Cerebral Cortex, October 2020;30: 5410–5419

doi: 10.1093/cercor/bhaa122
Advance Access Publication Date: 4 June 2020
Original Article

ORIGINAL ARTICLE

Neural Pattern Similarity Unveils the Integration of Social Information and Aversive Learning

Irem Undeger[1], Renée M. Visser[2] and Andreas Olsson[1]

[1]Section for Psychology, Department of Clinical Neuroscience, Karolinska Institute, Stockholm 171 77, Sweden and [2]Department of Clinical Psychology, University of Amsterdam, Amsterdam, 1018 WT, The Netherlands

Address correspondence to Irem Undeger, Section for Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Nobels väg 9, 171 77 Stockholm, Sweden. Email: irem.undeger@ki.se.

Abstract

Attributing intentions to others' actions is important for learning to avoid their potentially harmful consequences. Here, we used functional magnetic resonance imaging multivariate pattern analysis to investigate how the brain integrates information about others' intentions with the aversive outcome of their actions. In an interactive aversive learning task, participants ($n = 33$) were scanned while watching two alleged coparticipants (confederates)—one making choices intentionally and the other unintentionally—leading to aversive (a mild shock) or safe (no shock) outcomes to the participant. We assessed the trial-by-trial changes in participants' neural activation patterns related to observing the coparticipants and experiencing the outcome of their choices. Participants reported a higher number of shocks, more discomfort, and more anger to shocks given by the intentional player. Intentionality enhanced responses to aversive actions in the insula, anterior cingulate cortex, inferior frontal gyrus, dorsal medial prefrontal cortex, and the anterior superior temporal sulcus. Our findings indicate that neural pattern similarities index the integration of social and threat information across the cortex.

Key words: aversive learning, conditioning, intention, insula, MVPA, RSA

Introduction

To successfully manage our social interactions, we need to dynamically update our impressions of the people we interact with. This entails inferences about their thoughts, and emotions, and learning about the consequences of their actions. Research shows that the intentions behind an action can alter our impressions of both the action itself and the person performing the action (Ames and Fiske 2015; Zhang et al. 2016; Levine et al. 2018). For instance, a colleague who knowingly spills hot coffee on you might be remembered as a potential source of threat, but might be forgiven quickly if the spill was accidental. In support of this conclusion, clinical research has shown that harmful actions by intentional agents, such as torture or rape, are more likely to cause posttraumatic stress symptoms than experiences from non-interpersonal events, such as accidents or natural disaster (Fowler et al. 2013), highlighting how intentionality can affect the quality of threat learning. Although past imaging research has studied the neural representations of mental attributions (Koster-Hale and Saxe 2013; Wagner et al. 2016) and threat learning (Delgado et al. 2011; Spoormaker et al. 2011; Wheelock et al. 2014) in separation, little is known about how the brain integrates inferences about intentions with aversive outcomes. Here, we addressed this open question by examining how the brain integrates the perceived intentionality of an agent with the experience of the aversive outcomes of the agent's behavior. We asked how this integration affected 1) the formation of negative judgments about the individual and changed impressions about their actions, and 2) learning from these experiences.

Past research has identified a neural network that is involved in inferring the mind states of others, referred to as the "mentalizing network". This network includes the dorsal medial prefrontal cortex (dmPFC), the inferior frontal gyrus (IFG), the bilateral temporo-parietal junction (TPJ), and the insula (see Frith CD and Frith U 2006 and Koster-Hale and Saxe 2013 for a review). Activity in the mentalizing network is involved in forming impressions about others (Mitchell 2004; Mende-Siedlecki et al. 2013), and represents mental states and intentions of others (Young et al. 2010; Cushman et al. 2012; Tamir et al. 2016). A recent study has highlighted the role of anterior insula (AI) in the perception of intentional, compared to unintentional, choices that led to an aversive taste delivery (Liljeholm et al. 2014). Interestingly, certain regions of the mentalizing network overlap with regions implicated in threat learning, which includes regions responsive to threat, as well as those involved in safety processing and inhibition of threat responses (Fullana et al. 2015; Tovote et al. 2015). For example, the AI and the anterior cingulate cortex (ACC), both of which are often reported in threat learning tasks, seem to be involved in representing one's own and others' aversive experiences. For example, in a social fear learning task, in which participants learn the aversive value of a stimulus without direct experience, but through watching a demonstrator's aversive experiences, the AI and the ACC were recruited both during direct and social fear learning (Lindström et al. 2018). Importantly, the AI has been shown to represent the intensity of the aversive stimulus that the demonstrator experiences.

Pavlovian aversive conditioning is a standard paradigm to study how information about a stimulus is learned and updated by negative experiences (Pavlov 1927). In this paradigm, a neutral conditioned stimulus (CS+) is paired with a potentially harmful stimulus, such as an electrical shock, whereas a control stimulus (CS−) is unreinforced. Importantly, the perception of the aversive stimuli can also be altered by the specifics of the stimuli used. For example, fear-relevant stimuli are rated as more often paired with shocks than fear-irrelevant ones; causing so-called "illusory correlations" (Tomarken et al. 1989; Öhman and Mineka 2001). Relatedly, shocks following fear-relevant stimuli are rated as more painful than those following fear-irrelevant stimuli (Tomarken et al. 1989). Similarly, past research has shown that intentionally caused harms are perceived to be more painful than unintentional ones (Gray and Wegner 2008). Taken together, these findings raise the questions of if and how threat learning from others' actions can be altered by the intentionality of these actions, and how intentionality is neurally integrated with the value of threat over time.

Although traditional functional magnetic resonance imaging (fMRI) approaches have focused on averaging differences of responses between conditions, aversive learning is characterized by changes in responses to the CS as the participant accumulates information about the CSs. In a naturalistic social setting, the integration of harm and intentionality happen over time, as the individual learns about other individuals and their actions during social interaction. To understand when and how information about intentionality and harm is integrated, methods are needed that allow for the study of neural responses over time. A promising approach is multivoxel pattern analysis (MVPA), where instead of an average signal change, distributed (multivoxel) patterns of blood-oxygen-level–dependent (BOLD) signal are assessed to characterize the distinctive neural representation of a stimulus or condition (Haxby et al. 2001). Specifically, representational similarity analysis (RSA) (Kriegeskorte 2008) has been applied in a trial-by-trial manner to quantify the formation of fear associations during aversive conditioning (Visser et al. 2011, 2013, 2015, 2016a). Trial-by-trial RSA has shown learning about potential harm coincided with a "tuning" of neural activity patterns (Li et al. 2008), expressed as an increase in the similarity of response patterns related to threatening (CS+) stimuli compared to safe (CS−) stimuli (Visser et al. 2011, 2013, 2015, 2016b). Using a trial-by-trial RSA to assess how neural patterns evolve over time allows us to disentangle how we learn social and physical sources of threat, and answer outstanding questions on how the brain integrates knowledge about the intentionality of a social partner, and the harmful outcomes of their actions.

In this study, we used a modified aversive learning paradigm to investigate the development of neural activation patterns related to a participant's interaction with two alleged "coparticipants" (confederates). These coparticipants chose which stimulus should be presented to the participant: either a harmful stimulus (CS+) that was paired with shock, or a safe stimulus (CS−) never paired with a shock. Crucially, participants were told that one of the coparticipants was aware of the outcome of their choices, while the other was not, yielding a within-subject 2(harm/safety) × 2(intentional/unintentional) factorial design. To capture how the brain integrates knowledge about harm with social information, we assessed trial-by-trial changes in the similarity of neural activation patterns related to stimuli in each of the four conditions. We predicted a neural tuning effect in the mentalizing network to the intentional choice outcomes, compared to the unintentional ones, regardless of the valence of the chosen stimulus (hypothesis 1). Furthermore, based on prior research (Visser et al. 2011, 2013, 2015, 2016b; Dunsmoor et al. 2014), we hypothesized to see a selective increase in pattern similarity for harmful stimuli compared to safe stimuli, in areas of the aversive learning network, reflecting aversive learning of the CS+ stimulus (hypothesis 2). Finally, we predicted that regions previously reported in processing higher order social constructs and judgment making would integrate the outcome valence of an action and the mental state of the actor, as indexed by a selective increase in trial-by-trial similarity in patterns related to harmful outcomes that were intentionally delivered (hypothesis 3).

Materials and Methods

Participants

Forty healthy individuals were recruited via flyers and online recruitment systems for the initial aversive learning session. Data were discarded from the analyses if fMRI data had substantial head motion (>2 mm in any direction, $n = 6$), or the participant failed to understand the instructions, based on debriefing after the experiment ($n = 1$). The final sample of the fMRI analysis included 33 participants (18 males, all right-handed) between 18 and 34 years of age (mean: 25.07). Pupillometry data were used as an independent index of aversive learning, and data from participants who had more than 33% of trials of any condition missing (missing trial defined as > 50% missing sample for that trial) were discarded ($n = 5$) (Visser et al. 2013). Thus, when experimental phases were compared directly, pupillometry data are reported for 35 participants and fMRI data are reported for 33 participants. The participants that were excluded for each analysis do not overlap. All participants gave their written informed consent before participation and were naïve to the purpose of the experiment. The procedures were executed in compliance with relevant laws and institutional guidelines and were approved by the local ethical review board at Karolinska Institutet (dnr: 2017/138–31/2).

Apparatus and Materials

Stimuli

The face stimuli were photographs of the coparticipants (confederates). The CS were images drawn from four different categories: animals, fruits, tools, and buildings (Fig. 1B). The images were obtained from the website www.lifeonwhite.com and from publicly available resources on the internet, with their background removed (Dunsmoor et al. 2013). These categories are chosen as they are represented in different regions of the brain (Haxby et al. 2000). All stimuli were equalized to match luminance to assure differences in pupil dilation are not driven by luminosity, using the SHINE Toolbox (Willenbockel et al. 2010). The electrical stimulus was a 200 ms monopolar DC-pulse electric stimulation applied to the participant's left ankle.

Before the experiment, the intensity of the electric stimulus was individually adapted to be aversive but not painful.

Self-Report Questionnaires

Behavioral Measures

Each participant was asked to rate the likability of the two coparticipants, on a scale between 1 (not likable at all) and 5 (very likable), both before and after the experiment. After the experiment, each participant completed a contingency scale about the aversive learning stage, in which they report how many shocks they recall having received (if any) with each image, how many times each image was chosen by a coparticipant, and how much they expected to receive a shock when they saw each image (see Supplementary Material 1). Following this, each participant completed a questionnaire about their experiences through the aversive learning stage. This questionnaire included questions about the coparticipants, such as how angry the participant felt toward each of the coparticipants, how many electrical shocks they would like to deliver to them if given the chance, and what the participant thought the motivation of the intentionally harming coparticipant was (for the full questionnaires, see Supplementary Materials 1 and 2). Finally, participants were asked if they ever had any doubts about the experiment being a setup. This question was added to make sure all participants believed that the coparticipants were actually making choices and were recruited to be in the study just like the participant is, and participants were asked to give a "doubt rating" from 0 to 100%. Participants were invited to the behavioral lab 24 h after the experiment to do a memory test, followed by the personality measures: the autism-spectrum quotient (AQ) (Baron-Cohen et al. 2001) and the Liebowitz Social Anxiety Scale (LSAS) (Liebowitz 1987). The results of the additional tests on Day 2 are beyond the scope of this paper.

Pupil Dilation

During aversive learning pupil dilation responses and eye movements were recorded continuously, using an MR-compatible remote nonferromagnetic infrared Eyelink-1000 Long Range Mount eye tracker (SR Research Ltd, Mississauga, Ontario, Canada). Data were sampled at 250 Hz. The baseline pupil diameter was taken as the average response 500 ms preceding each trial. The response to each CS was calculated as the peak response during CS presentation (a window of 2.5 s) minus the baseline for that trial. Data that were obscured by blinks were discarded, and trials that suffer substantial signal loss (more than 50%) were eliminated and replaced using the linear trend at point (to a maximum of 33% of trials per condition).

Imaging Procedure

Acquisition and Preprocessing

Scanning was performed using a 3.0 T General Electrics MRI scanner using an 8-channel head-coil. Functional images are acquired using gradient echo-planar imaging (EPI) (repetition time = 2000 ms, time echo = 28 ms, flip angle = 80°, 42 (estimated) sagittal slices with interleaved acquisition, 3.0 × 3.0 × 3.0 mm) covering the whole brain. Higher order shimming was performed, and each session started with five dummy scans. Foam pads were used to minimize head motion. A high-dimensional T-1 weighted image (repetition time = 6.4 s, time to echo = 2.8 s, flip angle = 11°) was collected for anatomical visualization.

fMRI data for the RSA were analyzed using FEAT (FMRI Expert Analysis Tool) version 5.0, part of FSL (Oxford Centre for Functional MRI of the Brain (FMRIB) Software Library, http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). Preprocessing steps included: slice-time correction, motion correction, high-pass filtering in the temporal domain (SIGMA = 100 s), and prewhitening. Structural images were coregistered to the functional images and transformed to MNI (Montreal Neurological Institute) standard space using FNIRT (FMRIB's Nonlinear Image Registration Tool, FSL). The resulting normalization parameters were applied to the functional images.

Region of Interest Selection

The regions of interest (ROIs) in the aversive learning network were chosen based on the previous literature (Visser et al. 2011, 2013) and included the following: the ACC, the amygdala, the insula, hippocampus, and the ventromedial prefrontal cortex (vmPFC). For the mentalizing network, we chose left and right TPJ, dmPFC, and the anterior and the posterior superior temporal sulcus (arSTS, prSTS). Based on the previous literature on intentional harm IFGpt is additionally investigated (Yu et al. 2015). We used inferior temporal gyrus (ITGto) as a control region as we expect no effects of intentionality or CS's in this visual processing regions. We created anatomical ROIs using the Harvard–Juelich atlas and the regions unavailable in this tool were created via coordinates from the website Neurosynth, upon searching for the term "intention" (Yarkoni et al. 2011). For this, we used a 5 mm spherical mask at the coordinate and intersected them with masks from the anatomical atlas.

Experimental Paradigm

Instructions

Upon arrival to the experiment, the participant met the confederates (i.e., coparticipants) and was told that a lottery would

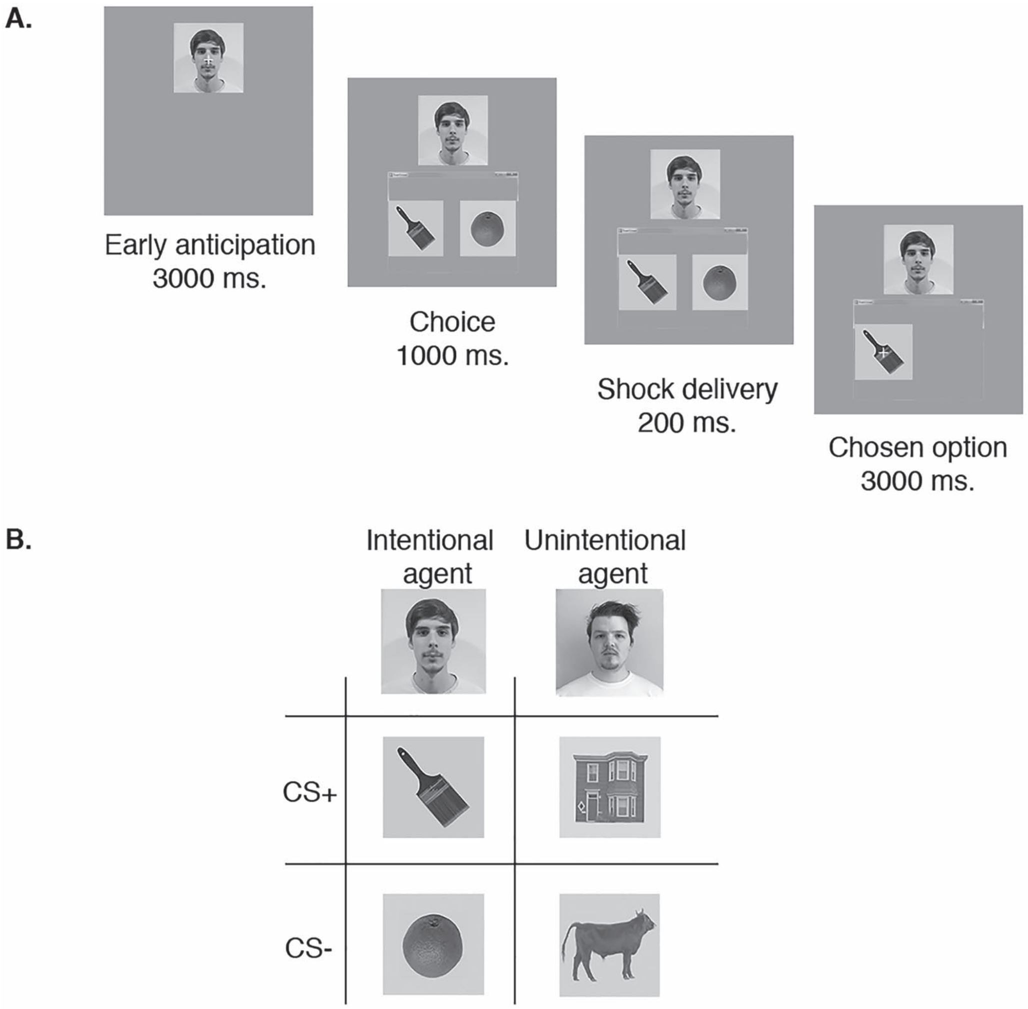

Figure 1. Trial illustration. (*A*) On each trial, the subject passively observed a coparticipant make a choice between two images. The coparticipant's face was always present on the screen and the window appeared when a connection has been made. A fixation cross was present on the coparticipant's face on the early anticipation phase and moved to the choice that is made during the choice period. This ensured that the participant viewed either the face or the choice during these periods, respectively. If the coparticipant chose an image that would be preceded with the delivery of a shock, the shock was delivered for 200 ms and ended before the choice image presentation. (*B*) The 2 × 2 design.

decide which one of the participants would conduct the experiment in the MR scanner (see Supplementary Material 3 for a verbatim account of the information that participants received before the start of the start of the experiment). Following this, the participant picked a paper from the lottery bag and found out that the paper indicated "MR camera". Upon revealing their own ballots allegedly stating "Outside", the coparticipants left the room. At this point, the participant repeated the instructions given during the initial instructions to the experimenter to make sure everything is clear.

The experiment included the following stages: functional localizer scan for each object category, a resting state scan, the aversive learning stage, a second resting state scan, and test. The next day, participants came back to the institute for a memory test outside the scanner. Only the results from the aversive learning phase will be discussedhere.

Aversive Learning

The aversive learning stage consisted of a modified aversive learning paradigm in which two coparticipants delivered shocks to the participant by means of choosing one out of two neutral images (Fig. 1A). For both coparticipants, choosing one of the images (CS+) caused the immediate delivery of shocks to the participant in 50% of the time during aversive learning (partial reinforcement). Choosing the other image (CS−) was never paired with the delivery of shocks to the participant. Importantly, participants believed that only one (the "intentional") coparticipant had information about which image choice led to the delivery of shocks, the participant believed that this coparticipant intentionally chose to deliver or not deliver shocks to the participant. In contrast, participants believed that the unintentional coparticipant had no information about the fact that they were delivering shocks, or the contingency between the shocks and the CS+ image.

Each coparticipant was assigned two images, as indicated above, counterbalanced across participants. The aversive learning stage consisted of 26 choices for each coparticipant, 13 for each CS type. Seven of the CS+'s were associated with the shock (unconditioned stimulus, USA). The participant was led to believe in the presence of an online screen-sharing system, which allowed them to watch decisions made by each coparticipant in real time. In each trial, the participant first viewed the photograph of the coparticipant, which was followed by a window frame (i.e., the screen of the coparticipant). Each trial consisted of three phases: 1) early anticipation (3 s): presentation of a facial photograph of the coparticipant, 2) choice (1 s): presentation of the photograph of the coparticipant together with the two alternative options (CS+ and CS− stimuli) presented just below the face, and 3) chosen option (3 s): presentation of the photograph of the coparticipant and the choice (CS+ or the CS−) made by the coparticipant, see Figure 1A for a more detailed overview. In the case of a reinforced CS+, shock delivery was for 200 ms at CS+ onset. Inter-trial intervals (ITI) were fixed to a length of 13 s, during which a fixation cross was presented. The onset of each trial was triggered by an fMRI pulse. The fixed and relatively long ITI (13 s) allowed us to directly compare two consecutive trials of each condition without the interference of temporal autocorrelations caused by temporal proximity. Indeed, this ITI length has been shown to reduce intrinsic noise correlations substantially compared to event-related designs (hisser et al. 2016a). The trial order was fixed (counterbalanced across all participants) and consisted of a repeated sequence of seven target trials, with filler trials of the same stimuli in between (Visser et al. 2013, 2015). Target trials consisted of nonreinforced trials, meaning the participant did not receive any electrical stimulation on those trials and responses would not be confounded by movement artifacts. Only target trials are used in fMRI analyses. For a more detailed explanation of the target and filler trials, see Supplementary Material 4. For pupil dilation analysis results are reported for all trials as movements artifacts are not an issue, and fast responses of the pupil allow us to disregard the time-period of response to the shock.

Data Analysis

Trial-by-Trial Similarity

For the trial-by-trial similarity analysis, each trial was modeled as a separate regressor in a voxel-wise whole-brain analysis using a single generalized linear model (GLM), with the US and motion parameters as nuisance regressors. We modeled the BOLD response using the onset and duration (3 s each) of the early anticipation and chosen option phases (Fig. 1A) of each trial, in separate GLM's. This yielded two sets of single trial regressors—one belonging to the early anticipation and the other to the chosen option phase. The resulting single-trial parameter estimates were further analyzed in Matlab by calculating pair-wise Pearson correlations between event-related spatial patterns of activation (vectors containing standardized parameter estimates per voxel). This results in a similarity matrix containing relations among trials for each participant and for each ROI (for early anticipation and chosen option phases separately). From this matrix, correlations of interest (e.g., between consecutive trials of the same stimulus) are selected. The strength of these correlations is used as a metric of similarity. Correlations are then Z-transformed. The matrix figures, however, display raw data as this facilitates the interpretation of the results. The correlations of interest were between two consecutive presentations of the same stimulus (within-stimulus), represented by the off-diagonal in the matrix (Fig. 3A).

Statistical Analysis

We performed repeated-measures analysis of variance (ANOVA) on preprocessed and Z-transformed pupil dilation, RSA correlations, and behavioral measures using Statistical package for the Social Sciences (SPSS, version: 25). We used paired *t*-tests to compare means between behavioral measures that compared intentional and unintentional confederates. We assumed that learning would be indexed by a main effect of aversiveness in the RSA correlations and pupil dilation responses, and an effect of intentionality would be indexed by a main effect of intentionality. We tested the predictions while correcting for multiple comparisons for the 12 ROIs, by limiting the false discovery rate (FDR) (Benjamini and Hochberg 1995).

Results

Behavioral Results

Evaluative ratings of each coparticipant before and after the aversive learning sessions confirmed that the manipulations were successful. Participants reported receiving more shocks when CS+'s (CS+ $_{\text{intent}}$ $M = 6.07$, standard deviation (SD) = 3.83; CS+ $_{\text{unintent}}$ $M = 3.41$, SD = 3.41; CS−$_{\text{intent}}$ $M = 0.71$, SD = 1.12; CS−$_{\text{unintent}}$ $M = 1.10$; SD = 2.28) were chosen $F(1,32) = 73.96$, $P < 0.001$, $\eta^2 = 0.69$. After the experiment, participants reported feeling greater discomfort to intentional shocks, compared to the unintentional ones ($t(31) = 2.56$, $P = 0.016$, $\eta^2 = 0.53$) (Fig. 2A). Additionally, participants reported receiving higher number of shocks from (t (31) = 2.41, $P = 0.022$, $d = 0.46$) (Fig. 2B), wanting to give more shocks back to (t (31) = 2.47, $P = 0.032$, $d = 0.33$) (Fig. 2C), and feeling more angry toward ($t(31) = 4.48$, $P < 0.001$, $d = 0.82$) (Fig. 2D), the intentional player. We tested if the participants overestimated the number of intentional shocks, by comparing the number of reported shocks from the intentional player to the actual number (i.e., 6). The results suggested that there was indeed a bias to overestimate the number of intentional shocks (mean = 1.65; SD = 3.88; $t(31) = 2.413$, $P = 0.022$). We also observed a greater decrease in likability ratings to the intentional player after aversive learning, compared to the unintentional player ($F(1,58) = 11.27$, $P = 0.004$, $\eta^2 = 0.13$) (Fig. 2E). Ratings completed after the experiment revealed that the participants expected to receive more shocks upon seeing the CS+ $_{\text{intent}}$, compared to the CS+ $_{\text{unintent}}$ (Interaction of CS type (2) and intentionality (2) ($F(1,31) = 4.38$, $P = 0.044$, $\eta^2 = 0.12$)). Participants also reported higher amount of intentional choices made, compared to unintentional ones (main effect of intentionality (2) $F(1,31) = 8.80$, $P = 0.006$, $\eta^2 = 0.22$)). At the end of the experiment we asked participants to indicate how much they doubted coplayers (confederates) involvement in the task in percentages. Eleven out of the 33 participants that are used in the final fMRI analysis indicated doubting the social interaction (e.g., doubted that the online connection was not real) more than 50% during the learning phase of the experiment. When repeating the analyses with believability ratings as a covariate the effects of anger ratings, expectancy ratings, and perceived number of shocks received from the intentional versus unintentional coplayer remained, but the effects of intentionality on discomfort, revenge, and likeability disappeared ($P > 0.05$) (Supplementary Table 1).

Pupil Dilation

We found greater pupil responses to the CS+ versus CS−, indicating that participants learned the contingencies ($F(1,33) = 21.95$, $P < 0.001$, $\eta^2 = 0.399$) (Fig. 2F). The intentionality of the choice had no effect on the pupil responses. There were no differences in

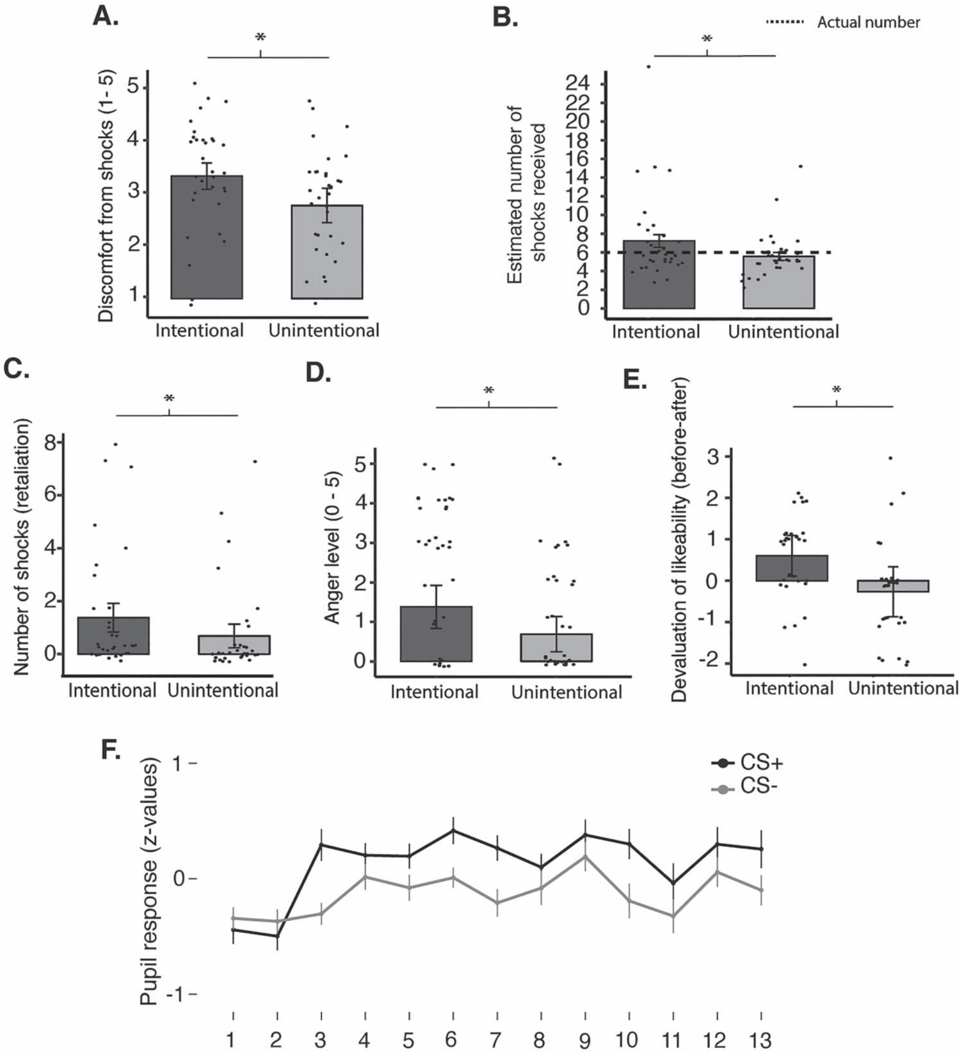

Figure 2. Pupil dilation and behavioral responses. Postexperimental questionnaire answers to questions regarding the social interaction: (*A*) discomfort of shocks received from each coplayer, (*B*) number of shocks the participant reported to receive via the presentation of each choice image, (*C*) how many shocks the participant would like to deliver back if given the chance, (*D*) how angry the participant felt toward the coparticipants. (*E*) Change in participants' evaluations of how likable each coparticipant was, from before the experiment to after the experiment. The number of shocks for both questions were open ended, the anger and likeability measures were reported out of a maximum point of 5, and a minimum 0. (*F*) Pupil dilation responses to CS+ (aversive) and CS− (safe) choice images during the social interaction. Error bars represent standard error of the mean (SEM).

pupil dilation responses to either of the social partners during the early anticipation phase (see Supplementary Fig. 1).

Trial-by-Trial Similarity

We analyzed fMRI data from the aversive learning phase of the experiment focusing on two distinct trial periods: i) early anticipation, during which the coparticipants face is shown on the screen, and ii) chosen option, during which the CS that was chosen is presented (see Fig. 1A). The early anticipation period allows us to investigate the responses to each individual the participant is interacting with, and the chosen option period allows us to investigate the responses to the action outcome. We focus our reporting of imaging results on findings that passed the statistical threshold in the main text and the results from all other ROIs are reported in the Supplementary Figure 2 and Supplementary Table 2.

Consistent with previous work (Visser et al. 2011, 2013), within stimulus trial-by-trial similarity analysis revealed learning curves (Fig. 3A–C) that index the formation of associative fear with differential increase of within-stimulus correlations to the CS+ (main effect of CS type (2)) in the insula (F (1,32) = 7.89, P = 0.008, η^2 = 0.198) (Fig. 3A), the ACC (F (1,32) = 6.98, P = 0.013, η^2 = 0.179) (Fig. 3B), the IFG (F (1,32) = 15.60, P < 0.001, η^2 = 0.328) (Fig. 3C), the dmPFC (F (1,32) = 5.874, P = 0.021, η^2 = 0.155), the arSTS (F (1,32) = 8.892, P = 0.005, η^2 = 0.217), the prSTS (F (1,32) = 5.443, P = 0.026, η^2 = 0.145), and the vmPFC (F (1,32) = 4.459, P = 0.043, η^2 = 0.122). We found significant main effects of the CS type also in the lTPJ (F (1,32) = 8.38, P = 0.003, η^2 = 0.241), which was not previously reported in other aversive learning studies. All of the reported values except for the vmPFC survived FDR correction. There were no significant differences between the conditions in the control region, ITGto. The regions in the aversive learning network had a preference for CS+ rather than

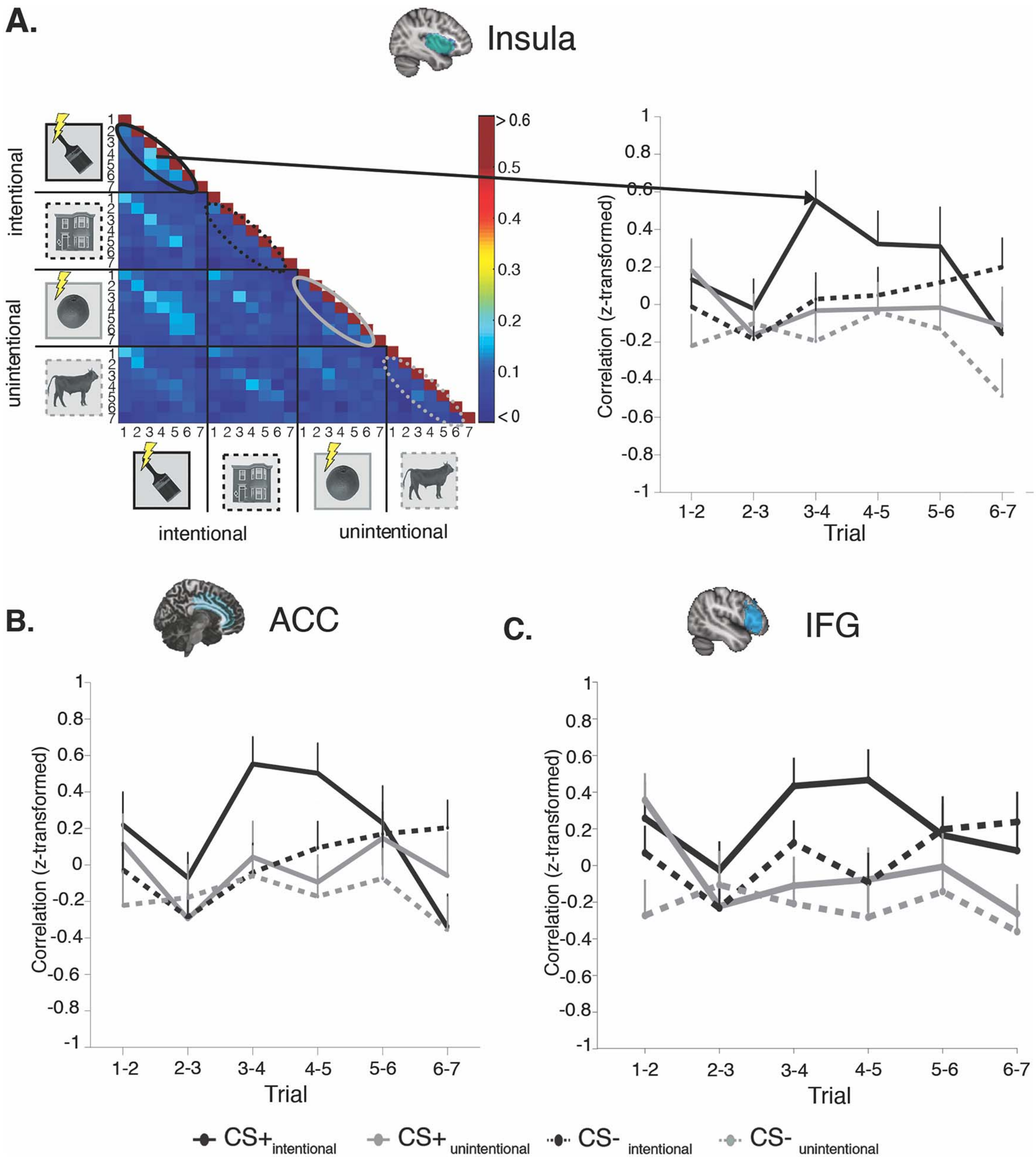

Figure 3. RSA. A. Trial-by-trial pattern similarity correlations in the Insula. The 24 × 24 correlation matrix represents correlations of neural patterns during learning. The off diagonal represents correlations between consecutive trials. The upper diagonal has been removed as it is a mirror image of the lower. B. Trial-by-trial similarity correlations in the ACC and C. in the IFG. Error bars represent SEM.

the CS−, which became more prominent after the habituation trials (Trials 1 and 2) (Fig. 3A–C).

There were main effects of intentionality (2) in within-stimuli trial-by-trial similarity in the insula (*F* (1,32) = 5.42, *P* = 0.02, η^2 = 0.145) (Fig. 3A), the ACC (*F* (1,32) = 5.97, *P* = 0.02, η^2 = 0.153) (Fig. 3B), the IFG (*F* (1,32) = 5.96, *P* = 0.02, η^2 = 0.157)(Fig. 3C), the dmPFC (*F* (1,32) = 4.70, *P* = 0.03, η^2 = 0.128), and the arSTS (*F* (1,32) = 4.06, *P* = 0.05, η^2 = 0.113), suggesting a preference in these regions to the intentional CS's. However, these effects did not survive FDR corrections. All ANOVA results are reported in Supplementary Table 2.

When we corrected for the believability of the intentionality manipulation, effects of intentionality remained significant for those ROIs that showed effects before, and became significant in the amygdala (*P* = 0.048) (Supplementary Table 3). These results highlight the additional effect of intentionality in the results presented, as the main effect of CS seems to be reliant on the believability of the experiment in certain ROIs.

To allow for comparisons with univariate analysis methods, we calculated average activation per trial in each of the ROIs (see Supplementary Fig. 3 and Supplementary Table 4 for the ANOVA results). Additionally, we conducted a whole-brain univariate GLM analysis. The results of the univariate analyses can be seen in Supplementary Figure 4, Supplementary Tables 5 and 6.

We explored effects outside our a priori ROIs by repeating the RSA analyses in 111 anatomical ROIs (all cortical and subcortical

ROIs provided with the Harvard–Oxford cortical and subcortical atlas, per hemisphere). All masks were thresholded at > 25% probability to limit overlap between neighboring regions. See Supplementary Table 7 for an overview of trial-by-trial similarity in all cortical and subcortical ROIs.

For a preliminary analysis of the relationship between the threat value in the insula and discomfort, see Supplementary Fig. 5.

Discussion

Our study aimed to investigate the neural mechanisms underlying the integration of knowledge about an individual and the outcome of their actions. Behaviorally, participants evaluated the intentional coparticipant to be less likable than before the interaction, were angrier toward him, and wanted to retaliate against him to a greater extent. The number of shocks received from the intentional coparticipant were overestimated but were reported close to correct for the unintentional coparticipant. Participants rated shocks received from the intentional coplayer as more uncomfortable than the unintentional ones. Neurally, we show that associative learning from intentional harmful actions increase the similarity of neural patterns throughout the cortex, specifically in regions related to understanding others' minds and aversive learning. Our results tentatively suggest that the regions involved in integrating information about the intentionality of an action overlaps with those of involved in aversive learning, and that specifically intentional harmful actions lead to a neural pattern representation even in the absence of shocks. Overall, we show no dissociation between aversive learning and mentalizing networks. Rather, it seems that intentionality of a harmful action can be observed as modulation of aversive processing.

Overlapping Network of Regions across the Cortex Represent Mentalizing and Aversive Learning

Our study integrates research in two different domains: i) mentalizing, and ii) aversive learning. We tentatively identified brain regions of intentionality information processing in the mentalizing network: the insula, the IFG, the dmPFC, and the arSTS. Additionally, we observed effects of intentionality in the ACC, which was not in the "mentalizing network" we outlined in our hypotheses. Most prior work on mentalizing used tasks involving third-party scenarios, typically asking the participant to first read a vignette about a fictitious character that harm another individual intentionally or unintentionally, and then asked to make judgments. Albeit important for understanding social cognition, such third-party tasks have limited validity in examining real-life social interactions. Second-party paradigms, in which the participant interacts with a coparticipant (e.g., a confederate) enhance the ecological validity of the tasks and thereby the generalizability of the results. We replicated findings from aversive learning research and have shown increased neural correlations in response to a stimulus associated with an aversive outcome in the insula, and the ACC (Visser et al. 2011, 2013, 2015, 2016a, 2016b; Dunsmoor et al. 2014), with the additions of the IFG, the dmPFC, the arSTS, and the prSTS—regions that have been reported in research on mentalizing (Koster-Hale and Saxe 2013; Yu et al. 2015). Additionally, the RSA captured changes related to safety responses in the vmPFC, as seen by the increase in similarity in the CS− responses. We observed that intentional harms are processed in a distributed network of regions, consisting of an overlap of regions from both the aversive learning and mentalizing networks. These findings support previous research demonstrating overlapping activity in a network, including the ACC and the insula that mediates both direct and social fear learning (Lindström et al. 2018). We show that the neural pattern similarities that develop during learning can be modified by social factors, which to our knowledge has not been documented before. Similarly, aversive learning literature shows enhanced arousal to fear-relevant CS+'s, compared to fear-irrelevant ones (Tomarken et al. 1989). The arousal effects are reflected in increased skin conductance responses but have not been replicated in neural pattern similarities.

The Role of ACC and the Insula in Representing the Self and Others

Here, we show that a neural pattern emerges in the insula while learning from other people's harmful actions that represents the integration of the aversive value of the action with the intentionality behind it (Fig. 3A–C). Past research has indeed shown that intentional harm is found to be more painful than nonintentional harms (Gray and Wegner 2008), but the neural representations of this effect has so far been unexplored. Our findings in the insula are in agreement with the current literature; the insula has been shown to be involved in alterations in the perception of a sensory experience caused by cognitive (Atlas and Wager 2012) and emotional (Orenius et al. 2017) states.

As delivery of electrical shocks lead to movement and multivariate analyses are very sensitive to motion, we used target trials in which the coparticipant chose a harmful option but resulted in no shocks (see Methods and Supplementary Materials for more details). All RSA correlations are reported for these target trials, meaning that the neural tuning in the insula relied on the learned value of the CS, and not the direct experience of a shock. Hence, the correlations we observe are representative of the learned value of an intentional harm, which has been directly experienced via a social interaction. This interpretation is in line with recent research that shows the ACC and the anterior part of the insula are part of a larger network that mediates the effects of social information on pain perception (Koban et al. 2019). Stimuli that were rated as "high pain" by others led to activity in ACC, the insula, dorsolateral prefrontal cortex, and parietal areas, compared to the "low pain" rated ones. Additionally, the application of univariate analyses in previous research has shown that the anterior part of the insula is responsive while waiting for an aversive taste delivery from an intentional agent, compared to an unintentional one (Liljeholm et al. 2014).

Illusory Correlations

Albeit receiving identical numbers of electrical shocks from each coparticipant, participants reported receiving more intentional shocks than they actually received (Fig. 2A). Hence, it seems that the participants have perceived the intentional coparticipant to make harmful choices more often than he actually did. As mentioned above, aversive learning studies reporting illusory correlations, participants have shown increased association between fear-related images and electrical shocks, than fear-irrelevant ones (for a review see Ohman and Mineka 2011). To our knowledge, there are no studies to date that report illusory correlations via a manipulation of knowledge about social partners. Research on illusory correlations typically shows enhanced skin conductance responses to the fear-relevant stimuli, however, we found no differences between intentional and unintentional harmful decisions in the pupil dilation analyses.

Limitations

The design of the study consisted of a manipulation involving actors, and we report our results based on a task involving allegedly real social interactions. This requires participants to believe that the coplayers and their online interactions were real. We noted that some participants were more doubtful than others (11 out of 33 doubted the experiment more than 50%, see Methods), although only three reported having 100% doubts. Reported believability effected discomfort, revenge, and likeability ratings (see Supplementary Table 1), but not the RSA results (Supplementary Table 3). The power of the electrical shocks serving as the US was set through calibrating the voltage based on the individuals' own report of being "uncomfortable". Information about others' mental states can, however, not be individually calibrated, and might be more elusive and subjectively variable. This might be the reason why the representation of the CS+ is stronger than the representation of intentionality in our RSA results.

Conclusions

Our study offers new insights into how the brain represents others' actions during social interactions. We introduced a naturalistic social interaction task to study learning from others' actions, and showed that learning can be indexed by neural pattern formation in time. Additionally, our results indicated that responses to the aversiveness of an action were regulated by the intentionality behind the action across the cortex. Taken together, we show that the intentionality of an action can enhance the perceived aversiveness of the outcome, which is reflected in neural pattern correlations during a social interaction.

Supplementary Material

Supplementary material can be found at *Cerebral Cortex* online.

Funding

The Knut and Alice Wallenberg Foundation (KAW 2014.0237), a European Research Council Starting grant (284366; Emotional Learning in Social Interaction project), and a Consolidator grant (2018–00877) from Swedish Research Foundation (VR). The Netherlands Organization for Scientific Research (NWO Veni) grant (016.195.246 to R.M.V.).

Notes

We thank Predrag Petrovic for comments on an earlier version, Rute Sequiera, Alexandros Kastrinogiannis, and Tess Collins for help during the data collection, and Amy Walsh for help with editing the manuscript. *Conflict of Interest*: None declared.

References

Ames DL, Fiske ST. 2015. Perceived intent motivates people to magnify observed harms. *Proc Natl Acad Sci*. 112(12):201501592. doi: 10.1073/pnas.1501592112.

Atlas LY, Wager TD. 2012. How expectations shape pain. *Neurosci Lett*. 520(2):140–148. doi: 10.1016/j.neulet.2012.03.039.

Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. 2001. The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J Autism Dev Disord*. 31(1):5–17. doi: 10.1023/A:1005653411471.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Cushman F, Gray K, Gaffey A, Mendes WB. 2012. Simulating murder: the aversion to harmful action. *Emotion*. 12(1):2–7. doi: 10.1037/a0025071.

Delgado MR, Jou RL, Phelps EA. 2011. Neural systems underlying aversive conditioning in humans with primary and secondary reinforcers. *Front Neurosci*. 5:71.doi: 10.3389/fnins.2011.00071.

Dunsmoor JE, Ahs F, Zielinski DJ, LaBar KS. 2014. Extinction in multiple virtual reality contexts diminishes fear reinstatement in humans. *Neurobiol Learn Mem*. 113:157–164. doi: 10.1016/j.nlm.2014.02.010.

Dunsmoor JE, Kragel PA, Martin A, Labar KS. 2013. Aversive learning modulates cortical representations of object categories. *Cereb Cortex*. doi: 10.1093/cercor/bht138. http://cercor.oxfordjournals.org/content/early/2013/05/23/cercor.bht138.full.

Fowler JC, Allen JG, Oldham JM, Frueh BC. 2013. Exposure to interpersonal trauma, attachment insecurity, and depression severity. *J Affect Disord*. 149(1–3):313–318. doi: 10.1016/j.jad.2013.01.045.

Frith CD, Frith U. 2006. The neural basis of mentalizing. *Neuron*. 50(4):531–534. doi: 10.1016/j.neuron.2006.05.001.

Fullana M, Harrison B, Soriano-Mas C, Vervliet B, Cardoner N, Àvila-Parcet A, Radua J. 2015. Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Mol Psychiatry*. 21:500–508. doi: 10.1038/mp.2015.88.

Gray K, Wegner DM. 2008. The sting of intentional pain. *Psychol Sci*. 19(12):1260–1262. doi: 10.1111/j.1467-9280.2008.02208.x.

Haxby JV, Hoffman EA, Gobbini MI. 2000. The distributed human neural system for face perception. *Trends Cogn Sci*. 4(6):223–233. doi: 10.1016/S1364-6613(00)01482-0.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*. 293:2425–2430. doi: 10.1126/science.1063736.

Koban L, Jepma M, López-Solà M, Wager TD. 2019. Different brain networks mediate the effects of social and conditioned expectations on pain. *Nat Commun*. 10(1):4096. doi: 10.1038/s41467-019-11934-y. http://www.ncbi.nlm.nih.gov/pubmed/31506426%0A http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6736972.

Koster-Hale J, Saxe R. 2013. Functional neuroimaging of theory of mind. In: Baron-Cohen S, Lombardo N, Tager-Flusberg, editors. *Understanding other minds*. 3rd edition. Oxford University Press, Oxford, United Kingdom. p. 132–163.

Kriegeskorte N. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. *Front Syst Neurosci*. 2:4. doi: 10.3389/neuro.06.004.2008. http://journal.frontiersin.org/article/10.3389/neuro.06.004.2008/abstract.

Levine S, Mikhail J, Leslie AM. 2018. Presumed innocent? How tacit assumptions of intentional structure shape moral judgment. *J Exp Psychol Gen*. 147(11):1728–1747. doi: 10.1037/xge0000459.

Li W, Howard JD, Parrish TB, Gottfried JA. 2008. Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. *Science*. 319(5871):1842–1845. doi: 10.1126/science.1152837.

Liebowitz MR. 1987. Social phobia. *Mod Probl Pharmacopsychiatry*. 22:141–173. doi: 10.1159/000414022. http://www.karger.com/?doi=10.1159/000414022.

Liljeholm M, Dunne S, O'Doherty JP. 2014. Anterior insula activity reflects the effects of intentionality on the anticipation of aversive stimulation. *J Neurosci*. 34(34):11339–11348. doi: 10.1523/JNEUROSCI.1126-14.2014. http://www.ncbi.nlm.nih.gov/pubmed/25143614.

Lindström B, Haaker J, Olsson A. 2018. A common neural network differentially mediates direct and social fear learning. *Neuroimage*. 167:121–129. doi: 10.1016/j.neuroimage.2017.11.039.

Mende-Siedlecki P, Cai Y, Todorov A. 2013. The neural dynamics of updating person impressions. *Soc Cogn Affect Neurosci*. 8(6):623–631. doi: 10.1093/scan/nss040.

Mitchell JP. 2004. Encoding-specific effects of social cognition on the neural correlates of subsequent memory. *J Neurosci*. 24:4912–4917.doi: 10.1523/jneurosci.0481-04.2004.

Ohman A, Mineka S. 2011. Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychol Rev*. 108(3):483–522. doi: 10.1037/0033-295X.108.3.483. http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.108.3.483.

Orenius TI, Raij TT, Nuortimo A, Näätänen P, Lipsanen J, Karlsson H. 2017. The interaction of emotion and pain in the insula and secondary somatosensory cortex. *Neuroscience*. 349:185–194. doi: 10.1016/j.neuroscience.2017.02.047.

Pavlov IP. 1927. *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Oxford, England: Oxford University Press.

Spoormaker VI, Andrade KC, Schröter MS, Sturm A, Goya-Maldonado R, Sämann PG, Czisch M. 2011. The neural correlates of negative prediction error signaling in human fear conditioning. *Neuroimage*. 54(3):2250–2256. doi: 10.1016/j.neuroimage.2010.09.042.

Tamir DI, Thornton MA, Contreras JM, Mitchell JP. 2016. Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. *Proc Natl Acad Sci*. 113(1):194–199. doi: 10.1073/pnas.1511905112.

Tomarken AJ, Mineka S, Cook M. 1989. Fear-relevant selective associations and covariation bias. *J Abnorm Psychol*. 98(4):381–394. doi: 10.1037/0021-843X.98.4.381. http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-843X.98.4.381.

Tovote P, Fadok JP, Lüthi A. 2015. Neuronal circuits for fear and anxiety. *Nat Rev Neurosci*. 16(6):317–331. doi: 10.1038/nrn3945. http://www.nature.com/doifinder/10.1038/nrn3945.

Visser RM, de Haan MIC, Beemsterboer T, Haver P, Kindt M, Scholte HS. 2016a. Quantifying learning-dependent changes in the brain: single-trial multivoxel pattern analysis requires slow event-related fMRI. *Psychophysiology*. 53(8):1117–1127. doi: 10.1111/psyp.12665. http://doi.wiley.com/10.1111/psyp.12665.

Visser RM, Haver P, Zwitser RJ, Scholte HS, Kindt M. 2016b. First steps in using multi-voxel pattern analysis to disentangle neural processes underlying generalization of spider fear. *Front Hum Neurosci*. 10(August):222. doi: 10.3389/fnhum.2016.00222. http://journal.frontiersin.org/article/10.3389/fnhum.2016.00222.

Visser RM, Kunze AE, Westhoff B, Scholte HS, Kindt M. 2015. Representational similarity analysis offers a preview of the noradrenergic modulation of long-term fear memory at the time of encoding. *Psychoneuroendocrinology*. 55:8–20. doi: 10.1016/j.psyneuen.2015.01.021.

Visser RM, Scholte HS, Beemsterboer T, Kindt M. 2013. Neural pattern similarity predicts long-term fear memory. *Nat Neurosci*. 16(4):388–390. doi: 10.1038/nn.3345.

Visser RM, Scholte HS, Kindt M. 2011. Associative learning increases trial-by-trial similarity of BOLD-MRI patterns. *J Neurosci*. 31(33):12021–12028. doi: 10.1523/JNEUROSCI.2178-11.2011. http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.2178-11.2011.

Wagner DD, Kelley WM, Haxby JV, Heatherton TF. 2016. The dorsal medial prefrontal cortex responds preferentially to social interactions during natural viewing. *J Neurosci*. 36(26):6917–6925. doi: 10.1523/jneurosci.4220-15.2016.

Wheelock MD, Sreenivasan KR, Wood KH, Ver Hoef LW, Deshpande G, Knight DC. 2014. Threat-related learning relies on distinct dorsal prefrontal cortex network connectivity. *Neuroimage*. 102:904–912. doi: 10.1016/j.neuroimage.2014.08.005.

Willenbockel V, Sadr J, Fiset D, Horne GO, Gosselin F, Tanaka JW. 2010. Controlling low-level image properties: the SHINE toolbox. *Behav Res Methods*. 42(3):671–684. doi: 10.3758/BRM.42.3.671. http://www.springerlink.com/index/10.3758/BRM.42.3.671.

Yarkoni T, Poldrack R, Nichols T, Van Essen D, Wager T. 2011. NeuroSynth: a new platform for large-scale automated synthesis of human functional neuroimaging data. *Frontiers in Neuroinformatics Conference Abstract: 4th INCF Congress of Neuroinformatics*; Boston, United States. doi: 10.3389/conf.fninf.2011.08.00058.

Young L, Bechara A, Tranel D, Damasio H, Hauser M, Damasio A. 2010. Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron*. 65(6):845–851. doi: 10.1016/j.neuron.2010.03.003. http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.1990.03450020118041.

Yu H, Li J, Zhou X. 2015. Neural substrates of intention-consequence integration and its impact on reactive punishment in interpersonal transgression. *J Neurosci*. 35(12):4917–4925. doi: 10.1523/JNEUROSCI.3536-14.2015. http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3536-14.2015.

Zhang Y, Yu H, Yin Y, Zhou X. 2016. Intention modulates the effect of punishment threat in norm enforcement via the lateral orbitofrontal cortex. *J Neurosci*. 36(35):9217–9226. doi: 10.1523/JNEUROSCI.0595-16.2016. http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0595-16.2016.