Neural representation of social expectations during interpersonal decisions

Authors: Paloma Díaz-Gutiérrez[1], Juan E. Arco[1], Sonia Alguacil[2], Carlos González-García[3] & María Ruz[1]

[1] Mind, Brain and Behavior Research Center, University of Granada

[2] University Isabel I

[3] Ghent University

Keywords: expectations, MVPA, social decisions, Ultimatum Game, valence

Word count: 9158

Contact information: Department of Experimental Psychology, University of Granada, 18071, Granada, Spain.

Telephone: +34 958 24 06 60. Facsimile number: +34 958 24 62 39

E-mail address: mruz@ugr.es (M. Ruz).

Abstract

Several studies highlight the relevance of prior personal information during social interactions. Such knowledge aids in the prediction of others, and it affects choices even when it is unrelated to their actual behaviour. In this investigation we aimed to study the neural representation of positive and negative personal social expectations, how these impact on subsequent choices, and the effect of mismatches between expectations and encountered behaviour. We employed functional Magnetic Resonance in combination with a version of the Ultimatum Game where, previous to fair and unfair offers, participants were provided with information about their partners' moral traits. Univariate and multivariate analyses evidenced a set of regions underlying expectations about the partners in the game, involving the supplementary motor area (SMA) and inferior frontal gyrus (IFG). Further, we were able to decode the valence of expectations in these regions, and also in the medial prefrontal cortex (mPFC). Additionally, performance of the classifier in these clusters correlated with a behavioural choice bias, manifesting the impact of the valence of the expectations on participants' decisions. These results suggest that expectations based on social information guide future interpersonal decisions, and that the neural representation of such expectations is related to their influence on behaviour.

Decision-making is a crucial constituent of our daily life. A great part of our decisions involve social contexts, where we constantly engage in interactions with others. To make choices that best fit our goals, we weight different sources of information. For instance, within the framework of prediction coding (Friston 2005), optimal decision-making combines sensory input (*evidence*) with predictions (*priors;* Summerfield and De Lange 2014). In this line, the role of these predictions, or expectations, has been thoroughly examined in perceptual decisions, where several studies have showed pre-activation of target-related brain areas during the expectation period, prior to target onset (e.g., Puri et al. 2009; Esterman and Yantis 2010; González-García et al. 2016). More recently, there have been proposals linking predictive coding and the representation of social traits in relation to social expectations (e.g., Tamir and Thornton 2018). Nonetheless, to this date, how prior social information influencing subsequent decision-making is represented in the brain is not well understood. In the current investigation, we aimed to study the neural representation of social expectations during an interpersonal ultimatum bargaining game.

The impact of priors has been consistently reported in non-social settings. For instance, in the field of visual processing, expectations have been shown to shape perception, determining the representation of targets (e.g., Kok et al. 2012). Here, even unconscious information may set a prior that fosters subsequent detection of perceptually ambiguous stimuli (Chang et al. 2016). Regarding the influence of priors on decision-making, Lopez-Persem et al. (2016) observed that, when making a binary choice between elements of the same category, the preferred item was chosen more often and faster than the non-preferred one, even when the value of the different alternatives was similar. Although this bias is more frequent when facing difficult or ambiguous choices (Samuelson and Zeckhauser 1988), it tends to lead to suboptimal decisions as people are guided by this “default” preference and

do not properly consider potential future outcomes (Fleming et al. 2010). Crucially, this is also the case for social decisions, which can be biased by several sources of information at different stages of processing (Díaz-Gutiérrez et al. 2017).

In social settings, one of the key aspects refers to the context in which the decision takes place, as shown for example in framing effects (De Martino et al. 2006), social categorization (e.g., race prejudice; see for example, Tortosa et al. 2013; Mendoza et al. 2014) or emotion-based expectations (Oosterhof and Todorov 2008; Alguacil et al. 2015). These biases in social decisions can be studied with a diverse set of interpersonal tasks. The Ultimatum bargaining Game (UG; Güth et al. 1982) where participants decide whether to accept or not monetary offers from different partners, is one of the paradigms most frequently used. Here, "rational" decisions from an economic point of view would always accept, since you can only earn money. However, choices are strongly influenced by the fairness of the offer (how balanced both halves of the split are). People often show high rejection rates towards unfair offers (Sanfey et al. 2003), which has been explained in terms of inequity-aversion tendencies (Fehr and Camerer 2007) and punishment (Brañas-Garza et al. 2014). Others have emphasized the importance of social norms, and how these impact the perception of fairness (Chang and Sanfey 2013). From this perspective, people have prior beliefs about the offers they expect, and in such way they are more prone to reject offers that violate these predictions.

Still, how the brain represents these socially relevant priors in interpersonal games is largely unknown. In relation to this, several studies have described a set of regions, known as the social cognition network, underlying the representation of knowledge that guides social predictions in a broad context (Frith and Frith 2008). This includes personal traits, stereotyping, semantic knowledge about people or inferences about others and their mental

states (Tamir et al. 2016; Tamir and Thornton 2018). This network includes the temporo-parietal junction (TPJ), superior temporal sulcus (STS), precuneus (PC), anterior temporal lobes (ATL), amygdala and the medial prefrontal cortex (mPFC; Frith and Frith 2001; Frith 2007; Mitchell et al. 2008; Contreras et al. 2013). This "mentalizing network" underlies processes such as Theory of Mind (ToM), that is, the understanding of other people's intentions (Saxe and Kanwisher 2003). More specifically, some of these regions have showed reduced activity but sharpened representations of stimuli, manifesting the predictive role of ToM processes (Koster-Hale and Saxe 2013). Similarly, in the context of decisions in social contexts, the mPFC has been related to expectations of others' behaviour, with a dorsal/orbital dissociation depending on whether predictions were about ruled-behaviour or mental states, respectively (Corradi-Dell'Acqua et al. 2015). On the other hand, there are also cases where priors are not followed by their predictable consequence. In this line, during social interaction, different studies (Ruz and Tudela 2011; Fouragnan et al. 2013; Gabay et al. 2014) have observed increased activation in the ACC and the anterior insula (aI) when expectations about partners do not match their subsequent behaviour. In this line, Chang and Sanfey (2013) found a relationship between the deviation of the expectations and increase of activity in the aI, ACC and SMA. Specifically, in the ultimatum game, an increase of activation in the dorsolateral PFC (dlPFC) and aI is related to participants' reaction to unfair offers (Sanfey et al. 2003; Knoch et al. 2006).

The personal traits of others are essential components of these social representations (Tamir and Thornton 2018) and their processing has a dynamic nature (Freeman and Ambady 2011; Stolier et al. 2018) where interactions between top-down and bottom-up processes evidence how priors routed in stereotypes modify and interact with more basic visual processing (Stolier and Freeman 2016; Stolier and Freeman 2017). Further, these personal traits can be decomposed in three different dimensions, being valence (positive vs. negative) a key one of

these (Thornton and Mitchell 2017; Tamir and Thornton 2018). The representation of the character of others in association with positive or negative information has been shown to be an important source of bias in interpersonal decisions (Díaz-Gutiérrez et al. 2017). For instance, Delgado et al. (2005), found that participants trusted more partners associated with positive moral traits, compared to those having negative ones. Furthermore, a variety of studies employing the UG paradigm have observed that participants tend to accept more offers from partners associated with positive descriptions, compared to negative ones (Gaertig et al. 2012). This tendency is steeper when participants navigate uncertain scenarios (Ruz et al. 2011). Moreover, in this context, the use of high density electroencephalography (EEG) has showed how negative descriptions of partners lead to a higher amplitude of the medial frontal negativity (MFN; associated with the evaluation of outcomes, Yeung and Sanfey 2004; Hajcak et al. 2006) when decisions are made (Moser et al. 2014). This data show how, regardless of fairness, people evaluate offers as more negative when they come from a disagreeable partner. This knowledge about different personal traits has been suggested to be integrated by the mPFC (Van Overwalle 2009) through coupling with regions responding to specific traits (Hassabis et al. 2014), and increasing its activation when a partner's behaviour violates previous trait implications (Ma et al. 2012).

Despite the key relevance of valence in psychological theories and its marked impact on social decision-making, its representation at the neural level and its effect on subsequent choices is not well understood (Barrett and Bliss-Moreau 2009). Results of a recent metaanalysis (Lindquist et al. 2015) provide evidence of a general recruitment of a set of regions for valenced versus neutral information, including the bilateral aI, the ventral and dorsal portions of the medial PFC (vm/dmPFC), the dorsal ACC, SMA, and lateral PFC, which are associated with the "salience network" (Seeley et al. 2007; Menon and Uddin 2010) and some of them also related to cognitive control (Brass and von Cramon 2004;

Dosenbach et al. 2008). When they carried out multivariate analysis (MVPA; Haxby et al. 2014) to search for distributed regions sensitive to valence discrimination (positive or negative vs. neutral), they did not observe any distinctive patterns, which suggests that affect is mainly represented in a valence-general set of brain regions. However, Lindquist et al. (2015), found that the vmPFC/ACC were more frequently activated in positive vs. negative than in positive vs. neutral contrasts. These results were taken to indicate that these regions may represent valence information along a single bipolar dimension (Lindquist et al., 2015). This seems to indicate, as the authors argue, that valence discrimination in the brain, given its complexity and specificity, may be better explained using more detailed analysis.

In the current functional Magnetic Resonance Imaging (fMRI) study, we employed a modified version of the UG to investigate how socially relevant priors represented by the valence of personal descriptions of partners, bias interpersonal economic choices during an Ultimatum Game (Gaertig et al. 2012). Specifically, we aimed to study which neural regions code for the generation and maintenance of positive and negative expectations about other people. Furthermore, we also wanted to assed how these expectations bias decisions as well as to replicate previous findings related to the nature of the offer. Last, we intended to ascertain which neural mechanisms might be engaged when there is a mismatch between personal expectations and the partners' behaviour. We expected to find specific neural representations underlying social priors, being here the expectations about the partners, with different patterns depending on the valence of these predictions (Lindquist et al. 2015). Specifically, we expected these patterns to be represented in regions previously related to social cognition and priors in decision-making (Saxe and Kanwisher 2003; Contreras et al. 2012; González-García et al. 2016). Finally, we predicted that control-related areas would be engaged when expectations did not match partners' behaviour.

Methods

Participants

Twenty-four volunteers were recruited from the University of Granada (M = 21.08, SD = 2.92, 12 men). All of them were right-handed with normal or corrected vision, and received economic remuneration (20-25 Euros, depending on their decisions). Participants signed a consent form approved by the Ethics Committee of the University of Granada.

Stimuli

We employed 16 adjectives used in previous studies (see Gaertig et al., 2012) as trait-valenced descriptions of the game proposers, extracted from the Spanish translation of the Affective Norms for English Words database (ANEW; Redondo et al. 2007). Half of the adjectives were positive (M= 7.65 valence, SD= 0.43), and the other half were negative (M= 2.3 valence, SD= 0.67). All words were matched in arousal (M= 5.69, SD= 0.76), number of letters (M= 6.19, SD= 1.42) and frequency of use (M= 20.19, SD= 18.47). In addition, we employed numbers from 1 to 9 (two in each trial) in black colour to represent different monetary offers. Stimuli were controlled and presented by E-Prime software (Schneider et al. 2002). Inside the scanner, the task was projected on a screen visible to participants through a set of mirrors placed on the radiofrequency coil.

Task and procedure

Participants played the role of the responder in a modified UG (e.g., Gaertig et al., 2012), deciding whether to accept or reject monetary offers made by different partners (proposers). If they accepted the offer, both parts earned their respective splits, whereas if they rejected it, neither of them earned money from that exchange (see Figure 1A). Offers consisted in splits of 10 Euros, which could be fair (5/5, 4/6) or unfair (3/7, 2/8, 1/9). The number on the

left was always the amount of money given to the participant, and the one on the right was the one proposed by the partners for themselves.

Personal information about the partners was included as adjectives with different valence. A third of these descriptions were positive, another third negative, and the last third was neutral, represented by text indicating the absence of information about that partner ("no test"). The valence of the adjectives was orthogonal to the fairness of offer. The order of the offers and adjectives was randomized, and each type of personal information (positive, negative, no information) preceded each offer equally within and across runs. Decision-response associations were counterbalanced across participants.

To add credibility to the interpersonal game setting, participants were told that they were about to receive offers made by other participants in a study of a previous collaboration with a foreign university. Furthermore, to engage participants in the game as a real social scenario, prior to the scanner they performed two tasks in which they had to make economic decisions that would affect prospective participants. In one of the tasks, participants acted as proposers, filling a questionnaire where they had to make offers for 16 different unknown partners, who would be involved in future experimental games. Here, they had to split 10 Euros into two parts, one for themselves and the other for their partners. Additionally, in a second task, they played a short version of the Dictator Game (Kahneman et al. 1986), where they decided how to divide another 10 Euros between themselves and an anonymous partner, who would have a mere passive role concerning the output of the offer. Moreover, participants were told that offers they were about to see in the scanner were each provided by a different partner. Each offer would be preceded by a word that described their partners in some way, information that had been obtained through personality and social

questionnaires. Participants were informed that their final payment would vary (20-25 Euros) according to their choices during the previous tasks and the game in the scanner.

In the scanner, participants performed a total of 192 trials, arranged in 8 runs (24 trials per run). In each run a start cue of 6 s was followed by 24 trials. Each trial (see Figure 1B) started with an adjective for 1 s (mean = 2.98º), preceding a jittered interval lasting 5.5 s on average (4-7 s, +/0.76º). Then, the offer appeared for 0.5 s (1.87º), followed by a second jittered interval (mean = 5.5 s; 4-7 s, +/0.76º). Overall, the task lasted 41 minutes approximately.

Image acquisition and preprocessing

MRI images were acquired using a Siemens Magnetom TrioTim 3T scanner, located at the Mind, Brain and Behaviour Research Centre in Granada. Functional images were obtained with a T2*-weighted echo planar imaging (EPI) sequence, with a TR of 2000 ms. Thirty-two descendent slices with a thickness of 3.5 mm (20% gap) were extracted (TE = 30 ms, flip angle = 80 º, voxel size of 3.5 mm^3). The sequence was divided in 8 runs, consisting of 166 volumes each. After the functional sessions, a structural image of each participant with a high-resolution T1-weighted sequence (TR = 1900 ms; TE = 2.52 ms; flip angle = 9º, voxel size of 1 mm^3) was acquired.

Data were preprocessed with SPM12 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/). The first three volumes of each run were discarded to allow the signal to stabilize. Images were realigned and unwarped to correct for head motion, followed by slice-timing correction. Afterwards, T1 images were coregistered with the realigned functional images. Then, functional images were spatially normalized according to the standard Montreal Neurological Institute (MNI) template, and

smoothed employing an 8 mm Gaussian kernel. Low-frequency artefacts were removed using a 128 high-pass filter. On the other hand, data for multivariate analyses was only head-motion and slice-time corrected and coregistered.

Univariate analyses

First-level analyses were conducted for each participant, following a General Linear Model in SPM12. We employed an event-related design, where activity was modelled using regressors for each valence type of adjective and for the offers. It included three regressors for the Words (positive, negative, no information) and six for the Offers (Fair offers_Positive, Fair offers_Negative, Fair offers_Neutral, Unfair offers_Positive, Unfair offers_Negative, Unfair offers_Neutral). Regressors were convolved with a standard hemodynamic response, with adjectives modelled with their duration (1 s + jitter), and offers modelled as events with zero duration.

At the second level of analysis, *t*-tests were conducted for comparisons related to the presence of expectations (information about the partner > no information), the valence of the information (positive > negative, negative > positive) and the fairness of the offer (fair > unfair, unfair > fair). We also carried out contrasts for the interaction between the events, where we had matched (congruent: positive descriptions followed by fair offers, negative descriptions followed by unfair offers) and mismatched (incongruent) trials (positive descriptions followed by unfair offers, negative descriptions followed by fair offers). To control for false positives at the group level, we employed permutations tests with statistical non-parametric mapping (SnPM13, http://warwick.ac.uk/snpm) and 5000 permutations. We performed cluster-wise inference on the resulting voxels with a cluster-forming threshold at 0.001, which was later used to obtain significant clusters (FWE corrected at $p<0.05$).

Multivariate analyses

First, we performed MVPA to examine the brain areas representing the valence of partners' information (positive vs. negative adjectives). To this end, we used realigned images (prior to normalization) on which we performed a whole brain searchlight (Kriegeskorte et al. 2006). This allowed us to explore neural pattern distribution with no prior assumptions about the regions representing the valence of the information. We employed The Decoding Toolbox (TDT; Hebart et al. 2015), with which we created 12-mm radius spheres, and for each sphere, a linear support vector machine classifier (C=1; Pereira et al. 2009) was trained and tested using a leave-one-out cross-validation scheme, employing the data from the 8 scanning runs (where training was performed with data from 7 runs and tested in the remaining run, in an iterative fashion). We used a Least-Squares Separate model (LSS; Turner 2010) to reduce collinearity between regressors (Abdulrahman and Henson 2016; Arco et al. 2018). This approach fits the standard hemodynamic response to two regressors: one for the current trial (positive/negative adjective followed by accepted/rejected offer) and a second one for all the remaining trials. As in the previous analyses, adjective regressors were modelled with their duration (1 s + jitter), and offers with zero duration. Consequently, the output of this model was one beta image per event (total = 128 images, 112 for training and 16 for testing in each iteration). Afterwards, non-parametrical statistical analyses were performed on the resulting accuracy maps following the method proposed by Stelzer et al. (2013) for MVPA data. We permuted the labels and trained the classifier 100 times for each subject. The resulting maps were then normalized to an MNI space. Afterwards, we randomly picked one of these maps per each participant and averaged them, obtaining a map of group accuracies. This procedure was repeated 50000 times, building an empirical chance distribution for each voxel position and selecting the 50th greatest value, which corresponds to the threshold that marks the statistical significance. Only the voxels that surpassed this were considered significant. The resulting map was FWE corrected at 0.05, previously

computing the cluster size that matched this value from the clusters obtained in the empirical distribution.

Relationship between decoding accuracy and choices

To examine the extent to which decoding accuracy between the two types of adjectives (positive vs. negative) related to decisions made by participants, we performed a correlation analysis between an individual bias index and mean decoding accuracy values from each significant cluster in the MVPA. To obtain this behavioural index, for each participant, we subtracted the average acceptance rate following negative descriptions to the average acceptance rate after positive descriptions (regardless of the nature of the offer). For each subject, we performed a one-tail (right) Spearman's correlation between the behavioural index and the decoding accuracy from each significant cluster (FDR-corrected for multiple comparisons).

Results

Behavioural data

Acceptance rates (AR) and reaction times (RTs) were analysed in a Repeated Measures ANOVA, with Offers (fair/unfair) and Valence of the descriptions (positive, negative, neutral) as factors. Greenhouse-Geisser correction was applied when sphericity was violated.

Acceptance rates

Participants responded on 100% of the trials. Data showed (see Figure 2) a main effect of Offer $F_{1,23} = 74.50$, $p < .001$, $\eta_p^2 = .764$, where fair offers were accepted more often (M = 84.09%; SD = 22.10) than unfair ones (M = 24.18%; SD = 24.10). Valence was also significant, $F_{2,22} = 13.735$, $p = .001$, $\eta_p^2 = .374$. Participants accepted more offers when they

were preceded by a positive description of the partner (*M* = 59.39%; *SD* = 23.09), than when there was no information (*M* = 56.31%; *SD* = 21.89) or this was negative (*M* = 46.70%; *SD* = 24.33). Planned comparisons revealed that these differences were significant between all pairs (all *p*s<.05). Finally, the Offer X Valence interaction was also significant, $F_{2,22}$ = 4.262, p = .033, η_p^2 = .156. Planned comparisons showed that for fair offers, there was no difference between positive and neutral information (p = .399), whereas for unfair offers, there was no difference in acceptance rates between negative and neutral information (p = .074).

Reaction times

Results showed (see Figure 2) a main effect of Offer $F_{1,23}$ = 22.489, p < .001, η_p^2 = .494, where participants took longer to respond to unfair (*M* = 1023.53 ms; *SD* = 373.10 ms) than to fair offers (*M* = 925.62 ms; *SD* = 309.57 ms). Nonetheless, neither Valence, $F_{2,22}$ = 1.05, p = .341, or their interaction, $F_{2,22}$ = 1.956, p = .168 were significant. In addition, to measure the influence of expectations on participant's responses (Ruz et al. 2011), we ran an ANOVA where we included the valence of the descriptions and the decision (accept, reject) made to the offers. Here, we did not find any effect of Valence, *F*<1, but we found significant effects of Decision, $F_{1,23}$ = 5.519, p = .028, η_p^2 = .194, since participants were faster to accept (*M* = 951.37 ms; *SD* = 356.01 ms) than to reject the offers (*M* = 988.97 ms; *SD* = 316.91 ms). Furthermore, data showed an interaction Valence X Decision, $F_{2,22}$ = 4.23, p = .025, η_p^2 = .155, replicating previous findings (Ruz et al. 2011; Gaertig et al. 2012). Planned comparisons indicated that these differences in RT for responses took place only after positive, $F_{1,23}$ = 13.997, p = .001, η_p^2 = .378 (Accept: *M* = 927.60 ms, *SD* = 297.37 ms; Reject: *M* = 993.91 ms, *SD* = 335.52 ms), and neutral descriptions, $F_{1,23}$ = 4.504, p = .045, η_p^2 = .165 (Accept: *M* = 955.8 ms, *SD* = 304.96 ms; Reject: *M* = 987.80 ms, *SD* = 328.48 ms), but not for negative descriptions, *F*<1.

Neuroimaging data

Univariate results

Expectations

When participants had personal information to **generate expectations** (Positive adjective & Negative adjective > No Information), we observed a cluster activity (see Figure 3) in the left dorsal aI (k = 109; -33, 21, 4) and bilateral Supplementary Motor Cortex (SMA; k = 138; -8, 11, 53; see Fig. 3). Additionally, there was higher activity (k = 264; 55, -35, 53) in the right inferior parietal lobe (right IPL) for **positive descriptions** compared to negative ones, but no cluster surpassed the statistical threshold (p>0.05) for the opposite contrast.

During *offer processing*, the previous presentation of **personal information** about the partner (Offer_Pos and Offer_Neg > Offer_Neu) yielded again significant activity involving the bilateral dorsal aI and right SMA (k = 23349; -33, 21, 4).

To check whether the regions related to personal information were the same during the presentation of the valenced adjectives and during the presentation of the offer (positive and negative > neutral in both cases), we ran a conjunction analysis (Nichols et al. 2005). Similarly to each contrast individually, we observed two clusters: one in the left IFG/aI (k = 93; -3, 21, 0) and one involving bilateral SMA (k = 126; -5, 18, 53), suggesting that both areas increased their activation during the expectation and the decision periods.

Offer fairness

Fair offers (Fair > Unfair) generated activity (see Figure 4) in the right medial frontal gyrus (mFG) and ACC (k = 171; 6, 39, -14), while the opposite contrast (unfair > fair) did not

yield any significant clusters (p>0.05). Furthermore, we examined neural responses depending on whether previous expectations were matched or not by the nature (fair vs. unfair) of the offer. Here, **congruence** (see Figure 4) between expectations and offer (Congruent > Neutral) showed a cluster of activity in right cerebellum (right Crus; k = 153; 17, -88, -32). Conversely, **incongruence** (see Figure 4) between expectations and offer (Incongruent > Neutral) yielded activations in the right medial Superior Frontal Gyrus (mSFG) and its lateral portion bilaterally (k = 401; 13, 39, 56), as well as in left IFG (k = 177; -54, 39, 0). Lastly, regarding possible conflict effects, a comparison between **mismatched vs. matched** trials (Incongruent > Congruent) showed (see Figure 5) clusters of bilateral activity in the IFG/aI (k = 232; -43, 25, -11/ k = 140; 34, 35, 4).

Multivariate results

Valence of expectations' classification

Expectations about the partner showed distinct patterns of neural activity depending of the valence of the description (see Figure 5), in a cluster including the left inferior and middle frontal gyrus (IFG/MFG) and aI (k = 319; -46.5, 28, -32.2), the bilateral medial frontal gyrus and ACC (k = 483; 6, 21, -19.6) and the bilateral middle cingulate cortex (MCC) and SMA (k = 339; -4.5, 14, 35).

Although the same comparisons (positive vs. negative) in univariate GLM only yielded a significant cluster activation in the IPL for positive > negative expectations, we ran a conjunction analysis (Nichols et al. 2005) to test whether the regions that increased their activation during the presentation of the adjectives (positive & negative > neutral) contained relevant information about the valence (as reflected by multivariate results). For this, we computed the intersection between the group maps from both contrasts. Results showed two

clusters (see Figure 5): one in the left IFG/aI (k = 56; -36, 25, 0) and one involving bilateral SMA (k = 69; -8, 18, 46).

Correlation between decoding accuracy and the bias index

To explore the link between behaviour and the decoding results, we correlated the mean decoding accuracies (positive vs. negative) for each significant cluster in the MVPA with the bias index for each participant, which represents how much influence the valence of the adjectives had on their choices. This analysis yielded significant positive correlations between the decoding accuracy for the descriptions' valence and the behavioural bias in all 3 significant clusters (see Figure 6): the left IFG/MFG and aI (r = .42; p = .02), bilateral mFG/ACC (r = .44; p = .015), and the left MCC/SMA (r = .53; p = .0038). Hence, the better these regions discriminated between the valence of the partners' information, the larger the effect of valenced information on subsequent choices.

Discussion

Our study investigated the neural basis of social valenced expectations during an interpersonal ultimatum bargaining game. Results revealed how social information about other people bias subsequent economic choices, as well as the brain regions increasing their activity during the maintenance of expectations with a later impact on behaviour. Furthermore, decoding analysis allowed us to observe the areas that represent the content of such expectations, and how the strength of these social valenced representations influence ulterior decisions.

The ultimatum game employed showed a clear behavioural effect of interpersonal expectations, where positive descriptions of others led participants to higher acceptance rates

compared to negative ones. Further, expectations influenced fair and unfair offers differently. Thus, for fair offers, there was no difference between having positive expectations and having no information about the partner, whereas for unfair offers there was no difference between lack of information and negative information. Additionally, the impact of the expectations was reflected on the speed of choices, where people needed more time to reject offers after positive (or neutral) expectations. This data replicates previous results, where valenced descriptions of partners in bargaining scenarios bias economic choices (Ruz et al. 2011; Gaertig et al. 2012; Moser et al. 2014), emphasizing the role of expectations (Sanfey 2009) and valenced morality in decision-making (Barrett and Bliss-Moreau 2009). Results also fit with evidence from non-social decision making (Fleming et al. 2010; Lopez-Persem et al. 2016), manifesting the impact of priors on subsequent decisions. Overall, the behavioural pattern of choices observed supports the utility of the experimental paradigm to induce interpersonal valenced expectations about others that bias subsequent choices made to the same set of objective behaviour (offers made by partners) of others.

Several regions increased their activation when participants held in mind social expectations about game partners. This information engaged the SMA and the dorsal aI, which were also active at the decision stage. These are regions previously related to preparation processes (Brass and von Cramon 2004), as well as sustained (Dosenbach et al. 2008) and transient (Sridharan et al. 2008; Menon and Uddin 2010) top-down control, in paradigms where participants use cue-related information to perform tasks of different nature on subsequent targets. In the current context, these areas may be involved in using the interpersonal information contained in the cue to guide or bias the action towards a certain choice, according to the valence of the expectation. This explanation fits with the pattern of behavioral results obtained where positive expectations increased the acceptance of the

offers, as well as speeded up choices to fair offers matching prior expectations. However, univariate contrasts between the words containing positive vs. negative information, in stark contrast with behavioral outcomes, showed effects restricted on a cluster in the IPL for the positive vs. negative contrast. This region has been related to the simulation of others' action in shared representations (Van Overwalle 2009), and a part of it is included in the TPJ, which plays a main role in ToM (Saxe and Kanwisher 2003). Although the peak cluster does not share coordinates with the peak of their TPJ, it is nearby (e.g., Scholz et al. 2009). The increase of activation in this region for positive expectations is intriguing, and perhaps it could indicate a higher reliance on positive descriptions by the ToM processes involved in our task. The pattern of speed of choices observed in RTs would fit with this explanation; only positive expectations speeded acceptance choices, whereas negative description did not speed rejections. Further research will be needed to replicate this imbalance of information and to better understand the nature of its underlying brain processes.

The use of a multivariate approach, based on classification analysis (i.e., MVPA), allowed us to observe which regions are sensitive to whether the expectations about the partners are positive or negative. These areas included the SMA/MCC, IFG/MPFC and mPFC/ACC. There was no difference in RT between positive and negative conditions (see Behavioural results section), and therefore, we can rule out the possibility that the classifier was mistakenly discriminating faster and slower conditions, and that it correctly differentiated patterns referring to valence instead.

The SMA has been previously associated with general preparatory processes (Brass and von Cramon 2004), although some studies have also been able to decode specific task sets in this region (Bode and Haynes 2009; Crittenden et al. 2015). In social scenarios, Chang and Sanfey (2013) found a relationship between the activity in the SMA and the deviation of

previous expectations. This region has also been linked to the unspecific representation of valence (Lindquist et al. 2015). Our conjunction analysis show that part of the SMA increases its activity during the expectation period and also shows different patterns depending on the valence of the expectation. This data could suggest that the SMA carries a role of general preparation, but it also contains specific fine information relevant to the task. The MCC, on the other hand, has been associated with the increase in efficiency in decision-making, being involved in the anticipation and consequent expectations of outcomes in a variety of non-social tasks (Vogt 2016). Further, it has also been related to prediction and monitoring of outcomes in social decisions (Apps et al. 2013), and it may play a similar role in our study.

On the other hand, the patterns of activity in a cluster involving the lateral prefrontal cortex (lPFC), including the IFG and MPFC, also discriminated the valence of the expectations. Interestingly, these areas were part of a large cluster that also increased their activation during the maintenance of social information, as revealed by univariate results. In non-social paradigms, the lateral PFC has been related to working memory maintenance (Sala et al. 2003; Morgan et al. 2013) and other forms of cognitive control (e.g., Reverberi et al. 2012). The IFG specifically has been also associated with the selection of semantic information (Wagner et al. 2001; Jefferies 2013), and it is also involved in the expectation to perform different non-social tasks employing verbal material (e.g., Sakai and Passingham 2006; González-García et al. 2017). Notably, our results extend this role to a social context (see also Van Overwalle 2009; Filkowski et al. 2016; Thye et al. 2018), where verbal information is used to generate positive or negative expectations about game partners, by showing that the pattern of activity in this frontal region differs depending on the nature of the information used to predict the proximal behaviour of others.

Interestingly, a region that did not increase its overall activation during the expectation period contained patterns related to the valence of the predictions, the mPFC/ACC. Crucially, this area overlaps with the region isolated in the metaanalysis by Lindquist et al. (2015), where they linked its activity with a bipolar representation of valence. On a broader context, this region is part of the social cognition network, associated with mentalizing processes (Koster-Hale and Saxe 2013; Tamir et al. 2016), and behaviour guided by social cues, along with the ACC. Previous studies relate the mPFC with predictions about others' desires (Corradi-Dell'Acqua et al. 2015), and priors during valued decisions (Lopez-Persem et al. 2016). Additionally, Van Overwalle (2009) linked this region to the integration of personal traits, and it has been extensively associated with the representation of intentions as well (Haynes et al. 2007). In our experiment, this area maintains valenced information about others, which is later employed to bias decisions about their economic offers, as corroborated by the correlation with choices discussed in the following paragraph.

The association between a brain region and a given behaviour is strengthened when a link can be observed between the fidelity of a pattern its activity and the behavioural outcome studied (Naselaris et al. 2011; Tong and Pratte 2012). To find this evidence we obtained, for each participant, a bias index representing how much the valence of the personal information influenced their choices, and correlated this index with the accuracy of the classifier in disentangling the patterns generated by positive and negative words. We observed a positive correlation between these two factors in the three clusters sensitive to the valence of expectations. Thus, the better the classifier distinguished between descriptions of different valence, the more people tended to accept offers preceded by positive compared to negative descriptions. These results strongly suggest that these valenced representations were used to weight posterior acceptance or rejection decisions to the same set of objective offers, biasing behaviour.

We could also observe the effect of expectations by studying the brain activity generated by offers that matched or mismatched previous expectations, that is, fair and unfair offers preceded by descriptions of the same or opposing valence. Here we found cerebellum activity when fair offers were preceded by positive descriptions and unfair ones followed negative adjectives. This region is associated with prediction in a variety of contexts, such as language (Lesage et al. 2017; Pleger and Timmann 2018) or social cognition (Van Overwalle et al. 2014), among others. In social scenarios, where people frequently anticipate others' needs or actions, the understanding of the role of the cerebellum in social predictions is particularly relevant (Sokolov et al. 2017). Although previous studies (see Berthoz 2002) found increased activity in the cerebellum when predictions (social norms) were violated, we observed the opposite. Hence, our data suggests that the in the current context the cerebellum may signal when predictions are matched by social observations. Conversely, when predictions are not met, we observed activation in the lPFC, including the IFG, a region previously associated with semantic cueing (González-García et al. 2016), semantic control (Jefferies 2013) and emotional regulation during social decisions (Grecucci et al. 2013). Therefore, this data also supports the relevance of expectations, even at decision stages, when participants face the outcome of the interaction. At this point, they may need to suppress the previous information to act in accordance with the offer.

The present study advances our knowledge about how prior information about others that we acquire through verbal means is represented in our brain and impacts economic choices. Unfortunately, the ecological validity of our study is limited by the context of fMRI scanning in a single location. However, we increased the credibility of the social scenario by means of instructions, where we recreated an actual delayed interaction between participants of different studies, and made actual earnings contingent on choices during the game. In

fact, none of the participants showed signs of susceptibility about the underlying fictitious nature of the study when debriefed at the end of the session. Futures studies could use some form of virtual reality during scanning (Mueller et al. 2012) together with more complex verbal descriptions of others to study whether similar brain regions represent this content and the way it is structured. Additionally, another interesting point would be to try to find if there is a sort of “common valence space” for the two stages of the paradigm. That is, to find out if there is shared information underlying the valence of the adjective (positive/negative) but also the "pleasantness" of the offer (fair-positive, unfair-negative). A future study designed to employ cross-classification decoding approaches (Kaplan et al. 2015) between the expectation and the evidence game periods could advance on this respect.

Conclusions

We adapted a classical bargaining Ultimatum Game to fMRI, where participants generated positive or negative expectations about partners who afterwards made fair or unfair economic offers. Choices to the same set of offers were biased by the valence of the information, highlighting the relevance of priors in social decisions. A set of regions previously related to preparatory processes and cognitive control, semantic cueing, social cognition and valued decisions increased their activity to hold social expectations, and some of them coded their valence content. Importantly, the fidelity of this information predicted the impact of social priors on posterior choices, stressing the relevance of the patterns of activity in these regions for observable behavior. Future research increasing the complexity of the information content of the priors could shed more light into how our brain predicts and reacts to the behavior of others in social contexts. Altogether, results suggest that expectations prepare for future decisions, using the social information about partners to predict the most likely outcome from the interpersonal exchange. Moreover, the neural

representation of these expectations turns out to be related to the extent to which participants are influenced by such predictions.

Funding

This work was supported by the Spanish Ministry of Science and Innovation (PSI2013-45567-P and PSI2016-78236-P to M.R.) and the Spanish Ministry of Education, Culture and Sport (FPU2014/04272 to P.D.G.). This research is part of P.D.G's activities for the Psychology Graduate Program of the University of Granada. mruz@ugr.es

References

Abdulrahman H, Henson RN. 2016. Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for Beta-series correlation and multi-voxel pattern analysis. Neuroimage. 125:756–766. doi:10.1016/j.neuroimage.2015.11.009.

Alguacil S, Tudela P, Ruz M. 2015. Ignoring facial emotion expressions does not eliminate their influence on cooperation decisions. Psicologica. 36:309–335.

Apps MAJ, Lockwood PL, Balsters JH. 2013. The role of the midcingulate cortex in monitoring others' decisions. Front Neurosci. 7:251. doi:10.3389/fnins.2013.00251.

Arco JE, González-García C, Díaz-Gutiérrez P, Ramírez J, Ruz M. 2018. Influence of activation pattern estimates and statistical significance tests in fMRI decoding analysis.

Barrett LF, Bliss-Moreau E. 2009. Affect as a psychological primative. Adv Exp Soc Psychol. 41:167–218. doi:10.1016/S0065-2601(08)00404-8.Affect.

Berthoz S. 2002. An fMRI study of intentional and unintentional (embarrassing) violations of social norms. Brain. 125:1696–1708. doi:10.1093/brain/awf190.

Bode S, Haynes JD. 2009. Decoding sequential stages of task preparation in the human brain. Neuroimage. 45:606–613. doi:10.1016/j.neuroimage.2008.11.031.

Brañas-Garza P, Espín AM, Exadaktylos F, Herrmann B. 2014. Fair and unfair punishers coexist in the Ultimatum Game. Sci Rep. 4:6025. doi:10.1038/srep06025.

Brass M, von Cramon DY. 2004. Decomposing Components of Task Preparation with Functional Magnetic Resonance Imaging. J Cogn Neurosci. 16:609–620. doi:10.1162/089892904323057335.

Chang LJ, Sanfey AG. 2013. Great expectations: Neural computations underlying the use of social norms in decision-making. Soc Cogn Affect Neurosci. 8:277–284. doi:10.1093/scan/nsr094.

Chang R, Baria AT, Flounders MW, He BJ. 2016. Unconsciously elicited perceptual prior. Neurosci Conscious. 2016:niw008. doi:10.1093/nc/niw008.

Contreras JM, Banaji MR, Mitchell JP. 2012. Dissociable neural correlates of stereotypes and other forms of semantic knowledge. Soc Cogn Affect Neurosci. 7:764–770. doi:10.1093/scan/nsr053.

Contreras JM, Banaji MR, Mitchell JP. 2013. Multivoxel Patterns in Fusiform Face Area Differentiate Faces by Sex and Race. PLoS One. 8: e69684. doi:10.1371/journal.pone.0069684.

Corradi-Dell'Acqua C, Turri F, Kaufmann L, Clément F, Schwartz S. 2015. How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation. Cortex. 70:21–34. doi:10.1016/j.cortex.2015.02.011.

Crittenden BM, Mitchell DJ, Duncan J. 2015. Recruitment of the default mode network during a demanding act of executive control. Elife. 4: e06481. doi:10.7554/eLife.06481.

Delgado MR, Frank RH, Phelps EA. 2005. Perceptions of moral character modulate the neural systems of reward during the trust game. Nat Neurosci. 8:1611–1618. doi:10.1038/nn1575.

Díaz-Gutiérrez P, Alguacil S, Ruz M. 2017. Bias and control in social decision-making. In: Ibáñez A, Sedeño L, Gacría A, editors. Neuroscience and Social Science: The Missing Link.

Cham, Switzerland: Springer. p. 47–68.

Dosenbach NUF, Fair DA, Cohen AL, Schlaggar BL, Petersen SE. 2008. A dual-networks architecture of top-down control. Trends Cogn Sci. 12:99–105. doi:10.1016/j.tics.2008.01.001.

Esterman M, Yantis S. 2010. Perceptual expectation evokes category-selective cortical activity. Cereb Cortex. 20:1245–1253. doi:10.1093/cercor/bhp188.

Fehr E, Camerer CF. 2007. Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci. 11:419–427. doi:10.1016/j.tics.2007.09.002.

Filkowski MM, Anderson IW, Haas BW. 2016. Trying to trust: Brain activity during interpersonal social attitude change. Cogn Affect Behav Neurosci. 16:325–338. doi:10.3758/s13415-015-0393-0.

Fleming SM, Thomas CL, Dolan RJ. 2010. Overcoming status quo bias in the human brain. Proc Natl Acad Sci. 107:6005–6009. doi:10.1073/pnas.0910380107.

Fouragnan E, Chierchia G, Greiner S, Neveu R, Avesani P, Coricelli G. 2013. Reputational Priors Magnify Striatal Responses to Violations of Trust. J Neurosci. 33:3602–3611. doi:10.1523/JNEUROSCI.3086-12.2013.

Freeman JB, Ambady N. 2011. A dynamic interactive theory of person construal. Psychol Rev. 118:247–279. doi:10.1037/a0022327.

Friston K. 2005. A theory of cortical responses. Philos Trans R Soc B Biol Sci. 360:815–836. doi:10.1098/rstb.2005.1622.

Frith CD. 2007. The social brain? Phil Trans R Soc B. 362:671–678. doi:10.1098/rstb.2006.2003.

Frith CD, Frith U. 2008. Implicit and Explicit Processes in Social Cognition. Neuron. 60:503–510. doi:10.1016/j.neuron.2008.10.032.

Frith U, Frith C. 2001. The biological basis of social interaction. Am Psychol Soc. 10:151–155. doi:https://doi.org/10.1111/1467-8721.00137.

Gabay AS, Radua J, Kempton MJ, Mehta MA. 2014. The Ultimatum Game and the brain: A meta-analysis of neuroimaging studies. Neurosci Biobehav Rev. 47:549–558. doi:10.1016/j.neubiorev.2014.10.014.

Gaertig C, Moser A, Alguacil S, Ruz M. 2012. Social information and economic decision-making in the ultimatum game. Front Neurosci. 6:103. doi:10.3389/fnins.2012.00103.

González-García C, Arco JE, Palenciano AF, Ramírez J, Ruz M. 2017. Encoding, preparation and implementation of novel complex verbal instructions. Neuroimage. 148:264–273. doi:10.1016/j.neuroimage.2017.01.037.

González-García C, Mas-Herrero E, de Diego-Balaguer R, Ruz M. 2016. Task-specific preparatory neural activations in low-interference contexts. Brain Struct Funct. 221:3997–4006. doi:10.1007/s00429-015-1141-5.

Grecucci A, Giorgetta C, Bonini N, Sanfey AG. 2013. Reappraising social emotions: the role of inferior frontal gyrus, temporo-parietal junction and insula in interpersonal emotion regulation. Front Hum Neurosci. 7:523. doi:10.3389/fnhum.2013.00523.

Güth W, Schmittberger R, Schwarze B. 1982. An experimental analysis of ultimatum bargaining. J Econ Behav Organ. 3:367–388. doi:10.1016/0167-2681(82)90011-7.

Hajcak G, Moser JS, Holroyd CB, Simons RF. 2006. The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. Biol Psychol. 71:148–154. doi:10.1016/j.biopsycho.2005.04.001.

Hassabis D, Spreng RN, Rusu AA, Robbins CA, Mar RA, Schacter DL. 2014. Imagine all the people: How the brain creates and uses personality models to predict behavior. Cereb Cortex. 24:1979–1987. doi:10.1093/cercor/bht042.

Haxby J V., Connolly AC, Guntupalli JS. 2014. Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. Annu Rev Neurosci. 37:435–456. doi:10.1146/annurev-neuro-062012-170325.

Haynes J, Sakai K, Rees G, Gilbert S, Frith C, Passingham RE. 2007. Reading Hidden

Intentions in the Human Brain. Curr Biol. 17:323–328. doi:10.1016/j.cub.2006.11.072.

Hebart MN, Görgen K, Haynes J-D. 2015. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front Neuroinform. 8:88. doi:10.3389/fninf.2014.00088.

Jefferies E. 2013. The neural basis of semantic cognition: Converging evidence from neuropsychology, neuroimaging and TMS. Cortex. 49:611–625. doi:10.1016/j.cortex.2012.10.008.

Kahneman D, Knetsch JL, Thaler RH. 1986. Fairness and Assumptions of Economics. J Bus. 59:S285-300.

Kaplan JT, Man K, Greening SG. 2015. Multivariate cross-classification: applying machine learning techniques to characterize abstraction in neural representations. Front Hum Neurosci. 9:151. doi:10.3389/fnhum.2015.00151.

Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E. 2006. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science. 314:829–832. doi:10.1126/science.1129156.

Kok P, Jehee JFM, de Lange FP. 2012. Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. Neuron. 75:265–270. doi:10.1016/j.neuron.2012.04.034.

Koster-Hale J, Saxe R. 2013. Theory of Mind: A Neural Prediction Problem. Neuron. 79:836–848. doi:10.1016/j.neuron.2013.08.020.

Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. Proc Natl Acad Sci U S A. 103:3863–3868. doi:10.1073/pnas.0600244103.

Lesage E, Hansen PC, Miall RC. 2017. Right Lateral Cerebellum Represents Linguistic Predictability. J Neurosci. 37:6231–6241. doi:10.1523/JNEUROSCI.3203-16.2017.

Lindquist KA, Satpute AB, Wager TD, Weber J, Barrett LF. 2015. The Brain Basis of Positive and Negative Affect: Evidence from a Meta-Analysis of the Human Neuroimaging Literature. Cereb Cortex. 26:1910–1922. doi:10.1093/cercor/bhv001.

Lopez-Persem A, Domenech P, Pessiglione M. 2016. How prior preferences determine decision-making frames and biases in the human brain. Elife. 5: e20317. doi:10.7554/eLife.20317.

Ma N, Vandekerckhove M, Baetens K, Overwalle F Van, Seurinck R, Fias W. 2012. Inconsistencies in spontaneous and intentional trait inferences. Soc Cogn Affect Neurosci. 7:937–950. doi:10.1093/scan/nsr064.

De Martino B, Kumaran D, Seymour B, Dolan RJ. 2006. Frames, biases, and rational decision-making in the human brain. Science. 313:684–7. doi:10.1126/science.1128356.

Mendoza SA, Lane SP, Amido DM, Amodio DM. 2014. For Members Only: Ingroup Punishment of Fairness Norm Violations in the Ultimatum Game. Soc Psychol Personal Sci. 5:662–670. doi:10.1177/1948550614527115.

Menon V, Uddin LQ. 2010. Saliency, switching, attention and control: a network model of insula function. Brain Struct Funct. 214:655–667. doi:10.1007/s00429-010-0262-0.

Mitchell TM, Shinkareva S V., Carlson A, Chang K-M, Malave VL, Mason RA, Just MA. 2008. Predicting Human Brain Activity Associated with the Meanings of Nouns. Science. 320:1191–1195. doi:10.1126/science.1152876.

Morgan HM, Jackson MC, Van Koningsbruggen MG, Shapiro KL, Linden DEJ. 2013. Frontal and parietal theta burst TMS impairs working memory for visual-spatial conjunctions. Brain Stimul. 6:122–129. doi:10.1016/j.brs.2012.03.001.

Moser A, Gaertig C, Ruz M. 2014. Social information and personal interests modulate neural activity during economic decision-making. Front Hum Neurosci. 8: 31. doi:10.3389/fnhum.2014.00031.

Mueller C, Luehrs M, Baecke S, Adolf D, Luetzkendorf R, Luchtmann M, Bernarding J. 2012. Building virtual reality fMRI paradigms: A framework for presenting immersive virtual environments. J Neurosci Methods. 209:290–298. doi:10.1016/j.jneumeth.2012.06.025.

Naselaris T, Kay KN, Nishimoto S, Gallant JL. 2011. Encoding and decoding in fMRI. Neuroimage. 56:400–410. doi:10.1016/j.neuroimage.2010.07.073.

Nichols T, Brett M, Andersson J, Wager T, Poline JB. 2005. Valid conjunction inference with the minimum statistic. Neuroimage. 25:653–660. doi:10.1016/j.neuroimage.2004.12.005.

Oosterhof NN, Todorov A. 2008. The functional basis of face evaluation. Proc Natl Acad Sci U S A. 105:11087–92. doi:10.1073/pnas.0805664105.

Van Overwalle F. 2009. Social cognition and the brain: A meta-analysis. Hum Brain Mapp. 30:829–858. doi:10.1002/hbm.20547.

Van Overwalle F, Baetens K, Mariën P, Vandekerckhove M. 2014. Social cognition and the cerebellum: A meta-analysis of over 350 fMRI studies. Neuroimage. 86:554–572. doi:10.1016/j.neuroimage.2013.09.033.

Pereira F, Mitchell T, Botvinick M. 2009. Machine learning classi!ers and fMRI: A tutorial overview. Neuroimage. 45:S199–S209. doi:10.1016/j.neuroimage.2008.11.007.Machine.

Pleger B, Timmann D. 2018. The role of the human cerebellum in linguistic prediction, word generation and verbal working memory: evidence from brain imaging, non-invasive cerebellar stimulation and lesion studies. Neuropsychologia. doi:10.1016/j.neuropsychologia.2018.03.012.

Puri AM, Wojciulik E, Ranganath C. 2009. Category expectation modulates baseline and stimulus-evoked activity in human inferotemporal cortex. Brain Res. 1301:89–99. doi:10.1016/j.brainres.2009.08.085.

Redondo J, Fraga I, Padrón I, Comesaña M. 2007. The Spanish adaptation of anew (Affective Norms for English Words). Behav Res Methods. 39:600–605. doi:10.3758/BF03193031.

Reverberi C, Görgen K, Haynes J-D. 2012. Compositionality of Rule Representations in Human Prefrontal Cortex. Cereb Cortex. 22:1237–1246. doi:10.1093/cercor/bhr200.

Ruz M, Moser A, Webster K. 2011. Social expectations bias decision-making in uncertain inter-personal situations. PLoS One. 6(2): e157. doi:10.1371/journal.pone.0015762.

Ruz M, Tudela P. 2011. Emotional conflict in interpersonal interactions. Neuroimage. 54:1685–1691. doi:10.1016/j.neuroimage.2010.08.039.

Sakai K, Passingham RE. 2006. Prefrontal Set Activity Predicts Rule-Specific Neural Processing during Subsequent Cognitive Performance. J Neurosci. 26:1211–1218. doi:10.1523/JNEUROSCI.3887-05.2006.

Sala JB, Rämä P, Courtney SM. 2003. Functional topography of a distributed neural system for spatial and nonspatial information maintenance in working memory. Neuropsychologia. 41:341–356. doi:10.1016/S0028-3932(02)00166-5.

Samuelson W, Zeckhauser R. 1988. Status quo bias in decision making. J Risk Uncertain. 1:7–59. doi:10.1007/BF00055564.

Sanfey AG. 2009. Expectations and social decision-making: Biasing effects of prior knowledge on Ultimatum responses. Mind Soc. 8:93–107. doi:10.1007/s11299-009-0053-6.

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. doi:10.1126/science.1082976.

Saxe R, Kanwisher N. 2003. People thinking about thinking peopleThe role of the temporo-parietal junction in “theory of mind.” Neuroimage. 19:1835–1842. doi:10.1016/S1053-8119(03)00230-1.

Schneider W, Eschman A, Zuccolotto A. 2002. E-Prime user’s guide.

Scholz J, Triantafyllou C, Whitfield-Gabrieli S, Brown EN, Saxe R. 2009. Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. PLoS One. 4: e4869. doi:10.1371/journal.pone.0004869.

Seeley WW, Menon V, Schatzberg AF, Keller J, Glover GH, Kenna H, Reiss AL, Greicius MD. 2007. Dissociable Intrinsic Connectivity Networks for Salience Processing and

Executive Control. J Neurosci. 27:2349–2356. doi:10.1523/JNEUROSCI.5587-06.2007.

Sokolov AA, Miall RC, Ivry RB. 2017. The Cerebellum: Adaptive Prediction for Movement and Cognition. Trends Cogn Sci. 21:313–332. doi:10.1016/j.tics.2017.02.005.

Sridharan D, Levitin DJ, Menon V. 2008. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. Proc Natl Acad Sci. 105:12569–12574. doi:10.1073/pnas.0800005105.

Stelzer J, Chen Y, Turner R. 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. Neuroimage. 65:69–82. doi:10.1016/j.neuroimage.2012.09.063.

Stolier RM, Freeman JB. 2016. Neural pattern similarity reveals the inherent intersection of social categories. Nat Neurosci. 19:795–797. doi:10.1038/nn.4296.

Stolier RM, Freeman JB. 2017. A Neural Mechanism of Social Categorization. J Neurosci. 37:5711–5721. doi:10.1523/JNEUROSCI.3334-16.2017.

Stolier RM, Hehman E, Freeman JB. 2018. A Dynamic Structure of Social Trait Space. Trends Cogn Sci. 22:197–200. doi:10.1016/j.tics.2017.12.003.

Summerfield C, De Lange FP. 2014. Expectation in perceptual decision making: Neural and computational mechanisms. Nat Rev Neurosci. 15:745–756. doi:10.1038/nrn3838.

Tamir DI, Thornton MA. 2018. Modeling the Predictive Social Mind. Trends Cogn Sci. 22:201–212. doi:10.1016/j.tics.2017.12.005.

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 U S A. 113:194–199. doi:10.1073/pnas.1511905112.

Thornton MA, Mitchell JP. 2017. Theories of Person Perception Predict Patterns of Neural Activity During Mentalizing. Cereb Cortex.:1–16. doi:10.1093/cercor/bhx216.

Thye MD, Murdaugh DL, Kana RK. 2018. Brain Mechanisms Underlying Reading the Mind from Eyes, Voice, and Actions. Neuroscience. 374:172–186.

doi:10.1016/j.neuroscience.2018.01.045.

Tong F, Pratte MS. 2012. Decoding Patterns of Human Brain Activity. Annu Rev Psychol. 63:483–509. doi:10.1146/annurev-psych-120710-100412.

Tortosa M, Lupiáñez J, Ruz M. 2013. Race, emotion and trust: An ERP study. Brain Res. 1494:44–55. doi:10.1016/j.biopsycho.2008.03.004.

Turner B. 2010. Comparison of methods for the use of pattern classificaion on rapid event-related fMRI data. Poster session presented at the Annual Meeting of the Society for Neuroscience, San Diego, CA.

Vogt BA. 2016. Midcingulate cortex: Structure, connections, homologies, functions and diseases. J Chem Neuroanat. 74:28–46. doi:10.1016/j.jchemneu.2016.01.010.

Wagner AD, Paré-Blagoev EJ, Clark J, Poldrack RA. 2001. Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval. Neuron. 31:329–338. doi:10.1016/S0896-6273(01)00359-2.

Yeung N, Sanfey AG. 2004. Independent Coding of Reward Magnitude and Valence in the Human Brain. J Neurosci. 24:6258–6264. doi:10.1523/JNEUROSCI.4537-03.2004.

Figure captions

Figure 1: A. Schematic display of the paradigm. **B.** Sequence of events in a trial. The task varied the Valence of the partner's information (Positive, Negative, No information) and the Fairness of the offer (Fair/Unfair).

Figure 2: Acceptance Rates (AR) and reaction times (RT) to fair and unfair offers preceded by positive, negative and neutral descriptions of the partner (error bars represent S.E.M).

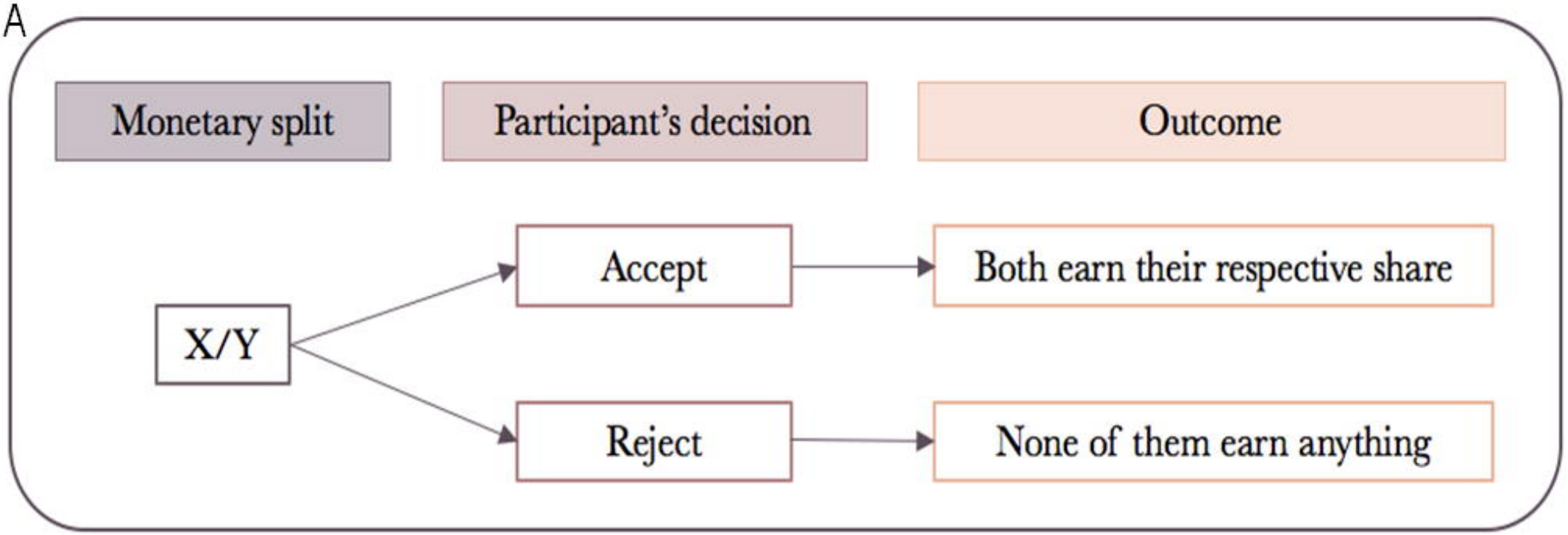
Figure 3: Univariate results during the expectation period. Scales reflect peaks of significant t-values ($p<.05$, FWE-corrected for multiple comparisons).

Figure 4: Univariate results for the offer. Scales reflect peaks of significant t-values ($p<.05$, FWE-corrected for multiple comparisons).

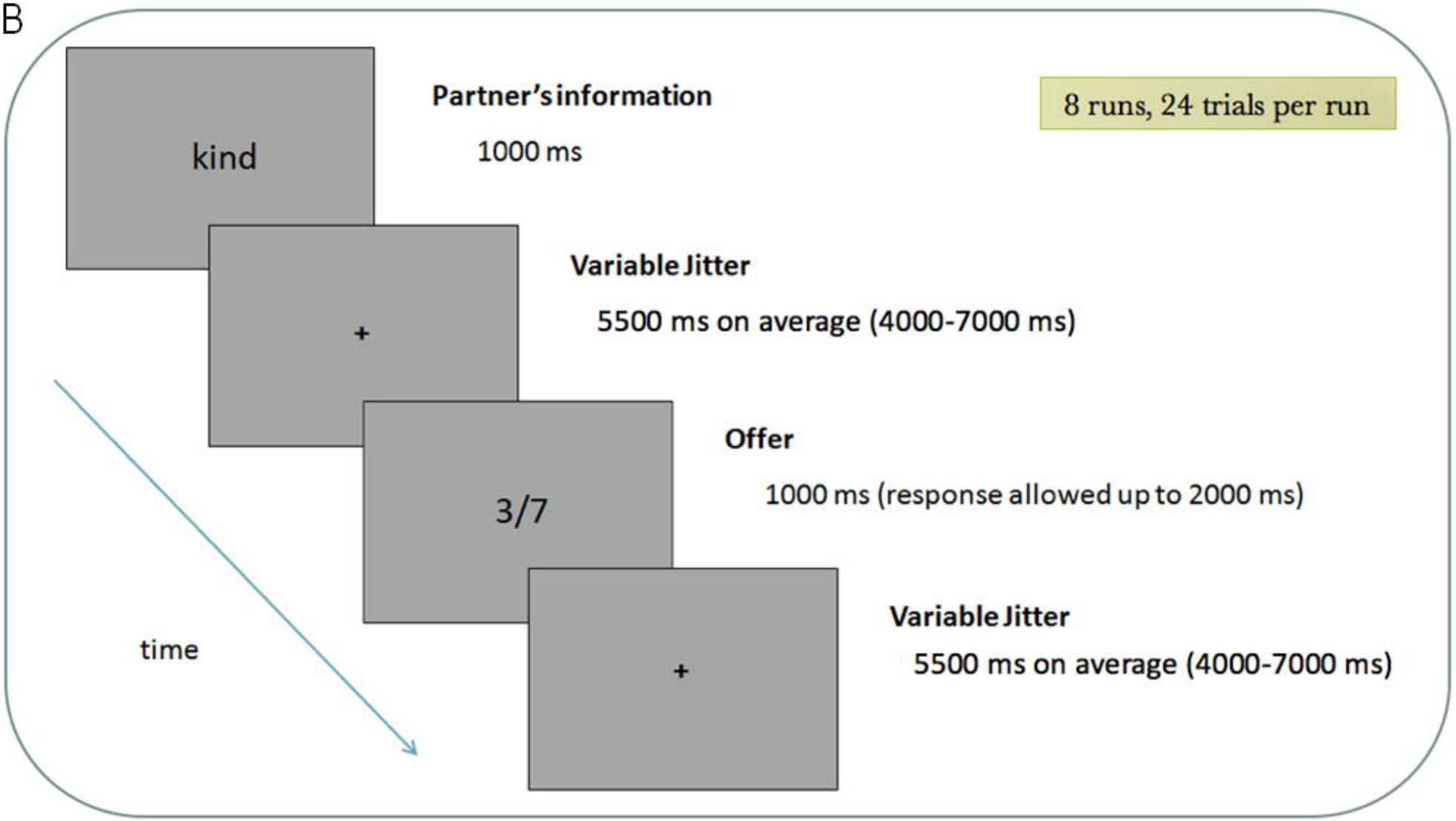
Figure 5: Multivariate results (violet). Different neural patterns for the valence (positive vs. negative) of the adjective during the expectation stage. Scales reflect corrected p-values (<.05). Regions significantly active both during univariate and multivariate analyses are highlighted in yellow.

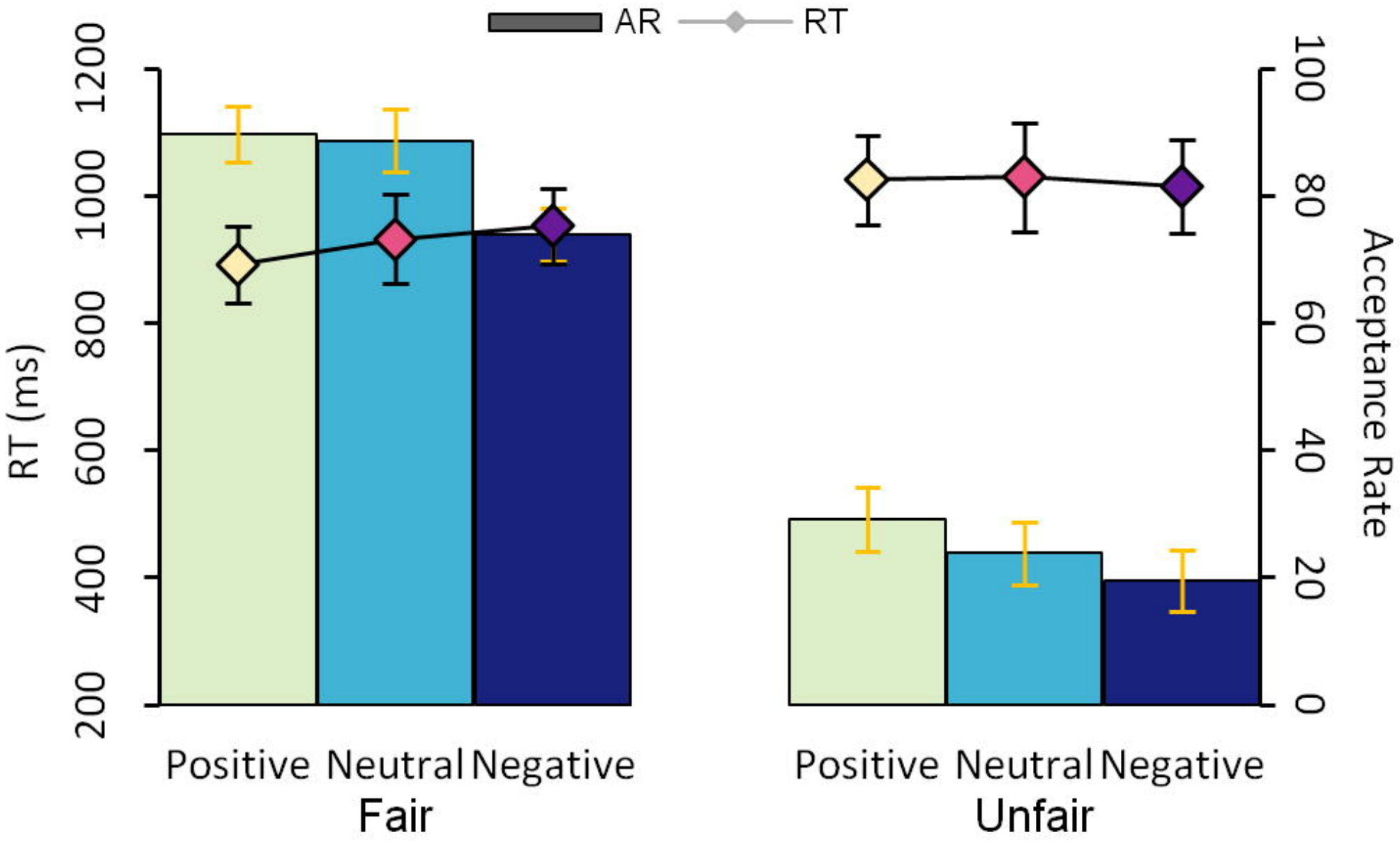
Figure 6: Scatter plots showing significant correlations between mean decoding accuracies in each cluster and the behavioural index. IFG: Inferior frontal gyrus. MFG: Middle frontal gyrus. ACC: Anterior Cingulate Cortex. MCC: Middle Cingulate Cortex. SMA: Supplementary Motor Area.

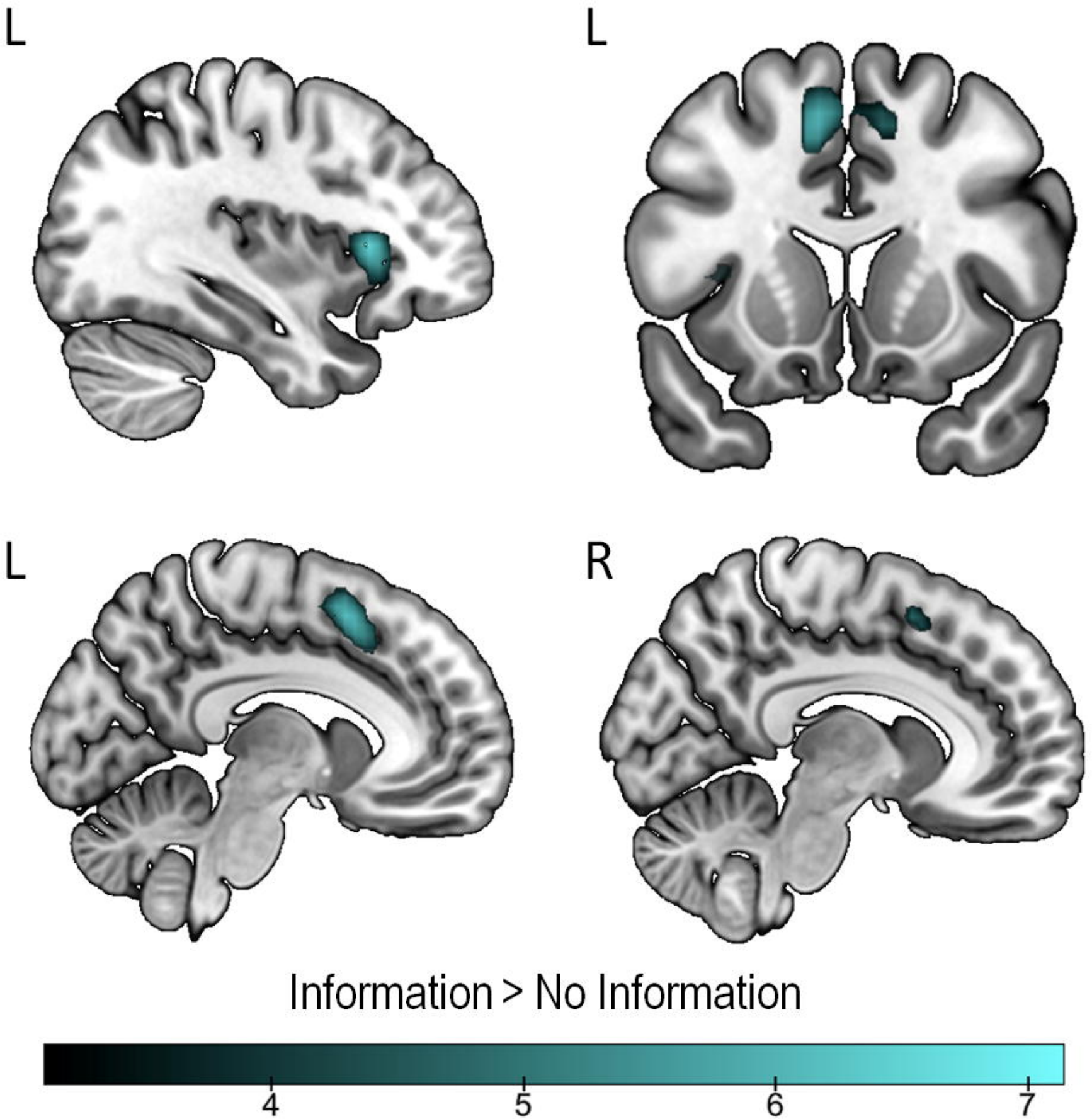
A

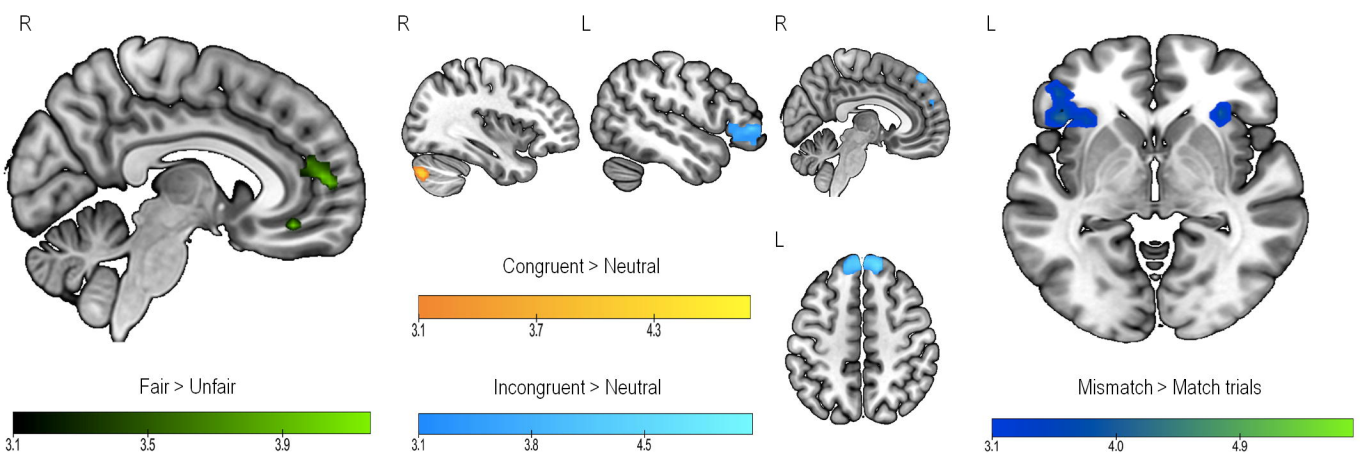

Monetary split
Participant's decision
Outcome
X/Y
Accept
Reject
Both earn their respective share
None of them earn anything

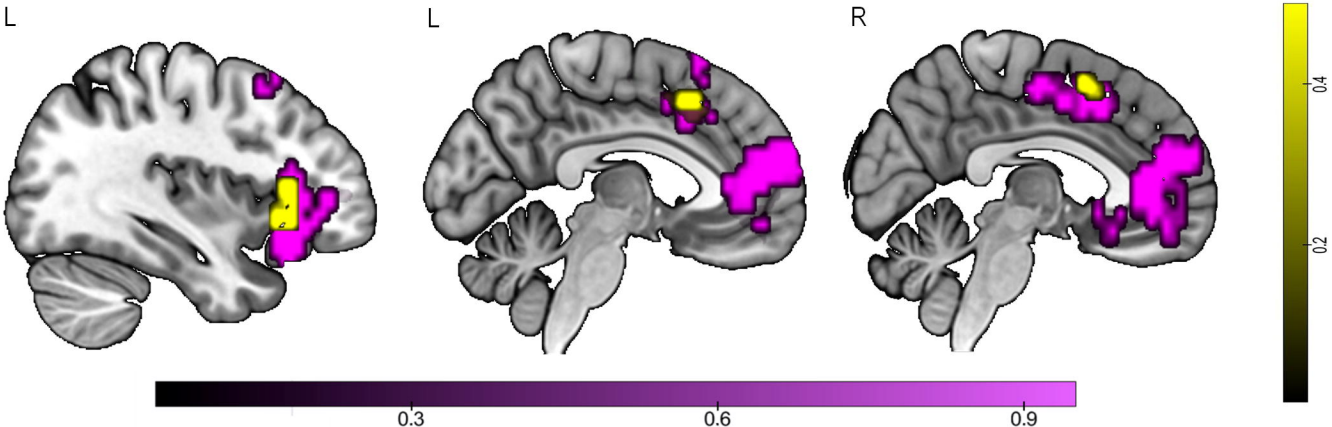
B

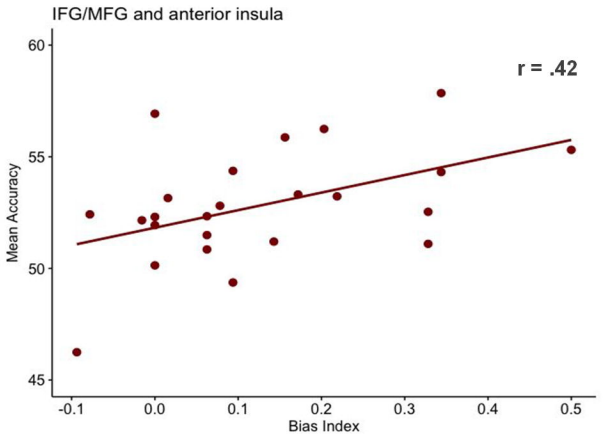

kind
Partner's information
1000 ms
8 runs, 24 trials per run
+
Variable Jitter
5500 ms on average (4000-7000 ms)
3/7
Offer
1000 ms (response allowed up to 2000 ms)
time
+
Variable Jitter
5500 ms on average (4000-7000 ms)

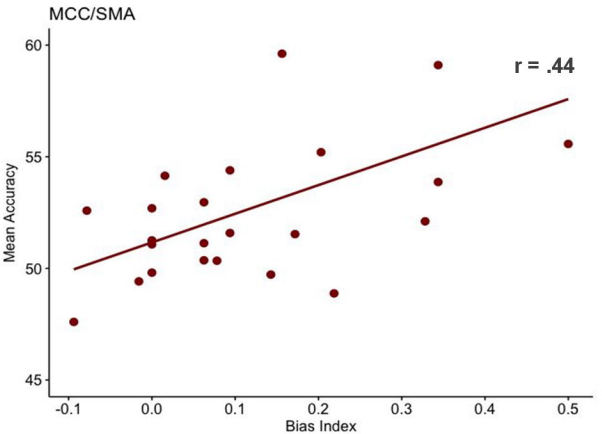

AR
RT
RT (ms)
1200
1000
800
600
400
200
Acceptance Rate
100
80
60
40
20
0
Positive
Neutral
Negative
Fair
Positive
Neutral
Negative
Unfair

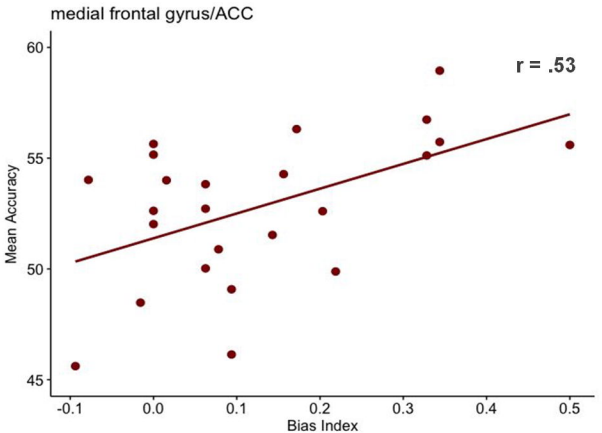

L
L
L
R
Information > No Information
4
5
6
7


R
R
L
R
L
L
Congruent > Neutral
3.1
3.7
4.3
Fair > Unfair
3.1
3.5
3.9
Incongruent > Neutral
3.1
3.8
4.5
Mismatch > Match trials
3.1
4.0
4.9


L
L
R
0.4
0.2
0.3
0.6
0.9


IFG/MFG and anterior insula
r = .42
Mean Accuracy
Bias Index
60
55
50
45
-0.1
0.0
0.1
0.2
0.3
0.4
0.5


MCC/SMA
r = .44
Mean Accuracy
Bias Index
60
55
50
45
-0.1
0.0
0.1
0.2
0.3
0.4
0.5


medial frontal gyrus/ACC
r = .53
Mean Accuracy
Bias Index
60
55
50
45
-0.1
0.0
0.1
0.2
0.3
0.4
0.5