Behavioral/Cognitive

Anterior Insula Activity Reflects the Effects of Intentionality on the Anticipation of Aversive Stimulation

Mimi Liljeholm,[1,2] **Simon Dunne,**[1,3] **and John P. O'Doherty**[1,3]

[1]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland, [2]Department of Cognitive Sciences, University of California, Irvine, California 92697, and [3]Division of the Humanities and Social Sciences and Computation and Neural Systems Program, California Institute of Technology, Pasadena, California 91125

If someone causes you harm, your affective reaction to that person might be profoundly influenced by your inferences about the intentionality of their actions. In the present study, we aimed to understand how affective responses to a biologically salient aversive outcome administered by others are modulated by the extent to which a given individual is judged to have deliberately or inadvertently delivered the outcome. Using fMRI, we examined how neural responses to anticipation and receipt of an aversive stimulus are modulated by this fundamental social judgment. We found that affective evaluations about an individual whose actions led to either noxious or neutral consequences for the subject did indeed depend on the perceived intentions of that individual. At the neural level, activity in the anterior insula correlated with the interaction between perceived intentionality and anticipated outcome valence, suggesting that this region reflects the influence of mental state attribution on aversive expectations

Key words: anterior insula; aversive stimuli; intentionality

Introduction

Imagine that you are working in an office and a colleague brushes past, knocking a scalding hot cup of coffee into your lap. Perhaps your reaction, and subsequent evaluative judgments regarding this colleague, would be very different depending on your inferences about their intent: if you think that the colleague spilled their coffee on you deliberately, you would likely develop a strong antipathy toward that individual. If, on the other hand, you noticed that the colleague tripped while going past your desk, thus attributing a lack of intentionality to their actions, you might not develop a strong aversion to that person. Inferences about an individual's intent play a critical role in judgments about moral conduct and criminal liability. It is not known, however, whether such abstract attributions also modulate affective responses to directly experienced biologically salient aversive stimuli. In this study, probing the limits of social cognition, we used fMRI to examine the influence of perceived intentionality on neural responses to individuals whose actions resulted in the immediate delivery of a noxious stimulus to the subject in the scanner.

While in the scanner, the subject observed four confederates, each of whom was choosing between two options that yielded slightly different monetary outcomes for the confederate. Each decision by a confederate resulted in the immediate delivery of one of two liquid outcomes, the aversive flavor (salty tea) or an affectively neutral taste (water), to the subject in the scanner via an electronic syringe pump positioned in the control room. In the initial setup, the subject was made to believe that two of the confederates knew about the consequences of their actions for the subject in the scanner (intentional), whereas the other two did not (nonintentional). One confederate in each intentionality condition routinely chose the option associated with the delivery of the aversive outcome, ostensibly for a minor monetary gain (1 cent per trial), whereas the other routinely delivered the neutral one for the same monetary gain (see Fig. 1*A*). We hypothesized that there would be an interaction between perceived intentionality and predicted outcome valence, such that the difference between evaluative ratings of a confederate that had delivered the aversive outcome and one that had delivered the neutral outcome would be greater in the intentional than in the nonintentional condition, and that this interaction would be mediated by areas implicated in both representing intentionality and in encoding predictions about aversive outcomes.

Materials and Methods

Participants

We ran two versions of the study: Experiment 1 and Experiment 2. Nineteen healthy normal volunteers (9 males, age 22.13 ± 4.04 years) participated in Experiment 1, and 18 healthy normal volunteers (9 males, age 22.74 ± 2.86 years) in Experiment 2. All participants were recruited locally from the city of Dublin, Ireland. In addition, for each participant in the scanner, another 4 confederates played the role of the individuals outside the scanner. To ensure availability, a total of 12 confederates (all male) were recruited from Trinity College Dublin, 4 of whom were assigned to participate in a given experimental session. The assignment of these 4 confederates to intentionality and outcome valence conditions

Received March 19, 2014; revised June 22, 2014; accepted July 14, 2014.

Author contributions: M.L. and J.P.O. designed research; M.L. and S.D. performed research; M.L. analyzed data; M.L. and J.P.O. wrote the paper.

This work was supported by a Wellcome Trust grant and by support from the NIMH Conte Center for the Neurobiology of Social Decision Making (P50MH094258-01A1) to J.O.D.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Mimi Liljeholm, Department of Cognitive Sciences, University of California, Irvine, CA 92697. E-mail: m.liljeholm@uci.edu.

DOI:10.1523/JNEUROSCI.1126-14.2014

was randomized across participants. Written informed consent was obtained from all participants, and the study was approved by the School of Psychology Research Ethics Committee at Trinity College Dublin.

Apparatus

The salty tea consisted of 0.5 M NaCl dissolved in water to which cold black tea has been added. This combination of salt and cold tea has been found in previous experiments to be aversive to human subjects on the basis of subjective pleasantness ratings (Kim et al., 2011). The affectively neutral stimulus was composed of bottled water. The liquids were delivered to the subjects using two computer-controlled syringe pumps placed in the scanner control room, which were attached to an SP220I electronic syringe pump (World Precision Instruments) in combination with electrically operated solenoid valves (100T3M, Biochem Valve). The liquids were delivered intraorally via a plastic tube connected to 60 ml Becton Dickinson syringes at one end and were placed in the subject's mouth at the other end. On each trial, 1 ml of the relevant liquid was delivered, a manageable quantity that can be easily swallowed while lying down (O'Doherty et al., 2001). In Experiment 1, a pressure pad transducer (MP-150, BIOPAC Systems) was taped near the participants' laryngeal prominence to measure motion due to swallowing. However, because these data were too noisy to be useful, the measure was abandoned in Experiment 2.

Experiment 1. At the start of the experiment, the participant was brought into a large room together with the four confederates. Four computer monitors had been set up in this room, each displaying the choice screen with which each confederate would interact (these choice screens were also displayed to the subject in the scanner on each trial; see Fig. 1*A*). The experimenter verified the names of all present and then read the following script:

"Each of you has been randomly assigned to one of three conditions. One of you will be in an observer condition, where you will observe and then make judgments about the decisions and actions of the others. The rest of you will be active conditions where you will be playing a simplified slot machine game. On each trial, you will have to decide between which of two buttons to press, red or green. Each option will have a monetary reward associated with it and one option will always be one cent greater than the other. Two of the slot machine players will be in an interactive condition where, in addition to the monetary reward, each button press will result in a liquid being delivered to the observer, either an aversive taste (salty tea), or water, as seen here. [At this point the experimenter brought all participants over to the two monitors that showed choice screens with options labeled according to liquid type]. The other two will be in a noninteractive condition, where you are only playing for money, so the buttons are just labeled Option A and Option B, and the button presses will not result in any liquid delivery. [The experimenter brought all participants over to the monitors showing arbitrarily labeled choice screens]. And the assignment to conditions is as follows: (participant name), you will be the observer, and you have also been selected for scanning today, so I will take you into a separate room in a minute. (confederate 1) and (confederate 2), you will be playing the slot machines that deliver liquid to (participant name), so you will be using these computers over here [the two confederates are directed toward the relevant computers], and (confederate 3) and (confederate 4), you will not be delivering any liquid to (participant), so you'll be playing the noninteractive slot machines over here. But before we begin, all of the slot-machine players need to have their photos taken, so that (participant) can see who is making the decision on each trial. And (participant) you can come with me."

At this point, the experimenter brought the participant into an adjacent room and gave them the following instructions:

"Recall that two of the other participants were told that their slot-machines were noninteractive: specifically, that they were only playing for monetary reward and that their actions did not result in any liquid being delivered to you. Well, that was not true! In reality, their choices will deliver liquid to you without them knowing about it. They will not be informed until after the experiment that their actions resulted in liquid being delivered to you, and whether that liquid was aversive or neutral. On each trial you will be shown a picture of the person who is making a decision on that trial, the choice screen viewed by that person, that you saw earlier, and a final screen showing their choice. Remember that participants whose options are simply labeled 'Option A' and 'Option B' do NOT know that their choices result in any liquid being delivered to you while participants whose options are labeled 'Aversive' and 'Water' do know about the liquids."

Once in the scanner, the subject was presented with a total of 120 trials, 30 with each confederate. The runs were broken into two sessions with 60 trials in each (for a depiction of the trial structure, see Fig. 1*A*). On each trial, the monetary rewards associated with the two options on each trial always differed by only 1 cent. In the intentional conditions, the greater reward was always listed under the option labeled aversive for one of the confederates and under the option labeled water for the other confederate. In the nonintentional conditions, the greater monetary reward was always associated with Option A, which always resulted in salty tea, for one confederate and always associated with Option B, which always resulted in water, for the other confederate. The color of the two options (i.e., red and green) and the position of the labels and associated rewards were randomized on each trial. The confederates always choose the option with the greatest monetary reward (in other words, the relationship between the confederates and the experienced liquid was deterministic). The order of trials was block-randomized, with each confederate appearing once in each block.

At the end of the experiment, participants provided ratings for the following questions about each confederate, while a picture of that confederate was displayed on the screen: "How likable is this person?" "How angry are you at this person?" "How immoral is this person?" "How much does this person remind you of salty tea?" and "How much did this person intend to give you salty tea?" All ratings scales assessing confederate traits ranged from 0 (not at all) to 100 (extremely). The order of the questions, and of the confederates, was randomized. Postscan ratings of the pleasantness of the two liquids were also collected, on a scale ranging from −10 (extremely unpleasant) to 10 (extremely pleasant).

After completing the postscan ratings, participants were asked whether they believed that the confederates' actions had in fact caused the delivery of the liquids during the experiment (see Results). They were then debriefed, being told by the experimenter that all liquid deliveries had been preprogrammed, and that the confederates had had absolutely no causal influence on, nor any moral responsibility for, any of the liquid deliveries. They were also asked whether they experienced residual anger toward any of the confederates (none did) and instructed that if at any future point they felt confused or concerned about what had happen during the course of the experiment, they should not hesitate to contact the experimenter to discuss these matters further.

Experiment 2. In Experiment 2, the design was essentially identical to Experiment 1, except for the inclusion of an additional control condition, which we call the "computer control." In addition to receiving the same instructions given for Experiment 1, the participants were instructed that, on some trials, the computer would randomly determine whether salty tea or water was to be delivered. It was emphasized that no confederate was involved on such trials and that this would be indicated by a fractal image taking the place of a confederates face. On those computer trials, the probability of receiving an aversive outcome was indeed 0.5; otherwise, a neutral outcome was delivered. 30 such computer trials (15 in each session) were added to the 120 confederate trials, for a total of 150 trials.

The purpose of this additional computer control condition was twofold: first, one possible explanation for any imaging effects in the intentional over the unintentional aversive condition is that participants experience more intense affective reactions to the intentionally aversive agent, and that this increase in the "affective" response is what is being reflected in the fMRI data. By including the nonintentional computer condition in which the aversive outcome is delivered only 50% of the time, we introduce a "weaker" nonintentional aversive condition (with very low intentionality, and with a lower probability of receiving an aversive outcome than in both aversive confederate conditions). The affective response to this condition should be even weaker than that to the nonintentional aversive condition. If a brain area found to respond more to the intentional compared with nonintentional aversive condition also responds more to the nonintentional aversive condition than to the computer control condition, then this would support the "affective intensity" argument.

The second purpose of including the computer control condition was to rule out an outcome uncertainty explanation for any effects of intentionality. In particular, it is possible that participants view the intentionally aversive agent as having two competing objectives: one being to maximize monetary reward, and the other being to not hurt the participant, whereas such competing objectives might not be presumed to be present in the other confederate conditions. This may lead to greater uncertainty about the expected outcome in the intentional aversive condition potentially accounting for any differences in neural responses. The computer control condition addresses this concern because the 50% aversive outcome in this condition yields maximal uncertain from the perspective of the participant (because in all other conditions the outcome is always reliably given whether neutral or aversive). Thus, if activity in a brain area responds more in the computer condition compared with the other conditions, then this likely reflects an effect of uncertainty. To further address this latter concern, we also solicited an additional subjective rating from the participants at the end of Experiment 2 asking how confident they were that a given confederate would chose to deliver salty tea for a minor monetary gain. If reported levels of confidence are lower in the intentional than nonintentional aversive condition, then this would support an uncertainty explanation of imaging effects.

Assessment of credibility of experimental procedure

When asked, before debriefing, whether they believed that the confederate's actions had caused the delivery of the liquids during the experiment, the vast majority of participants stated that they did indeed believe that this was the case in both experiments. A few remaining participants (3 in Experiment 1 and 2 in Experiment 2) reported being vaguely suspicious of the possibility of deception, although none of these indicated that they were certain that this was the case. Of course, simply raising the possibility of deception might bias a participant's perception of the veracity of the experimental instructions; in this sense, the more indirect measure of how much the participant thought that a confederate had "intended to deliver salty tea" might provide a closer estimate of actual suspicions about deception. Regardless, because exclusion of those participants who expressed doubts did not change our results in any substantive way, they were included in all reported analyses.

Imaging procedure

A 3 Tesla scanner (Phillips Achieva) was used to acquire structural T1-weighted images and T2*-weighted echoplanar images (repetition time = 2.65 s; echo time = 30 ms; flip angle = 90°; 45 transverse slices; matrix = 64 × 64; field of view = 192 mm; thickness = 3 mm; slice gap = 0 mm) with BOLD contrast. To recover signal loss from dropout in the medial orbitofrontal cortex (mOFC) (O'Doherty et al., 2002), each horizontal section was acquired at 30° to the anterior commissure–posterior commissure axis.

Behavioral analysis

In each experiment, planned comparisons of "intentional aversive" and "nonintentional aversive" conditions were performed on all ratings of confederates traits, and *t* tests were performed on the ratings of liquid pleasantness. In addition, for each experiment, each type of judgment about confederate traits was entered into a separate repeated-measures ANOVA, with intentionality and outcome valence as within-subject factors.

Imaging analysis

Image processing and statistical analyses were performed using SPM5 (http://www.fil.ion.ucl.ac.uk/spm). The first four volumes of images were discarded to avoid T1 equilibrium effects. All remaining volumes were corrected for differences in slice acquisition, realigned to the first volume, spatially normalized to the MNI echoplanar imaging template, and spatially smoothed with a Gaussian kernel (8 mm, full width at half-maximum). We used high-pass filter with cutoff = 128 s. Each trial was divided into three periods, based on the stimuli presented on the screen: an early anticipatory period, during which only the confederates face was presented (see Fig. 1*A*, first screen), a late anticipatory period, during which a depiction of the choice screen, ostensibly viewed by the confederate, was presented together with the face (see second screen in Fig. 1*A*), and a liquid delivery period which, again, only showed the confederates face (see fourth screen in Fig. 1*A*). We specified a separate linear model for each subject, with 24 regressors; one for each condition, in each trial period, and for each session. Six regressors accounting for the residual effects of head motion were also included. No orthogonalization was applied to any of these regressors. All regressors were convolved with a canonical hemodynamic response function and, for each subject, contrasts were calculated for simple effects, main effects, and the interaction of our intentionality and outcome variables, in each trial period and each session. Group-level random-effects statistics were generated by entering the contrasts of intentionality and outcome valence from each subject into a between-subjects analysis. One such group-level model was generated for each study, for each of the relevant condition contrasts (i.e., main effects and the interaction) and for each of the specified trial periods (anticipation and liquid delivery).

Small-volume corrections (SVCs) were performed on several *a priori* regions of interest using an 8 mm sphere. We used coordinates identified in previous studies of theory of mind [(Young and Saxe, 2009), left (−58, −58, 24) and right (56, −52, 22) temporal parietal junctions, medial prefrontal cortex (−2, 52, 22), and posterior cingulate (0, −54, 40); (Sanfey et al., 2003), left (−33,14, −1) and right (35, 15, 3) anterior insula; (Kampe et al., 2003), left (46, 4, −46) and right (−46, 2, −42) temporal pole] as well as studies of aversive conditioning with social stimuli [(Gottfried and Dolan, 2004), left (−18, −9, −18) and right (15, −3, −20) amygdala; (Jensen et al., 2003), left (−8, 10, −2) and right (12, 6, −4) ventral striatum]. Unless otherwise indicated, all other effects were reported at $p < 0.05$, using cluster size thresholding (cst) to adjust for multiple comparisons (Forman et al., 1995). AlphaSim, a Monte Carlo simulation was used to determine cluster size and significance. Using an individual voxel probability threshold of $p = 0.005$ indicated that using a minimum cluster size of 111 MNI-transformed voxels resulted in an overall significance of $p < 0.05$. For display purposes, statistical maps in all figures are shown at an uncorrected threshold of $p < 0.005$.

To eliminate nonindependence bias for plots of parameter estimates, a leave-one-subject-out (LOSO) (Esterman et al., 2010) approach was used, in which 19 GLMs were run with one subject left out in each, and with each GLM defining the voxel cluster for the left out subject. Spheres (10 mm) centered on the LOSO peaks (identified within ROIs for small volume corrections) were then used to extract mean β weights for each of the four conditions and these were averaged across subjects to plot overall effect sizes.

Results

Behavioral results

Postscan evaluative ratings of each confederate (Fig. 1*B*) confirmed that the manipulations were successful. In each experiment, planned comparisons revealed a significant difference between intentional aversive and nonintentional aversive conditions for ratings of how angry the participant was at a confederate, how likable a confederate was, and how much a confederate had intended to deliver salty tea to the participant, as well as for the rating collected only in Experiment 2, of how confident a participant felt that a given confederate would chose to deliver salty tea for a minor monetary gain, all *p* values <0.05. In Experiment 1, significant differences between intentional and nonintentional aversive conditions also emerged for ratings of how immoral a confederate was, and how much the confederate reminded the participant of salty tea, *p* values <0.05: although similar trends were seen for these ratings in Experiment 2, they did not reach significance.

In Experiment 1, significant intentionality by outcome valence interactions were found for ratings of how much each confederate had intended to deliver the aversive outcome, how angry the subject was at the confederate and how much a confederate

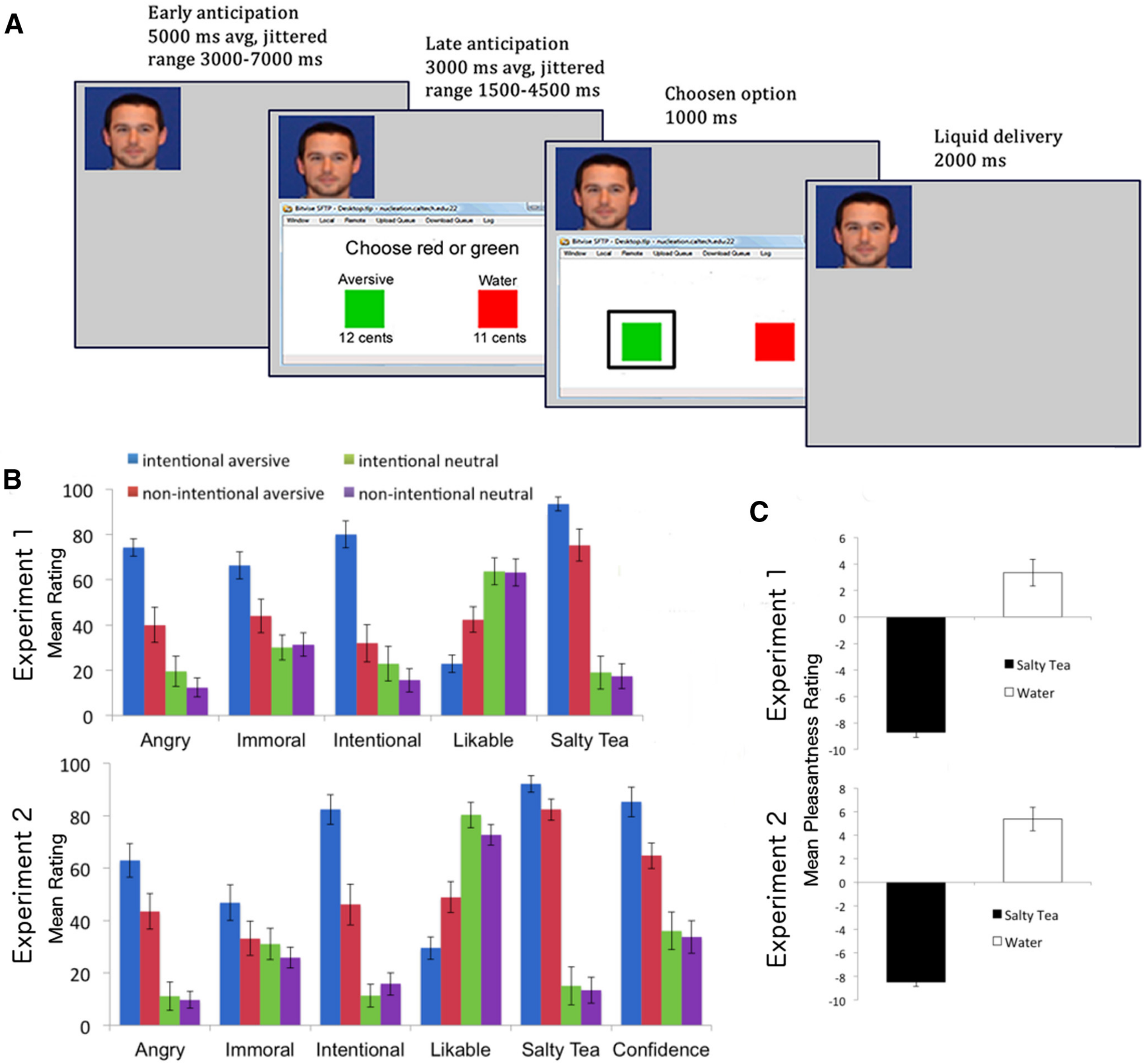

Figure 1. Trial illustration and behavioral results. ***A***, Trial illustration. On each trial, the subject passively observed a confederate choosing between two options based on a small monetary incentive, with the confederate's face and the choice screen viewed by that confederate displayed to the subject. Each choice by a confederate resulted in the delivery of either aversive salty tea or water to the subject in the scanner. In the intentional conditions, the two options on the confederates choice screen were labeled with the liquids as in the figure, whereas in the nonintentional conditions the confederate's options were simply labeled "Option A" and "Option B." Trials were separated by a jittered ITI (average 8000 ms). In the computer condition of Experiment 2, trials were identical to those with confederates, except for the replacement of a confederate's face with a fractal image and the probabilistic (i.e., 0.5) delivery of salty tea versus water. ***B***, Mean ratings of confederate traits (*y*-axis), from Experiment 1 (top) and Experiment 2 (bottom). Error bars indicate SEM. ***C***, Mean ratings (*y*-axis) of the pleasantness of salty tea and water outcomes, from Experiment 1 (top) and Experiment 2 (bottom). Error bars indicate SEM.

reminded the participant of salty tea, all p values <0.05. Although there were clear trends, interactions did not reach significance for ratings of confederate likability or immorality. In Experiment 2, significant interactions again emerged for ratings of anger toward a confederate, and for the confederate's intent to deliver salty tea, as well as for confederate likability and for the participants confidence that a confederate would chose to deliver salty tea for a minor monetary gain, all p values <0.05. No interaction was found in Experiment 2 for the rating of a confederates immorality, nor for how much a confederate reminded the participant of salty tea. Nonetheless, as can be seen in Figure 1*B*, there were clear trends toward interactions for each type of rating in each experiment, suggesting that the behavioral effects of our manipulations were highly consistent across experiments.

Participants also provided postscan ratings of the subjective pleasantness of the two liquids (Fig. 1*C*), on a scale ranging from -10 (extremely unpleasant) to 10 (extremely pleasant), with the salty tea being rated as significantly less pleasant than the water in each study, p values <0.001. Planned comparisons revealed that, in both experiments, ratings of both liquids also deviated significantly from the neutral 0-point (two-tailed p values <0.01), suggesting that each of the two outcomes had affective valence, in opposite directions.

Table 1. Significant imaging effects from both studies: one peak per cluster

	Trial period	Test	Area	*x*, *y*, *z*	Correction	Cluster size at $p < 0.005$
Study 1	Face only	Intent > no intent	L insula	−36, 18, 3	SVC	12
			L TPJ	−57, −48, 30	SVC	28
		Salty tea > water	Lingual gyrus	21, −51, −9	CST	125
		Intent × liquid	R cOFC	42, 45, −9	CST	138
	Choice screen	Intent > no intent	mPFC	0, 42, −3	CST	430
			L FO	−48, 27, 3	CST	498
			R FO	39, 24, −15	CST	476
			L TPJ	−57, −60, 24	SVC	50
			R TPJ	63, −45, 15	CST	338
			Thalamus	12, −9, 9	CST	148
		Salty tea > water	L VS	−3, 9, −6	SVC	16
		Water > salty tea	Cerebellum	27, 57, −9	CST	203
			R OFC	−3, −60, 36	CST	507
		Intent × liquid	L insula	−42, 33, 6	CST	107
	Outcome	Intent > no intent	PC	−6, −42, 9	CST	175
			mPFC	6, 66, 21	CST	349
			Cerebellum	−30, −81, −33	CST	828
		Salty tea > water	R insula	45, 15, −3	CST	137
			Cerebellum	−39, −60, −33	CST	265
		Water > salty tea	R IPC	42, −57, 54	CST	211
			R cOFC	21, −60, 3	CST	163
			R MFG	45, 39, 21	CST	139
Study 2	Choice screen	Intent > no intent	mPFC	−9, 53, 22	CST	544
			L FO	−27, 17, −20	CST	820
			R FO	33, 20, −17	CST	419
			L TPJ	−48, −58, 13	SVC	115
			R TPJ	60, −55, 16	SVC	11
		Salty tea > water	L VS	−3, 11, −2	SVC	22
			R VS	9, 11, −5	SVC	15
		Water > salty tea	Cerebellum	0, −52, −26	CST	1028
		Intent × liquid	L insula	−42, 8, −5	CST	332
			R IFG	48, 32, 1	CST	176
	Outcome	Intent > no intent	PC	12, −48, 7	CST	321
			mPFC	−6, 62, 25	SVC	41
			L amygdala/hippocampus	−24, −4, −17	CST	183
			R amygdala/hippocampus	21, −10, −14	CST	283
		Salty tea > water	Supplemental motor	6, −7, 64	CST	134
			Rolandic operculum	54, −7, 19	CST	894
			L postcentral sulcus	−54, −4, 22	CST	283
			Cerebellum	15, −64, −20	CST	2715
		Water > salty tea	R IPC	48, −49, 46	CST	589
			R cOFC	24, 65, 1	CST	342
			R MFG	51, 20, 31	CST	526

Neuroimaging results

We analyzed the fMRI data from the two experiments focusing on three distinct trial periods: (1) An early anticipatory period (first screen in Fig. 1*A*), in which only a confederates face was shown on the screen; (2) a subsequent late anticipatory period (second screen in Fig. 1*A*), during which a depiction of the choice screen, ostensibly viewed by the confederate, was presented together with the face; and (3) the liquid delivery period (fourth screen in Fig. 1*A*). We focus our reporting of imaging results on those findings that were statistically significant in both Experiment 1 and Experiment 2 (all of which emerged during the late anticipatory and liquid delivery periods), as these are likely the most robust. For completeness, we also report findings that became significant when the data were pooled across the two experiments. A full list of significant effects from each study and each trial period is provided in Table 1.

Main effects of intentionality

During the late anticipatory period, in each study, our test for a main effect of intentionality revealed significant activity throughout a so-called "theory of mind" network (Kampe et al., 2003; Saxe and Kanwisher, 2003; Rilling et al., 2004; den Ouden et al., 2005; Young and Saxe, 2008, 2009), including the left and right temporal-parietal junctions (TPJ), the medial prefrontal cortex (mPFC), and bilaterally in the frontal operculum (FO), extending into the inferior frontal gyrus (IFG), ventral anterior insula, and temporal poles (TP) (see Fig. 2*A*). During the liquid delivery period, significant effects of intentionality emerged again in mPFC, as well as in the posterior cingulate (PC), in each experiment (Fig. 2*B*).

Main effects of anticipated outcomes

In each study, activity in the ventral striatum (VS) during the late anticipatory period was greater for confederates that had routinely delivered the aversive liquid than for those that had delivered the neutral liquid (Fig. 3*A*). These results suggest that, although they had no knowledge beforehand about which confederate would deliver which liquid, participants learned to expect either salty tea or water, given only a confederate's face and choice screen, based on the experienced contingencies. The results

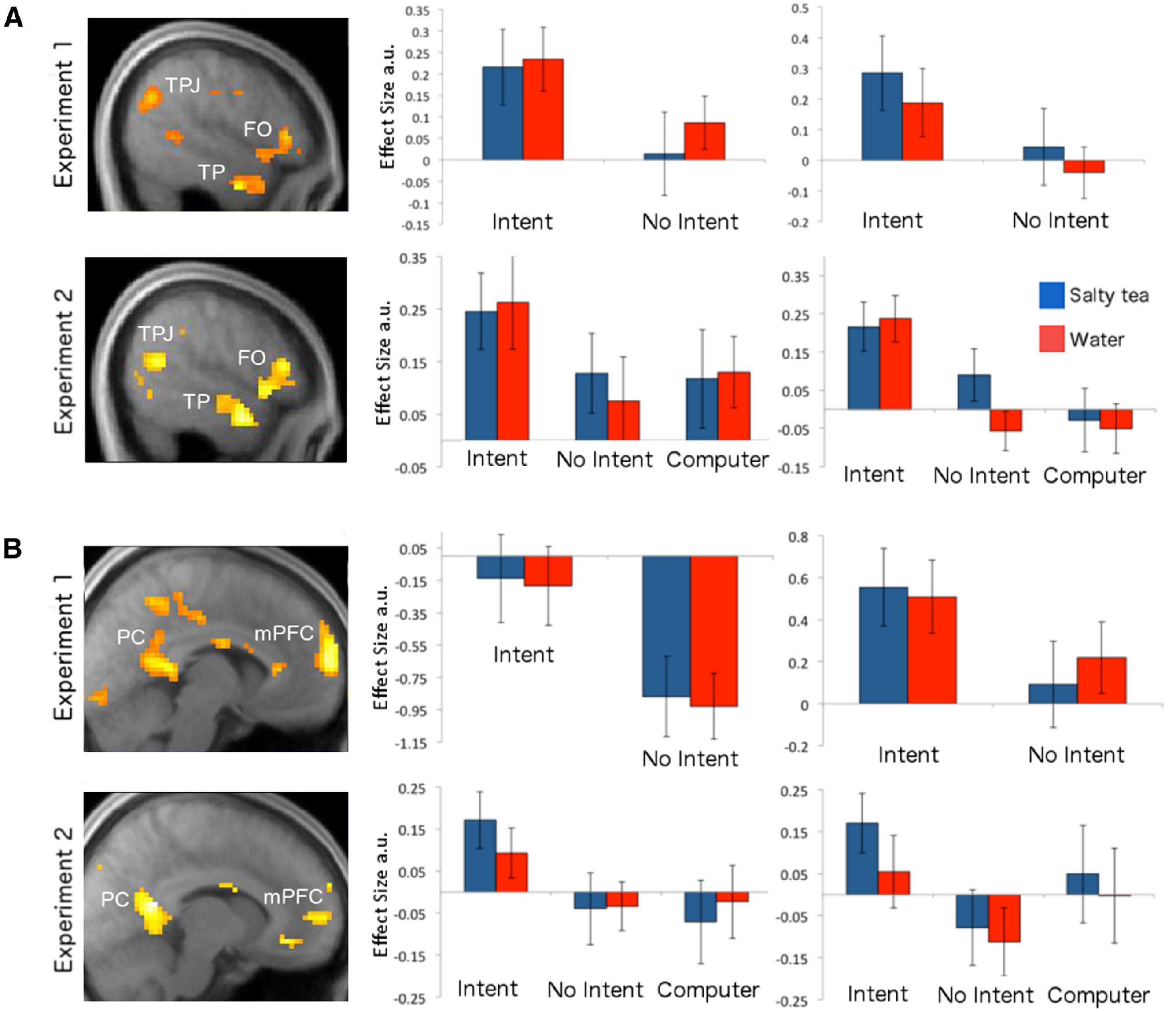

Figure 2. BOLD correlates of intentionality. Betas for each condition were extracted at LOSO coordinates (see Materials and Methods). Error bars indicate SEM. ***A***, Neural activation during the late anticipatory period in the first (top) and replication study (bottom). Effects are shown in the temporal parietal junction (betas plotted in middle panel), the temporal pole (betas plotted in right panel), and the frontal operculum. ***B***, Neural activation during the liquid delivery period in the first (top) and replication study (bottom). Effects are shown in the mPFC (betas plotted in middle panel) and PC (betas plotted in right panel).

are also consistent with a growing literature showing increased ventral striatal activity during anticipation and receipt of aversive stimuli (Jensen et al., 2003; Wrase et al., 2007; Delgado et al., 2011).

During the liquid delivery period, greater activity in response to the aversive than the neutral liquid was found in the right dorsal anterior insula (Fig. 3*B*), whereas the reverse contrast, assessing greater responses to the neutral than the aversive liquid, yielded effects in the right central orbitofrontal cortex (cOFC), inferior parietal cortex (IPC), and middle frontal gyrus (MFG).

Integration of intentionality and anticipated aversive outcomes

We were particularly interested in assessing whether neural activity correlated with an interaction between intentionality and anticipated outcome valence similar to that observed for our behavioral measures, such that responses to a confederate that had delivered the aversive, relative to the neutral, outcome would be greater in the intentional than in the nonintentional condition. During the late anticipatory trial period, in each study, this test revealed effects in the left dorsal anterior insula (Fig. 4). In the replication study, planned comparisons performed on betas extracted at insular LOSO coordinates (see Materials and Methods) revealed that, in addition to a significant difference between intentional and nonintentional aversive conditions (two-tailed $p < 0.001$), each of these conditions also differed significantly, and in opposite directions, from the computer control condition (both two-tailed p values <0.05).

A test was also performed to assess whether the interaction effects in the anterior insula were correlated with the degree to which an interaction was reflected in subjective ratings of how much each confederate had intended to deliver salt tea. Using SVC on our ROI in the left dorsal anterior insula, this test did reveal a significant correlation in the first study; however, the effect did not reach significance in the second study, nor did it reach significance when data were pooled across studies, although it was apparent at an uncorrected threshold of $p < 0.005$.

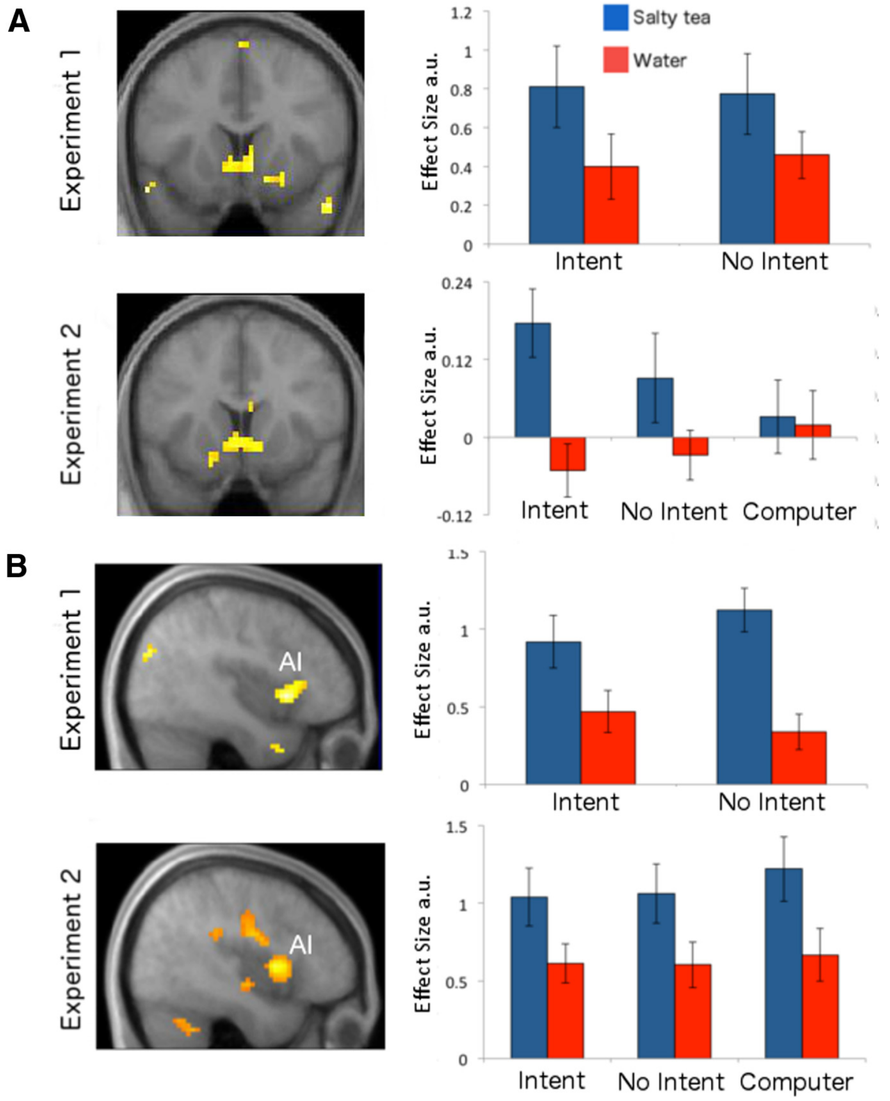

Figure 3. BOLD effects of aversive outcome anticipation. Betas for each condition were extracted at LOSO coordinates. Error bars indicate SEM. ***A***, Effects during the late anticipatory period in Experiment 1 (top) and Experiment 2 (bottom). Effects are shown in the ventral striatum. ***B***, Neural activation during the liquid delivery period in Experiment 1 (top) and Experiment 2 (bottom). Effects are shown in the right dorsal anterior insula (AI).

As this effect failed to reach significance, we refrain from discussing it further.

Additional effects when pooling the data across both experiments

No effects survived our strict two-experiment replication criterion during the earliest anticipatory period of each trial, in which only a confederates face was shown on the screen. However, when the data were pooled across experiments, significant effects did emerge during this trial period: A main effect of intentionality, such that activity was greater in intentional than in nonintentional conditions, was observed in the left dorsal anterior insula (SVC, −36, 23, 1) and left TPJ (SVC, −54, −52, 28). Furthermore, a main effect of anticipated outcome valence, such that activity was greater during early anticipation of salty tea than of water, was found in the lingual gyrus (CST, 18, −55, −2).

Discussion

In this study, we sought to explore how perceived intentionality influences the acquisition of an aversion toward an individual whose actions result in actual physical discomfort for the subject, and to identify the brain regions involved in mediating integration of socio-cognitive and affective processes. We found that the difference between rated likeability of, and anger toward, an agent whose actions had resulted in a noxious experience for the subject and one whose actions had resulted in a neutral outcome was greater when the agents were believed to know about the consequences of their actions for the subject. At a neural level, activity in the anterior insula, an area previously implicated in both the anticipation and receipt of aversive outcomes and mentalizing, correlated with an intentionality by anticipated outcome interaction.

Previous research indicates that inferences about intent profoundly influence judgments of culpability (Borg et al., 2006; Cushman, 2008; Lagnado and Channon, 2008; Young and Saxe, 2008, 2009). Generally, in such studies, participants are presented with brief fictitious scenarios in which the protagonist either intentionally or unintentionally harms another individual, with judgments of the protagonists blameworthiness being consistently higher for intentional than unintentional actions. At the neural level, a network of structures has been shown to support inferences about others' mental state, including the TPJ and mPFC (den Ouden et al., 2005; Young and Saxe, 2008, 2009). Inferences about intent have also been shown to influence economic decision-making in strategic games (Sanfey et al., 2003; Rilling et al., 2004). For example, scanning participants as they played an ultimatum game with opponents who they believed to be either human or a computer, Sanfey et al. (2003) found that unfair offers were rejected more often on human trials than on computer trials, suggesting that inequitable humans elicited strong negative emotions, and this behavioral effect correlated with activity in the anterior insula.

Of course, judgments about the culpability of fictitious individuals have an intrinsic socio-cognitive component: perceiving the outcome as negative to begin with requires some form of mentalizing about the experience of the harmed individual. Likewise, monetary currency, and the equitability of its distribution, obtains significance entirely from an interpersonal agreement about the valence of arbitrary symbols. It is not surprising, therefore, that tasks such as those described in the previous paragraph appear to strongly recruit processes of mental state attribution. In contrast, no previous study has addressed the question of whether such abstract attributions also modulate affective responses to directly experienced biologically salient aversive stimuli. Here, we demonstrate an influence of perceived intentionality on neural responses to individuals whose actions resulted in the immediate delivery of a noxious stimulus to the subject in the scanner, suggesting that abstract inferences about

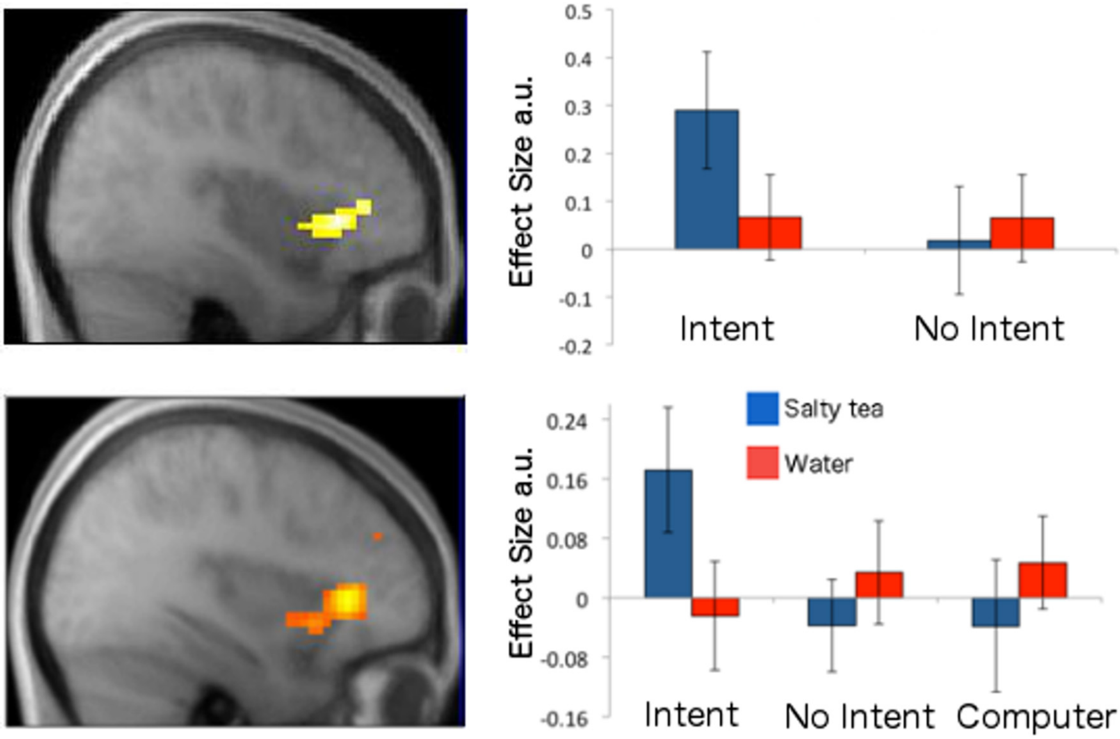

Figure 4. BOLD effects of the interaction between intentionality and aversive outcome anticipation. Betas for each condition were extracted at LOSO coordinates. Error bars indicate SEM. Effects during the anticipatory period in Experiment 1 (top) and Experiment 2 (bottom) are shown in the anterior insula.

others' mental states can regulate the visceral anticipation of aversive stimulation.

One important question is whether the interaction effect in the anterior insula merely reflects a general enhancement of aversive affect, or greater uncertainty about the trial outcome, for the confederate that knowingly delivered the aversive liquid, such that any (nonsocial) manipulation that resulted in this type of enhancement would yield the same pattern of activity in this area. For example, greater uncertainty might be due to the fact that the intentionally aversive confederate knowingly choose between delivering an aversive outcome and a personal monetary loss. Behavioral ratings, however, make both of these alternative explanations unlikely. As can be seen in Figure 1*B*, mean ratings of how much each confederate reminded the participant of salty tea, as well as those of how confident the participant was that, if given the opportunity, a confederate would choose to deliver salty tea for a minor monetary gain, are dramatically greater for the nonintentional aversive condition than for both of the two neutral conditions: in contrast, as shown in Figure 4, anterior insular responses did not differ at all across neutral and nonintentionally aversive conditions.

To directly assess anterior insular activity in response to uncertainty, and to anticipation of salty tea, a replication experiment included a control condition, in which participants were told that a computer algorithm selected between delivery of salty tea and water with a probability of 0.5. Thus, this condition entailed maximum uncertainty about the liquid outcome, as well as a much weaker salty tea contingency than in either of the aversive confederate conditions. If insular cortex responses reflect uncertainty about the decisions of the intentionally aversive confederate, they should be as great or greater in the computer control condition. Conversely, if they reflect attenuated aversive encoding in the nonintentional condition, they should be even lower in the computer control condition. Contrary to these predictions, we found that anterior insular activity in the control condition was both significantly lower than that in the intentional aversive condition, and greater than that in the nonintentional aversive condition, suggesting that neither uncertainty nor simple differences in the overall magnitude of aversive affect can account for the effects in this area. Instead, we interpret these effects as being indicative of a unique role of this region in discriminating between intentional and nonintentionally harmful individuals.

It is notable that the anterior insula did not exhibit robust responses during anticipation of the aversive outcome in the nonintentional condition, given previous studies reporting effects in this area during anticipation of aversive outcomes (Gottfried and Dolan, 2004; Delgado et al., 2011) and in light of recent findings, suggesting that this region acts a general "saliency hub," allocating attentional recourses to salient external stimuli (Menon and Uddin, 2010; Deen et al., 2011; Touroutoglou et al., 2012). One possible explanation for this result is that the presence of the affectively significant intentionally aversive condition produced a relative contrast effect, resulting in a much weaker affective response during anticipation of the nonintentionally delivered aversive outcome. Indeed, anterior insular responses to both rewarding and punishing events have been shown to be modulated by the overall context in which those events occur (Elliott et al., 2000), and it is well established that the behavioral reaction to a particular stimulus depends on the affective properties of other stimuli that have recently been experienced (e.g., Mellers et al., 1997).

Another important consideration is that, although the deviation of pleasantness ratings from neutrality was greater for salty tea than for water, both liquids deviate significantly from 0, suggesting that the water was rewarding rather than neutral, perhaps because it allowed participants to rinse out any residual taste of salty tea. It is interesting to note that, despite this affectively positive response to the water outcome, the response in the dorsal anterior insula during the late anticipatory period was specific to the intentional aversive confederate, with no differences observed among the other three conditions. Importantly, while this selective response in the anterior insula suggests that "intent to do harm" may be a uniquely salient event, it does not imply that only aversive stimuli are subject to the modulatory influence of mentalizing processes.

The anterior insula has been implicated in a wide range of cognitive, social, and affective processes, including pain perception (Baliki et al., 2009), processing of facial expressions (Morris et al., 2008), empathy (Jackson et al., 2005), and the attribution of agency (Farrer and Frith, 2002). More pertinently, effects in this area have been reported by studies assessing the role of mental state attributions in strategic game interactions (Sanfey et al., 2003; Singer et al., 2004), as well as studies of Pavlovian fear conditioning to social (Gottfried and Dolan, 2004) and nonsocial (Delgado et al., 2008, 2011) stimuli. Together, these studies suggest that the anterior insula may bridge neural systems involved in social attribution and those that support the encoding of predictions about aversive outcomes. Our results go further, providing direct evidence for the involvement of this region in the influence of mentalizing computations on affective encoding.

Our manipulation of intentionality elicited activity in a network of structures, including the TPJ, temporal poles and MFC, frequently identified in studies on the neural basis of theory of mind (Kampe et al., 2003; Saxe and Kanwisher, 2003; Rilling et al., 2004; den Ouden et al., 2005; Young and Saxe, 2008, 2009). Some such studies contrasted neural responses to short fictitious stories that require inferences either about mental states or about mechanical and physical events (den Ouden et al., 2005; Young and Saxe, 2008, 2009). Others have looked at interpersonal interactions, contrasting responses to human and computer opponents in strategic games (Rilling et al., 2004), and yet others have investigated brain activity while subjects view animated characters perform object-oriented actions (Pelphrey et al., 2004; Carter et al., 2011). Although operational definitions of mental state representations differ greatly across these tasks, they share the element of perceived intent: that is, inferences about an agent's knowledge of, and desire to bring about, the outcome of their action. One possibility, therefore, is that the implicated areas serve specifically to detect, and assess the consequences of, goal-directed behavior.

In conclusion, our results show, at both behavioral and neural levels, that responses to visceral aversive stimulation are subject to the ameliorating effects of socio-cognitive representations of "mentalizing." What remains is to explain why abstract cognitive evaluations of intent would modulate evaluations of aversive stimuli. Obviously, to obtain favorable outcomes in social and personal interactions, we have to learn to predict the actions of other individuals. How does intentionality relate to this need? One possibility is that intent is a critical component of the causal structure underlying observable behavior (Lagnado and Channon, 2008) and, as such, integral to the generality of action predictions. Specifically, an individual who intentionally and voluntarily acts in a harmful way may be more likely to do so in future encounters, and in other contexts, than an individual that accidentally inflicts harm. Alternatively, intentionally harmful individuals may elicit distinctly "social" emotions (Eisenberger et al., 2003) that are dissociable from those due to primary appetitive and aversive events, and that serve specifically to facilitate social interactions and communication. Further research is needed to pinpoint the relationship between inferred intent and action prediction, and to explore the role of mentalizing in affective learning. For now, our results suggest that abstract inferences about mental states do indeed modulate fundamental processes of aversive learning, and that this cognitive and affective integration is mediated by the anterior insula.

References

Baliki MN, Geha PY, Apkarian AV (2009) Parsing pain perception between nociceptive representation and magnitude estimation. J Neurophysiol 101:875–887. CrossRef Medline

Borg JS, Hynes C, Van Horn J, Grafton S, Sinnott-Armstrong W (2006) Consequences, action, and intention as factors in moral judgments: an fMRI investigation. J Cogn Neurosci 18:803–817. CrossRef Medline

Carter EJ, Hodgins JK, Rakison DH (2011) Exploring the neural correlates of goal-directed action and intention understanding. Neuroimage 54: 1634–1642. CrossRef Medline

Cushman F (2008) Crime and punishment: differential reliance on causal and intentional information for different classes of moral judgment. Cognition 108:353–380. CrossRef Medline

Deen B, Pitskel NB, Pelphrey KA (2011) Three systems of insular functional connectivity identified with cluster analysis. Cereb Cortex 21:1498–1506. CrossRef Medline

Delgado MR, Nearing KI, Ledoux JE, Phelps EA (2008) Neural circuitry underlying the regulation of conditioned fear and its relation to extinction. Neuron 59:829–838. CrossRef Medline

Delgado MR, Jou RL, Phelps EA (2011) Neural systems underlying aversive conditioning in humans with primary and secondary reinforcers. Front Neurosci 5:71. CrossRef Medline

den Ouden HE, Frith U, Frith C, Blakemore SJ (2005) Thinking about intentions. Neuroimage 28:787–796. CrossRef Medline

Eisenberger NI, Lieberman MD, Williams KD (2003) Does rejection hurt? An FMRI study of social exclusion. Science 302:290–292. CrossRef Medline

Elliott R, Friston KJ, Dolan RJ (2000) Dissociable neural responses in human reward systems. J Neurosci 20:6159–6165. Medline

Esterman M, Tamber-Rosenau BJ, Chiu YC, Yantis S (2010) Avoiding non-independence in fMRI data analysis: leave one subject out. Neuroimage 50:572–576. CrossRef Medline

Farrer C, Frith CD (2002) Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. Neuroimage 15:596–603. CrossRef Medline

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (1995) Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med 33:636–647. CrossRef Medline

Gottfried JA, Dolan RJ (2004) Human orbitofrontal cortex mediates extinction learning while accessing conditioned representations of value. Nat Neurosci 7:1144–1152. CrossRef Medline

Jackson PL, Meltzoff AN, Decety J (2005) How do we perceive the pain of others? A window into the neural processes involved in empathy. Neuroimage 24:771–779. CrossRef Medline

Jensen J, McIntosh AR, Crawley AP, Mikulis DJ, Remington G, Kapur S (2003) Direct activation of the ventral striatum in anticipation of aversive stimuli. Neuron 40:1251–1257. CrossRef Medline

Kampe KK, Frith CD, Frith U (2003) "Hey John": signals conveying communicative intention toward the self activate brain regions associated with "mentalizing," regardless of modality. J Neurosci 23:5258–5263. Medline

Kim H, Shimojo S, O'Doherty JP (2011) Overlapping responses for the expectation of juice and money rewards in human ventromedial prefrontal cortex. Cereb Cortex 21:769–776. CrossRef Medline

Lagnado DA, Channon S (2008) Judgments of cause and blame: the effects of intentionality and foreseeability. Cognition 108:754–770. CrossRef Medline

Mellers BA, Schwartz A, Ho K, Ritov I (1997) Decision affect theory: emotional reactions to the outcomes of risky options. Psychol Sci 8:423–429. CrossRef

Menon V, Uddin LQ (2010) Saliency, switching, attention and control: a network model of insula function. Brain Struct Funct 214:655–667. CrossRef Medline

Morris JP, Pelphrey KA, McCarthy G (2008) Perceived causality influences brain activity evoked by biological motion. Soc Neurosci 3:16 –25. CrossRef Medline

O'Doherty JP, Deichmann R, Critchley HD, Dolan RJ (2002) Neural responses during anticipation of a primary taste reward. Neuron 33:815–826. CrossRef Medline

O'Doherty J, Rolls ET, Francis S, Bowtell R, McGlone F (2001) Representation of pleasant and aversive taste in the human brain. J Neurophysiol 85:1315–1321. Medline

Pelphrey KA, Morris JP, McCarthy G (2004) Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. J Cogn Neurosci 16: 1706 –1716. CrossRef Medline

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, Cohen JD (2004) The neural correlates of theory of mind within interpersonal interactions. Neuroimage 22:1694 –1703. CrossRef Medline

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. Science 300:1755–1758. CrossRef Medline

Saxe R, Kanwisher N (2003) People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind." Neuroimage 19: 1835–1842. CrossRef

Singer T, Kiebel SJ, Winston JS, Dolan RJ, Frith CD (2004) Brain responses to the acquired moral status of faces. Neuron 41:653– 662. CrossRef Medline

Touroutoglou A, Hollenbeck M, Dickerson BC, Feldman Barrett LF (2012) Dissociable large-scale networks anchored in the right anterior insula subserve affective experience and attention. Neuroimage 60:1947–1958. CrossRef Medline

Wrase J, Kahnt T, Schlagenhauf F, Beck A, Cohen MX, Knutson B, Heinz A (2007) Different neural systems adjust motor behavior in response to reward and punishment. Neuroimage 36:1253–1262. CrossRef Medline

Young L, Saxe R (2008) The neural basis of belief encoding and integration in moral judgment. Neuroimage 40:1912–1920. CrossRef Medline

Young L, Saxe R (2009) Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. Neuropsychologia 47:2065–2072. CrossRef Medline