Neuron, Volume *78*

Supplemental Information

Social Manipulation

# of Preference in the Human Brain

Keise Izuma and Ralph Adolphs

Supplemental Figures

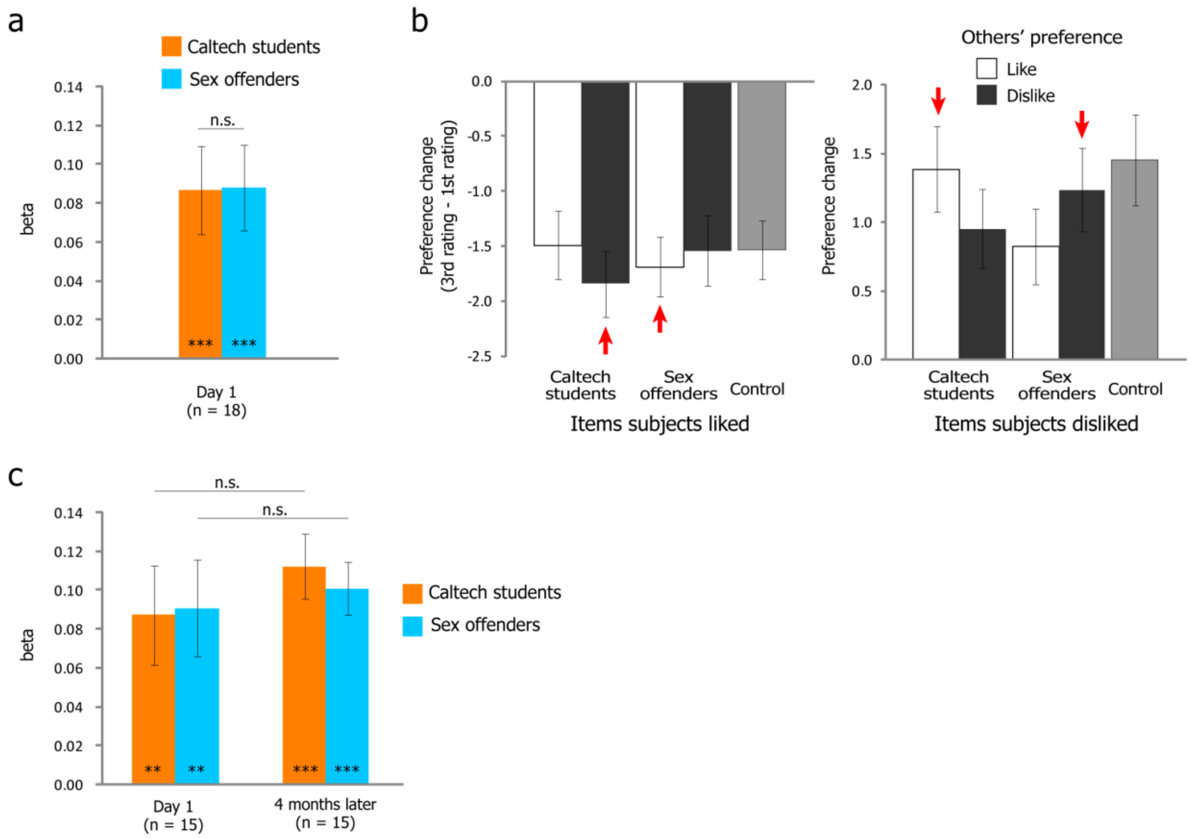

Figure S1, Related to Figure 2A. Behavioural results in the preference rating task

(**a**) Average beta values for the effect of the CII on preference change (2nd ratings – 1st ratings). The degree of cognitive imbalance (CII) significantly predicted subjects' preference changes. Note that since we expect negative preference changes (preference decrease) when Caltech students disliked or sex offenders liked items, the CII was multiplied by –1 in these cases before entering into a regression. (**b**) The effect of others' opinion on preference change after four months. Red arrows indicate imbalanced condition according to balance theory. Note that preferences showed considerable change even without any social influence ("control") after this time interval. Nonetheless, the significant 3-way interaction we found confirms that subjects' preferences remained socially influenced even after 4 months. (**c**) Average beta values for the effect of the CII on preference change in Day 1 (2nd ratings – 1st ratings) and four months later (3rd ratings – 1st ratings). The analyses are based on the data of 15 subjects who participated the third preference rating task. For panels **a** and **c**, asterisks inside each bar are based on one-sample t-test (one-tailed). ** $p < 0.01$, *** $p < 0.001$. Means and S.E.M. shown.

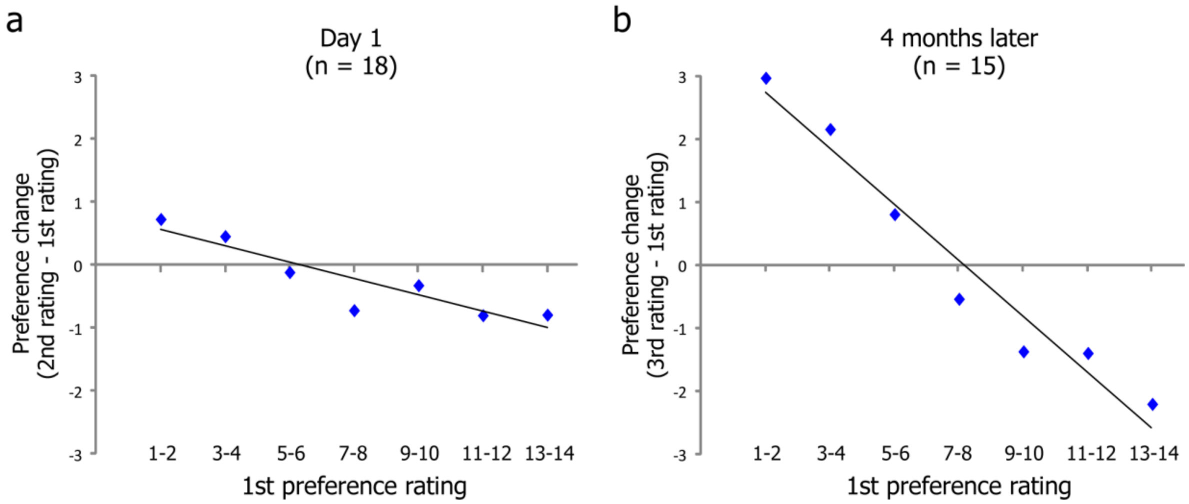

Figure S2, Related to Figure 2A. Regression-to-the-mean effect

Regression-to-the-mean effect in the control conditions (30 t-shirts) seen in the second preference rating task (**a**) and the third preference rating task (**b**). Preference changes in the control conditions were plotted as a function of subjects' first preference ratings with a linear regression line.

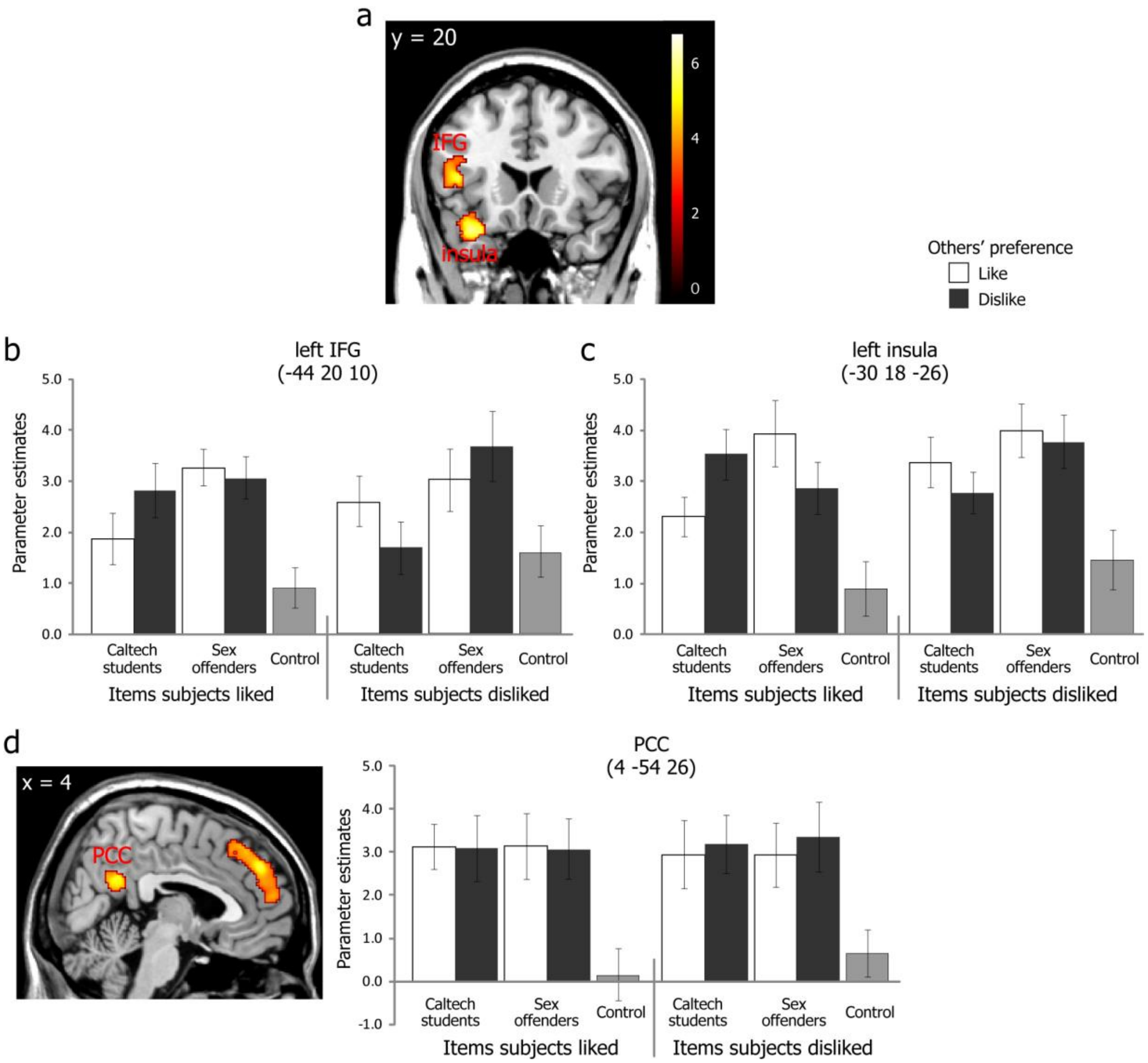

Figure S3, Related to Figure 2B. Activation patterns of the areas significantly correlated with CII.

(**a**) Left IFG and left insula regions significantly correlated with the CII for both groups (conjunction analysis). Activation patterns in left IFG (**b**) and left insula (**c**). The 3-way (Group X Self preference X Other preference) interactions were significant in both regions. (**d**) The PCC region significantly correlated with CIIs and its activation pattern. PCC did not show a significant 3-way interaction. Beta values are extracted using a LOSO cross-validation procedure (see Methods for more details). Means and S.E.M. shown.

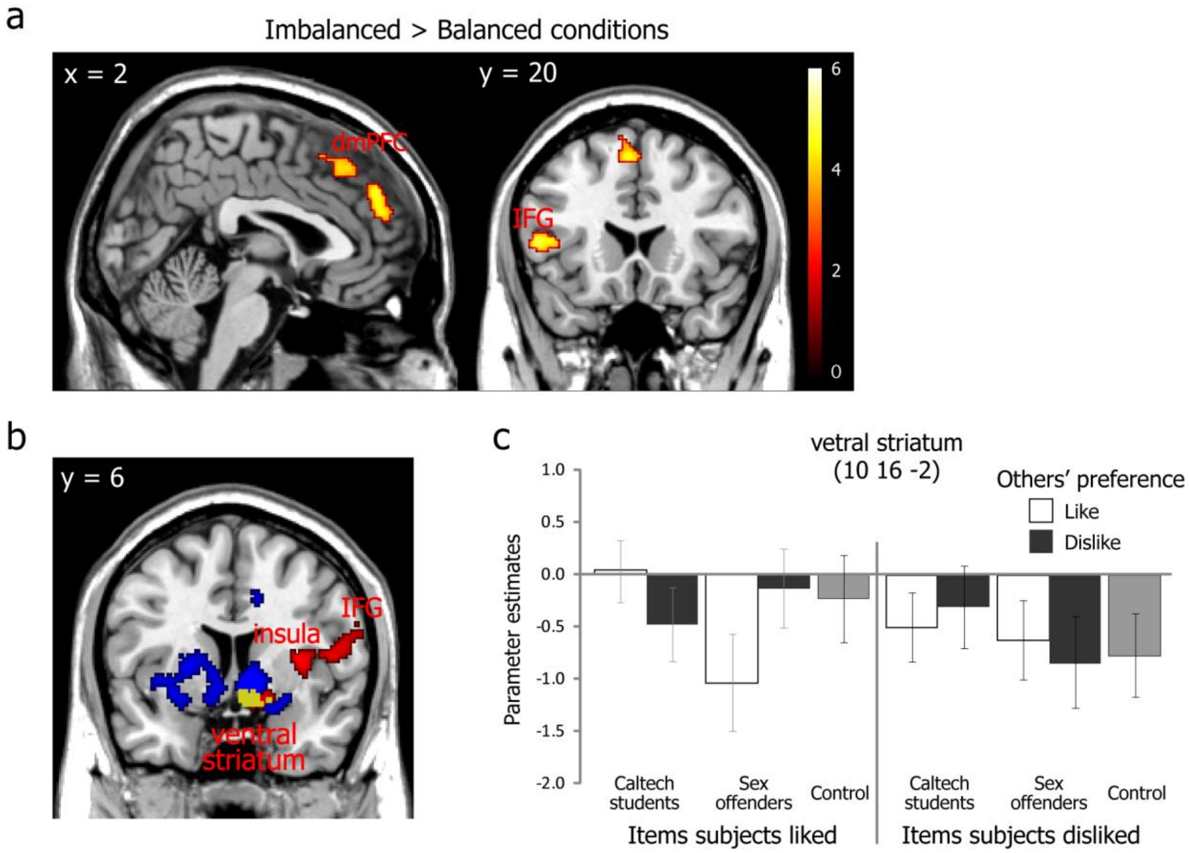

Figure S4, Related to Figure 2B. Brain areas activated by the imbalanced vs. balanced contrast and the reverse contrast

(**a**) Brain areas activated by the contrast of four imbalanced conditions vs. four balanced conditions (i.e., exploring the 3-way interaction contrast) (see Figure 1d). dmPFC, and left IFG showed significant activations. No other activation was found in this contrast. (**b**) Brain areas activated by the contrast of four balanced conditions vs. four imbalanced conditions (red), brain areas sensitive to reward cues (blue; area whose activity are parametrically modulated by reward level), and their overlap (yellow). Right ventral striatum (x = 10, y = 16, z = -2), right insula (x = 34, y = 4, z = 8) and right IFG (x = 48, y = 4, z = 10) showed higher activations in the four balanced conditions compared to imbalanced conditions, and the striatal region was also sensitive to reward. Because large regions were significantly associated with monetary reward in the MIDT, the more stringent threshold of $p < 0.05$ (whole brain FWE corrected) was used to display reward-related activation (blue). (**c**) Activation patterns in right ventral striatum. Beta values are extracted using a LOSO cross-validation procedure (see Methods for more details). Means and S.E.M. shown.

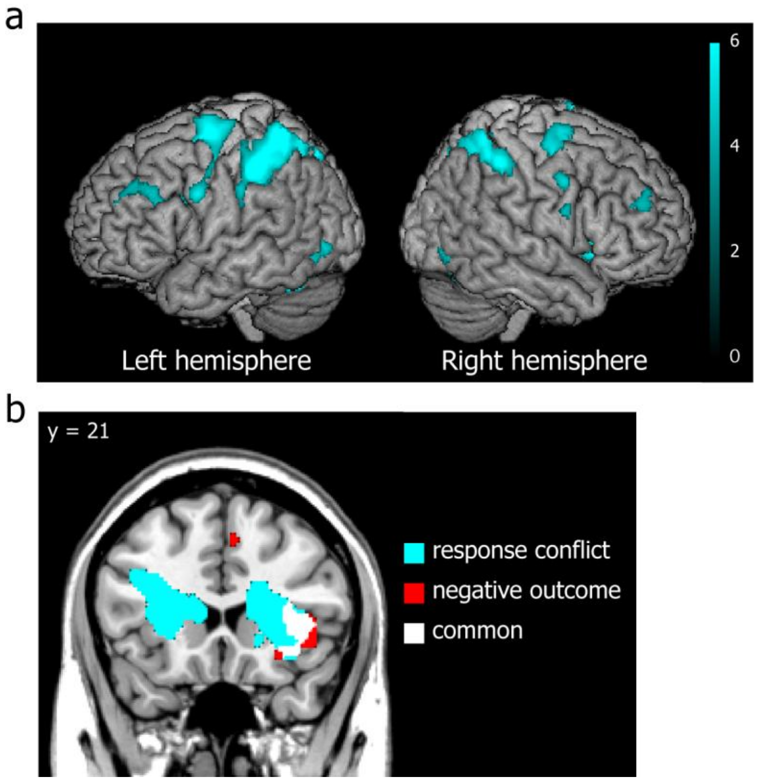

Figure S5, Related to Figure 4A. Brain areas activated by two localizer tasks

(**a**) Brain areas activated by the Interference vs. Control contrast in MSIT. (**b**) Coronal slice showing the overlapped areas in right insula between response conflict related areas in MSIT and negative feedback related areas in MIDT.

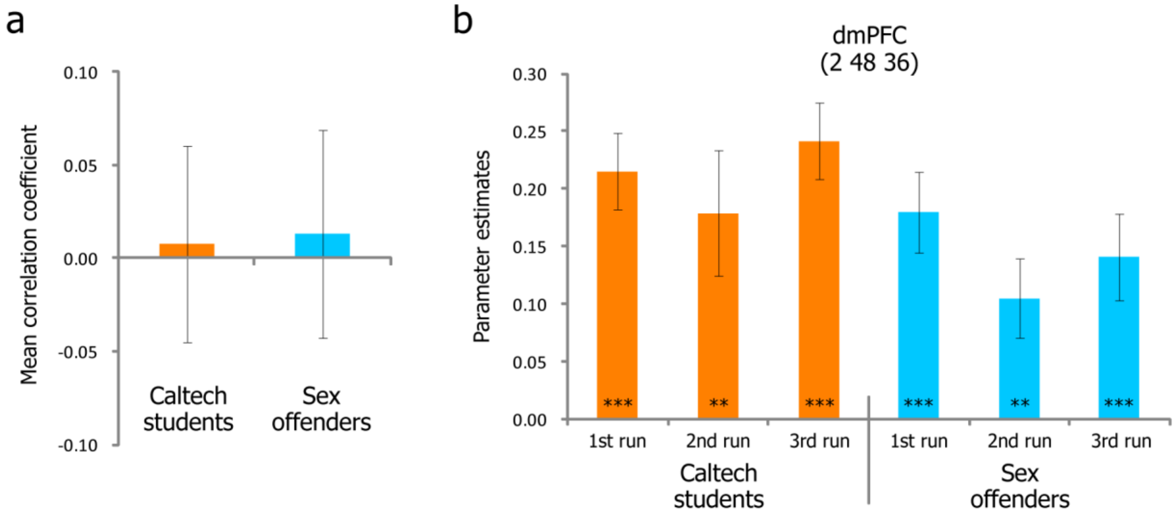

Figure S6, Related to Figure 2. Relation between subject's own preference and others' preference, and the dmPFC activation over time

(**a**) Mean correlation coefficients between subjects' own preference and their expectation of others' preferences for each group (Caltech students or sex offenders). (**b**) Association between the dmPFC activity and the CII separately for each of 3 fMRI runs in the first preference rating task. Asterisks inside each bar are based on one-sample t-test (one-tailed). ** $p < 0.01$, *** $p < 0.001$. Means and S.E.M. shown.

Table S1, Related to Figure 2B. Cognitive Imbalance Index (CII)

	CII				
Subject's preference	Caltech students Like	Caltech students Dislike	Sex offenders Like	Sex offenders Dislike	Control
1	13	0	0	13	0
2	12	1	1	12	0
3	11	2	2	11	0
4	10	3	3	10	0
5	9	4	4	9	0
6	8	5	5	8	0
7	7	6	6	7	0
8	6	7	7	6	0
9	5	8	8	5	0
10	4	9	9	4	0
11	3	10	10	3	0
12	2	11	11	2	0
13	1	12	12	1	0
14	0	13	13	0	0

Higher CII indicates greater social imbalance.

Supplemental Experimental Procedures

Preference ratings in the first preference rating task and effects of regression-to-the-mean

On average, subjects' ratings of the t-shirts were fairly equally distributed between liked (rating >= 8; 56%) and disliked (rating <=7; 44%) items. The 2 (Group) x 2 (Self preference) x 2 (Other preference) repeated measure ANOVA on the first preference ratings showed no significant effects (except for the obvious main effect of self preference). Furthermore, the first preference rating in each of our eight conditions (Figure 1d) did not differ significantly from the corresponding control condition (Self-Like control condition or Self-Dislike control condition) (ps > 0.18), suggesting that subjects' preference changes found in the present study were largely driven by the social feedback they received between first and second ratings.

There was a small effect of regression to the mean, as would be expected from any slightly noisy data: as seen in Figure S2, preferences for items highly liked by subjects in the first rating tended to decrease in the second and third ratings even in the control condition (i.e., without social influence). Similarly, preferences for items highly disliked by subjects in the first rating tend to increase in subsequent ratings even in the control condition. However, the pattern of preference changes we found in the second and third preference rating tasks cannot be explained simply by regression towards the mean since the first preference ratings were fairly even distributed across conditions to begin with (i.e., the strength of any effects of regression-to-the-mean effect should not differ between conditions).

We confirmed that the CII predicted self-reported preference change even after controlling for any regression-to-the-mean. As the strength of any regression-to-the-mean effect depends on the first preference rating for each t-shirt (Figure S2), we ran a linear regression model predicting preference changes with subject's first ratings as an additional regressor. Even so, the CII significantly predicted subjects' preference changes for both student and offender groups in the second rating (Day 1) as well as third rating tasks (4 months later) (all ps < 0.031).

Memory test after third preference rating task.

After their third preference rating task, 15/18 participants performed a memory task asking them to rate what they remembered the other groups' (students, offenders) preferences had been when they were revealed four months prior. Subjects' performances (26.1%) were no different from chance level (25%) ($t_{(14)} = 1.14$, $p = 0.14$, n.s.). Thus, after four months, subjects did not explicitly remember how each t-shirt had been rated by Caltech students or sex offenders. Individual differences in the effect of the CII on preference change after four months were not correlated with their memory performance (Caltech students $r = –0.25$, $p = 0.38$ n.s. and sex offenders $r = 0.14$, $p = 0.62$, n.s.) nor interval (number of days elapsed) between the first and third preference rating tasks (Caltech students $r = 0.04$, $p = 0.90$ n.s. and sex offenders $r = –0.12$, $p = 0.68$, n.s.). Thus, participants' memory for others' preferences were not related to how strongly their own preferences remained influenced after this extended period of time.

Areas activated by the imbalance vs. balance contrast (and the reverse contrast).

The dmPFC and left IFG activations were also found when all four imbalanced conditions were contrasted against all four balanced conditions regardless of group (Figure S4a). No other area was activated in this contrast.

When all four balanced conditions were contrasted with all four imbalanced conditions, right ventral striatum, right IFG and right insula showed significant activations (Figure S4b). Further ROI analysis with a leave-one-subject-out (LOSO) cross-validation procedure (see Methods for details) showed that the 3-way (group X self preference X other preference) interaction was significant in right ventral striatum (p = 0.009; see Figure S4c). The ventral striatum, an area involved in reward processing, was previously reported to show higher activation when one's opinion agreed with an expert's (Campbell-Meiklejohn et al., 2010) or the majority's (Klucharev et al., 2009) opinion. Extending those previous studies, our data showed that the ventral striatum is activated not only when a participant's opinion agrees with that of the Caltech students (liked group) but also when a participant's opinion disagrees with that of the sex offenders (disliked group), consistent with the interpretation that subjects experienced balanced states as more rewarding than imbalanced states. However, the activity in ventral striatum (also right IFG and right insula) did not track the degree of cognitive imbalance (CII).

dmPFC activity is not explained by expectation violation or surprise signal.

Several past studies reported that dmPFC regions are related to expectation violation (Somerville et al., 2006) or a "surprise" signal (Hayden et al., 2011). If we assume that subjects expect the preferences of other Caltech students to be the same as their own, but the preferences of sex offenders to be the opposite of theirs, the CII might be confounded with the degree of expectation violation.

To address this issue, we asked subjects to guess others' preference for t-shirts presented in the two control conditions (so that they had no other information about the preferences of others for these t-shirts) at the end of the experiment. Their expectations of others' preferences were unrelated to their own preference, with average correlation coefficients between these two sets of preference ratings not significantly different from zero for both groups (both p > 0.80) (Figure S6a). This suggests that participants did not necessarily think that their preference should be similar to other students' preferences or completely opposite to those of the sex offenders.

A final possibility might be that as we gave participants feedback about others' preferences which were by design decorrelated with the participant's own preferences, participants learned, over the course of the experiment, that other students' or sex offenders' preferences cannot be predicted from their own preference. If so, the correlation between the CII and expectation violation might be attenuated over time. However, when we extracted dmPFC beta values separately for each of the three fMRI runs in the first preference rating task, the data showed no sign of decreasing correlation between the CII and the dmPFC activity, and the CII significantly predicted the dmPFC activity for both groups even during the last run (ps < 0.001) (Figure S6b). Furthermore, individual differences in dmPFC beta values in the last fMRI run were not related to the strength of perceived association between a participant's own preference and their expectation of others' preferences (ps > 0.32). Taken together, these data suggest that the dmPFC activity found in the present study is highly unlikely to be explained by simple violation of expectation about another person's preference.

Multi-Source Interference Task (MSIT)

Behavioural results

As expected, our data showed that the MSIT produced a robust reaction time (RT) interference effect. RTs during the Interference condition (mean RT = 719 ms) were much slower than the Control condition (mean RT = 470 ms), and the difference was highly significant ($t_{(17)} = 20.2$, $p < 0.001$). There was also a significant difference in performance (Control = 99.6%, Interference = 94.6%; $t_{(17)} = 2.53$, $p = 0.011$).

fMRI results

As seen in Figure 4a, the contrast of Interference vs. Control conditions revealed significant activations in the pre-SMA. Bilateral dorsolateral prefrontal cortex (DLPFC), superior parietal lobule (SPL) and insula were also activated (Figure S5a), results largely consistent with prior reports (Bush and Shin, 2006). We also found that the strength of the activation in the pre-SMA (Interference > Control) was positively correlated with individual differences in the RT interference effect ($r = 0.57$, $p = 0.007$), but not with the performance difference between the two conditions ($r = –0.04$, $p = 0.43$, n.s.).

Monetary Incentive Delay Task (MIDT)

Behavioral results

The mean RT during the MIDT was 243 ms. RTs were significantly modulated by incentive. Subjects were significantly faster to respond in $2.0 trials (176 ms) compared to $0 trials (209 ms; $t_{(17)} = 5.45$, $p < 0.001$) and $0.2 trials (199 ms; $t_{(17)} = 4.43$, $p < 0.001$). There was also a significant difference in RTs between $0 and $0.2 trials ($t_{(17)} = 3.29$, $p = 0.002$). Similar results were obtained for hit rate. Overall mean hit rate was 65.5 %. Mean hit rates in $2 (71.5 %) were significantly higher than in $0 trials (59.3 %; $t_{(17)} = 4.41$, $p < 0.001$) and $0.2 (66.7 %; $t_{(17)} = 2.06$, $p = 0.028$) trials. There was also a significant difference between $0 and $0.2 trials ($t_{(17)} = 2.27$, $p = 0.019$).

We also tested if subjects adjusted their behavior depending on the feedback they received in a previous trial. We expected that if subjects use feedback information for behavioral adjustment, their RTs should be faster after they received a miss feedback in a previous trial than after receiving a hit feedback. To test this idea, we ran a linear regression analysis with the RT difference between each trial n and a previous trial $n – 1$ as a dependent variable. Then we analyzed that difference as a function of whether subjects had received hit or miss feedback on trial $n – 1$ (dummy coded as hit = 1, miss = 0). The incentive level of each trial *n* ($0, $0.2 or $2) and RT in a previous trial *n* – 1 were also included as regressors. We excluded RTs faster than 100 ms, slower than 600 ms, or deviating from the subject's mean by more than 3 standard deviations.

The result revealed that, as expected, the feedback subjects received had a significant influence on their RTs in the next trial (i.e., RTs were faster after receiving a miss feedback; mean $\beta = 4.10$, $t_{(17)} = 2.15$, $p = 0.024$) even when the incentive level and RT in a previous trial were taken into consideration. Not surprisingly, the incentive level and RT in a previous trial also had significant effects (both ps < 0.001).

fMRI results

To identify areas especially sensitive to negative feedback, we compared all miss feedback vs. all hit feedback regardless of incentive level. As seen in Figure 4a, the posterior part of dmPFC was activated by this contrast. The only other activated area was the right insula, and the right insula was also the only region showing overlap between response conflict-related areas in the MSIT and negative feedback-related areas in the MIDT (Figure S5b).

It should be noted however that when the statistical threshold was lowered (p < 0.01 uncorrected), a common activated area was found in the area between pre-SMA and posterior dmPFC activations reported in Figure 4)

Furthermore, as widely reported previously in fMRI studies using the MIDT (Knutson et al., 2000), the activity in ventral striatum was significantly modulated by reward magnitude (see Figure S4b; blue).

Supplemental References

Bush, G., and Shin, L.M. (2006). The Multi-Source Interference Task: an fMRI task that reliably activates the cingulo-frontal-parietal cognitive/attention network. Nature Protoc. *1*, 308-313.

Campbell-Meiklejohn, D.K., Bach, D.R., Roepstorff, A., Dolan, R.J., and Frith, C.D. (2010). How the opinion of others affects our valuation of objects. Curr. Biol. *20*, 1165-1170.

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

Klucharev, V., Hytonen, K., Rijpkema, M., Smidts, A., and Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. Neuron *61*, 140-151.

Knutson, B., Westdorp, A., Kaiser, E., and Hommer, D. (2000). FMRI visualization of brain activity during a monetary incentive delay task. Neuroimage *12*, 20-27.

Somerville, L.H., Heatherton, T.F., and Kelley, W.M. (2006). Anterior cingulate cortex responds differentially to expectancy violation and social rejection. Nature Neurosci. *9*, 1007-1008.