Excerpt from:

Shared States: Using MVPA to test neural overlap between self-focused emotion imagery and other-focused emotion understanding.

Suzanne Oosterwijk[1,3*], Lukas Snoek[2*], Mark Rotteveel[1,3], Lisa F. Barrett[4], & H. Steven Scholte[2,3]

[*] These authors contributed equally to this work.

Method

Subjects

In total, we tested twenty-two Dutch undergraduate students from the University of Amsterdam (14 females; M_{age} = 21.48, SD_{age} = 1.75). Of those twenty-two subjects, thirteen subjects were tested twice in two sessions about one week apart. Half of those sessions were used for the model optimization-procedure. The other half of the sessions, combined with an additional nine subjects (who were tested only once), constituted the model validation set (see Analysis and model optimization procedure section). In total, two subjects were excluded from the model validation dataset: one subject was excluded because there was not enough time to complete the experimental protocol and another subject was excluded due to excessive movement (> 3 mm within data acquisition runs).

All subjects signed informed consent prior to the experiment. The experiment was approved by the University of Amsterdam's ethical review board. Subjects received 22.50 euro per session. Standard exclusion criteria regarding MRI safety were applied and people who were on psychopharmacological medication were excluded a priori.

Experimental design

Self-focused emotion imagery task. The self-focused emotion imagery task (SF-task) was created to preferentially elicit *self-focused* processing of action, interoceptive or situational information associated with emotion. Subjects processed short linguistic cues that described actions (e.g., "*pushing* *someone away*"; "*making a fist*"), interoceptive sensations (e.g., "*being out of breath*"; "*an increased heart rate*"), or situations (e.g., "*alone in a park at night*"; "*being falsely accused*") and were instructed to imagine performing or experiencing the content. The complete instruction is presented in the Supplementary Materials; all stimuli used in the SF-task are presented in Supplementary Table 1. Linguistic cues were selected from a pilot study performed on an independent sample of subjects (n = 24). Details about this pilot study are available on request. The descriptions generated in this pilot study were used as qualitative input to create short sentences that described actions, sensations or situations that were associated with negative emotions, without including discrete emotion terms. The cues did not differ in number of words, nor in number of characters ($F < 1$).

The SF-task was performed in two runs subsequent to the other-focused task using the software package Presentation (Version 16.4, www.neurobs.com). Each run presented sixty sentences on a black background (20 per condition) in a fully-randomized event-related fashion, with a different randomization for each subject. Note that implementing a separate randomization for each subject prevents inflated false positive pattern correlations between trials of the same condition, which may occur in single-trial designs with short inter-stimulus intervals (Mumford, Davis, & Poldrack, 2014). A fixed inter-trial-interval of two seconds separated trials; twelve null-trials (i.e. a black screen for eight seconds) were mixed with the experimental trials at random positions during each run (see Figure 1).

Other-focused emotion understanding task. The other-focused emotion understanding task (OF-task) was created to preferentially elicit *other-focused* processing of action, interoceptive or situational information associated with emotion. Subjects viewed images of people in negative situations (e.g., a woman screaming at a man, a man held at gunpoint). A red rectangle highlighted the face of the person that the subjects should focus on to avoid ambiguity in images depicting more than one person. Image blocks were preceded by a cue indicating the strategy subjects should use in perceiving the emotional state of the people in the images (Spunt & Lieberman, 2011). The cue "*How*" instructed the subjects to identify actions that were informative about the person's emotional state (i.e., *How* does this person express his/her emotions?). The cue "*What*" instructed subjects to identify interoceptive sensations that the person could experience (i.e., *What* does this person feel in his/her body). The cue "*Why*" instructed subjects to identify reasons or explanations for the person's emotional state (i.e., *Why* does this person feel an emotion?). The complete instruction is presented in the Supplementary Materials

Stimuli for the OF-task were selected from the International Affective Picture System database (IAPS; Lang, Bradley, & Cuthbert, 2008), the image set developed by the Kveraga lab (http://www.kveragalab.org/stimuli.html; Kveraga, Boshyan, Adams, Mote, Betz, Ward, Hadjikhani, Bar & Barret, 2015) and the internet (Google images). We selected images based on a pilot study, performed on an independent sample of subjects (n = 22). Details about this pilot study are available on request.

The OF-task was presented using the software package Presentation. The task presented thirty images on a black background in blocked fashion, with each block starting with a what/why or how cue (see Figure 1). Each image was shown three times, once for each cue type. Images were presented in blocks of six, each lasting six seconds, followed by a fixed inter trial interval of two seconds. Null-trials were inserted at random positions within the blocks. Both the order of the blocks and the specific stimuli within and across blocks were fully randomized, with a different randomization for each subject.

Procedure

Each experimental session lasted about two hours. Subjects who underwent two sessions had them on different days within a time-span of one week. On arrival, subjects gave informed consent and received thorough task instructions, including practice trials (see the Supplementary Materials for a translation of the task instructions). The actual time in the scanner was 55 minutes, and included a rough 3D scout image, shimming sequence, three-minute structural T1-weighted scan, one functional run for the OF-task and two functional runs for the SF-task. We deliberately chose to present the SF-task after the OF-task to exclude the possibility that the SF-task affected the OF-task, thereby influencing the success of the decoding procedure.

After each scanning session, subjects rated their success rate for the SF-task and OF-task (see Supplementary Figure 1). In session 2, subjects filled out three personality questionnaires that will not be further discussed in this paper and were debriefed about the purpose of the study.

Image acquisition

Subjects were tested using a Philips Achieva 3T MRI scanner and a 32-channel SENSE headcoil. A survey scan was made for spatial planning of the subsequent scans. Following the survey scan, a 3-minute structural T1-weighted scan was acquired using 3D fast field echo (TR: 82 ms, TE: 38 ms, flip angle: 8°, FOV: 240 × 188 mm, 220 slices acquired using single-shot ascending slice order and a voxel size of 1.0 × 1.0 × 1.0 mm). After the T1-weighted scan, functional T2* weighted sequences were acquired using single shot gradient echo, echo planar imaging (TR=2000 ms, TE=27.63 ms, flip angle: 76.1°, FOV: 240 × 240 mm, in-plane resolution 64 x 64, 37 slices (with ascending acquisition), slice thickness 3 mm, slice gap 0.3 mm, voxel size 3 × 3 × 3 mm), covering the entire brain. For the SF-task, 301 volumes were acquired; for the OF-task 523 volumes were acquired.

Analysis and model optimization procedure

As MVPA is a fairly novel technique, no consistent, optimal analysis pipeline has been established (Etzel, Valchev, & Keysers, 2011). Therefore we adopted a validation strategy in the present study that is advised in the pattern classification field (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Kay, Naselaris, Prenger, & Gallant, 2008). We generated an optimization and validation data-set by running the SF-task and OF-task twice, in two identical experimental sessions for a set of thirteen subjects. The sessions were equally split between the optimization and validation set; first and second sessions were counterbalanced between the two sets. Based on a request received during the review process, we added an additional nine subjects to the validation data-set. Ultimately, the optimization-set held thirteen sessions and the validation-set, after exclusion of two subjects (see Subjects section), held twenty sessions.

In the optimization set, we explored various preprocessing parameters (i.e. smoothing kernel, low-pass filter, and ICA-based denoising strategies) and MVPA hyperparameter values (i.e. univariate feature selection thresholds and train/test size ratio during cross-validation) to find the optimal combination of hyperparameters and preprocessing settings in terms of classification performance. We measured classification performance by using repeated random subsampling cross-validation with 1000 iterations within the optimization-set (see Multi-voxel pattern analysis subsection for details); we determined the combination of optimal preprocessing and model hyperparameters based on mean classification performance (see Supplementary Table 2 and Supplementary Figure 2). The optimal parameters were then applied to the validation-set, in which the findings from the optimization-set were replicated. All findings discussed in the results section follow from the validation-set (see Supplementary Figure 3 for an overview of the findings from the optimization-set).

Preprocessing and single-trial modeling

Functional and structural data were preprocessed and analyzed using FSL 5.0 (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012) and MATLAB (2012b; www.mathworks.com/products/matlab), using an in-house developed preprocessing pipeline and the parameters established in the optimization procedure. Functional data was corrected for motion (using FSL MCFLIRT) and slice timing and was spatially smoothed (5 mm isotropic kernel). After preprocessing, individual time series were modeled using a double gamma hemodynamic response function in a single-trial GLM design using FSL's FEAT. Resulting beta-values were converted to *t*-values (Misaki, Kim, Bandettini, & Kriegeskorte, 2010), constituting a whole-brain pattern of *t*-values per trial. Subsequently, the data was indexed by a gray-matter mask (excluding most white-matter, CSF, and brainstem voxels). Thus, the data points for the multi-voxel pattern analysis consist of whole-brain (gray matter) *t*-value patterns per trial. For the optimization analyses, the data was transformed to standard space (MNI152, 2 mm), but to reduce computation time for the validation data, and in particular its corresponding permutation analysis, analyses on the validation dataset were performed on data in native (functional) space.

Multi-voxel pattern analysis

MVPA pipeline. Before fitting the classifier on the train-data in each iteration of the cross-validation scheme (which is described in more detail in the next section), standardization and voxel selection were estimated and applied to the train-set. Standardization ensured that each feature (i.e. voxel) had zero mean and unit variance across trials. After standardization, voxel selection was performed in each iteration on the train-trials by extracting the voxels with the highest average pairwise Euclidian distance across classes, which will be subsequently referred to as a voxel's differentiation score. More specifically, differentiation scores were calculated by subtracting the mean value across trials per class from each other (i.e. action – interoception, action – situation, interoception – situation), normalizing these values across voxels (yielding 'z-scores'), and taking their absolute value. The three resulting values per voxel were averaged and the most differentiating voxels (*z*-score threshold: 2.3, as determined by the optimization-procedure) were extracted and used as features when fitting the classifier. Importantly, the standardization parameters (voxel mean and variance) and voxel indices (i.e. which voxels had differentiation scores above threshold) were estimated from the train-data only and subsequently applied to the test-data to ensure independence between the train- and test-set. After standardization and voxel selection in each iteration, a support vector classifier was fit on the training data and cross-validated on the test-set, generating a class probability for each trial in the test-set (see Supplementary Figure 4A for a schematic overview of the MVPA pipeline fitting and cross-validation). Our classifier of choice was the support vector classifier (SVC) implementation from the scikit-learn 'svm' module (Pedregosa et al., 2011) with a linear kernel, fixed regularization parameter ('C') of 1.0, one-vs-one multiclass strategy, estimation of class probability output (instead of discrete class prediction), and otherwise default parameters.

Cross-validation scheme and bagging procedure. Cross-validation of the classification analysis was implemented using a repeated random subsampling cross-validation paradigm (also known as Monte Carlo cross-validation), meaning that, for each iteration of the analysis, the classification pipeline (i.e. standardization, voxel selection, and SVM fitting) was applied on a random subset of data points (i.e. the train-set) and cross-validated on the remaining data (i.e. the test-set). Each trial belonged to one out of three classes: action, interoception, or situation. Following the results from the parameter optimization process, we selected four trials per class for testing, amounting to twelve test-trials per iteration.

Per iteration, the classification pipeline was fit on the train-trials from the SF-data. Subsequently, this classifier was cross-validated on twelve test SF-trials ('self-analysis') and twelve test OF-trials ('cross-analysis'; see Figure 2). This process was subsequently iterated 100,000 times to generate a set of class distributions for each trial. After all iterations, the final predicted class of each trial was determined by its highest summed class probability across iterations (also known as 'soft voting'; see Supplementary Figure 4B). This strategy of a random sub-sampling cross-validation scheme in combination with majority (soft) voting is more commonly known as 'bagging' (Breiman, 1996). An important advantage of bagging is that it reduces model overfitting by averaging over an ensemble of models, which is especially useful for multi-voxel pattern analyses because fMRI data is known to display high variance (Varoquaux, Raamana, Engemann, Hoyos-Idrobo, Schwartz, & Thirion, preprint).

After generating a final prediction for all trials using this soft voting method, we constructed confusion matrices for both the self- and cross-analysis. In each raw confusion matrix with prediction counts per class, cells were normalized by dividing prediction counts by the sum over rows (i.e. the total amount of predictions per class), yielding precision-scores (also known as Positive Predictive Value). In other words, this metric represents the ratio of true positives to the sum of true positives and false positives (see Supplementary Figure 5 for a description of the results expressed as *recall* estimates, or the ratio of true positives to the total number of samples in that class). This classification pipeline generated subject-specific confusion matrices that were subsequently averaged to generate the final classification scores.

Statistical evaluation. To evaluate the statistical significance of the observed average precision-scores in the confusion matrices, we permuted the original self- and cross-analysis 1300 times per subject with randomly shuffled class labels, yielding 1300 confusion matrices (with precision-scores). We then averaged the confusion matrices across subjects, yielding 1300 permuted confusion matrices reflecting the null-distribution of each cell of the matrix (which is centered around chance level classification, i.e. 33%). For each cell in the diagonal of the observed confusion matrix, *p*-values were calculated as the proportion of instances of values in the *permuted* matrix which were higher than the values in the *observed* matrix (Nichols & Holmes, 2001). To correct for multiple comparisons, *p*-values were tested against a Bonferroni-corrected threshold. The distribution of precision-scores and the relationship between precision-scores in the self- and cross-analysis is reported in Supplementary Figure 6.

Spatial representation. To visualize the classifier feature weights, we plotted the absolute feature weights averaged over iterations, subjects, and pairwise classifiers (action vs. interoception, action vs. situation, interoception vs. situation) that underlie our multiclass classification analysis. We chose to visualize the spatial representation of our model by plotting the average absolute feature weights, because the absolute value of feature weights in linear SVMs can be interpreted as how important the weights are in constructing the model's decision hyperplane (Guyon, Weston, Barnhill, & Vapnik, 2002; Ethofer, Van De Ville, Scherer, & Vuilleumier, 2009; Stelzer, Buschmann, Lohmann, Margulies, Trampel, & Turner, 2014). To correct for a positive bias in plotting *absolute* weights, we ran the main classification analysis again with permuted labels to extract the average absolute feature weights that one would expect by chance. Subsequently, a voxel-wise independent *t*-test was performed for all feature weights across subjects, using the average permuted feature weights as the nullhypothesis, yielding an interpretable *t*-value map (see the supplementary code notebook on our Github repository for computational details).