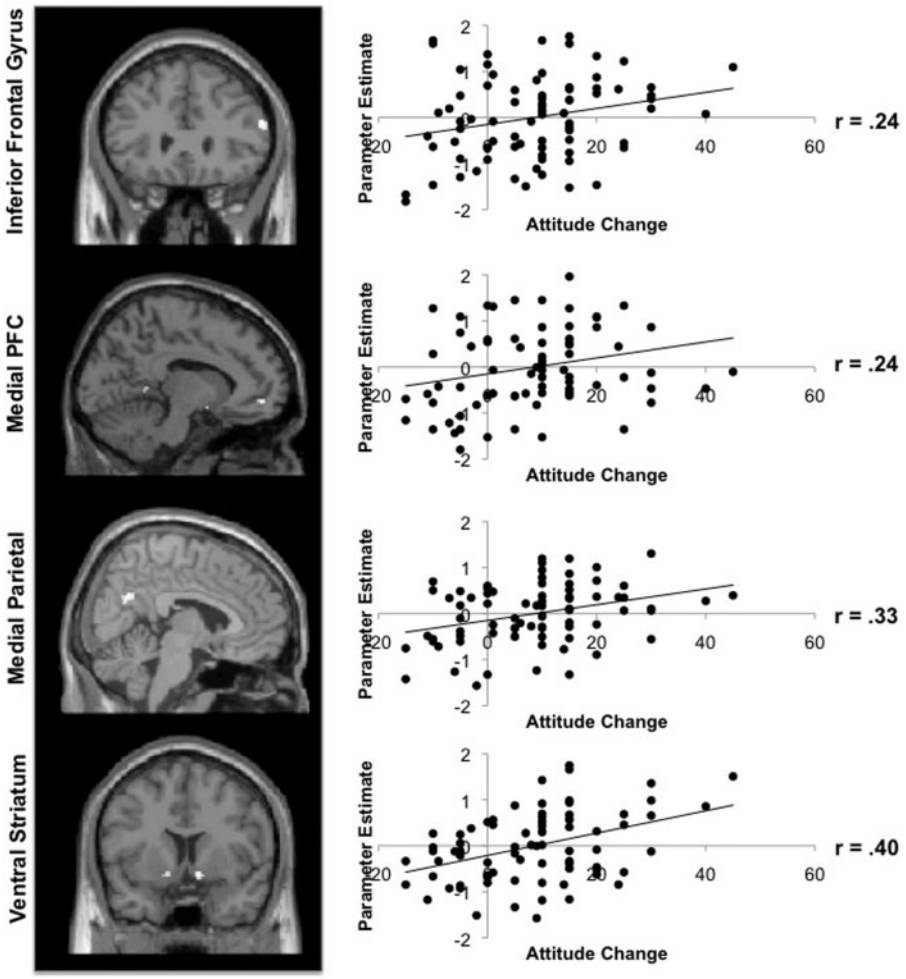

Fig. 2 Neural correlates of post-decision attitude change. The left column shows clusters of activation identified in group level analyses in right IFG, medial prefrontal cortex, medial parietal cortex and ventral striatum, whose time course of activation correlated positively with the attitude change measure from trial to trial. The right column shows scatterplots of attitude change and neural activation for a typical single subject in these brain regions for graphical purpose. Each point on the plot represents a single trial.

Table 1 Regional brain activity identified with event-related regression analyses, associated with magnitude of decision-related attitude change

Regions	BA	Coordinates			Voxels	*t*-values
		x	*y*	*z*	ke	
Positive correlation with attitude change						
Medial PFC	10/32	−8	50	−8	12	4.08
Inferior-frontal gyrus	45	56	28	18	24	5.19
Parahippocampal gyrus	29	−16	−50	4	41	4.18
Precuneus	7	6	−62	28	43	5.91
Ventral striatum		−12	4	−12	15	4.47
		10	8	−10	21	4.9
Negative correlation with attitude change						
Anterior insula		−30	22	−8	101	4.89
		32	26	−10	37	3.75
Inferior parietal lobule	40	−54	−36	44	28	3.8

Coordinates are in MNI space, significance based on $P < 0.005$ for magnitude of activation with an extent threshold of 10 voxels. PFC, prefrontal cortex.

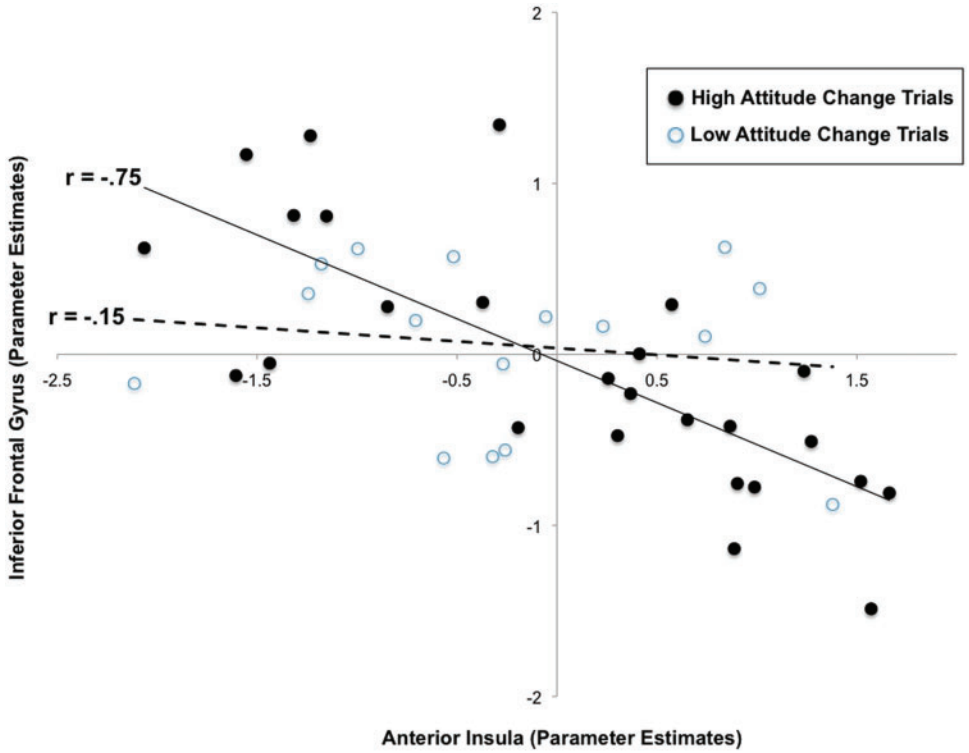

Fig. 3 Negative connectivity between right-IFG and left-anterior insula as a function of magnitude of attitude change. Data are representative of a typical subject. Solid circles represent trials resulting in large amounts of attitude change (>10%). Open circles represent trials resulting in small amounts of attitude change (from −10 to 10%). Derived from PPI analysis using a 5-mm sphere centered at the peak voxel of activity in anterior insula identified during initial regression analyses as a source region.

the conflict aroused during counter-attitudinal behavior in that study may have ultimately been resolved by processes engaged at a later point in time, perhaps during re-evaluation of attitudes. In contrast, decision making may involve a more rapid deployment of processes aimed at resolution of discomfort or distress generated early in the decision-making process, as demonstrated by the inhibitory relationship between right IFG and regions such as anterior insula, associated with attitude change in the current study.

The link between activity in medial frontoparietal regions and attitude change is also consistent with numerous models of cognitive dissonance reduction. Apart from conflict resolution, the most commonly invoked construct in dissonance-related attitude change is self-relevance. It has been suggested that conflicts of greater self-relevance arouse more dissonance and thus lead to greater attitude change (Aronson, 1968; Elliot and Devine, 1994). The fact that the medial fronto-parietal areas observed here are the regions most commonly activated in studies of self-reflection and self-reference (Lieberman, 2007a) suggests that more self-relevant decisions produce greater attitude change. In addition to its role in self-referential processing, medial prefrontal cortex has also been implicated in evaluative processes engaged while imagining positively valanced stimuli (Cunningham *et al.*, 2011), as well as generating goals directed at obtaining positive outcomes (Packer and Cunningham, 2009). Taken together, it seems medial prefrontal cortex may have the capacity to evaluate positively valenced, personally relevant stimuli and promote action toward goals directed at obtaining positive outcomes.

Activity in ventral striatum was also associated with attitude change. The striatum plays an important role in processing hedonic information related to reward outcomes (Delgado *et al.*, 2000; O'Doherty *et al.*, 2002), and tracks subjective value of those potential outcomes (Knutson *et al.*, 2001). The relationship between striatal activity and attitude change suggests preferences for stimuli are potentially being assessed and updated throughout the decision-making process. Indeed, a study that utilized a similar experimental paradigm described here, but focused on brain activity during pre- and post-decision stimulus ratings, found striatal activity during initial ratings predicted subsequent selection of items, and change in striatal activity during ratings following decision making was highly correlated with attitude change (Sharot *et al.*, 2009). Although brain activity during decision making was not assessed in that study, these data provide further evidence that attitudes are likely adjusted throughout the decision-making process.

While the current data are consistent with an emotion regulation account of decision-related attitude change, a similar inverse relationship between activity right-IFG and limbic regions has also been linked with cognitive processes associated with evaluation of stimulus characteristics more generally (Cunningham and Zelazo, 2007). For example,

affect-based appraisal of stimuli (Lieberman *et al.*, 2007), inhibition or promotion of specific stimulus features during evaluation (Ochsner and Gross, 2005) and appraisal of stimuli as a function of contextual factors (Lieberman *et al.*, 2004; Wager *et al.*, 2004) are each associated with increased activity in right IFG and decreased activity in limbic regions of the brain. Thus, deliberative evaluative processes may also facilitate attitude change in the context of the decision-making process. Additional studies are needed to better clarify the role of evaluative processes, conflict and emotion regulation in the context of decision-related attitude change and cognitive dissonance more generally.

Although this investigation was not designed to address mechanisms underlying decision-making processes *per se*, the decision-making literature provides further support for the idea that decision-making processes may in and of themselves promote attitude change. For example, when making value-based purchasing decisions, rejecting an attractive item for purchase is associated with increased activity in bilateral insula, whereas accepting such an item for purchase results in decreased activity in the same regions, accompanied by increased activity in IFG (Knutson *et al.*, 2007). In the current study, a strikingly similar pattern of activity was related to the attitude change associated with making decisions. Likewise, subjective value of items involved in a purchasing decision is often associated with activity in ventromedial prefrontal cortex (Chib *et al.*, 2009), as well as ventral striatum (Knutson *et al.*, 2001) and activity in these same regions relate to attitude change in the current study. Together, these data support the idea that attitude change may be a natural byproduct of decision-making processes, and may not always require extended, deliberative thought to occur. This does not, however, preclude the idea that deliberative processes activated sometime after decision making may also contribute to attitude change, but instead, allows for the possibility that processes engaged much more quickly and without extended deliberation have the capacity to influence attitudes as well.

The ability to set resolution processes in motion quickly when faced with affective distress has important implications. Unresolved affective conflict or distress interferes with other on-going cognitive processes, while resolution of this conflict, often via top-down inhibition of activity in limbic regions of the brain, rapidly restores processing capacity (Etkin *et al.*, 2006). This may allow cognitive and attentional resources to be directed toward more relevant emotional stimuli in the current context. Additionally, resolving affective distress may also facilitate subsequent decision-promoting behavior (Harmon-Jones *et al.*, 2008). If conflict or distress is unresolved, future actions toward decision-related targets may be encumbered by recurrent affective distress, eliciting re-engagement of processes directed at attempting resolve this distress. Thus, activating processes associated with attitude change during decision making may be an adaptive strategy for both engaging in on-going cognitive processes beyond a specific decision, and allow subsequent behavior toward decision targets to be less effortful.

This investigation of cognitive dissonance processes with fMRI provides new insights into the neuro-cognitive mechanisms by which making a difficult choice between equally attractive alternatives can produce attitude change (i.e. rationalization). These results suggest that processes driving attitude change may be engaged quickly, and involve conflict resolution mechanisms associated with coping with affective distress. The fact that neuro-cognitive processes engaged during a few seconds of decision making are associated with attitude change suggests that, though attitude change may appear like disingenuous rationalization from the outside, the processes driving it may in fact be engaged quite quickly, and without the individual's explicit intention.

SUPPLEMENTARY DATA

Supplementary data are available at *SCAN* online.

Conflict of Interest

None declared.

REFERNCES

Aron, A.R., Robbins, T.W., Poldrack, R.A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, *8*, 170–7.

Aronson, E. (1968). Dissonance theory: progress and problems. In: Abelson, R.P., editor. *Theories of Cognitive Consistency: A Sourcebook*. Chicago: Rand McNally.

Berkman, E.T., Lieberman, M.D. (2009). Using neuroscience to broaden emotion regulation: theoretical and methodological considerations. *Social and Personality Psychology Compass*, *3*, 475–93.

Brehm, J.W. (1956). Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology*, *52*, 384–9.

Chib, V.S., Rangel, A., Shimojo, S., O'Doherty, J.P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience*, *29*, 12315–20.

Cunningham, W.A., Zelazo, A. (2007). Attitudes and evaluations: a social cognitive neuroscience perspective. *Trends in Cognitive Science*, *11*, 97–104.

Cunningham, W.A., Johnsen, I.R., Waggoner, A.S. (2011). Orbitofrontal cortex provides cross-modal valuation of self-generated stimuli. *Social Cognitive and Affective Neuroscience*, *6*, 286–93.

Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C., Fiez, J.A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, *84*, 3072–7.

Egan, L.C., Santos, L.R., Bloom, P. (2007). The origins of cognitive dissonance: evidence from children and monkeys. *Psychological Science*, *18*, 978–83.

Elliot, A.J., Devine, P.G. (1994). On the motivational nature of cognitive dissonance: dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, *67*, 382–94.

Etkin, A., Egner, T., Peraza, D.M., Kandel, E.R., Hirsch, J. (2006). Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. *Neuron*, *51*, 871–82.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Evanston: Row, Peterson & Company.

Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage*, *6*, 218–29.

Goel, V., Dolan, R.J. (2003). Explaining modulation of reasoning by belief. *Cognition*, *87*, B11–22.

Harmon-Jones, E., Harmon-Jones, C. (2002). Testing the action-based model of cognitive dissonance: the effect of action orientation on postdecisional attitudes. *Personality and Social Psychology Bulletin*, *28*, 711–23.

Harmon-Jones, E., Harmon-Jones, C., Fearn, M., Sigelman, J.D., Johnson, P. (2008). Left frontal cortical activation and spreading of alternatives: Tests of the action-based model of dissonance. *Journal of Personality and Social Psychology*, *94*, 1–15.

Kalisch, R., Wiech, K., Critchley, H.D., et al. (2005). Anxiety reduction through detachment: subjective, physiological, and neural effects. *Journal of Cognitive Neuroscience*, *17*, 874–83.

Knutson, B., Adams, C.M., Fong, G.W., Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, *21*, 1–5.

Knutson, B., Rick, S., Wimmer, G.E., Prelec, D., Loewenstein, G. (2007). Neural predictors of purchases. *Neuron*, *53*, 147–56.

Lieberman, M.D. (2007). Social cognitive neuroscience: a review of core processes. *Annual Review of Psychology*, *58*, 259–89.

Liberman, M.D., Cunningham, W. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social Cognitive and Affective Neuroscience*, *4*, 423–8.

Lieberman, M.D., Eisenberger, N.I., Crockett, M.J., Tom, S.M., Pfeifer, J.H., Way, B.M. (2007). Putting feelings into words: Affect labeling disrupts amygdala activity in response to affective stimuli. *Psychological Science*, *18*, 421–8.

Lieberman, M.D., Jarcho, J.M., Berman, S., et al. (2004). The neural correlates of placebo effects: A disruption account. *NeuroImage*, *22*, 447–55.

Lieberman, M.D., Ochsner, K.N., Gilbert, D.T., Schacter, D.L. (2001). Do amnesics exhibit cognitive dissonance reduction? The role of explicit memory and attention in attitude change. *Psychological Science*, *12*, 135–40.

Maldjian, J.A., Laurienti, P.J., Kraft, R.A., Burdette, J.H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, *19*, 1233–9.

Mayberg, H.S., Silva, J.A., Brannan, S.K., et al. (2002). The functional neuroanatomy of the placebo effect. *The American Journal of Psychiatry*, *159*, 728–37.

Ochsner, K.N., Gross, J.J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences*, *9*, 242–9.

Ochsner, K.N., Ray, R.D., Cooper, J.C., et al. (2004). For better or for worse: neural systems supporting the cognitive down- and up-regulation of negative emotion. *NeuroImage*, *23*, 483–99.

O'Doherty, J.P., Deichmann, R., Critchley, H.D., Dolan, R.J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron*, *33*, 815–26.

Packer, D.J., Cunningham, W.A. (2009). Neural correlates of reflection on goal states: The role of regulatory focus and temporal distance. *Social Neuroscience*, *4*, 412–25.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*, 1755–8.

Sarinopoulos, I., Dixon, G.E., Short, S.J., Davidson, R.J., Nitschke, J.B. (2006). Brain mechanisms of expectation associated with insula and amygdala response to aversive taste: implications for placebo. *Brain, Behavior, and Immunity*, *20*, 120–32.

Sharot, T., De Martino, B., Dolan, R.J. (2009). How choice reveals and shapes expected hedonic outcome. *Journal of Neuroscience*, *29*, 3760–5.

Shultz, T.R., Lepper, M.R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, *103*, 219–40.

Simon, D., Krawczyk, D.C., Holyoak, K.J. (2004). Construction of preferences by constraint satisfaction. *Psychological Science*, *15*, 331–6.

Tabibnia, G., Satpute, A.B., Lieberman, M.D. (2008). The sunny side of fairness: Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science*, *19*, 339–47.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*, 273–89.

van Veen, V., Krug, M.K., Schooler, J.W., Carter, C. (2009). Neural activity predicts attitude change in cognitive dissonance. *Nature Neuroscience*, *12*, 1469–74.

Wager, T.D., Nichols, T.E. (2003). Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *NeuroImage*, *18*, 293–309.

Wager, T.D., Rilling, J.K., Smith, E.E., et al. (2004). Placebo-induced changes in fMRI in the anticipation and experience of pain. *Science*, *303*, 1162–7.

Zanna, M.P., Cooper, J. (1974). Dissonance and the pill: an attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, *29*, 703–9.

Supplementary Material for "The neural basis of rationalization: Cognitive dissonance reduction during decision-making"

Johanna M. Jarcho[1,2]

Elliot T. Berkman[3]

Matthew D. Lieberman[3]

[1]Department of Psychiatry & Biobehavioral Science, University of California, Los Angeles, CA 90025, USA.

[2]Center for Neurobiology of Stress, University of California, Los Angeles, CA 90025, USA

[3]Department of Psychology, University of California, Los Angeles, CA 90025, USA

Pilot Study

Because decision-related attitude change had not been examined with fMRI, we conducted a pilot study to determine the feasibility of such a study. Constraints of event-related fMRI studies differ sufficiently from those involved in classic behavioral studies of decision-related attitude change (Brehm, 1956) that the standard paradigm needed to be modified. In behavioral studies, subjects have traditionally made a small number of decisions between similarly rated items, with statistical power generated by including dozens of subjects. In contrast, fMRI samples are typically small by comparison, with each subject being exposed multiple times to each type of trial. Thus, we designed a decision-related attitude change paradigm in which subjects made many more decisions than in a typical behavioral study.

Our goals in this pilot study were twofold. First, we wanted to determine whether a modified paradigm with multiple decision-making trials would produce group-level effects similar to studies using a classic paradigm. It is possible that repeated decisions by individual subjects might contaminate or interfere with the typically observed effect of decision-related attitude change. Second, we wanted to determine the proportion of subjects that demonstrated significant within-subject effects, and thus exhibited reliable attitude change across many trials. This is critical because only those subjects who produce reliable effects will be used in the event-related fMRI analyses conducted in the primary study. Just as neuroimaging studies of the placebo response typically exclude non-responders from analyses (Mayberg et al., 2002; Sarinopoulos et al., 2006), we also planned to exclude those who did not demonstrate significant levels of post-decision attitude change ('non-responders') from our neuroimaging analyses. Determining the

responder rate in this pilot study will inform our sample size in the main neuroimaging study.

METHODS

Subjects

Thirteen right-handed, English-speaking subjects (4 male; mean age 20 ± 1.58 years) were recruited from the UCLA campus. All subjects gave informed consent to participate according to a protocol approved by the University of California, Los Angeles Institutional Review Board, and received 1 hour of class credit for their time, plus an additional $10 of compensation. Three additional subjects were eliminated from analyses because of technical difficulties.

Procedures

Except when noted, behavioral methods are the same as those used in the study reported in the main text (depicted in Supplementary Fig. S1). All portions of the study were completed in a behavioral testing room. Once subjects provided their initial ratings of how much they liked 140 names and 140 paintings, they exited the testing room and completed filler questionnaires while the experimenter prepared stimuli for the decision-making phase of the study. This delay between initial ratings and decision-making was designed to be similar in duration to the delay subjects would experience in an fMRI study (i.e., the positioning of the subject and structural scanning that precede functional scans). Subjects were then invited back into the testing room, where they were told they would be presented with pairs of names and pairs of paintings, and they were to choose

the item in each pair they preferred. Just as in the primary study, when subjects were presented with pairs of names, they were instructed to imagine they were choosing a name for their own future child, and to select which name in each pair they would rather name that child. When choosing between paintings, subjects were led to believe that they would receive posters of two paintings they indicated they preferred.

The pairs of similarly rated items were presented over two counterbalanced, 5-minute runs, with one run of names and the other of paintings. Each run consisted of 40 5-second trials during which subjects indicated by button press which item in each pair they preferred. Once subjects made this choice a blank screen replaced the stimuli for the duration of the 5-second trial.

After decision-making, subjects once again rated how much they liked all 280 items on a 1-100 scale. At the end of the study, subjects learned they would not receive posters and were instead compensated an additional $10.

Quantification of attitude change scores

An 'attitude change score' was calculated for each item by subtracting the initial rating from the final rating of an item, such that a positive score indicated attitudes toward an item had become more positive during the course of the study. Stimuli that comprised each of the 80 pairs presented during decision-making were classified, post hoc, as a selected or rejected item. The 120 items that were rated twice, but not included in the decision-making phase were classified as 'no choice' items. Results are depicted in Supplementary Fig. S2.

Quantification of post-decision attitude change scores at the subject level

Previous behavioral studies find that subjects demonstrate decision-related attitude change effect *on average*. It remains unknown what proportion of subjects consistently show the effect from trial to trial because there have not been enough trials in past studies to measure this within-subjects effect. To assess this, we computed a paired samples t-test for each subject to determine if attitude changes scores were greater for selected compared with rejected items across all trials. Significant differences indicate the subject demonstrated a general tendency to produce decision-related attitude change. To control for the possibility that attitude change was simply due to repeated ratings, attitude change scores were also computed for 120 items that were rated twice, without an intervening choice.

RESULTS

Magnitude of attitude change did not differ by the order in which decision-making runs occurred, nor was it affected by whether stimuli were presented during the first or second half of initial and final rating phases. Initial ratings of names and paintings were not significantly different, and attitude change scores did not vary based on whether stimulus pairs were comprised of paintings or names. Given this, data were pooled together across runs and across paintings and names for the remainder of analyses.

The first goal of the study was to determine whether the adapted paradigm produced decision-related attitude change at the group level. Attitude change scores were significantly more positive for selected items (M=2.27, SD=5.02) than rejected items (M=-3.61, SD=6.71; $t(12)=3.08$, $p<.01$). Since items were paired based on the similarity

of their initial ratings, this difference indicates that following decision-making, selected items came to be viewed more positively, relative to rejected items, (Supplementary Fig. S2). Attitudes about no choice items did not change (M=-1.39, SD=3.97; t(12)=1.26, p=ns).

The second goal of the pilot study was to determine what percent of subjects produced reliable decision-related attitude change across trials. Paired samples t-tests compared the attitude change score for selected and rejected items on a subject-by-subject basis, and revealed that the majority of subjects (69.3%) demonstrated a significant level of attitude change across trials. The mean difference in attitude change scores between selected and rejected items for subjects with significant attitude change ("responders") was 7.90 (SD=7.40), and the mean for other subjects ("non-responders") was 1.34 (SD=1.90). Based on these data, we expected 60%-70% of subjects sampled from a similar undergraduate population to be responders in the fMRI study.

DISCUSSION

We sought to develop an experimental paradigm that was able to produce decision-related attitude change while conforming to methodological constraints posed by fMRI. A classic experimental paradigm known to produce decision-supporting attitude change (Brehm, 1956) was modified such that the number of items subjects rated, as well as the number of subsequent decisions made about those items, was increased dramatically, while the amount of time provided for decision-making was limited to 5 seconds. Additionally, the stimuli that subjects rated and made choices about were drawn from multiple domains. Despite these modifications, significant levels of attitude change

occurred, regardless of the order in which choices and ratings were made, and the type of stimuli that were presented. The ability to replicate previous behavioral results, even with a relatively small sample size, suggests the modified paradigm utilized in the pilot study is a valid paradigm that can be implemented in the context of fMRI studies decision-related attitude change. Additionally, we observed that 69% of subjects produced significant within subject attitude changes effects. This suggests that in our main fMRI study, 60%-70% of subjects should have enough trials with attitude change to conduct within subject event-related analyses.

Supplementary data from fMRI study

Both classic and current models of cognitive dissonance suggest conflict sets cognitive dissonance reduction, and thus attitude change, in motion (Festinger, 1957). As such, the absence of a relationship between activity in dorsal anterior cingulate, a brain region most commonly associated with conflict (Botvinick et al., 2004), is somewhat puzzling. Also somewhat puzzling is the negative relationship between attitude change and activity in the anterior insula, a brain region associated with negative affect and distress (Sanfey et al., 2003; Tabibnia et al., 2008), which is often considered the other factor necessary to drives cognitive dissonance reduction (Festinger, 1957). Exploratory analyses altering the time window investigated for each trial may help account for these apparent discrepancies.

Exploratory analysis of the relationship between attitude change and functional brain activity preceding decision-making

Exploratory event-related regression analyses focused on brain activity that occurred during the initial period of each trial. Specifically during the time that elapsed between initial presentation of each pair of stimuli, and the subject's button press indicating which item in the pair they preferred. Duration of this initial period was automatically recorded via computer, and was used to define the time frame of analysis for each trial.

Each trial was modeled as an event with the attitude change score entered as a parametric modulator. Each trial, which varied in duration based on response time, was convolved with the hemodynamic response function; the attitude change score for that trial was then used as a scaling factor for the hemodynamic response function. This analysis produced an activation map of regions that were significantly associated with attitude change for each subject during the time between initial presentation of stimuli, and the button press that signified a decision had been made. These maps were then entered into a random-effects analysis at the group level.

RESULTS AND DISCUSSION

Although absent in analyses that included the entire 5-second trial, when analyses were limited to the time prior to subjects' decision on each trial (M=2.19 sec, SD=.38), left dorsal anterior cingulate was positively associated with attitude change (-2, 32, 40; k=180; t(11)=4.35). Whereas activity in anterior insula was negatively correlated with attitude change across the entire 5-second trial, when analyses were limited to the time

that elapsed prior to subjects' decision on each trial, left anterior insula activity was also positively correlated with attitude change (-30, 32, -2; k=20; t(11)=3.52, p<.005).

Thus, activity in dorsal anterior cingulate, which is often associated with conflict, is positively associated with attitude change, but only early during the decision-making process. Activity in the anterior insula, which occurred very early in the decision-making process, was also positively associated with attitude change, but the extent to which it was ultimately dampened down over the entire trial, was negatively associated with attitude change. Although exploratory in nature, these data provide support for the idea that greater conflict or discomfort may be associated with greater attitude change. Indeed, in a recent study, a similar relationship between activity in anterior insula, dorsal anterior cingulate, and attitude change was demonstrated in the context of performing counter-attitudinal behavior, a situation also thought to elicit cognitive dissonance (van Veen et al., 2009), further supporting the theory that conflict or distress set attitude change in motion. However, unlike the current study, van Veen and colleagues did not identify potential neural mechanisms activated during counter-attitudinal behavior associated with the resolution of distress that presumably facilitate attitude change. This suggests that the conflict aroused during counter-attitudinal behavior may ultimately be resolved at a later point in time, perhaps during re-evaluation of attitudes. In contrast, decision-making may involve a more rapid resolution of discomfort or distress generated early in the decision-making process, as demonstrated by the inhibitory relationship between right IFG and regions such as anterior insula, associated with attitude change.

SUPPLEMENTARY REFERENCES

Botvinick, M.M., Cohen, J.D., Carter, C., 2004. Conflict monitoring and anterior cingulate cortex: an update. Trends in Cognitive Sciences 8, 539-546.

Brehm, J.W., 1956. Postdecision changes in the desirability of alternatives. The Journal of Abnormal and Social Psychology 52, 384-389.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Oxford, England: Row, Peterson 291.

Mayberg, H.S., Silva, J.A., Brannan, S.K., Tekell, J.L., Mahurin, R.K., McGinnis, S., Jerabek, P.A., 2002. The functional neuroanatomy of the placebo effect. The American Journal of Psychiatry 159, 728-737.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D., 2003. The neural basis of economic decision-making in the Ultimatum Game. Science 300, 1755-1758.

Sarinopoulos, I., Dixon, G.E., Short, S.J., Davidson, R.J., Nitschke, J.B., 2006. Brain mechanisms of expectation associated with insula and amygdala response to aversive taste: implications for placebo. Brain, Behavior, and Immunity 20, 120-132.

Tabibnia, G., Satpute, A.B., Lieberman, M.D., 2008. The sunny side of fairness: Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychological Science 19, 339-347.

van Veen, V., Krug, M.K., Schooler, J.W., Carter, C., 2009. Neural activity predicts attitude change in cognitive dissonance. Nat Neurosci 12, 1469-1474.

Supplementary Figure Captions

Supplementary Figure S1. Diagram of experimental design. (a) Initial Rating: Outside the fMRI scanner, subjects rated how much they liked paintings and names on a 1-100 point scale. (b) Decision-Making: Once inside the scanner, subjects indicated by button press which one of two similarly rated items they would rather hang in their home (paintings) or name their child (names). (c) Final Rating: After exiting the scanner, subjects re-rated all items.

Supplementary Figure S2. Pilot Study: Average change in attitudes from pre-choice to post-choice ratings. Items were categorized post hoc based on whether the item was 'selected' or 'rejected' during decision-making, or was rated twice but excluded from the decision-making phase of the study ('No Choice').

Supplementary Figure S1

Supplementary Figure S2