TOWARDS PREDICTION

Studing the mind and brain in
the age of machine learning

Lukas Snoek

Towards prediction

This thesis was typeset using (R) Markdown, LaTeX and the `bookdown` R-package
ISBN: xxx-xx-xxxx-xxx-x
Printing: Acme Press, Inc.

Towards prediction

Studying the mind and brain in the age of machine learning

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op maandag 21 oktober 2021, te 14 uur

door

Lukas Snoek

geboren te Hoevelaken

Promotiecommissie:

Promotor:	dr. H.S. Scholte	Universiteit van Amsterdam
Copromotor:	dr. S. Oosterwijk	Universiteit van Amsterdam
Overige leden:	prof. dr. R.E. Jack	University of Glasgow
	prof. dr. R.W. Goebel	Maastricht University
	prof. dr. B.U. Forstmann	Universiteit van Amsterdam
	prof. dr. A.H. Fischer	Universiteit van Amsterdam
	prof. dr. D. Borsboom	Universiteit van Amsterdam
	prof. dr. A.G. Sanfey	Radboud University

Faculteit: Faculteit der Maatschappij- en Gedragswetenschappen

Contents

Chapter 1

Introduction

The first chapter of the thesis, which introduces your PhD project. The filler-text below was created with the postmodernism generator[1].

Something about the coming about of this thesis. More of a "lessons learned" rather than a coherent research topic.

1.1 The brain is not a dictionary

Something about looking at populations of neurons/voxels/areas rather than simple one-to-one relationships. Shared states.

1.2 The brain (probably) does not care about your hypothesis

Facial expression models.

[1] http://www.elsewhere.org/journal/pomo

1.3 Interpretability and prediction are a trade-off (for now)

A plea for prediction but a cautionary tale for interpreting predictive models (confounds)

1.4 Exploration should be embraced more

Something about the "context of discovery" (cf. TCM), preregistration, and confirmation vs. exploration (Morbid curiosity.)

1.5 Proper generalization is hard

Within and between subject variance is not noise, but unexplained variance (AU limitations).

1.6 Psychology is complex, so it needs complex models

Which need to be fit on complex, large datasets. (AOMIC)

Chapter 2

Shared states: using MVPA to test neural overlap between self-focused emotion imagery and other-focused emotion understanding

---

This chapter has been published as: Oosterwijk, S.*, Snoek, L.*, Rotteveel, M., Barrett, L. F., & Scholte, H. S. (2017). Shared states: using MVPA to test neural overlap between self-focused emotion imagery and other-focused emotion understanding. *Social cognitive and affective neuroscience, 12*(7), 1025-1035.

* Shared first authorship

Abstract

The present study tested whether the neural patterns that support imagining "performing an action", "feeling a bodily sensation" or "being in a situation" are directly involved in understanding *other people's* actions, bodily sensations and situations. Subjects imagined the content of short sentences describing emotional actions, interoceptive sensations and situations (self-focused task), and processed scenes and focused on *how* the target person was expressing an emotion, *what* this person was feeling, and *why* this person was feeling an emotion (other-focused task). Using a linear support vector machine classifier on brain-wide multi-voxel patterns, we accurately decoded each individual class in the self-focused task. When generalizing the classifier from the self-focused task to the other-focused task, we also accurately decoded whether subjects focused on the emotional actions, interoceptive sensations and situations of *others*. These results show that the neural patterns that underlie self-imagined experience are involved in understanding the experience of other people. This supports the theoretical assumption that the basic components of emotion experience and understanding share resources in the brain.

2.1 Introduction

To navigate the social world successfully it is crucial to understand other people. But how do people generate meaningful representations of other people's actions, sensations, thoughts and emotions? The dominant view assumes that representations of other people's experiences are supported by the same neural systems as those that are involved in generating experience in the self (e.g., Gallese et al., 2004; see for an overview Singer, 2012). We tested this principle of self-other neural overlap directly, using multi-voxel pattern analysis (MVPA), across three different

aspects of experience that are central to emotions: actions, sensations from the body and situational knowledge.

In recent years, evidence has accumulated that suggests a similarity between the neural patterns representing the self and others. For example, a great variety of studies have shown that observing actions and sensations in other people engages similar neural circuits as acting and feeling in the self (see for an overview Bastiaansen et al., 2009). Moreover, an extensive research program on pain has demonstrated an overlap between the experience of physical pain and the observation of pain in other people, utilizing both neuroimaging techniques (e.g., Lamm et al., 2011) and analgesic interventions (e.g., Rütgen et al., 2015; Mischkowski et al., 2016). This process of "vicarious experience" or "simulation" is viewed as an important component of empathy (Carr et al., 2003; Decety, 2011; Keysers & Gazzola, 2014). In addition, it is argued that mentalizing (e.g. understanding the mental states of other people) involves the same brain networks as those involved in self-generated thoughts (Uddin et al., 2007; Waytz & Mitchell, 2011). Specifying this idea further, a constructionist view on emotion proposes that both emotion experience and interpersonal emotion understanding are produced by the same large-scale distributed brain networks that support the processing of sensorimotor, interoceptive and situationally relevant information (Barrett & Satpute, 2013; Oosterwijk & Barrett, 2014). An implication of these views is that the representation of self- and other-focused emotional actions, interoceptive sensations and situations overlap in the brain.

Although there is experimental and theoretical support for the idea of self-other neural overlap, the present study is the first to directly test this process using MVPA across three different aspects of experience (i.e. actions, interoceptive sensations and situational knowledge). Our experimental design consisted of two

different tasks aimed at generating self- and other-focused representations with a relatively large weight given to either action information, interoceptive information or situational information.

In the *self-focused* emotion imagery task (SF-task) subjects imagined performing or experiencing actions (e.g., *pushing someone away*), interoceptive sensations (e.g., *increased heart rate*) and situations (e.g., *alone in a park at night*) associated with emotion. Previous research has demonstrated that processing linguistic descriptions of (emotional) actions and feeling states can result in neural patterns of activation associated with, respectively, the representation and generation of actions and internal states (Oosterwijk et al., 2015; Pulvermüller & Fadiga, 2010). Furthermore, imagery-based inductions of emotion have been successfully used in the MRI scanner before (Oosterwijk et al., 2012; Wilson-Mendenhall et al., 2011), and are seen as robust inducers of emotional experience (Lench et al., 2011). In the *other-focused* emotion understanding task (OF-task), subjects viewed images of people in emotional situations and focused on actions (i.e., *How* does this person express his/her emotions?), interoceptive sensations (i.e., *What* does this person feel in his/her body) or the situation (i.e., *Why* does this person feel an emotion?). This task is based on previous research studying the neural basis of emotion oriented mentalizing (Spunt & Lieberman, 2012).

With MVPA, we examined to what extent the SF- and OF-task evoked similar neural patterns. MVPA allows researchers to assess whether the neural pattern associated with one set of experimental conditions can be used to distinguish between another set of experimental conditions. This relatively novel technique has been successfully applied to the field of social neuroscience in general (e.g., Gilbert et al., 2012; Brosch et al., 2013; Parkinson et al., 2014), and the field of self-other neural overlap in particular. For example, several MVPA studies recently assessed

whether experiencing pain and observing pain in others involved similar neural patterns (Corradi-Dell'Acqua et al., 2016; Krishnan et al., 2016). Although there is an ongoing discussion about the specifics of shared representation in pain based on these MVPA results (see for an overview Zaki et al., 2016), many authors emphasize the importance of this technique in the scientific study of self-other neural overlap (e.g., Corradi-Dell'Acqua et al., 2016; Krishnan et al., 2016).

MVPA is an analysis technique that decodes latent categories from fMRI data in terms of multi-voxel patterns of activity (Norman et al., 2006a). This technique is particularly suited for our research question for several reasons. First of all, although univariate techniques can demonstrate that tasks activate the same brain regions, only MVPA can statistically test for shared representation (Lamm & Majdandžić, 2015). We will evaluate whether multivariate brain patterns that distinguish between mental events in the SF-task can be used to distinguish, above chance level, between mental events in the OF-task. Second, MVPA analyses are particularly useful in research that is aimed at examining distributed representations (Singer, 2012). Based on our constructionist framework, we indeed hypothesize that the neural patterns that will represent self- and other focused mental events are distributed across large-scale brain networks. To capture these distributed patterns, we used MVPA in combination with data-driven univariate feature selection on whole-brain voxel patterns, instead of limiting our analysis to specific regions-of-interest (Haynes, 2015). And third, in contrast to univariate analyses that aggregate data across subjects, MVPA can be performed within-subjects and is thus able to incorporate individual variation in the representational content of multivariate brain patterns. In that aspect within-subject MVPA is sensitive to individual differences in how people imagine actions, sensations and situations, and how they understand others. In short, for our purpose to explicitly test the assumption that self

and other focused processes share neural resources, MVPA is the designated method.

We tested the following two hypotheses. First, we tested whether we could classify *self-imagined* actions, interoceptive sensations and situations above chance level. Second, we tested whether the multivariate pattern underlying this classification could also be used to classify the how, what and why condition in the *other-focused* task.

2.2 Methods

Subjects

In total, we tested 22 Dutch undergraduate students from the University of Amsterdam (14 females; M_{age} = 21.48, $s.d._{age}$ = 1.75). Of those 22 subjects, 13 subjects were tested twice in 2 sessions about 1 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 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 A.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 60 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 et al., 2014). A fixed inter-trial–interval of 2 seconds separating trials; 12 null-trials (i.e. a black screen for 8 seconds) were mixed with the experimental trials at random positions during each run (see Figure 2.1).

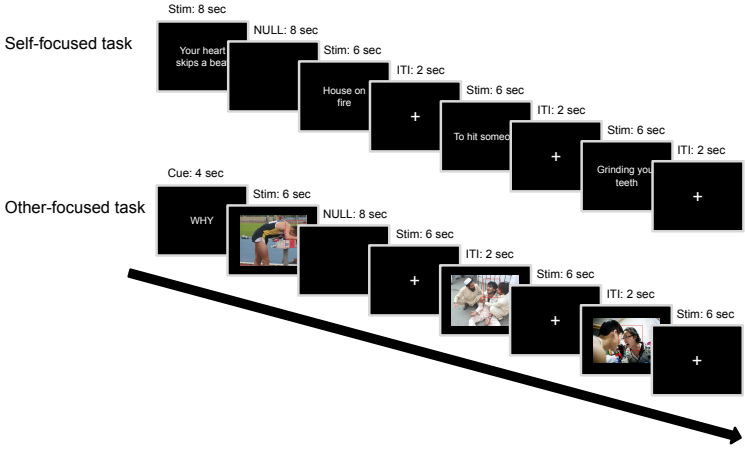


FIGURE 2.1 Overview of the self-focused and other-focused task.

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, 2012). 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, 2005; Lang et al., 1997), the image set developed by the Kveraga lab (http://www.kveragalab.org/stimuli.html; Kveraga et al., 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 2.1). Each image was shown three times, once for each cue type. Images were presented in blocks of six, each lasting 6 seconds, followed by a fixed inter trial interval of 2 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 2 hours. Subjects who underwent two sessions had them on different days within a time span of 1 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, 3-min 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 A.1). In the second session, 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-min 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 × 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.

2.3 Model optimization procedure

As MVPA is a fairly novel technique, no consistent, optimal MVPA pipeline has been established (Etzel et al., 2011). Therefore, we adopted a validation strategy in the present study that is advised in the pattern classification field (Kay et al., 2008; Kriegeskorte, Simmons, Bellgowan, et al., 2009a). This strategy entailed that we separated our data into an optimization dataset to find the most optimal parameters for preprocessing and analysis, and a validation dataset to independently verify classification success with those optimal parameters. We generated an

optimization and validation dataset 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 (see Figure 2A); first and second sessions were counterbalanced between the two sets. Based on a request received during the review process, we added nine new subjects to the validation dataset. Ultimately, the optimization-set held 13 sessions and the validation-set, after exclusion of 2 subjects (see Subjects section), held 20 sessions.

In the optimization-set, we explored how different preprocessing options and the so-called 'hyperparameters' in the MVPA pipeline affected the performance of the (multivariate) analyses (visualized in Figure 2.2B; see MVPA pipeline subsection for more details). Thus, we performed the self- and cross-analyses *on the data of the optimization set* multiple times with different preprocessing options (i.e., smoothing kernel, low-pass filter and ICA-based denoising strategies) and MVPA hyperparameter values (i.e., univariate feature selection *threshold* and train/test size ratio during cross-validation). We determined the optimal parameters on the basis of classification performance, which was operationalized as the mean precision value after a repeated random subsampling procedure with 1000 iterations. A list with the results from the optimization procedure can be found in Supplementary Table A.2 and Supplementary Figure A.2. The optimal parameters were then used for preprocessing and the self- and cross-analysis within the validation-set, in which the findings from the optimization-set were replicated. All findings discussed in the 2.4 section follow from the validation-set (see Supplementary Figure A.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 et al., 2012) and MATLAB (2012b;

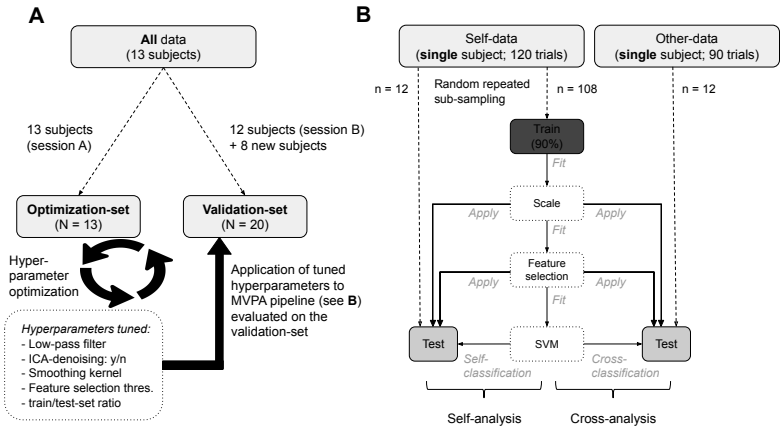


FIGURE 2.2 Schematic overview of the cross-validation procedures. **A**) The partitioning of the dataset into an optimization-set (used for tuning of preprocessing and MVPA hyperparameters) and a validation-set (used to get a fully cross-validated, unbiased estimate of classification performance). The preprocessing and MVPA hyperparameters yielded from the optimization procedure were subsequently applied to the preprocessing and MVPA pipeline of the validation-set. **B**) The within-subject MVPA pipeline of the self- and cross-analysis implemented in a repeated random subsampling scheme with 100,000 iterations. In each iteration, 90% of the self-data trials (i.e. train-set) were used for estimating the scaling parameters, performing feature selection and fitting the SVM. These steps of the pipeline (i.e. scaling, feature selection, SVM fitting) were subsequently applied to the independent test-set of both the self-data trials and the other-data trials.

www.mathworks.com/products/matlab), using an in-house developed preprocessing pipeline and the parameters established in the optimization procedure. Functional data were 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 et al., 2010), constituting a whole-brain pattern of *t*-values per trial. Subsequently, the data were indexed by a gray-matter mask (excluding most white-matter, CSF and brainstem voxels). Thus, the data points for the MVPA consist of whole-brain (gray matter) *t*-value patterns per trial. For the optimization analyses, the data were transformed to standard space (MNI152, 2 mm) using FSL's FNIRT. 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

Within the optimization and validation dataset, we implemented an iterated cross-validation scheme that separated the data into a train-set and a test-set (this procedure is described in more detail in the next section). Before fitting the classifier on the train-set in each iteration of the cross-validation scheme, 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-set 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; see Model optimization procedure section) 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-set only and subsequently applied to the test-set to ensure independence between the train- and test-set (see Figure 2B). After standardization and voxel selection in each iteration, a support vector classifier (SVC) was fit on the train-set and cross-validated on the test-set, generating a class probability for each trial in the test-set. Our classifier of choice was the SVC implementation from the scikit-learn `svm` module (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, & others, 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 scheme (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 be-

longed 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 12 test-trials per iteration.

Per iteration, the classifier was fit on the train-set from the SF-data. Subsequently, this classifier was cross-validated on 12 test SF-trials (test-set "self-analysis") and 12 test OF-trials (test-set "cross-analysis"; see Figure 2B). 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 A.4). 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, 2018).

After generating a final prediction for all trials using the 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 A.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, 2002). 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 A.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 (Ethofer et al., 2009; Guyon et al., 2002; Stelzer et al., 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 null-hypothesis, yielding an interpretable *t*-value map (see the supplementary code notebook on our Github repository for computational details).

Additional analyses

In addition to the self-analysis and the self-to-other cross-analysis presented in the main text, we also performed a within-subjects other-to-self cross-analysis (see for a similar approach Corradi-Dell'Acqua et al., 2016) and a between-subjects self-analysis and self-to-other cross-analysis. These analyses forward largely similar results as the analyses presented in the main text. Due to space constraints, we present these additional analyses in the Supplementary Materials. Supplementary Figure A.7 represents confusion matrices with precision and recall estimates for the other-to-self cross-analysis. Supplementary Figure A.8 presents the results of MVPA analyses using condition-average voxel patterns across subjects instead of single-trial patterns within subjects.

Univariate analysis

To be complete, we also report a set of univariate analyses performed on the SF-task and the OF-task data. The univariate analyses were performed on the validation dataset, and were subject to the same preprocessing steps as the MVPA analysis, except that we did not model each trial, but each condition as a separate regressor. The group-level analysis was performed with FSL's FLAME1 option. To examine differences in neural activity between conditions, we calculated contrasts between the three classes in the SF-task (self-action *vs* self-interoception; self-action *vs* self-situation and self-interoception *vs* self-situation) and the three classes in the OF-task (other-action *vs* other-interoception; other-action *vs* other-situation

and other-interoception *vs* other-situation). We report clusters that were corrected using cluster-correction with a voxel-wise threshold of 0.005 ($z = 2.7$) and a cluster-wise p-value threshold of 0.05.

Code availability

The MVPA-analysis and subsequent (statistical) analyses were implemented using custom Python scripts, which depend heavily on the skbold package, a set of tools for machine learning analyses of fMRI data developed in-house (see https://github.com/lukassnoek/skbold). The original scripts were documented and are hosted at the following Github repository: https://github.com/lukassnoek/SharedStates.

2.4 Results

Multi-voxel pattern analysis

The analyses of the SF-task demonstrated that voxel patterns reflecting imagined self-focused actions, interoceptive sensations and situations associated with emotion could be decoded accurately for each individual class (all $p < 0.001$, see Figure 2.3). Furthermore, when we generalized the classifier based on the SF-task to the data from the OF-task (i.e. cross-analysis), we found that neural representations of emotional actions, interoceptive sensations and situations of others could also be reliably decoded above chance (all $p < 0.001$; see Figure 2.3). Supplementary Table A.3 presents mean precision-scores across classes for each subject separately. As predicted, our findings demonstrate that *self-imagined* actions, interoceptive sensations and situations are associated with distinct neural patterns. Furthermore, and as predicted, our findings demonstrate that the patterns associated with self-imagined actions, sensations and situations can be used to decode *other-focused* actions, interoceptive

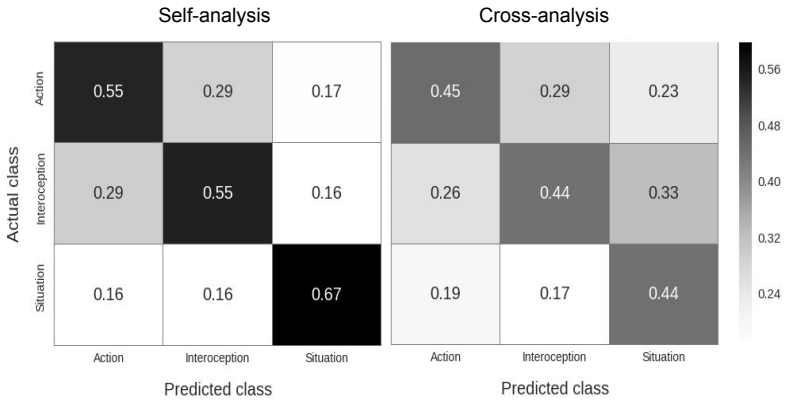


FIGURE 2.3 Confusion matrices for the self- (left diagram) and cross-analysis (right diagram). Values indicate precision-scores, representing the proportion of true positives given all predictions for a certain class. Note that action and interoception columns in the cross-analysis confusion matrix do not add up to 1, which is caused by the fact that, for some subjects, no trials were predicted as action or interoception, rendering the calculation of precision ill-defined (i.e., division by zero). In this case, precision scores were set to zero.

sensations and situations (see Supplementary Figure A.7 for the complementary other-to-self cross-analysis).

To visualize which neural regions were involved in the successful decoding of the three classes in the OF-task and SF-task, we display in Figure 2.4 the averaged absolute values of the SVM feature weights. Note that Figure 2.4 only displays one feature map, as both the self and cross-analysis depend on the same model. Regions displaying high and consistent feature weights across subjects were frontal pole (including parts of the dorsomedial prefrontal cortex and ventromedial prefrontal cortex), orbitofrontal cortex (OFC), inferior frontal gyrus (IFG), superior frontal gyrus (SFG), middle frontal gyrus (MFG), insular cortex, precentral gyrus, postcentral gyrus, posterior cingulate cortex/precuneus, superior parietal lobule (SPL), supramarginal gyrus (SMG), angular gyrus (AG), middle temporal

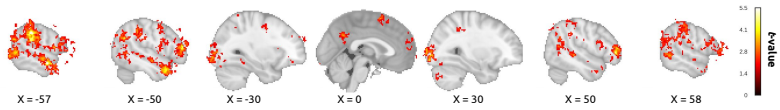


FIGURE 2.4 Uncorrected *t*-value map of average feature weights across subjects; *t*-values were calculated by dividing the average absolute feature weights, which was corrected for positive bias by subtracting the mean permuted absolute weight across all iterations, by the standard error across subjects. Only voxels belonging to clusters of 20 or more voxels are shown.

gyrus (MTG), temporal pole (TP), lateral occipital cortex (lOC) and occipital pole (see Supplementary Table A.4 for an overview of all involved regions).

Univariate analyses

Figure 2.5 displays the pattern of neural activity revealed by univariate contrasts between the three different classes in the SF-task and the OF-task. For the sake of brevity, we summarize the most relevant univariate results here. Please see the Supplementary Materials and the study's Github repository for an overview of all clusters.

In the SF-task, action was associated with increased involvement of the MFG, SFG, AG, SMG, lOC and middle temporal gyrus (temporo-occipital) when compared with interoception, and increased involvement of the IFG, MFG, SFG, anterior cingulate cortex (ACC), supplementary motor area (SMA), precentral gyrus, postcentral gyrus, insular cortex, SMG, SPL, lOC and middle temporal gyrus (temporo-occipital) when compared with situation. Interoception was associated with increased involvement of the insular cortex, precentral gyrus, postcentral gyrus and central operculum when compared with action, and increased involvement of the insular cortex, central operculum, parietal operculum, IFG, frontal pole, ACC, SMA, precentral

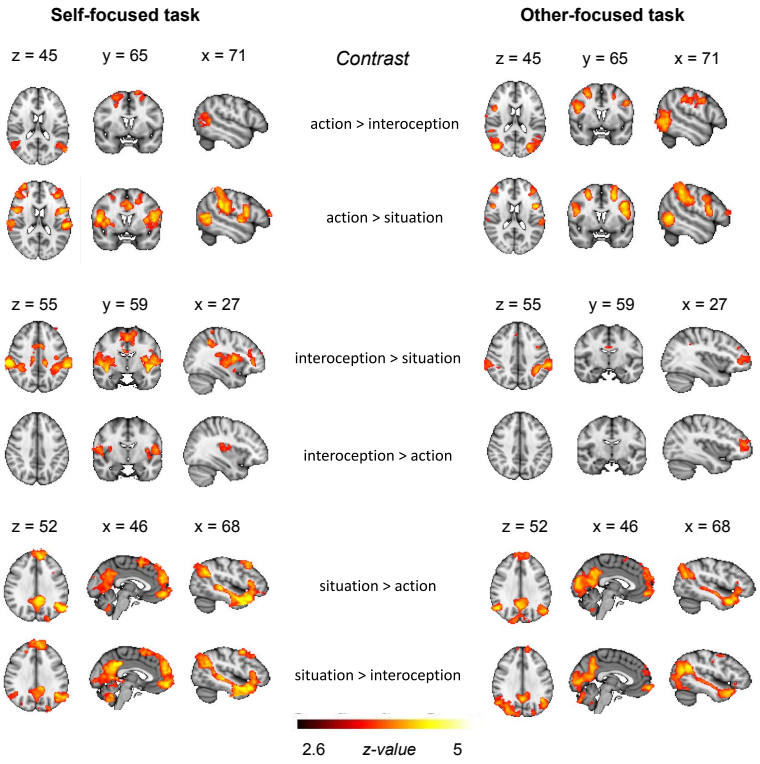


FIGURE 2.5 Univariate contrasts for the self-focused and other-focused task.

gyrus, postcentral gyrus, SMG, SPL and putamen when compared with situation. The situation *vs* action contrast and the situation *vs* interoception contrast forwarded clusters in similar regions, including the temporal pole, superior/middle temporal gyrus, IFG, SFG, frontal pole, medial prefrontal cortex (mPFC), OFC, precuneus, posterior cingulate cortex (PCC), lOC, fusiform gyrus, hippocampus and lingual gyrus.

In the OF-task, action was associated with increased involvement of the IFG, MFG, SFG, precentral gyrus, postcentral gyrus,

SMG, SPL, middle/inferior temporal gyrus (temporo-occipital), lOC and fusiform gyrus, when compared with interoception, and increased involvement of the IFG, MFG, SFG, frontal pole, precentral gyrus, postcentral gyrus, SMG, SPL, middle/inferior temporal gyrus (temporo-occipital) and lOC, when compared with situation. Interoception was associated with increased involvement of the left frontal pole when compared with action, and increased involvement of the SMG, SPL, precentral gyrus, postcentral gyrus, PCC, IFG and frontal pole, when compared with situation. The situation *vs* action contrast and the situation *vs* interoception contrast forwarded clusters in similar regions, including the temporal pole, superior/middle temporal gyrus, frontal pole, mPFC, PCC, precuneus, AG, lOC, occipital pole, fusiform gyrus and lingual gyrus.

2.5 Discussion

In this study, we investigated the neural overlap between self-focused emotion imagery and other-focused emotion understanding using a decoding approach. The results confirmed our hypothesis that other-focused representations of emotion-related actions, bodily sensations and situations can be decoded from neural patterns associated with accessing similar sources of information in a self-focused task. This cross-classification was successful even though the tasks employed different stimulus materials and instructions. Thus, the observed neural overlap between the underlying processes in the SF-task and OF-task cannot be attributed to similarities in stimulus dimensions or task instructions. Rather, we conclude from our findings that emotion experience and emotion understanding have basic psychological processes in common.

Although we could successfully classify the interoception class in the SF-task (across both datasets), and in the OF-task in

the validation dataset, we were not able to successfully classify the interoception class in the OF-task in the optimization dataset. Furthermore, although precision and recall metrics demonstrated similar results for the action and situation cross-classification in the validation dataset, these metrics demonstrated different results for the classification of the interoception class (see Supplementary Figure A.5). This difference was partly driven by the fact that trials were very infrequently classified as interoception in the cross-classification analysis. The finding that subjects reported lower success rates for the *what* trials in which they were asked to identify interoceptive sensations in other people than for the *how* (action) and *why* (situation) trials may point to a possible explanation for the inconsistent findings regarding interoception. Although speculative, it may be relatively easy to recognize (and represent) interoceptive sensations when they are described in words (as in the SF-task), but relatively hard to deduce these sensations when only diffuse cues about someone's internal state are available (e.g. posture, frowning facial expression, as in the OF-task).

An exploration of the spatial characteristics of the distributed neural pattern associated with successful decoding of the SF-task and OF-task revealed regions that are commonly active during self- and other-focused processing. First, we found that successful classification was associated with voxels in the precentral gyrus, IFG, SMA and SPL. These same regions were also revealed by the univariate analyses, in particular for the action and interoception classes. These regions are part of the so-called "mirror" network, which is argued to support both action planning and action understanding (Bastiaansen et al., 2009; Gallese et al., 2004; Spunt & Lieberman, 2012; Van Overwalle & Baetens, 2009). Furthermore, we found that successful classification was associated with voxels in the lateral occipital cortex and fusiform gyrus, which have been linked in the literature to the processing of both concrete and abstract action (Wurm & Lingnau, 2015)

and the (visual) processing of emotional scenes, faces and bodies (Gelder et al., 2010; Sabatinelli et al., 2011). The univariate analyses demonstrated activity in the lOC and the fusiform gyrus in particular for the situation class, both when subjects viewed images of other people in emotional situations, and when subjects imagined being in an emotional situation themselves.

Second, we found that successful classification was associated with voxels in regions associated with somatosensory processing (postcentral gyrus) and the representation of interoceptive sensations (insular cortex, see Craig & Craig, 2009; Medford & Critchley, 2010). Univariate analyses of the SF-task also demonstrated involvement of these regions for both the action and interoception classes. This pattern of activation is consistent with embodied cognition views that propose that thinking about or imagining bodily states is grounded in simulations of somatosensory and interoceptive sensations (Barsalou, 2009). In contrast to previous work on interoceptive simulation when observing pain or disgust in other people (cf. Bastiaansen et al., 2009; Lamm et al., 2011), the univariate analyses of the OF-task did not demonstrate insular cortex activation for the interoception class.

And third, we found that successful classification was associated with voxels in the middle temporal gyrus (including the temporal pole), PCC/precuneus, dmPFC and vmPFC. These regions are part of the so-called "mentalizing" network (or "default" network). This same network was also revealed by the univariate analyses, in particular for the situation class. Meta-analyses have demonstrated that the mentalizing network is commonly active during tasks involving emotion experience and perception (Lindquist et al., 2012), mentalizing/theory of mind (Spreng et al., 2009; Van Overwalle & Baetens, 2009), judgments about the self and others (Denny et al., 2012) and semantic/conceptual processing in general (Binder et al., 2009). Moreover, this network contributes to the representation of emotion knowledge

(Peelen et al., 2010) and is involved in both empathy (Keysers & Gazzola, 2014; Zaki & Ochsner, 2012) and self-generated thought (Andrews-Hanna et al., 2014). We propose that this network supports the implementation of situated knowledge and personal experience that is necessary to generate rich mental models of emotional situations, both when experienced individually, and when understood in someone else (cf. Barrett & Satpute, 2013; Oosterwijk & Barrett, 2014).

The most important contribution of our study is that it provides direct evidence for the idea of shared neural resources between self-and other focused processes. It is important, however, to specify what we think this "sharedness" entails. In research on pain, there is an ongoing discussion about whether experiencing pain and observing pain in others are distinct processes (Krishnan et al., 2016), or whether experiencing and observing pain involve a shared domain-specific representation (e.g., a discrete pain-specific brain state; Corradi-Dell'Acqua et al., 2016) and/or the sharing of domain-general processes (e.g. general negative affect; Zaki et al., 2016). Connecting to this discussion, we think that it is unlikely that our decoding success reflects the sharing of discrete experiential states between the SF-task and OF-task. After all, unlike in studies on pain, the stimuli in our tasks referred to a large variety of different actions, sensations and situations. Instead, decoding success in our study is most likely due to shared brain state configurations, reflecting the similar engagement of domain-general processes evoked by self- and other-focused instances of action (or interoceptive sensation or situation). This interpretation is consistent with views that suggests that global processes are shared between pain experience and pain observation (Lamm et al., 2011; Zaki et al., 2016) or between self- and other-focused tasks in general (e.g., Legrand & Ruby, 2009). Moreover, this interpretation is consistent with the suggestion that neural re-use is a general principle of brain functioning (e.g., Anderson, 2016).

In our constructionist view, we posit that emotion imagery and understanding share basic psychological processes (cf. Oosterwijk & Barrett, 2014). More specifically, both emotion imagery and understanding are "conceptual acts" in which the brain generates predictions based on concept knowledge (including sensorimotor and interoceptive predictions) that are meaningful within a particular situational context (Barrett, 2012; Barrett & Simmons, 2015). Based on accumulating evidence, we propose that these predictions are implemented in domain-general brain networks (cf. Oosterwijk et al., 2012; Barrett & Satpute, 2013). The relative contribution of these networks depends on the demands of the situational context. Specifically, in contexts where people are focused on actions and expressions (their own or someone else's) a network that supports the representation of sensorimotor states (i.e., the mirror system) may contribute relatively heavily; in contexts where people are focused on bodily states (their own or someone else's) a network that supports the representation of interoceptive states (i.e., the salience network) may contribute relatively heavily; and in contexts where people are focused on interpreting a situation (their own or someone else's) a network that supports a general inferential meaning-making function (i.e., the mentalizing network) may contribute relatively heavily (see also Oosterwijk et al., 2015). We believe that it is likely that our ability to successfully distinguish between classes in the self-task relies on the relatively *different* patterns of activity across these networks for actions, interoceptive sensations and situations. Regarding our ability to successfully generalize from the self- to the other-focused task, we believe that this relies on the relatively *similar* pattern of activity across these networks when people generate self-focused or other-focused instances of action (or interoceptive sensation or situation).

Our explicit manipulation of the weight of action, interoceptive and situational information in the SF-task and the OF-task tests the possibility of shared representation in a novel way. Although

this procedure may seem artificial, social neuroscience studies support the notion that there is contextual variety in the contribution of action, interoceptive, and situation information when understanding other people (Oosterwijk et al., 2015; Van Overwalle & Baetens, 2009). Moreover, this weighting may mimic the variability with which these sources of information contribute to different instances of subjective emotional experience in reality (Barrett, 2012). In future directions, it may be relevant to apply the current paradigm to the study of individuals in which access to these sources of information is disturbed (e.g., individuals with different types of psychopathology) or facilitated (e.g., individuals with high interoceptive sensitivity).

In short, the present study demonstrates that the neural patterns that support imagining "performing an action", "feeling a bodily sensation" or "being in a situation" are directly involved in understanding other people's actions, sensations and situations. This supports our prediction that self- and other-focused emotion processes share resources in the brain.

2.6 Acknowledgements

The authors would like to thank David Amodio for his helpful comments on a previous draft of this manuscript.

Chapter 3

How to control for confounds in decoding analyses of neuroimaging data

---

This chapter has been published as: Snoek, L.*, Miletić, S.*, & Scholte, H.S. (2019). How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*, 184, 741-760.

* Shared first authorship

Abstract

Over the past decade, multivariate "decoding analyses" have become a popular alternative to traditional mass-univariate analyses in neuroimaging research. However, a fundamental limitation of using decoding analyses is that it remains ambiguous which source of information drives decoding performance, which becomes problematic when the to-be-decoded variable is confounded by variables that are not of primary interest. In this study, we use a comprehensive set of simulations as well as analyses of empirical data to evaluate two methods that were previously proposed and used to control for confounding variables in decoding analyses: post hoc counterbalancing and confound regression. In our empirical analyses, we attempt to decode gender from structural MRI data while controlling for the confound "brain size". We show that both methods introduce strong biases in decoding performance: post hoc counterbalancing leads to better performance than expected (i.e., positive bias), which we show in our simulations is due to the subsampling process that tends to remove samples that are hard to classify or would be wrongly classified; confound regression, on the other hand, leads to worse performance than expected (i.e., negative bias), even resulting in significant below chance performance in some realistic scenarios. In our simulations, we show that below chance accuracy can be predicted by the variance of the distribution of correlations between the features and the target. Importantly, we show that this negative bias disappears in both the empirical analyses and simulations when the confound regression procedure is performed in every fold of the cross-validation routine, yielding plausible (above chance) model performance. We conclude that, from the various methods tested, cross-validated confound regression is the only method that appears to appropriately control for confounds which thus can be used to gain more insight into the exact source(s) of information driving one's decoding analysis.

3.1 Introduction

In the past decade, multivariate pattern analysis (MVPA) has emerged as a popular alternative to traditional univariate analyses of neuroimaging data (Haxby, 2012; Norman et al., 2006b). The defining feature of MVPA is that it considers patterns of brain activation instead of single units of activation (i.e., voxels in MRI, sensors in MEG/EEG). One of the most-often used type of MVPA is "decoding", in which machine learning algorithms are applied to neuroimaging data to predict a particular stimulus, task, or psychometric feature. For example, decoding analyses have been used to successfully predict various experimental conditions within subjects, such as object category from fMRI activity patterns (Haxby et al., 2001) and working memory representations from EEG data (LaRocque et al., 2013), as well between-subject factors such as Alzheimer's disease (vs. healthy controls) from structural MRI data (Cuingnet et al., 2011) and major depressive disorder (vs. healthy controls) from resting-state functional connectivity (Craddock et al., 2009). One reason for the popularity of MVPA, and especially decoding, is that these methods appear to be more sensitive than traditional mass-univariate methods in detecting effects of interest. This increased sensitivity is often attributed to the ability to pick up multidimensional, spatially distributed representations which univariate methods, by definition, cannot do (Jimura & Poldrack, 2012). A second important reason to use decoding analyses is that they allow researchers to make predictions about samples beyond the original dataset, which is more difficult using traditional univariate analyses (Hebart & Baker, 2017).

In the past years, however, the use of MVPA has been criticized for a number of reasons, both statistical (Allefeld et al., 2016; Davis et al., 2014; Gilron et al., 2017; Haufe et al., 2014) and more conceptual (Naselaris & Kay, 2015; Weichwald et al., 2015)

in nature. For the purposes of the current study, we focus on the specific criticism put forward by Naselaris & Kay (2015) , who argue that decoding analyses are inherently ambiguous in terms of what information they use (see Popov et al., 2018 for a similar argument in the context of encoding analyses). This type of ambiguity arises when the classes of the to-be-decoded variable systematically vary in more than one source of information (see also Carlson & Wardle, 2015; Ritchie et al., 2017; Weichwald et al., 2015). The current study aims to investigate how decoding analyses can be made more interpretable by reducing this type of "source ambiguity".

To illustrate the problem of source ambiguity, consider, for example, the scenario in which a researcher wants to decode gender.[1] (male/female) from structural MRI with the aim of contributing to the understanding of gender differences — an endeavor that generated considerable interest and controversy (Chekroud et al., 2016; Del Giudice et al., 2016; Glezerman, 2016; Joel & Fausto-Sterling, 2016; Rosenblatt, 2016). By performing a decoding analysis on the MRI data, the researcher hopes to capture meaningful patterns of variation in the data of male and female participants that are predictive of the participant's gender. The literature suggests that gender dimorphism in the brain is manifested in two major ways (Good et al., 2001; O'Brien et al., 2011). First, there is a *global* difference between male and female brains: men have on average about 15% larger intracranial volume than women, which falls in the range of mean gender differences in height (8.2%) and weight (18.7%; Gur et al., 1999; Lüders et al., 2002).[2] Second, brains of men and women are known to differ *locally*: some specific brain areas are

[1]The terms "gender" and "sex" are both used in the relevant research literature. Here, we use the term gender because we refer to self-reported identity in the data described below.

[2]Note that information related to global brain size persists when researchers analyze the structural MRI data in a common, normalized brain

on average larger in women than in men (e.g., in superior and middle temporal cortex; Good et al., 2001) and vice versa (e.g., in frontomedial cortex; Goldstein et al., 2001). One could argue that, given that one is interested in explaining behavioral or mental gender differences, global differences are relatively uninformative, as it reflects the fact than male *bodies* are on average larger than female bodies (Gur et al., 1999; Sepehrband et al., 2018). As such, our hypothetical researcher is likely primarily interested in the *local* sources of variation in the neuroanatomy of male and female brains.

Now, supposing that the researcher is able to decode gender from the MRI data significantly above chance, it remains unclear on which source of information the decoder is capitalizing: the (arguably meaningful) local difference in brain structure or the (in the context of this question arguably uninteresting) global difference in brain size? In other words, the data contain more than one source of information that may be used to predict gender. In the current study, we aim to evaluate methods that improve the interpretability of decoding analyses by controlling for "uninteresting" sources of information.

Partitioning effects into *true* signal and *confounded* signal

Are multiple sources of information necessarily problematic? And what makes a source of information interesting or uninteresting? The answers to these questions depend on the particular goal of the researcher using the decoding analysis (Hebart & Baker, 2017). In principle, multiple sources of information in

space, because spatial registration "squeezes" relatively large brains into a smaller template, increasing voxel statistics (e.g., gray matter density in VBM analyses), and vice versa (Douaud et al., 2007). This effect of global brain size similarly affects functional MRI analyses (Brodtmann et al., 2009).

the data do not pose a problem if a researcher is only interested in accurate *prediction*, but not in *interpretability* of the model (Bzdok, 2017; Haufe et al., 2014; Hebart & Baker, 2017). In brain-computer interfaces (BCI), for example, accurate prediction is arguably more important than interpretability, i.e., knowing which sources of information are driving the decoder. Similarly, if the researcher from our gender decoding example is only interested in accurately predicting gender regardless of model interpretability, source ambiguity is not a problem.[3] In most scientific applications of decoding analyses, however, model interpretability is important, because researchers are often interested in the relative contributions of different sources of information to decoding performance. Specifically, in most decoding analyses, researchers often (implicitly) assume that the decoder is *only* using information in the neuroimaging data that is related to the variable that is being decoded (Ritchie et al., 2017). In this scenario, source ambiguity (i.e., the presence of *multiple* sources of information) *is* problematic as it violates this (implicit) assumption. Another way to conceptualize the problem of source ambiguity is that, using the aforementioned example, (global) brain size is *confounding* the decoding analysis of gender. Here, we define a confound as *a variable that is not of primary interest, correlates with the to-be-decoded variable (the target), and is encoded in the neuroimaging data.*

To illustrate the issue of confounding variables in the context of decoding clinical disorders, suppose one is interested in building a classifier that is able to predict whether subjects are suffering from schizophrenia or not based on the subjects' gray matter data. Here, the variable "schizophrenia-or-not" is the variable of interest, which is assumed to be encoded in the neuroimaging

[3]However, if accurate prediction is the only goal in this scenario, we would argue that there are probably easier and less expensive methods than neuroimaging to predict a participant's gender.

data (i.e., the gray matter) and can thus be decoded. However, there are multiple factors known to covary with schizophrenia, such as gender (i.e., men are more often diagnosed with schizophrenia than women; McGrath et al., 2008) and substance abuse (Dixon, 1999), which are also known to affect gray matter (Bangalore et al., 2008; Gur et al., 1999; Van Haren et al., 2013). As such, the variables gender and substance abuse can be considered confounds according to our definition, because they are both correlated with the target (schizophrenia or not) and are known to be encoded in the neuroimaging data (i.e., the effect of these variables is present in the gray matter data). Now, if one is able to classify schizophrenia with above-chance accuracy from gray matter data, one cannot be sure which source of information within the data is picked up by the decoder: information (uniquely) associated with schizophrenia or (additionally) information associated with gender or substance abuse? If one is interested in more than mere accurate *prediction* of schizophrenia, then this ambiguity due to confounding sources of information is problematic.

Importantly, as our definition suggests, what *is* or *is not* regarded as a confound is relative — it depends on whether the researchers deems it of (primary) interest or not. In the aforementioned hypothetical schizophrenia decoding study, for example, one may equally well define the severity of substance abuse as the to-be-decoded variable, in which the variable "schizophrenia-or-no"" becomes the confounding variable. In other words, one researcher's signal is another researcher's confound. Regardless, if decoding analyses of neuroimaging data are affected by confounds, the data thus contain two types of information: the "true signal" (i.e., variance in the neuroimaging data related to the target, but unrelated to the confound) and the "confounded signal" (i.e., variance in the neuroimaging data related to the target that is also related to the confound; see Figure 3.1). In other words,

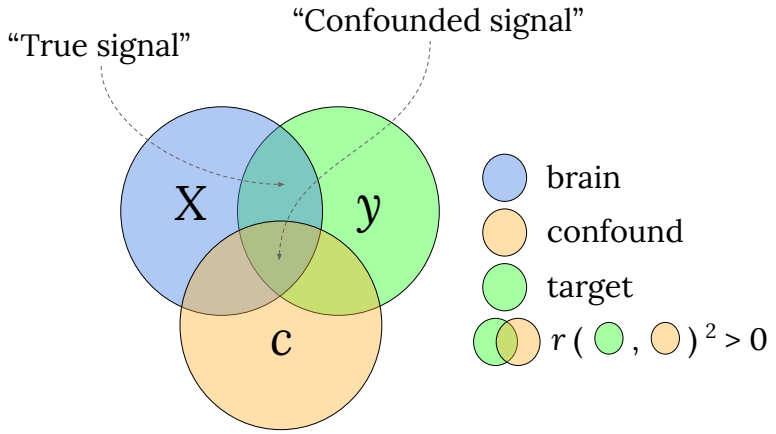


FIGURE 3.1 Visualization of how variance in brain data (X) can partitioned into "True signal" and "Confounded signal", depending on the correlation structure between the brain data (X), the confound (C), and the target (y). Overlapping circles indicate a non-zero (squared) correlation between the two variables.

source ambiguity arises due to the presence of both true signal and confounded signal and, thus, controlling for confounds (by removing the confounded signal) provides a crucial methodological step forward in improving the interpretability of decoding analyses.

In the decoding literature, various methods have been applied to control for confounds. We next provide an overview of these methods, highlight their advantages and disadvantages, and discuss their rationale and the types of research settings they can be applied in. Subsequently, we focus on two of these methods to test whether these methods succeed in controlling for the influence of confounds.

Methods for confound control

In decoding analyses, one aims to predict a certain target variable from patterns of neuroimaging data. Various methods discussed in this section are supplemented with a mathematical formalization; for consistency and readability, we define the notation we will use in Table 3.1.

TABLE 3.1 Notation.

Symbol	Dims.	Description
N	-	Number of samples (usually subjects or trials)
K	-	Number of neuroimaging features (e.g., voxels or sensors)
P	-	Number of confound variables (e.g., age, reaction time, or brain size)
X_{ij}	$N \times K$	The neuroimaging patterns (often called the "data" in the current article), where the subescript $i \in 1 \ldots N$ refers to the individual samples (rows), and the subscript $j \in 1 \ldots K$ to individual features (columns)
y	$N \times 1$	The target variable (i.e., what is to be decoded)
C	$N \times P$	The confound variable(s)
$\hat{\beta}$	$K + 1$	The parameters estimated in a general linear model (GLM)
w	$K + 1$	The parameters estimated in a decoding model
r_{Cy}	-	Sample Pearson correlation coefficient between C and y
$r_{y(X.C)}$	-	Sample semipartial Pearson correlation coefficient between X and y, controlled for C
$p(r_{Cy})$	-	p-value of sample Pearson correlation between C and y

Note: Format based on Diedrichsen and Kriegeskorte (2017). For the correlations (r), we assume that $P = 1$ and thus that the correlations in the table reduce to a scalar.

A priori counterbalancing

Ideally, one would prevent confounding variables from influencing the results as much as possible before the acquisition of the neuroimaging data.[4] One common way do this (in both traditional "activation-based" and decoding analyses) is to make sure that potential confounding variables are *counterbalanced* in the experimental design (Görgen et al., 2017). In experimental research, this would entail randomly assigning subjects to design cells (e.g., treatment groups) such that there is no structural correlation between characteristics of the subjects and design cells. In observational designs (e.g., in the gender/brain size example described earlier), it means that the sample is chosen such that there is no correlation between the confound (brain size) and *observed* target variable (gender). That is, given that men on average have larger brains than women, this would entail including only men with relatively small brains and women with relatively large brains.[5] The distinction between experimental and observational studies is important because the former allow the researcher to randomly draw samples from the population, while the latter require the researcher to choose a sample that is not representative of the population, which limits the conclusions that can be drawn about the population (we will revisit this issue in the Discussion section).

Formally, in decoding analyses, a design is counterbalanced when the confound C and the target y are statistically independent. In practice, this often means that the sample is chosen

[4] In the context of behavioral data, a priori counterbalancing is often called "matching" or a employing a "case-control design" @[Cook2002-hb].

[5] Note that the counterbalancing process is the same for both traditional univariate (activation-based) studies and decoding studies, but the direction of analysis is reversed in univariate (e.g., gender → brain) and decoding studies (e.g., brain → gender). As such, in univariate studies the confound (e.g., brain size) is counterbalanced with respect to the predictor(s) (e.g., gender) while in decoding studies the confound (e.g., brain size) is counterbalanced with respect to the target (e.g., gender).

so that there is no significant correlation coefficient between C and y (although this does not necessarily imply that C and y are actually independent). To illustrate the process of counterbalancing, let's consider another hypothetical experiment: suppose one wants to set up an fMRI experiment in which the goal is to decode abstract object category (e.g., faces vs. houses) from the corresponding fMRI patterns (cf. Haxby et al., 2001), while controlling for the potential confounding influence of low-level or mid-level stimulus features, such as luminance, spatial frequency, or texture (Long et al., 2017). Proper counterbalancing would entail making sure that the images used for this particular experiments have similar values for these low-level and mid-level features across object categories (see for details Görgen et al., 2017). Thus, in this example, low-level and mid-level stimulus features should be counterbalanced with respect to object category, such that above chance decoding of object category cannot be attributed to differences in low-level or mid-level stimulus features (i.e., the confounds).

A priori counterbalancing of potential confounds is, however, not always feasible. For one, the exact measurement of a potentially confounding variable may be impossible until data acquisition. For example, the brain size of a participant is only known after data collection. Similarly, Todd et al. (2013) found that their decoding analysis of rule representations was confounded by response times of to the to-be-decoded trials. Another example of a "data-driven" confound is participant motion during data acquisition (important in, for example, decoding analyses applied to data from clinical populations such as ADHD; Yu-Feng et al., 2007). In addition, a priori counterbalancing of confounds may be challenging because of the limited size of populations of interest. Especially in clinical research settings, researchers may not have the luxury of selecting a counterbalanced sample due to the small number of patient subjects available for

testing. Lastly, researchers may simply discover confounds after data acquisition.

Given that a priori counterbalancing is not possible or undesirable in many situations, it is paramount to explore the possibilities of controlling for confounding variables after data acquisition for the sake of model interpretability, which we discuss next.

Include confounds in the data

One perhaps intuitive method to control for confounds in decoding analyses is to include the confound(s) in the data (i.e., the neuroimaging data, X; see, e.g., Sepehrband et al., 2018) used by decoding model. That is, when applying a decoding analysis to neuroimaging data, the confound is added to the data as if it were another voxel (or sensor, in electrophysiology). This intuition may stem from the analogous situation in univariate (activation-based) analyses of neuroimaging data, in which confounding variables are controlled for by including them in the design matrix together with the stimulus/task regressors. For example, in univariate analyses of functional MRI, movement of the participant is often controlled for by including motion estimates in the design matrix of first-level analyses (Johnstone et al., 2006); in EEG, some control for activity due to eye-movements by including activity measured by concurrent electro-oculography as covariates in the design-matrix (Parra et al., 2005). Usually, the general linear model is then used to estimate each predictor's influence on the neuroimaging data. Importantly, the parameter estimates ($\hat{\beta}$) are often interpreted as reflecting the unique contribution[6] of each predictor variable, independent from the

[6]However, parameter estimates only reflect unique variance when ordinary, weighted, or generalized least squares is used to find the model parameters. Other (regularized) linear models, such as ridge regression or LASSO, are not guaranteed to yield parameters that explain unique proportions of variance.

influence of the confound.

Contrary to general linear models as employed in univariate (activation-based) analyses, including confound variables in the data as predictors for *decoding* models is arguably problematic. If a confound is included in the data in the context of decoding models, the parameter estimates of the features (often called "feature weights", *w*, in decoding models) may be corrected for the influence of the confound, but the *model performance* (usually measured as explained variance, R^2, or classification accuracy; Hebart & Baker, 2017) is not. That is, rather than providing an estimate of decoding performance "controlled for" a confound, one obtains a measure of performance when explicitly *including* the confound as an interesting source of variance that the decoder is allowed to use. This is problematic because research using decoding analyses generally does not focus on parameter estimates but on statistics of model performance. Model performance statistics (e.g., R^2, classification accuracy) alone cannot disentangle the contribution of different sources of information as they only represent a single summary statistic of model fit (Ritchie et al., 2017). One might, then, argue that additionally inspecting feature weights of decoding models may help in disambiguating different sources of information (Sepehrband et al., 2018). However, it has been shown that feature weights cannot be reliably mapped to specific sources of information, i.e., as being task-related or confound-related (e.g., features with large weights may be completely uncorrelated with the target variable; Haufe et al., 2014; Hebart & Baker, 2017). As such, it does not make sense to include confounds in the set of predictors when the goal is to disambiguate the different sources of information in decoding analyses.

Recently, another approach similar to including confounds in the data has been proposed, which is based on the idea of a dose-response curve (Alizadeh et al., 2017). In this method, in-

stead of adding the confound(s) to the model directly, the relative contribution of true and confounded signal is systematically controlled. The authors show that this approach is able to directly quantify the unique contribution of each source of information, thus effectively controlling for confounded signal. However, while sophisticated in its approach, this method only seems to work for categorical confounds, as it is difficult (if not impossible) to systematically vary the proportion of confound-related information when dealing with continuous confounds or when dealing with more than one confound.

Control for confounds during pattern estimation

Another method that was used in some decoding studies on functional MRI data aims to control for confounds in the initial procedure of estimating activity patterns of the to-be-decoded events, by leveraging the ability of the GLM to yield parameter estimates reflecting unique variance (Woolgar et al., 2014). In this method, an initial "first-level" (univariate) analysis models the fMRI time series (s) as a function of both predictors-of-interest (X) and the confounds (C), often using the GLM[7]:

$$s = X\beta_x + C\beta_c + \varepsilon \tag{3.1}$$

Then, only the estimated parameters ($\hat{\beta}$, or normalized parameters, such as t-values or z-values) corresponding to the predictors-of-interest ($\hat{\beta}_x$) are used as activity estimates (i.e., the used for predicting the target y) in the subsequent decoding analyses. This method thus takes advantage of the shared variance partitioning in the pattern estimation step to control for potential confounding variables. However, while elegant

[7] Note that X and C, here, refer to (usually HRF-convolved) predictors of the time series signal (s) for a single voxel. In the rest of the article, X and C refer to features that are defined across samples (not time).

in principle, this method is not applicable in between-subject decoding studies (e.g., clinical decoding studies; Waarde et al., 2014; Cuingnet et al., 2011), in which confounding variables are defined across subjects, or in electrophysiology studies, in which activity patterns do not have to be[8] estimated in a first-level model, thus limiting the applicability of this method.

Post hoc counterbalancing of confounds

When a priori counterbalancing is not possible, some have argued that post hoc counterbalancing might control for the influence of confounds (Rao et al., 2017, pp. 24, 38). In this method, given that there is some sample correlation between the target and confound ($r_{Cy} \neq 0$) in the entire dataset, one takes a subset of samples in which there is no empirical relation between the confound and the target (e.g., when $r_{Cy} \approx 0$). In other words, post hoc counterbalancing is a way to *decorrelate* the confound and the target by subsampling the data. Then, subsequent decoding analysis on the subsampled data can only capitalize on true signal, as there is no confounded signal anymore (see Figure 3.2). While intuitive in principle, we are not aware of whether this method has been evaluated before and whether it yields unbiased performance estimates.

Confound regression

The last and perhaps most common method to control for confounds is removing the variance that can be explained by the confound (i.e., the confounded signal) from the neuroimaging

[8]Note that, technically, one could use the "Control for confounds during pattern estimation" method in electrophysiology as well, by first fitting a univariate model explaining the neuroimaging data (X_j for $j = 1 \ldots K$) as a function of both the target (y) and the confound (C) and subsequently only using the parameter estimates of the target-predictor ($\hat{\beta}_x$) as patterns in the subsequent decoding analysis.

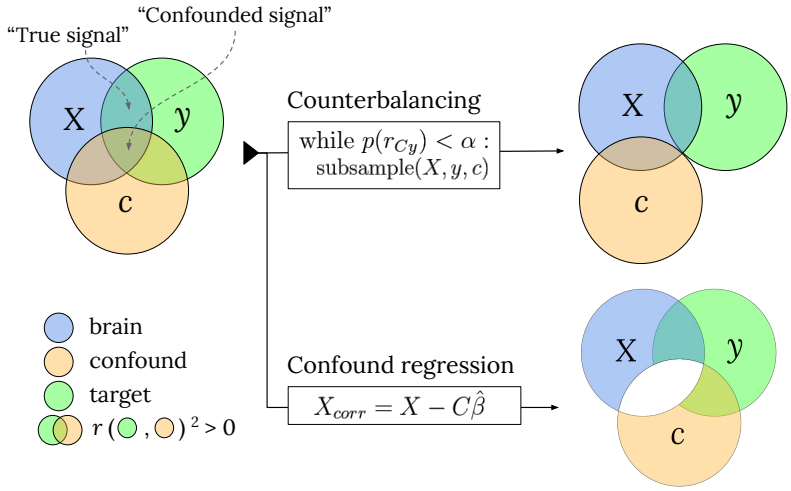


FIGURE 3.2 A schematic visualization how the main two confound control methods evaluated in this article deal with the "confounded signal", making sure decoding models only capitalize on the "true signal".

data directly (Abdulkadir et al., 2014; Dukart et al., 2011; Kostro et al., 2014; Rao et al., 2017; Todd et al., 2013) — a process we refer to as *confound regression* (also known as "image correction"; Rao et al., 2017). In this method, a (usually linear) regression model is fitted on each feature in the neuroimaging data (i.e., a single voxel or sensor) with the confound(s) as predictor(s). Thus, each feature in the neuroimaging data X is modelled as a linear function of the confounding variable(s), C:

$$X_j = C\beta + \varepsilon \tag{3.2}$$

We can estimate the parameter(s) for feature using, for example, ordinary least squares as follows (for an example using a different model, see Abdulkadir et al., 2014):

$$\hat{\beta}_j = (C^T C)^{-1} C^T X_j \tag{3.3}$$

Then, to remove the variance of (or "regress out") the confound from the neuroimaging data, we can subtract the variance in the data associated with confound ($C\hat{\beta}_j$) from the original data:

$$X_{j,\text{corr}} = X_j - C\hat{\beta}_j \tag{3.4}$$

In which $X_{j,\text{corr}}$ represents the neuroimaging feature X_j from which all variance of the confound is removed (including the variance shared with y, i.e., the confounded signal; see Figure 3.2). When subsequently applying a decoding analysis on this corrected data, one can be sure that the decoder is not capitalizing on signal that is correlated with the confound, which thus improves interpretability of the decoding analysis.

Confound regression has been applied in several decoding studies. Todd et al. (2013) were, as far as the current authors are aware, the first to use this method to control for a confound (in their case, reaction time) that was shown to correlate with their target variable (rule A vs. rule B). Notably, they both regressed out reaction time from the first-level time series data (similar to the "Control for confounds during pattern estimation" method) *and* regressed out reaction time from the trial-by-trial activity estimates (i.e., confound regression as described in this section). They showed that controlling for reaction time in this way completely eliminated the above chance decoding performance. Similarly, Kostro et al. (2014) observe a substantial drop in classification accuracy when controlling for scanner site in the decoding analysis of Huntington's disease, but only when scanner site and disease status were actually correlated. Lastly, Rao et al. (2017) found that, in contrast to Kostro et al. and Todd et al., confound regression yielded similar (or slightly lower, but still significant) performance compared to the model without confound control, but it should be noted that this study used a regression model (instead of a classification model) and evaluated confound control in the specific situation when the training

set is confounded, but the test set is not.[9] In sum, while confound regression has been used before, it has yielded variable results, possibly due to slightly different approaches and differences in the correlation between the confounding variable and the target.

Current study

In summary, multiple methods have been proposed to deal with confounds in decoding analyses. Often, these methods have specific assumptions about the nature or format of the data (such as "A priori counterbalancing" and "Confound control during pattern estimation"), differ in their objective (e.g., *prediction* vs. *interpretation*, such as in "Include confounds in the data"), or have yielded variable results (such as "Confound regression"). Therefore, given that we are specifically interested in interpreting decoding analyses, the current study evaluates the two methods that are applicable in most contexts: post hoc counterbalancing and confound regression (but see Supplementary Materials for a tentative evaluation of this method based on simulated functional MRI data). In addition to these two methods, we propose a third method — a modified version of confound regression —– which we show yields plausible, seemingly unbiased, and interpretable results.

To test whether these methods are able to effectively control for confounds and whether they yield plausible results, we apply

[9]Note that we did not discuss studies that implement a different confound regression procedure (e.g., Abdulkadir et al., 2014; Dukart et al., 2011), in which confound regression is only estimated on the samples from a single class of the target variable (e.g., in our gender decoding example, this would mean that confound regression models are only estimated on the data from male, or female, subjects). As this form of confound regression does not disambiguate the sources of information driving the decoder, it is not discussed further in this article.

them to empirical data, as well as to simulated data in which the ground truth with respect to the signal in the data (i.e., the proportion of true signal and confounded signal) is known. For our empirical data, we enact the previously mentioned hypothetical study in which participant gender is decoded from structural MRI data. We use a large dataset ($N = 217$) of structural MRI data and try to predict subjects' gender (male/female) from gray and white matter patterns while controlling for the confound of "brain size" using the aforementioned methods, which we compare to a baseline model in which confounds are not controlled for. Given the previously reported high correlations between brain size and gender (Barnes et al., 2010; Smith & Nichols, 2018), we expect that successfully controlling for brain size yields lower decoding performance than using uncorrected data, but not below chance level. Note that higher decoding performance after controlling for confounds is theoretically possible when the correlation between the confound and variance in the data *unrelated* to the target (e.g., noise) is sufficiently high to cause suppressor effects (see Figure 1 in Haufe et al., 2014; Hebart & Baker, 2017). However, because our confound, brain size, is known to correlate strongly with our target gender (approx. $r = 0.63$; Smith & Nichols, 2018), it is improbable that it also correlates highly with variance in brain data that is unrelated to gender. It follows then that classical suppression effects are unlikely and we thus expect lower model performance after controlling for brain size.

However, shown in detail below, both post hoc counterbalancing and confound regression lead to unexpected results in our empirical analyses: counterbalancing fails to reduce model performance while confound regression consistently yields low model performance up to the point of significant below chance accuracy. In subsequent analyses of simulated data, we show that both methods lead to *biased* results: post hoc counterbalancing yields inflated model performance (i.e., positive bias)

because subsampling selectively selects a subset of samples in which features correlate more strongly with the target variable, suggesting (indirect) circularity in the analysis (Kriegeskorte, Simmons, Bellgowan, et al., 2009b). Furthermore, our simulations show that negative bias (including significant below chance classification) after confound regression on the entire dataset is due to reducing the signal below what is expected by chance (Jamalabadi et al., 2016), which we show is related to and can be predicted by the standard deviation of the empirical distribution of correlations between the features in the data and the target. We propose a minor but crucial addition to the confound regression procedure, in which we cross-validate the confound regression models (which we call "cross-validated confound regression", CVCR), which solves the below chance accuracy issue and yields plausible model performance in both our empirical and simulated data.

3.2 Methods

Data

For the empirical analyses, we used voxel-based morphometry (VBM) data based on T1-weighted scans and tract-based spatial statistics (TBSS) data based on diffusion tensor images from 217 participants (122 women, 95 men), acquired with a Philips Achieva 3T MRI-scanner and a 32-channel head coil at the Spinoza Centre for Neuroimaging (Amsterdam, The Netherlands).

VBM acquisition & analysis

The T1-weighted scans with a voxel size of 1.0 × 1.0 × 1.0 mm were acquired using 3D fast field echo (TR: 8.1 ms, TE: 3.7 ms, flip angle: 8°, FOV: 240 × 188 mm, 220 slices). We used

"FSL-VBM" protocol (Douaud et al., 2007) from the FSL software package (version 5.0.9; Smith et al., 2004); using default and recommended parameters (including non-linear registration to standard space). The resulting VBM-maps were spatially smoothed using a Gaussian kernel (3 mm FWHM). Subsequently, we organized the data in the standard pattern-analysis format of a 2D ($N \times K$) array of shape 217 (subjects) × 412473 (non-zero voxels).

TBSS acquisition & analysis

Diffusion tensor images with a voxel size of 2.0 × 2.0 × 2.0 mm were acquired using a spin-echo echo-planar imaging (SE-EPI) protocol (TR: 7476 ms, TE: 86 ms, flip angle: 90°, FOV: 224 × 224 mm, 60 slices), which acquired a single b = 0 (non-diffusion-weighted) image and 32 (diffusion-weighted) b = 1000 images. All volumes were corrected for eddy-currents and motion (using the FSL command "eddy_correct") and the non-diffusion-weighted image was skullstripped (using FSL-BET with the fractional intensity threshold set to 0.3) to create a mask that was subsequently used in the fractional anisotropy (FA) estimation. The FA-images resulting from the diffusion tensor fitting procedure were subsequently processed by FSL's tract-based spatial statistics (TBSS) pipeline (Smith et al., 2006), using the recommended parameters (i.e., non-linear registration to FSL's 1 mm FA image, construction of mean FA-image and skeletonized mean FA-image based on the data from all subjects, and a threshold of 0.2 for the skeletonized FA-mask). Subsequently, we organized the resulting skeletonized FA-maps into a 2D ($N \times K$) array of shape 217 (subjects) × 128340 (non-zero voxels).

Brain size estimation

To estimate the values for our confound, global brain size, we calculated for each subject the total number of non-zero voxels

in the gray matter and white matter map resulting from the segmentation step in the FSL-VBM pipeline (using FSL's segmentation algorithm "FAST"; Zhang et al., 2001). The number of non-zero voxels from the gray matter map was used as the confound for the VBM-based analyses and the number of non-zero voxels from the white matter map was used as the confound for the TBSS-based analyses. Note that brain size estimates from total white matter volume and total gray matter volume correlated strongly, $r(216) = 0.93$, $p < 0.001$.

Data and code availability

In the Github repository corresponding to this article (https://github.com/lukassnoek/MVCA), we included a script (`download_data.py`) to download the data (the 4D VBM and TBSS nifti-images as well as the non-zero 2D samples × features arrays). The repository also contains detailed Jupyter notebooks with the annotated empirical analyses and simulations reported in this article.

Decoding pipeline

All empirical analyses and simulations used a common decoding pipeline, implemented using functionality from the *scikit-learn* Python package for machine learning (Abraham et al., 2014; Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, et al., 2011). This pipeline included univariate feature selection (based on a prespecified amount of voxels with highest univariate difference in terms of the ANOVA F-statistic), feature-scaling (ensuring zero mean and unit standard deviation for each feature), and a support vector classifier (SVC) with a linear kernel, fixed regularization parameter ($C = 1$), and sample weights set to be inversely proportional to class frequency (to account for class im-

balance). In our empirical analyses, we evaluated model performance for different numbers of voxels (as selected by the univariate feature selection). For our empirical analyses, we report model performance as the F_1 score, which is insensitive to class imbalance (which, in addition to adjusted sample weights, prevents the classifier from learning the relative probabilities of target classes instead of representative information in the features; see also Supplementary Figure B.14 for a replication of part of the results using AUROC, another metric that is insensitive to class imbalance). At chance level classification, the F_1 score is expected to be 0.5. For our simulations, in which there is no class imbalance, we report model performance using accuracy scores. In figures showing error bars around the average model performance scores, the error bars represent 95% confidence intervals estimated using the "bias-corrected and accelerated" (BCA) bootstrap method using 10,000 bootstrap replications (Efron, 1987). For calculating BCA bootstrap confidence intervals, we used the implementation from the open source "scikits.bootstrap" Python package (https://github.com/cgevans/scikits-bootstrap). Statistical significance was calculated using non-parametric permutation tests, as implemented in scikit-learn, with 1000 permutations (Ojala & Garriga, 2010).

Evaluated methods for confound control

Post hoc counterbalancing

We implemented post hoc counterbalancing in two steps. First, to quantify the strength of relation between the confound and the target in our dataset, we estimated the point-biserial correlation coefficient between the confound, C (brain size), and the target, y (gender) across the entire dataset (including all samples $i = 1 \ldots N$). Because of both sampling noise and measurement noise, sample correlation coefficients vary around the population correlation coefficient and are thus improbable to be 0 *ex-*

actly.[10] Therefore, in the next step, we subsampled the data until the correlation coefficient between and becomes non-significant at some significance threshold α: $p(r_{Cy}) > \alpha$.

In our analyses, we used an α of 0.1. Note that this is more "strict"[11] than the conventionally used threshold ($\alpha = 0.05$), but given that decoding analyses are often more sensitive to signal in the data (whether it is confounded or true signal), we chose to err on the safe side and counterbalance the data using a relatively strict threshold of $\alpha = 0.1$.

Subsampling was done by iteratively removing samples that contribute most to the correlation between the confound and the target until the correlation becomes non-significant. In our empirical data in which brain size is positively correlated with gender (coded as male = 1, female = 0) this amounted to iteratively removing the male subject with the largest brain size and the female subject with the smallest brain size. This procedure optimally balances (1) minimizing the correlation between target and confound and (2) maximizing sample size. As an alternative to this "targeted subsampling", we additionally implemented a procedure which draws random subsamples of a given sample size until it finds a subsample with a non-significant correlation coefficient. If such a subsample cannot be found after 10,000 random draws, sample size is decreased by 1, which is repeated

[10] For continuous confounds, it is practically impossible to achieve a correlation with the target of *exactly* zero, which is the reason we subsample until it is smaller than a prespecified threshold. For categorical confounds, however, a correlation between the confound and the target of exactly zero is possible (this amounts to equal proportions of levels of c within each class of y; Görgen et al., 2017), even *necessary*, because it is impossible to find a (K-fold) cross-validation partitioning in which each split is counterbalanced w.r.t. the confound if the correlation *in the entire dataset* between the target and the confound is not zero.

[11] We refer to a relatively high α as "strict", here, because we use it here for the purpose of demonstrating no effect.

until a subsample is found. This procedure resulted in much smaller subsamples than the targeted subsampling procedure (i.e., a larger power loss) since the optimal subsample is hard to find randomly.[12] In the analyses below, therefore, we used the targeted subsampling procedure. Importantly, even with extreme power loss, random subsampling can cause the same biases as will be described for the targeted subsampling method below (cf. Figure 3.8 and Figure 3.10 and Supplementary Figures B.13 and B.14).

Then, given that the subsampled dataset is counterbalanced with respect to the confound, a random stratified K-fold cross-validation scheme is repeatedly initialized until a scheme is found in which *all* splits are counterbalanced as well (cf. Görgen et al., 2017). This particular counterbalanced cross-validation scheme is subsequently used to cross-validate the MVPA pipeline. We implemented this post hoc counterbalancing method as a scikit-learn-style cross-validator class, available from the aforementioned Github repository (in the `counterbalance.py` module).

Confound regression

In our empirical analyses and simulations, we tested two different versions of confound regression, which we call "whole-dataset confound regression" (WDCR) and "cross-validated confound regression" (CVCR). In WDCR, we regressed out the confounds from the predictors *from the entire dataset at once*, i.e., before entering the iterative cross-validated MVPA pipeline (the approach taken by Abdulkadir et al., 2014; Dubois et al., 2018; Kostro et al., 2014; Todd et al., 2013). Note that we can do

[12] One could run the "random subsampling" procedure with more than 10,000 draws in order to reduce the aforementioned power loss; but in the extreme, this would result in the same optimal subsample that can be found much faster by targeted subsampling.

this for all K voxels at once using the closed-form OLS solution, in which we first estimated the parameters $\hat{\beta}_C$:

$$\hat{\beta}_C = (C^T C)^{-1} C^T X \tag{3.5}$$

where C is an array in which the first column contained an intercept and the second column contained the confound brain size. Accordingly, $\hat{\beta}_C$ is an $2 \times K$ array. We then removed the variance associated with the confound from our neuroimaging data as follows:

$$X_{\text{corr}} = X - C\hat{\beta}_C \tag{3.6}$$

Now, X_{corr} is an array with the same shape as the original X array, but without any variance that can be explained by confound, C (i.e., X is residualized with regard to C).

In our proposed cross-validated version of confound regression (which was mentioned but not evaluated by Rao et al., 2017, p. 25), "CVCR", we similarly regressed out the confounds from the neuroimaging data, but instead of estimating $\hat{\beta}_C$ on the entire dataset, we estimated this within each fold of training data (X_{train}):

$$\hat{\beta}_{C,\text{train}} = (C_{\text{train}}^T C_{\text{train}})^{-1} C_{\text{train}}^T X_{\text{train}} \tag{3.7}$$

And we subsequently used these parameters ($\hat{\beta}_{C,\text{train}}$) to remove the variance related to the confound from both the train set (X_{train} and C_{train}):

$$X_{\text{train,corr}} = X_{\text{train}} - C_{\text{train}}\hat{\beta}_{C,\text{train}} \tag{3.8}$$

and the test test (X_{test} and C_{test}):

$$X_{\text{test,corr}} = X_{\text{test}} - C_{\text{test}}\hat{\beta}_{C,\text{test}} \tag{3.9}$$

Thus, essentially, CVCR is the cross-validated version of WDCR. One might argue that regressing the confound from the train set only, i.e., implementing only equation (3.8), not equation (3.9), is sufficient to control for confounds as it prevents the decoding model from relying on signal related to the confound. We evaluated this method and report the corresponding results in Supplementary Figure B.10.

We implemented these confound regression techniques as a *scikit-learn* compatible transformer class, available in the open-source *skbold* Python package (https://github.com/lukassnoek/skbold) and in the aforementioned Github repository.

Control for confounds during pattern estimation

In addition to post hoc counterbalancing and confound regression, we also evaluated how well the "control for confounds during pattern estimation" method controls for the influence of confounds in decoding analyses of (simulated) fMRI data. The simulation methods and results can be found in the Supplementary Materials.

Analyses of simulated data

In addition to the empirical evaluation of counterbalancing and confound regression in the gender decoding example, we ran three additional analyses on simulated data. First, we investigated the efficacy of the three confound control methods on synthetic data with known quantities of "true signal" and "confounded signal", in order to detect potential biases. Second, we ran additional analyses on simulated data to investigate the positive bias in model performance observed after post hoc counterbalancing. Third, we ran additional analyses on simulated data

to investigate the negative bias in model performance observed after WDCR. In the Supplementary Materials, we investigate whether the confound regression results generalize to (simulated) functional MRI data (Supplementary Figure B.1 and B.2).

Efficacy analyses

In this simulation, we evaluated the efficacy of the three methods for confound control on synthetic data with a prespecified correlation between the confound and the target, r_{Cy}, and varying amounts of "confounded signal" (i.e., the explained variance in y driven by shared variance between X and y). These simulations allowed us to have full control over (and knowledge of) the influence of both signal and confound in the data, and thereby help us diagnose biases associated with post hoc counterbalancing and confound regression.

Specifically, in this efficacy analysis, we generated hypothetical data sets holding the correlation coefficient between C and y constant, while varying the amount of true signal and confounded signal. We operationalized true signal as the squared semipartial Pearson correlation between y and each feature in X, controlled for C. As such, we will refer to this term as signal R^2:

$$\text{signal } R^2 = r^2_{y(X.C)} \tag{3.10}$$

In the simulations reported and shown in the main article, we used $r_{Cy} = 0.65$, which corresponds to the observed correlation between brain size and gender in our dataset. To generate synthetic data with this prespecified structure, we generated (1) a data matrix X of shape $N \times K$, (2) a target variable y of shape $N \times 1$, and (3) a confound variable C of shape $N \times P$. For all simulations, we used the following parameters: $N = 200$, $K = 5$, and $P = 1$ (i.e., a single confound variable). We generated y as a categorical variable with binary values, $y \in \{0, 1\}$, with equal class

probabilities (i.e., 50%), given that most decoding studies focus on binary classification. We generated C as a continuous random variable drawn from a standard normal distribution. We generated each feature X_j as a linear combination of y and C plus Gaussian noise. Thus, for each predictor $j = 1 \ldots K$ in X_j:

$$X_j = \beta_y y + \beta_C C + \varepsilon, \varepsilon \sim \mathcal{N}(0, \gamma) \tag{3.11}$$

in which β_y represented the weight given to y, and β_C represented the weight given to C in the generation of the feature X_j, and $\mathcal{N}(0, \gamma)$ is the normal distribution with zero mean and standard deviation γ. The parameters β_y and β_C were both initialized with a value of 1. First, if the difference between the total variance explained and the sum of the desired signal R^2 and confound R^2 values was larger than 0.01, the standard deviation of the normal distribution from which the errors were drawn (i.e., γ) was adjusted (decreased with 0.01 when the total R^2 is too low, increased with 0.01 when the total R^2 is too high), after which was generated again. This process was iterated until the target total R^2 value is found. Then, the total variance explained was partitioned into confound R^2 and signal R^2. If one or both of these values differed from the targeted values by more than 0.01, the generative parameters β_y and β_C were adjusted: if signal R^2 is too low, was increased with 0.01, and decreased with 0.01 otherwise. If confound R^2 is too low, β_C was increased with 0.01, and decreased with 0.01 otherwise. After adjusting these parameters, X_j was generated again. This process was iterated until the data contain the desired "true signal" and "confounded signal".

We evaluated the different methods for confound control for two values of signal R^2 (0.004, representing plausible null data,[13] and

[13]Note that plausible null data do not reflect a signal R^2 of 0, because this statistic is biased towards values larger than 0 (because it represents a squared number) when dealing with noisy data, hence our choice of signal $R^2 = 0.004$.

0.1, representing a plausible true effect) and a range of confound R^2 values (in steps of 0.05: $0.00, 0.05, 0.10, \ldots, 0.35$). This simulation was iterated 10 times (with different partitions of the folds) to ensure the results were not influenced by random noise. Importantly, the specific scenario in which confound R^2 equals 0, which represents data without any confounded signal (r^2_{yX}), served as "reference model performance" to which we can compare the efficacy the confound control methods. This comparison allowed us to detect potential biases.

After the data were generated, a baseline model (no confound control) and the three methods outlined above (post hoc counterbalancing, WDCR, and CVCR) were applied to the simulated data using the standard pipeline described in the Decoding pipeline section (but without univariate feature selection) and compared to the reference performance.

Analysis of positive bias after post hoc counterbalancing

As detailed below, post hoc counterbalancing did not lead to the expected decrease in model performance; instead, there appeared to be a trend towards an *increase* in model performance. To further investigate the cause of this unexpected result, we simulated a multivariate normal dataset with three variables, reflecting our data (X), target (y), and confound (C), with 1000 samples (N) and a single feature ($K = 1$). We iterated this data generation process 1000 times and subsequently selected the dataset which yielded the largest (positive) difference between model performance after post hoc counterbalancing versus no confound control. In other words, we used the dataset in which the counterbalancing issue was most apparent. While not necessarily representative of typical (neuroimaging) datasets, this process allowed us to explain and visualize how it is possible that model performance increases after counterbalancing the data.

To generate data from a multivariate normal distribution, we first generated variance-covariance matrices with unit variance for all variables, so that covariances can be interpreted as correlations. The covariances in the matrix were generated as pairwise correlations (r_{yX}, r_{Cy}, $r_{C}X$), each sampled from a uniform distribution with range $[-0.65, 0.65]$. We generated data using such prespecified correlation structure because the relative increase in model performance after counterbalancing did not appear to occur when generating completely random (normally distributed) data. Moreover, we restricted the range of the uniform distribution from which the pairwise correlations are drawn to $[-0.65, 0.65]$ because a larger range can result in covariance matrices that are not positive-semidefinite. After generating the three variables, we binarized the target variable (y) using a mean-split ($y = 0$ if $y < \bar{y}$, $y = 1$ otherwise) to frame the analysis as a classification problem rather than a regression problem.

We then subsampled the selected dataset using our post hoc counterbalancing algorithm and subsequently ran the decoding pipeline (without univariate feature selection) on the subsampled ("retained") data in a 10-fold stratified cross-validation scheme. Notably, we cross-validated our fitted pipeline not only to the left-out *retained* data, but also to the data that did not survive the subsampling procedure (the *rejected* data; see Figure 3.3). Across the 10 folds, we kept track of two statistics from the retained and rejected samples: (1) the classification performance, and (2) the signed distance to the decision boundary. Negative distances in binary classification (in simple binary classification with $y \in \{0, 1\}$) reflect a prediction of the sample as $y = 0$, while positive distances reflect a prediction of the sample as $y = 1$. As such, a correctly classified sample of class 0 has a negative distance from the decision boundary, while a correctly classified sample of class 1 has a positive distance from the decision boundary. Here, however, we wanted to count the distance

of samples that are on the "incorrect" side of the decision boundary as *negative* distances, while counting the distance of samples that are on the "correct" side of the decision boundary as positive distances. To this end, we used a "re-coded" version of the target variable ($y^* = -1$ if $y = 0$, $y^* = 1$ otherwise) and multiplied it with the distance. Consequently, negative distances of *correct* samples of condition 0 become positive and positive distances of *incorrect* samples of condition 0 become negative (by multiplying them by -1). As such, we calculated the signed distance from the decision boundary (δ_i) for any sample i as:

$$\delta_i = y^*(w^T X_i + b) \tag{3.12}$$

in which w refers to the feature weights (coefficients) and b refers to the intercept term. Any differences in these two statistics (proportion correctly classified and signed distance to the classification boundary) between the retained and rejected samples may signify biases in model performance estimates (i.e., better cross-validated model performance on the retained data than on the rejected data would confirm positive bias, as it indicates that subsampling tends to reject hard-to-classify samples). We applied this analysis also to the empirical data (separately for the different values of K) to show that the effect of counterbalancing, as demonstrated using simulated data, also occurs in the empirical data.

Analysis of negative bias after WDCR

As also detailed below, WDCR can lead to significantly below chance accuracy. To investigate the cause of this below chance performance (and to demonstrate that CVCR does not lead to such results), we performed two follow-up simulations. The first follow-up simulation shows that the occurrence of below chance

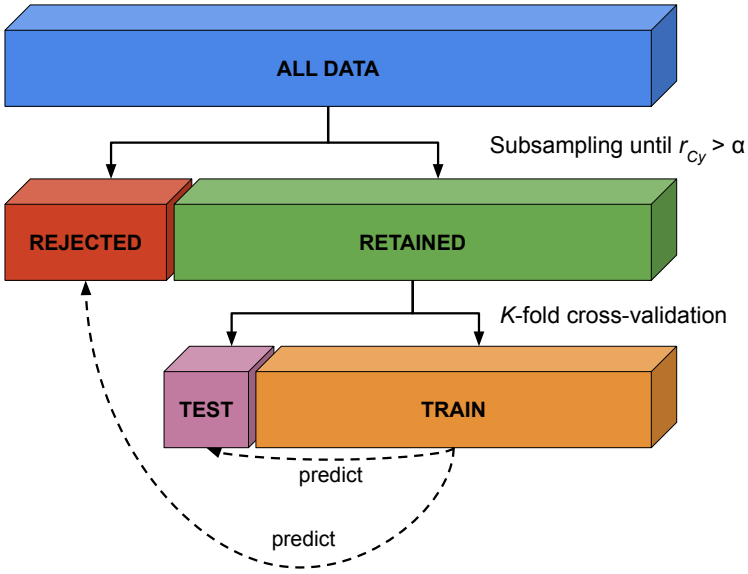


FIGURE 3.3 Visualization of method to evaluate whether counterbalancing yields unbiased cross-validated model performance estimates.

accuracy depends on the distribution of feature-target correlations (r_{yX}; see for a similar argument Jamalabadi et al., 2016), and the second follow-up simulation shows that WDCR artificially narrows this distribution. This artificial narrowing of the distribution is exacerbated both by an increasing number of features (K), as well as higher correlations between the target and confound (r_{Cy}).

In the first simulation, we simulated random null data (drawn from a standard normal distribution) with 100 samples (N) and 200 features (K), as well as a binary target feature ($y \in \{0, 1\}$). We then calculated the cross-validated prediction accuracy using the standard pipeline (without univariate feature selection) described in the Decoding pipeline section; we iterate this pro-

cess 500 times. Then, we show that the variance of the cross-validated accuracy is accurately predicted by the standard deviation (i.e., "width") of the distribution of correlations between the features and the target (r_{yX_j} with $j = 1 \ldots K$), which we will denote by $sd(r_{yX})$. Importantly, we show that below chance accuracy likely occurs when the standard deviation of the feature-target correlation distribution is lower than the standard deviation of the sampling distribution of the Pearson correlation coefficient parameterized with the same number of samples ($N = 200$) and the same effect size (i.e., $\rho = 0$, because we simulated random null data). The sampling distribution of the Pearson correlation coefficient is described by Kendall & Stuart (1973). When $\rho = 0$ (as in our simulations), the equation is as follows:

$$f(r; N) = (1 - r^2)^{\frac{N-4}{2}} [\mathcal{B}](\frac{1}{2}, \frac{N-2}{2})^{-1} \tag{3.13}$$

where $\mathcal{B}(a, b)$ represents the Beta-function.

Then, in a second simulation, we similarly simulated null data as in the previous simulation, but now we also generate a continuous confound (C) with a varying correlation with the target ($r_{Cy} \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$). Before subjecting the data to the decoding pipeline, we regressed out the confound from the data (i.e., WDCR). We did this for different numbers of features ($K \in \{1, 5, 10, 50, 100, 500, 1000\}$). Then, we applied CVCR on the simulated data as well for comparison.

3.3 Results

Influence of brain size

Before evaluating the different methods for confound control, we determined whether brain size is truly a confound given our

proposed definition ("a variable that is not of primary interest, correlates with the target, and is encoded in the neuroimaging data"). We evaluated the relationship between the target and the confound in two ways. First, we calculated the (point-biserial) correlation between gender and brain size, which was significant for both the estimation based on white matter, $r(216) = .645, p < 0.001$, and the estimation based on grey matter, $r(216) = .588, p < 0.001$, corroborating the findings by Smith & Nichols (2018). Second, as recommended by Görgen et al. (2017), who argue that the potential influence of confounds can be discovered by running a classification analysis using the confound as the (single) feature predicting the target, we ran our decoding pipeline (without univariate feature selection) using brain size as a single feature to predict gender. This analysis yielded a mean classification performance (F_1 score) of 0.78 (*SD* = .10) when using brain size estimated from white matter and 0.81 (*SD* = .09) when using brain size estimated from gray matter, which are both significant with $p < 0.001$ (see Figure 3.4A).

To estimate whether brain size is encoded in the neuroimaging data, we compared the distribution of bivariate correlation coefficients (of each voxel with brain size) with the sampling distribution of correlation coefficients when $\rho = 0$ and $N = 217$ (see section Analysis of negative bias after WDCR for details). Under the null hypothesis that there are no correlations between brain size and voxel intensities, each individual correlation coefficient between a voxel and the confound can be regarded as an independent sample with $N = 217$ (ignoring correlations between voxels for simplicity). Because K is very large for both the VBM and TBSS data, the empirical distribution of correlation coefficients should, under the null hypothesis, approach the analytic distribution of correlation coefficients parametrized by $\rho = 0$ and $N = 217$. Contrarily, the density plots in Fig. 3.4B clearly show that the observed correlation coefficients distribution does not follow the sampling distribution (with both an in-

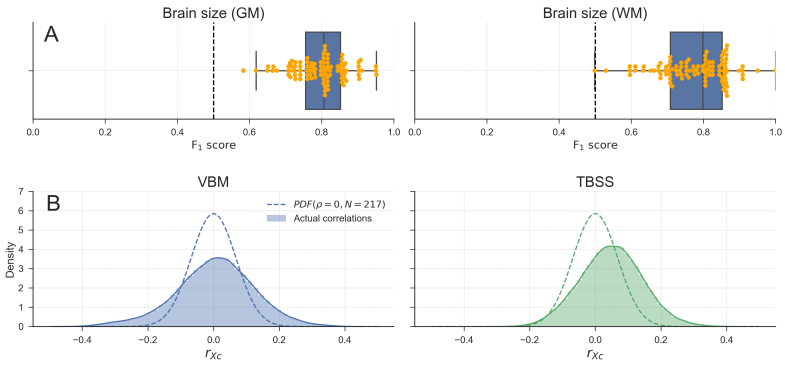


FIGURE 3.4 A) Model performance when using brain size to predict gender for both brain-size estimated from grey matter (left) and from white matter (right). Points in yellow depict individual F_1 scores per fold in the 10-fold cross-validation scheme. Whiskers of the box plot are 1.5x the interquartile range. B) Distributions of observed correlations between brain size and voxels (r_{XC}), overlayed with the analytic sampling distribution of correlation coefficients when $\rho = 0$ and $N = 217$, for both the VBM data (left) and TBSS data (right). Density estimates are obtained by kernel density estimation with a Gaussian kernel and Scott's rule (Scott, 1979) for bandwidth selection.

crease in variance and a shift of the mode). This indicates that at least some of the correlation coefficients between voxel intensities and brain size are extremely unlikely under the null hypothesis. Note that this interpretation is contingent on the assumption that the relation between brain size and VBM/TBSS data is linear. In the Supplementary Materials and Results (Supplementary Figures B.7-B.9), we provide some evidence for the validity of this assumption.

Baseline model: no confound control

In our baseline model on the empirical data, for different numbers of voxels, we predicted gender from structural MRI data

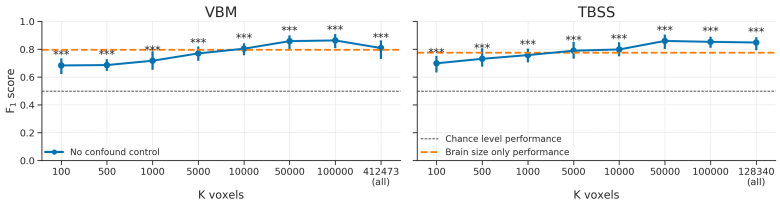


FIGURE 3.5 Baseline scores using the VBM (left) and TBSS (right) data without any confound control. Scores reflect the average F_1 score across 10 folds; error bars reflect 95% confidence intervals. The dashed black line reflect theoretical chance-level performance and the dashed orange line reflects the average model performance when only brain size is used as a predictor for reference; Asterisks indicates significant performance above chance: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$.

(VBM and TBSS) without controlling for brain size (see Figure 3.5). The results show significant above chance performance of the MVPA pipeline based on both the VBM data and the TBSS data. All performance scores averaged across folds were significant ($p < 0.001$).

These above chance performance estimates replicate previous studies on gender decoding using structural MRI data (Del Giudice et al., 2016; Rosenblatt, 2016; Sepehrband et al., 2018) and will serve as a baseline estimate of model performance to which the confound control methods will be compared.

In the next three subsections, we will report the results from the three discussed methods to control for confounds: post hoc counterbalancing, whole-dataset confound regression (WDCR), and cross-validated confound regression (CVCR).

Post hoc counterbalancing

Empirical results

In order to decorrelate brain size and gender (i.e., $r_{Cy} > 0.1$), our subsampling algorithm selected 117 samples in the VBM data

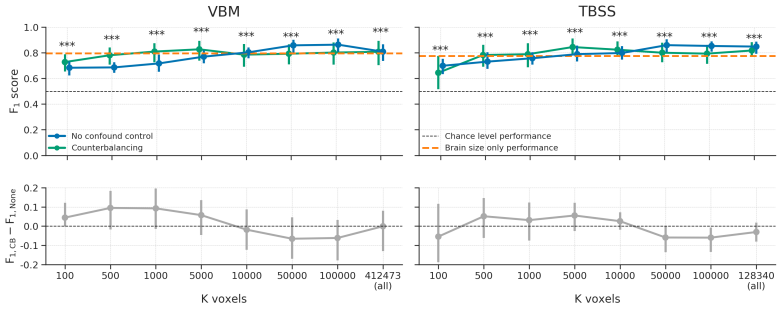


FIGURE 3.6 Model performance after counterbalancing (green) versus the baseline performance (blue) for both the VBM (left) and TBSS (right) data (upper row) and the difference in performance between the methods (lower row). Performance reflects the average (difference) F_1 score across 10 folds; error bars reflect 95% confidence intervals. The dashed black line reflect theoretical chance-level performance (0.5) and the dashed orange line reflects the average model performance when only brain size is used as a predictor. Asterisks indicates significant performance above chance: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$.

(i.e., a sample size reduction of 46.1%) and 131 samples in the TBSS data (i.e., a reduction of 39.6%). The model performance for different values of (number of voxels) are shown in Figure 3.6. Contrary to our expectations, the predictive accuracy of our decoding pipeline after counterbalancing was similar to baseline performance. This is particularly surprising in light of the large reductions in sample size, which results in a substantial loss in power, which in turn is expected to lead to lower model performance.

One could argue that the lack of expected decrease in model performance after counterbalancing can be explained by the possibility that the subsampling and counterbalancing procedure just leads to the selection of different features during univariate feature selection compared to the baseline model. In other words,

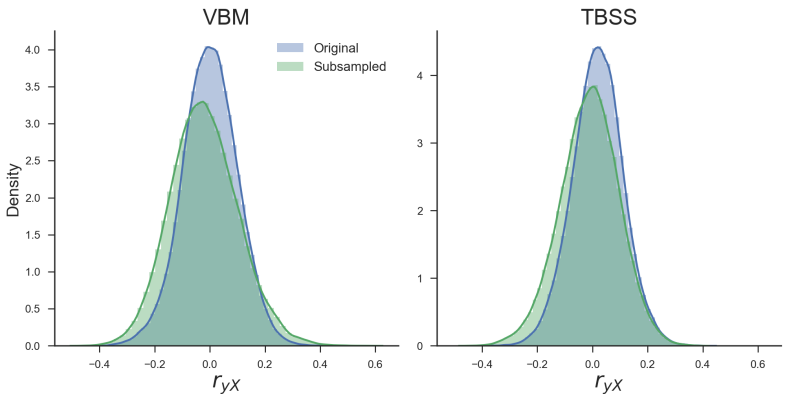


FIGURE 3.7 Density plots of the correlations between the target and voxels across all voxels before (blue) and after (green) subsampling for both the VBM and TBSS data. Density estimates are obtained by kernel density estimation with a Gaussian kernel and Scott's rule (Scott, 1979) for bandwidth selection.

the increase in model performance may be caused by the feature selection function, which selects "better" voxels (i.e., containing more "robust" signal), resulting in similar model performance in spite of the reduction in sample size. However, this does not explain the similar scores for counterbalancing and the baseline model when using all voxels (the data points at $K\text{voxels} = \ldots(\text{all})$ in Figure 3.6). Another possibility for the relative increase in model performance based on the counterbalanced data versus the baseline model is that counterbalancing increased the amount of signal in the data. Indeed, counterbalancing appeared to increase the (absolute) correlations between the data and the target (r_{yX}), which is visualized in Figure 3.7, suggesting an increase in signal.

This apparent increase in the correlations between the target and neuroimaging data goes against the intuition that removing the influence of a confound that is highly correlated with the target

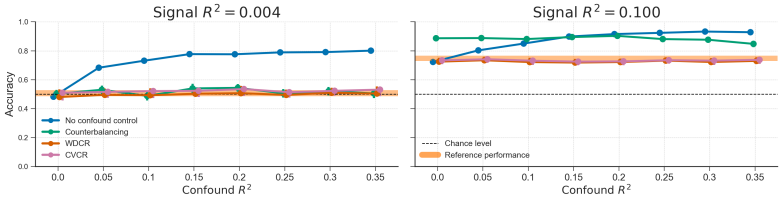


FIGURE 3.8 Results from the different confound control methods on simulated data without any experimental effect (signal $R^2 = 0.004$; left graph) and with some experimental effect (signal $R^2 = 0.1$; right graph) for different values of confound R^2. The orange line represents the average performance (±1 SD) when confound $R^2 = 0$, which serves as a "reference performance" for when there is no confounded signal in the data. For both graphs, the correlation between the target and the confound, r_{yC}, is fixed at 0.65. The results from the WDCR and CVCR methods are explained later.

will reduce decoding performance. To further investigate this, we replicated this effect of post hoc counterbalancing on simulated data, as described in the next section (Efficacy analyses), and additionally investigated the cause of the negative bias observed after WDCR using a separate set of simulations.

Efficacy analysis

To evaluate the efficacy of the three confound control methods, we simulated data in which we varied the strength of confound R^2 and signal R^2, after which we applied the three confound control methods to the data. The results of this analysis show that counterbalancing maintains chance-level model performance when there is almost no signal in the data (i.e., signal $R^2 = 0.004$; Figure 3.8, left graph, green line). However, when there is some signal (i.e., signal $R^2 = 0.1$; Fig. 8, right graph), we observed that counterbalancing yields similar or even higher scores than the baseline model, replicating the effects observed in the empirical analyses.

As is apparent from Figure 3.8 (right panel), when there is some signal, the counterbalanced data seem to yield better performance than the baseline model only for relatively low confound R^2 values (confound $R^2 < 0.15$). As suggested by our findings in the empirical data (see Figure @(ref:fig-confounds-decoding-7)), we hypothesized that the observed improvement in model performance after counterbalancing was caused by the increase in correlations between the target and features in the neuroimaging data. In support of this hypothesis, Figure 3.9 illustrates the relations between the strength of the confound (confound R^2, color coded), the increase in correlations after post hoc counterbalancing ($\delta r_{yX} = r_{yX}^{\text{after}} - r_{yX}^{\text{after}}$; x-axis) for each confound R^2, and the resulting difference in model performance ($\text{ACC}_{\text{CB}} - \text{ACC}_{\text{baseline}}$; y-axis). The figure shows that the increase or decrease in accuracy after counterbalancing (compared to baseline) depends on δr_{yX} ($r(79) = .922$, $p < 0.001$), which in turn depends on confound R^2 ($r(79) = -0.987$, $p < 0.001$). To reiterate, these differences in model performance are only due to the post hoc counterbalancing procedure and not due to varying signal in the simulated data. The effect of post hoc counterbalancing on model performance thus seems to depend on the strength of the confound.

While this relationship in Figure 3.9 might be statistically interesting, it does not explain why post hoc counterbalancing tends to increase the correlations between neuroimaging data and target, and even outperforms the baseline model when confound R^2 is low and some signal is present. More importantly, it does not tell us whether the post hoc counterbalancing procedure uncovers signal that is truly related to the target — in which case the procedure suppresses noise — or inflates performance estimates and thereby introduces positive bias. Therefore, in the next section, we report and discuss results from a follow-up simulation that intuitively shows why post hoc counterbalancing leads to

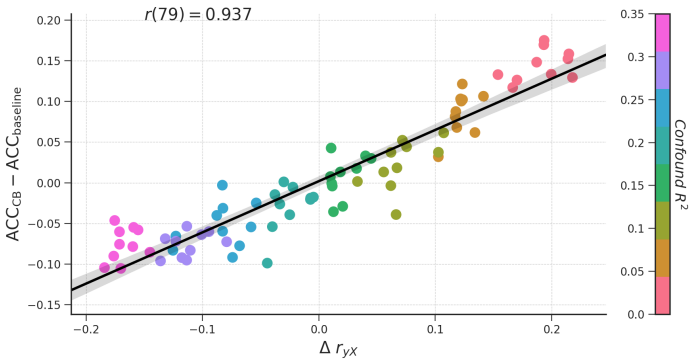


FIGURE 3.9 The relationship between the increase in correlations between target and data (r_{yX}) after subsampling, confound R^2, difference in model performance (here: accuracy) between the counterbalance model and baseline model ($\mathrm{ACC_{CB}} - \mathrm{ACC_{baseline}}$).

an increase in performance, and furthermore shows that this increase is in fact a positive bias.

Analysis of positive bias after post hoc counterbalancing

With this follow-up analysis, we aimed to visualize the scenario in which post hoc counterbalancing leads to a clearly better performance than model performance without confound control. As such, we generated 1000 data sets using a covariance matrix that we knew leads to a large difference between the baseline model and model performance after counterbalancing (i.e., data with a low confound R^2). From these 1000 datasets, we selected the dataset that yielded the largest difference for our visualization (see the Analysis of positive bias after post hoc counterbalancing section in the Methods for details).

The data that yielded the largest difference (i.e., a performance increase from 0.613 to 0.804, a 31% increase) are visualized in Figure 3.10. Each sample's confound value (C) is plotted against

its feature value (X), both before subsampling (upper scatter plot) and after subsampling (lower scatter plot). From visual inspection, it appears that the samples rejected by the subsampling procedure (i.e., the samples with the white border) have relatively large absolute values of the confound variable, which tend to lie close to or on the "wrong" side of the classification boundary (i.e., the dashed black line) in this specific configuration of the data. In other words, subsampling seems to reject samples that are harder to classify or would be incorrectly classified based on the data (here, the single feature of X). The density plots in Figure 3.10 show the same effect in a different way: while the difference in the modes of the distributions of the confound (C) between classes is reduced after subsampling (i.e., the density plots parallel to the y-axis), the difference in the modes of the distributions of the data (X) between classes is actually increased after subsampling (i.e., the density plots parallel to the x-axis).

We quantified this effect of subsampling by comparing the signed distance from the decision boundary (i.e., the dashed line in the upper scatter plot) between the retained samples and the rejected (subsampled) samples, in which a larger distance from the decision boundary reflects a higher confidence of the classifier's prediction (see Figure 3.3 for a visualization of this method). Indeed, we found that samples that are removed by subsampling lie significantly closer to (or on the "wrong" side of) the decision boundary (M = -.358, SD = .619) than samples that are retained after subsampling (M = .506, SD = .580), as indicated by a independent samples t-test, $t(998) = 22.32$, $p < 0.001$. Also (which follows from the previous observation), samples that would have been removed by subsampling are more often classified incorrectly (75% incorrect) than the samples that would have been retained by subsampling (20% incorrect), as indicated by a chi-squared test, $\chi^2 = 270.29$, $p < 0.001$.

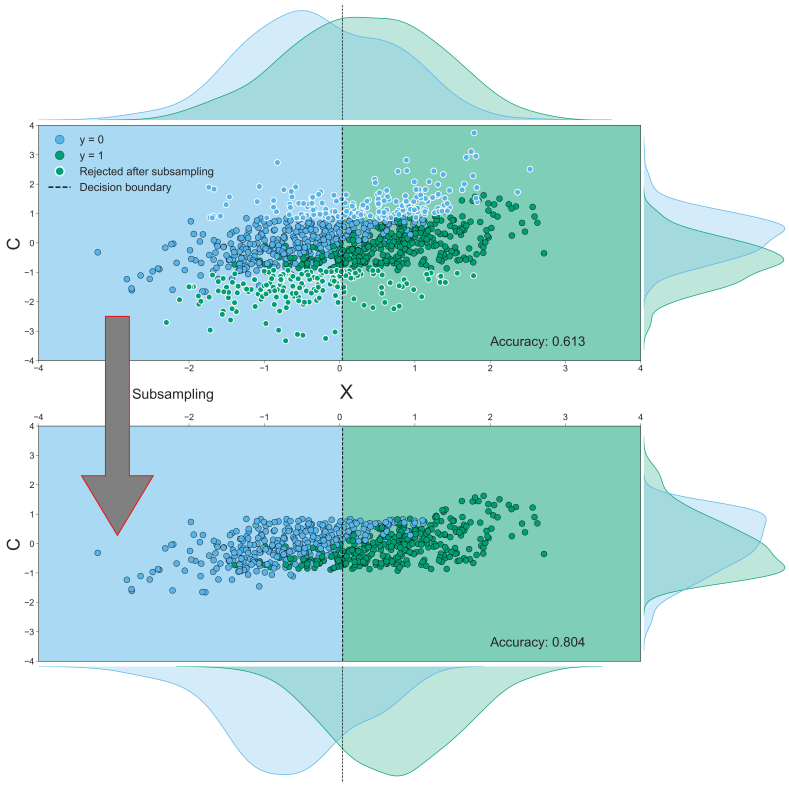


FIGURE 3.10 Both scatterplots visualize the relationship between the data (X with $K = 1$, on the x-axis), the confound (C, on the y-axis) and the target (y). Dots with a white border in the upper scatterplot indicate samples that are rejected in the subsampling process; the lower scatterplot visualizes the data without these rejected samples. The dashed black lines in the scatterplot represent the decision boundary of the SVM classifier; the color of the background shows how samples in that area are classified (a blue background means a prediction of $y = 0$ and a green background means a prediction of $y = 1$). The density plots parallel to the y-axis depict the distribution of the confound (C) for the samples in which $y = 0$ (blue) and in which $y = 1$ (green). The density plots parallel to x-axis depict the distribution of the data (X) for the samples in which $y = 0$ (blue) and in which $y = 1$ (green). Density estimates are obtained by kernel density estimation with a Gaussian kernel and Scott's rule (Scott, 1979) for bandwidth selection.

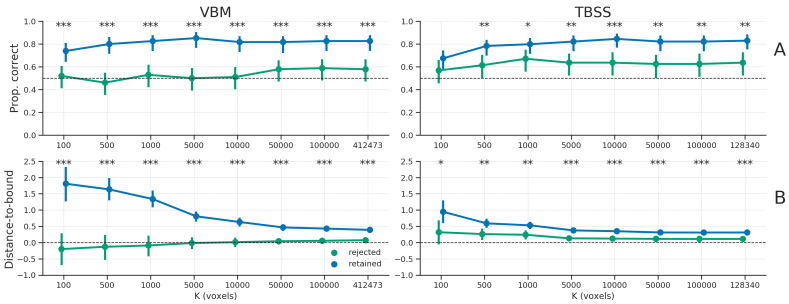


FIGURE 3.11 **A**) The proportion of samples classified correctly, separately for the "retained" samples (blue line) and "rejected" samples (green line); the dashed line represents chance level (0.5). **B**) The average distance to the classification boundary for the retained and rejected samples; the dashed line represents the decision boundary, with values below the line representing samples on the "wrong" side of the boundary (and vice versa). Asterisks indicates a significant difference between the retained and rejected samples: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$.

To show that the same effect (i.e., removing samples that tend to be hard to classify or would be wrongly classified) occurred in the empirical data after counterbalancing as well, we applied the same analysis of comparing model performance and distance to boundary between the retained and rejected samples to the empirical data. Indeed, across all different numbers of voxels (K), the retained samples were significantly more often classified correctly (Figure 3.11A) and had a significantly larger distance to the classification boundary (Figure 3.11B) than the rejected samples. This demonstrates that the same effect of post hoc counterbalancing, as shown in the simulated data, likely underlies the increase in model performance of the counterbalanced data relative to the baseline model in the empirical data.

One can wonder how much the occurrence of these observed biases in post hoc counterbalancing depends on the specific method of subsampling used. Random subsampling led to qual-

itatively similar results as targeted subsampling (cf. Supplementary Figures B.13 and B.14 with random subsampling). Instead, the bias is introduced through features that weakly correlate with the target in the whole sample, but strongly in subsamples where there is no correlation between target and the confound (features which, as our results show, exist in the neuroimaging data). That is, the bias is an indirect result of decorrelating target and confound in the sample, which is an essential step in post hoc counterbalancing (in fact, it is the *goal* of counterbalancing). For this reason, we consider it unlikely (but not impossible) that there exists a way to subsample data without introducing biases.

In summary, removing a subset of observations to correct for the influence of a confound can induce substantial bias by removing samples that are harder to classify using the available data. The bias itself can be subtle (e.g., in our empirical results, the predictive performance falls in a realistic range of predictive performances), and could remain undetected when present. Therefore, we believe that post hoc counterbalancing by subsampling the data is an inappropriate method to control for confounds.

Whole-dataset confound regression (WDCR)

Empirical results

In addition to post hoc counterbalancing, we evaluated the efficacy of "whole-dataset confound regression" (WDCR), i.e. regressing out the confound from each feature separately using all samples from the dataset to control for confounds. Compared to the baseline model, WDCR yielded a strong decrease in performance, even dropping (significantly) below chance for all TBSS analyses and a subset of the VBM analyses (see Figure 3.12).

This strong (and implausible) reduction in model performance after WDCR is investigated in more detail in the next two sections on the results from the simulations.

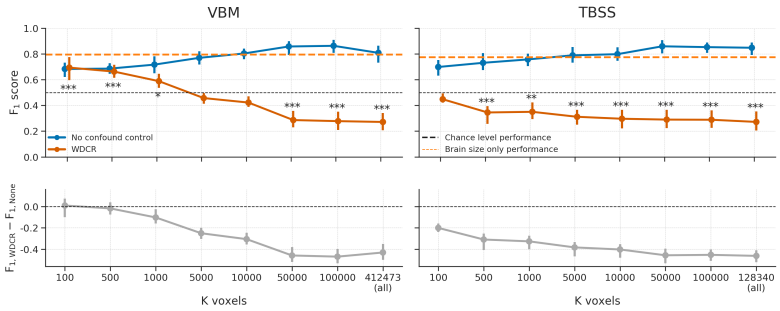


FIGURE 3.12 Model performance after WDCR (orange) versus the baseline performance (blue) for both the VBM (left) and TBSS (right) data. Performance reflects the average F_1 score across 10 folds; error bars reflect 95% confidence intervals. The dashed black line reflect theoretical chance-level performance (0.5) and the dashed orange line reflects the average model performance when only brain size is used as a predictor. Asterisks indicates performance of the WDCR model that is significantly above or below chance: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$.

Efficacy analysis

The results from the analyses investigating the efficacy of the confound control methods (see Figure 3.8) show that WDCR accurately corrects for the confound in both in data without signal (i.e., when signal $R^2 = 0.004$) and in data with some signal (i.e., when signal $R^2 = 0.1$), as evident from the fact that the performance after WDCR is similar to the reference performance. This result (i.e., plausible performance after confound control) stands in contrast to the results from the empirical analyses, which is why we ran a follow-up analysis on simulated data to investigate this specific issue.

Analysis of negative bias after WDCR

Inspired by the work of Jamalabadi et al. (2016) on below chance accuracy in decoding analyses, we ran several follow-up analyses to get insight into why WDCR leads to below chance model

performance. As Jamalabadi et al. show, below chance model performance occurs when the data contain little signal. In our first follow-up simulation, we sought to refine the explanation of the cause of below chance model performance by linking it to the observed standard deviation of the empirical distribution of correlations between the data (X) and the target (y). To do so, we simulated random data (X) and a binary target ($y \in \{0,1\}$) and estimated (per fold) the cross-validated classification accuracy using the standard pipeline described in the methods section. We repeated this process 500 times, yielding 500 data sets. The expected *average* predictive accuracy for each dataset is 0.5, but this varies randomly across folds and iterations. We hypothesized that this variance can be explained by the standard deviation ("width") of the initial feature-target correlation distribution, $sd(r_{Xy})$: narrower distributions may yield relatively lower cross-validated classification accuracy than relatively wider feature-target correlation distributions. Indeed, we find that the initial standard deviation of this distribution is significantly correlated with the cross-validated accuracy, $r(499) = 0.73$, $p < 0.001$ (Figure 3.13A). Importantly, we find that this relationship holds for different values of N (see Supplementary Figure B.15, for different sizes of the test set (see Supplementary Figure B.16), and for different sizes of K (see Supplementary Figure B.17).

This observation, then, begs the question: *why* do narrower-than-chance correlation distributions lead to below chance accuracy? One potential explanation of below chance accuracy is that the classifier may learn a particular (linear) relationship between features and the target in the train set (e.g., $r_{Xy} = 0.05$), while the sign of this relationship is "flipped" in the test set (e.g., $r_{Xy} = -0.03$; see Jamalabadi et al., 2016), which is known in the machine learning literature as "dataset shift" (Quionero-Candela et al., 2009). This situation would lead classifiers to predict the exact opposite classes for samples in the test set, leading

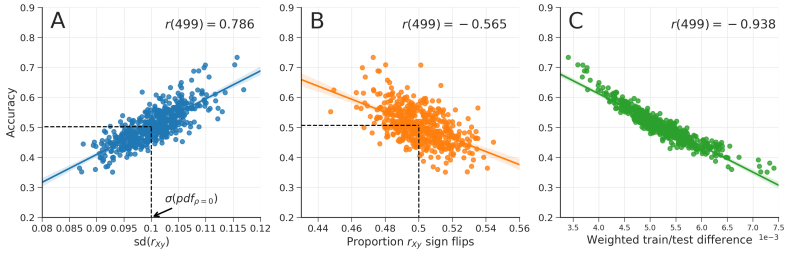


FIGURE 3.13 **A**) The relationship between the standard deviation of the distribution of feature-target correlations, $sd(r_{yX})$, and accuracy across iterations of cross-validated classification analyses of null data. The vertical dashed line represents the standard deviation from the sampling distribution parameterized with $\rho = 0$ and $N = 100$ (i.e., the same parameters used to generate the null data); the horizontal dashed line represents the expected accuracy for data with this standard deviation based on the regression line estimated from the data across simulations (see Supplementary Figure B.15 for the same plot with different values for N). **B**) The relationship between the proportion of features of which the sign of their correlation with the target (r_{Xy}) "flips" between the train-set and the test-set and accuracy. The vertical dashed line represents a proportion of 0.5., i.e., 50% of the features flip their correlation sign, which corresponds approximately with an accuracy of 0.5. **C**) The relationship between the weighted difference between feature-target correlations in the train and test set (see equation (3.14)) and accuracy.

to below chance accuracy. In the results of our simulated data, the standard deviation of the feature-target distribution was indeed significantly negatively correlated with the proportion of features that flipped the sign of their correlation between the train set and test set, $r(499) = -.687$, $p < 0.001$. This means that a higher density of feature-target correlations around 0 (i.e., a narrower width of the corresponding distribution) leads to more "sign flips". This phenomenon of "sign flipping" has been reported before in the context of (a priori) counterbalancing of categorical variables (X) with respect to the target (y), where it

was observed that complete counterbalancing led to consistent "sign flipping" and consequently 0% accuracy (Görgen et al., 2017). Similarly, we found that the proportion of features that flip sign was significantly negatively correlated with accuracy, $r = -.565$, $p < 0.001$, indicating that larger proportions of features that flip sign leads to lower accuracy (see Figure 3.13B). Interestingly, at a proportion of 0.5, accuracy is approximately at chance level (0.5; dashed lines in Figure 3.13B).

This relationship between "sign flipping" and accuracy, however, leaves room for improvement in terms of explaining the variance of accuracy scores. Therefore, we sought to further refine our "model" of accuracy by defining dataset shift not by the proportion of sign flips, but by the average *difference* between the feature-target correlations between the train set and test set. Moreover, because not all features contribute equally strongly to a classifier's prediction (i.e., they are weighted), we furthermore weighed each feature's "shift" by the associated classifier weight (w_j). Formally, we estimated dataset shift ($\hat{ds}$) thus as follows:

$$\hat{ds} = \frac{1}{K} \sum_{j=1}^{K} (r_{X_{j,\text{train}}, y_{\text{train}}} - r_{X_{j,\text{test}}, y_{\text{test}}}) w_j \tag{3.14}$$

Indeed, the correlation between this particular operationalization of "dataset shift" and accuracy across simulations was much higher than just the proportion of sign flips, $r(499) = 0.934$ (Figure 3.13B).

Having established the relation between the standard deviation of the initial feature-target correlation distribution and accuracy, we followed up our simulation by investigating specifically the effect of WDCR on the standard deviation of the correlation distribution. We investigated this by simulating data with different strengths of the correlation between the confound and the target (r_{Cy}) and the number of features (K). From Figure 3.14A,

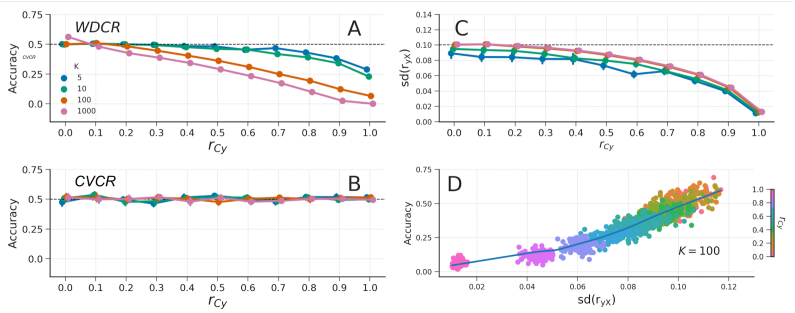


FIGURE 3.14 **A**) The effect of WDCR on data varying in the correlation of the confound with the target (r_{Cy}; x-axis) and the number of features (K; different lines). **B**) The effect of CVCR on data varying in the correlation of the confound with the target and the number of features. The dashed black line represents chance model performance in subplots A and B. **C**) The relation between the correlation of the confound with the target (r_{Cy}) and the standard deviation of the feature-target correlation distribution, $sd(r_{yX})$ for the WDCR data. The dashed black line represents the standard deviation of the correlation distribution predicted by the sampling distribution. **D**) The relation of the standard deviation of the correlation distribution and accuracy for the WDCR data (only shown for the data when $K = 100$; see Supplementary Figure B.18 for visualizations of this effect for different values of K). The data depicted in all panels are null data.

it is clear that, while the expected chance level is 0.5 in all cases, model performance quickly drops below chance for increasing correlations between the target and the confound, as well as for increasing numbers of features; even leading to a model performance of 0% when the confound is perfectly correlated with the target and when using 1000 features. Figure 3.14C shows that, indeed, higher values lead to narrower correlation distributions, which is shown in Figure 3.14D to yield relatively lower accuracy scores.

In summary, our simulations show that below chance accuracy is accurately predicted by the standard deviation (i.e., "width")

of the distribution of empirical feature-target correlations and that WDCR reduces this standard deviation, which explains why the empirical analyses yielded below chance model performance (especially for larger numbers of voxels).

Cross-validated confound regression (CVCR)

Empirical results

As the results from the empirical analyses and simulations suggest, the use of WDCR is problematic because of the partitioning of the dataset into a separate train set and test set after confound regression. As such, our proposed cross-validated confound regression (CVCR) methods suggests to move the confound regression procedure inside the cross-validation loop, thereby also cross-validating this step. As expected, compared to the baseline model (i.e., no confound control), the results from the empirical analyses using CVCR show reduced (but not below chance) model performance for both VBM and TBSS data, and all different numbers of voxels (see Figure 3.15). Notably, for some numbers of voxels, model performance was not significantly above chance level.

We also evaluated whether regressing the confound from the train set only was sufficient to control for confounds, but found that it does not effectively control for confounds when there is no true signal (i.e., there is positive bias), which is visualized in more detail in Supplementary Figure B.10 (cf. Figure 3.8).

Efficacy analysis

Similar to WDCR, CVCR yielded plausible and unbiased model performance (see Figure 3.8, pink line). Moreover, when applied to the simulated null data, CVCR yielded model performance scores at chance level across all levels of the confound-target correlation and numbers of features (see Figure 3.14B).

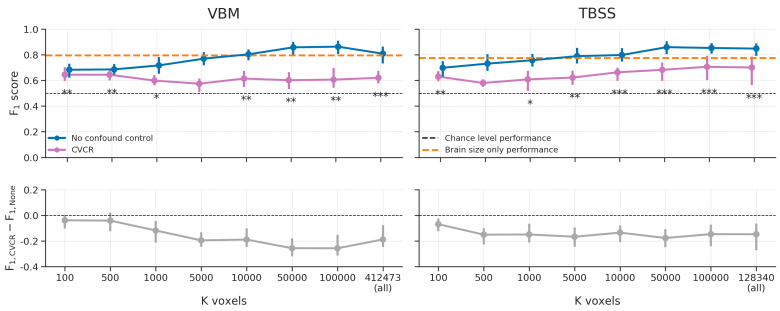


FIGURE 3.15 Model performance after CVCR (pink) versus the baseline performance (blue) for both the VBM (left) and TBSS (right) data. Performance reflects the average F_1 score across 10 folds; error bars reflect 95% confidence intervals across 1000 bootstrap replications. The dashed black line reflect theoretical chance level performance (0.5) and the dashed orange line reflects the average model performance when only brain size is used as a predictor. Asterisks indicates performance of the CVCR model that is significantly above or below chance: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$.

Summary methods for confound control

In this section, we investigated the effects of different method to control confounds (post hoc counterbalancing, WDCR, and CVCR) on empirical MRI data and simulated data (see Figure 3.16 for a summary of the empirical results). Post hoc counterbalancing was, at least using the subsampling method described, clearly unable to effectively control for confounding influences, which is putatively caused by indirect circularity in the analysis process due to subsampling. Confound regression showed an expected drop in model performance (but not below chance level), but only when the confound regression step is properly cross-validated (i.e., the CVCR version).

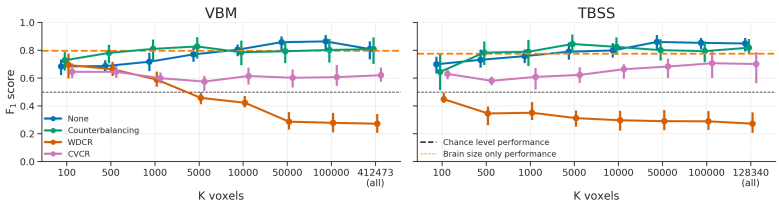


FIGURE 3.16 An overview of the empirical results on the four different confound methods: None, post hoc counterbalancing, WDCR, and CVCR.

3.4 Discussion

Decoding analyses have become a popular alternative to univariate analyses of neuroimaging data. This analysis approach, however, inherently suffers from ambiguity in terms of which source of information is picked up by the decoder (Naselaris & Kay, 2015). Given that one is often interested in model interpretability rather than merely accurate prediction (Hebart & Baker, 2017), one should strive to control for alternative sources of information (i.e., other than the target of interest) that might drive decoding. Effectively controlling for these alternative sources of information, or *confounds*, helps in disambiguating decoding models. In this article, we reviewed and tested two generic, broadly applicable methods that aim to control for confounds in decoding analyses: post hoc counterbalancing and confound regression. Additionally, we proposed a third method that, unlike the other two methods, has shown to effectively control for confounds.

Both when applied to empirical and simulated data, we found that neither post hoc counterbalancing nor (whole-dataset) confound regression yielded plausible and unbiased model performance estimates. First, we found that post hoc counterbalancing leads to *optimistic* (i.e., positively biased) model performance estimates, which is a result of removing samples that are hard

to classify or would be wrongly classified, during the subsampling process. Because this subsampling process is applied to the entire dataset at once (i.e., it is not cross-validated), it can be seen as a form of indirect circular analysis (Kriegeskorte, Simmons, Bellgowan, et al., 2009b), in which the data themselves are used to inform analysis decisions, which can lead to biased generalization estimates. Second, our initial evaluation of confound regression, which was applied on the entire dataset ("WDCR"), yielded pessimistic (i.e., negatively biased) and even significantly below chance model performance estimates. Extending previous research (Jamalabadi et al., 2016), we show that this negative bias occurs when the "signal" in the data (operationalized as the width of the feature-target correlation distribution) is lower than would be expected by chance, which we link to the sampling distribution of the Pearson correlation coefficient. Importantly, we show that WDCR systematically narrows the width of the correlation distribution — and thus leads to lower model performance — which is exacerbated by both higher correlations between target and confound, as well as by a larger number of features.

The negative bias observed in WDCR is caused by the fact that it is performed on the whole dataset at once, leading to statistical dependencies between subsequent train and test partitions. To overcome this negative bias, we propose to cross-validate the confound regression procedure (which we call "Cross-Validated Confound Regression", CVCR). We show that this method yields plausible model performance in the empirical analyses (i.e., significantly above chance model performance) and nearly unbiased model performance in the simulations, for different datasets varying in the amount of features (K) and the strength of the confound (r_{Cy}). Moreover, initial supplementary simulations suggest that these results generalize to (simulated) fMRI data (Supplementary Figure B.1), seemingly demonstrating effective control of confounds across different degrees of autocor-

relation (Supplementary Figure B.2). The method may show some negative bias in some scenarios due to the fact that, in the train set, CVCR will remove all variance associated with the confound (even variance *spuriously* correlated with the confound). However, this bias seems, at least in the simulated scenarios, very small. Overall, we believe that our results demonstrate that CVCR is a flexible and effective method to control for confounds in decoding analyses of neuroimaging data.

Relevance and consequences for previous and future research

A priori and post hoc counterbalancing

We believe our results have implications not only for post hoc counterbalancing, but a priori counterbalancing in observational designs in general. In both behavioral research (Wacholder et al., 1992) and neuroimaging research [Gorgen2017-sy], a priori counterbalancing (or case-control "matching") is a common strategy to avoid effects of confounds. However, as we show in the current study, this may unintentionally remove samples that are harder to predict, especially when there is little shared variance between the confound and the other predictors (i.e., when there is low confound R^2). Because, conceptually, this represents a form of circular analysis, counterbalancing — regardless of whether it is applied a priori or post hoc — can yield biased model performance estimates. To some extent, the bias in the post hoc counterbalancing results should not come as a surprise: as noted in the Methods section, counterbalancing in observational research requires the researcher to choose a sample that is not representative of the population (see also Sedgwick, 2013). As a result, out-of-sample predictive performance drops significantly, in our case even to chance level.

Since post hoc counterbalancing does not show any positive bias in model performance when there is no signal at all (i.e., signal

R^2), one could argue that any observed significant above chance effect, while positively biased in terms of effect magnitude, can be interpreted as evidence that there must be signal in the data in the first place. However, we argue against this interpretation for two reasons. First, any above chance predictive performance of models fitted after subsampling is not only positively biased, but also does not cross-validate to the rejected samples (see Figure 3.11). That is, the model picks up relations between features and target that are only present in the subsample, and not in the samples left out of the analysis. As a result, it is questionable whether (and if so, how) the model should be interpreted — after all, we assume that the rejected samples were drawn from the population of interest in a valid way. Second, any possible absence of above chance model performance after subsampling can neither be interpreted as evidence for an *absence* of a true effect, since the subsampling procedure necessarily leads to a (often substantial) power loss. It could still well be that in the original sample there was a true relation between features and target. Thus, interpretation of modelling efforts after subsampling is problematic in case of both *presence* and *absence* of above chance model performances.

Confound regression

In contrast to post hoc counterbalancing, confound regression in its uncross-validated form (i.e., WDCR) has been applied widely in the context of decoding analyses (Dubois et al., 2018; Kostro et al., 2014; Rao et al., 2017; Todd et al., 2013). Indeed, the first study that systematically investigated the effect of confounds in decoding analyses (Todd et al., 2013) used WDCR to account for the confounding effect of reaction times (RT) on decoding of rule representations and found that WDCR completely eliminated the predictive performance that was found without controlling for RT. This observation, however, can potentially be explained by the negative bias induced by WDCR.

This possible explanation is corroborated by a follow-up study that similarly looked into RT confounding the decoding of rule representations (Woolgar et al., 2014), who did not use WDCR but accounted for RT confounding by including it as a covariate during the pattern estimation procedure (see Supplementary Materials for a tentative evaluation of this method), which in contrast to the study by Todd et al. yielded significant decoding performance. Moreover, while not specifically investigated here, we expect a similar negative bias to occur when a confound is removed from a continuous target variable using WDCR — which may offer an explanation for the null finding of Dubois et al. (2018), who fail to decode personality characteristics from resting-state fMRI.

Relevance to other analysis methods

While this article focuses on controlling for confounds in decoding analyses specifically, we believe that our findings may be relevant for analysis methods beyond decoding analyses as well. In fact, methods for controlling for confounds (or alternative sources of information) have previously been investigated and applied in another type of MVPA named "representational similarity analysis" (RSA; Kriegeskorte et al., 2008). In the context of RSA, the explained variance in the neural data is often partitioned into different (model-based) feature sets (i.e., sources of information), which allows one to draw conclusions about the unique influence of each source of information (see, e.g., Groen et al., 2018; Hebart et al., 2018; Ramakrishnan et al., 2014). Specifically, variance partitioning in RSA is done by removing the variance from the representational dissimilarity matrix (RDM) based on the feature set that needs to be controlled for. Notably, the variance of the RDMs that are not of interest can be removed from only the neural RDM (Hebart et al., 2018; Ramakrishnan et al., 2014) or both from the neural RDM

and the RDM of interest (Groen et al., 2018). While the analysis context is different, the underlying technique is identical to confound regression as described and evaluated in this article. Importantly, the studies employing this variance partitioning technique (Groen et al., 2018; Hebart et al., 2018; Ramakrishnan et al., 2014) similarly report plausible model performances after confound regression (i.e., relatively lower but not below chance performance), corroborating our results with (cross-validated) confound regression. Note that the distinction between WDCR and CVCR in the context of most RSA studies (including the aforementioned studies) is largely irrelevant, as representational similarity analyses are not commonly cross-validated. However, recently, some have proposed to use cross-validated distance measures (such as the cross-validated Mahalanobis distance; Guggenmos et al., 2018; Walther et al., 2016) in RSA, which could suffer from negative bias when combined with (not cross-validated) variance partitioning similar to what we observed with WDCR in the context of decoding analyses.

We believe that especially our findings with regard to WDCR and CVCR may be relevant for any cross-validated analysis, regardless of the "direction" of analysis (encoding vs. decoding) and the dimensionality of the neural data (univariate vs. multivariate approaches). In general, our findings with respect to negative bias after WDCR were to be expected, as it introduces dependence between the train set and the test set which violates the crucial assumption of independence of any cross-validated analysis. While a violation of the independence assumption often leads to positive bias such as in "double dipping" (Kriegeskorte, Simmons, Bellgowan, et al., 2009b), we show here that it may also lead to negative bias. Either way, our findings reinforce the idea that data analysis operations should *never* be applied to the entire dataset before subjecting the data to a cross-validated analysis. Therefore, we believe that our findings with respect to

WDCR and CVCR will generalize to any cross-validated analysis (such as cross-validated MANOVA, Allefeld & Haynes, 2014; or cross-validated encoding models, Naselaris et al., 2011), but future research is necessary to substantiate this claim.

Importance for gender decoding studies

The importance of proper confound control is moreover highlighted by the empirical question we address. Without any optimization of the prediction pipeline, we were able to predict gender with a model performance up to approximately 0.85 without confound control. This is in line with reports from various other studies (Del Giudice et al., 2016; Rosenblatt, 2016; Sepehrband et al., 2018). However, this predictive performance is driven by a mixture two sources of information: global and local differences in brain structure. With confound control, however, we show that predictive performance using only local differences lies around 0.6 for VBM data and 0.7 for TBSS data — a substantial drop in performance. Especially because the remaining predictive performance is lower than predictive performance using only brain size, we argue that the use of proper confound control may lead one to draw substantially different conclusions about the differences in brain structure between men and women. For the debate on gender dimorphism, it is thus extremely important to take global brain size into account in the context of decoding analyses (as has been similarly recommended for mass-univariate analyses; Barnes et al., 2010).

Choosing a confound model: linear vs. nonlinear models

In the present paper, we focused on the use of linear models for confound control. It is crucial to note that the efficacy of confound control depends on the suitability of the confound regression model employed. Removing variance associated with

a confound using a linear model removes only the variance of data (features) that is linearly related to the confound. When a confound is nonlinearly related to the data, some variance associated with the confound can remain in the data after a linear confound model is used to regress out variance. It is possible that the decoding model subsequently applied still picks up this residual "confounded" variance. In other words, an unsuitable confound model may control for confounds imperfectly.

The exact relation between confound and (brain) data is hardly ever known a priori. However, it is possible to explore the nature of this relation using the data at hand. For example, a researcher can apply a cross-validated prediction pipeline to predict a feature (e.g., VBM voxel intensity) from the confound. The researcher can then test what type of model (linear or nonlinear) describes the relation between confound and data best. In the Supplementary Materials (section "Linear vs nonlinear confound models: predicting VBM and TBSS data based on brain size"), we provide an example of this approach. We used linear, quadratic, and cubic regression models to predict VBM and TBSS voxel intensity using brain size as feature. In the Supplementary Results, we show that linear models perform equally well as or better than polynomial models for the majority of voxels (Supplementary Figures B.7 and B.9). Further, for voxels where polynomials outperform linear models, the difference between model performances is minimal (Supplementary Figure B.8). Thus, in the empirical research question explored in this paper, a linear confound model seems to suit the data very well.

Practical recommendations

As indicated by the title of this article, we will now outline some practical recommendations for dealing with confounds in decoding analyses of neuroimaging data. First, one needs to obtain an accurate measurement of potential confounds (Westfall

& Yarkoni, 2016). While we assumed the availability of such a measure in this article, this is not always trivial. In experimental settings, for example, reaction times can potentially be identified as a confound (Todd et al., 2013; Woolgar et al., 2014), but arguably, it is not reaction time but rather an unobserved variable related to reaction time (e.g., difficulty or attention) that confounds the analysis. In such scenarios, the best one can do is measure reaction time as a proxy, and be aware that any subsequent confound control method is limited by how well this proxy corresponds to the actual confound. Second, one needs to identify which variables actually confound a decoding analysis. To detect confounds, we recommend using the "same analysis approach" outlined by Görgen et al. (2017). In short, this method involves trying to predict the target variable using your confound(s) as predictive features (for example, when using only brain size to predict gender). In case of significant above chance decoding performance, and assuming the confounds are actually encoded in the neuroimaging data, the hypothesized confounds will most likely influence the actual decoding analysis. While in the current article we focused on simple univariate confounding effects (i.e., confounding by a single variable), the same analysis approach is not limited to detecting univariate confounds — it facilitates detecting multivariate (i.e., confounding by multiple variables) or interaction effects (i.e., confounding by interaction effects between variables) as well. For example, if one hypothesizes that the target variable is related to the interaction between confound C_1 and C_2 (i.e., $C_1 \times C_2$), one can simply use the interaction term as the potential confound in the same analysis approach to evaluate the potential confounding influence.

Once the specific confound terms have been identified, we recommend regressing out the confound from the data in a cross-validated manner (i.e., using CVCR). Specifically, we recommend including confound regression as the first step in your decoding pipeline to avoid the effect of confounds on other opera-

tions in the pipeline (such as univariate feature selection; Chu et al., 2012). In this article, we used ordinary least squares (OLS) regression to remove the influence of confounds from the data, because a linear model describes the relation between brain size and VBM/TBSS voxel intensities well (see Supplementary Figures B.7-B.9). However, not only linear models can be used to remove variance associated with a confound from the data — it is possible to use nonlinear models (potentially with multiple confounds and interactions between them) if it is clear that the relation between confounds and neuroimaging features is nonlinear (see previous section for details on choosing a confound model). However, as a limitation to the presented results, we did not test whether CVCR also leads to (nearly) unbiased results when used with nonlinear models. We advise, therefore, in such cases, to first test in a simulation study whether CVCR provides an unbiased confound control method with nonlinear models before use with actual data.

3.5 Conclusions

In general, we believe that the contributions of the current study are twofold. First and foremost, it provides a systematic evaluation of widely applicable methods to control for confounds and shows that, of the methods investigated, only one ("cross-validated confound regression") appears to yield plausible and almost unbiased results. The results from this evaluation hopefully prevents researchers from using post hoc counterbalancing and whole-dataset confound regression, which we show may introduce (unintended) biases. Moreover, we made all analyses and preprocessed data openly available (https://github.com/lukassnoek/MVCA) and provide a simple implementation for cross-validated confound regression that interfaces with the popular scikit-learn package in the Python programming lan-

guage. Second, we believe that this study improves understanding of the elusive phenomenon of below chance accuracy (building on previous work by Jamalabadi et al., 2016). In general, we hope that this study helps researchers in gaining more insight into their decoding analyses by providing a method that disentangles the contributions of different sources of information that may be encoded in their data.

Chapter 4

The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses

Chapter 5

Choosing to view morbid information involves reward circuitry

Chapter 6

Using predictive modeling to quantify the importance and limitations of action units in emotion perception

Chapter 7

Comparing models of dynamic facial expression perception

Chapter 8

Summary and general discussion

My view on going forward.

8.1 Explore!

Theories are like toothbrushes, no one likes to use someone else's.

8.2 Think *big*

Big, complex datasets to train big, complex models.

8.3 Rethink psychology education

Embrace and teach interdisciplinary.

Appendices

Appendix A

Supplement to Chapter 2

A.1 Stimuli used for SF-task

Table A.1 Stimuli used for SF-task

Class	Dutch	English translation
Action	Hard wegrennen	Running away fast
	Iemand wegduwen	Pushing someone away
	Iemand stevig vastpakken	Holding someone tightly
	Je hoofd schudden	Shaking your head
	Heftige armgebaren maken	Making big arm gestures
	Ergens voor terugdeinzen	Recoiling from something
	Je ogen dichtknijpen	Closing your eyes tightly
	Je ogen wijd open sperren	Opening your eyes widely
	Je wenkbrauwen fronsen	Frowning with your eyebrows
	Je schouders ophalen	Raising your shoulders
	Op de vloer stampen	Stamping on the floor
	In elkaar duiken	Cowering
	Je schouders laten hangen	Slumping your shoulders
	Je vuisten ballen	Tighten your fists

	Je borst vooruit duwen	Push your chest forward
	Je tanden op elkaar zetten	Clench your teeth
	Je hand voor je mond slaan	Put your hand in front of your mouth
	Onrustig bewegen	Moving restlessly
	Heen en weer lopen	Walking back and forth
	Je hoofd afkeren	Turning your head away
Interoception	Een brok in je keel	A lump in your throat
	Buiten adem zijn	Being out of breath
	Een versnelde hartslag	A fast beating heart
	Je hart klopt in de keel	You heart is beating in your throat
	Een benauwd gevoel	An oppressed feeling
	Een misselijk gevoel	Being nauseous
	Druk op je borst	A pressure on your chest
	Strak aangespannen spieren	Tense muscles
	Een droge keel	A dry throat
	Koude rillingen hebben	Cold shivers
	Bloed stroomt naar je hoofd	Blood is going to your head
	Een verdoofd gevoel	A numb feeling
	Je hebt tintelende ledenmaten	Tingling limbs
	Een verlaagde hartslag	A slow heartbeat
	Je hebt zware ledematen	Heavy limbs
	Een versnelde ademhaling	Fast breathing
	Je hebt hoofdpijn	Headache
	Je hebt buikpijn	Stomachache

	Zweet staat in je handen	Sweaty palms
	Je maag keert zich om	Your stomach churns
Situation	Vals beschuldigd worden	Being falsely accused
	Dierbare overlijdt	A loved one dies
	Vlees is bedorven	Meat that has gone off
	Je wordt bijna aangereden	You are almost hit by a car
	Iemand naast je braakt	Someone next to you vomits
	Huis staat in brand	House is on fire
	Zonder reden ontslagen worden	Being fired for no reason
	Een ongemakkelijke stilte	An uncomfortable silence
	Alleen in donker park	Alone in a dark park
	Inbraak in je huis	A house burglary
	Een gewond dier zien	Seeing a wounded animal
	Tentamen verknallen	Messing up your exam
	Je partner bedriegt je	You partner cheats on you
	Dierbare is vermist	A loved one is missing
	Belangrijke sollicitatie vergeten	Forgot a job interview
	Onvoorbereid presentatie geven	Giving a presentation unprepared
	Je baas beledigt je	Your boss offends you
	Goede vriend negeert je	A good friend neglects you
	Slecht nieuws bij arts	Bad news at the doctor
	Bommelding in metro	A bomb alarm in the metro

Note:

The stimulus materials presented in Table S1 were selected from a pilot study. In this pilot study we asked an independent sample of twenty-four subjects to describe how they would express an emotion in their behavior, body posture or facial expression (action information), what specific sensations they would feel inside their body when they would experience an emotion (interoceptive information), and for what reason or in what situation they would experience an emotion (situational information). These three questions were asked in random order for twenty-eight different negative emotional states, including anger, fear, disgust, sadness, contempt, worry, disappointment, regret and shame. The descriptions generated by these subjects were used as qualitative input in order to create our stimulus set of twenty short sentences that described emotional actions, sensations or situations. With this procedure, we ensured that our stimulus set held sentences that were validated and ecologically appropriate for our sample.

A.2 Instructions

Full instruction for the other-focused emotion understanding task

Translated from Dutch; task presented first.

In this study we are interested in how the brain responds when people understand the emotions of others in different ways. In the scanner you will see images that display emotional situations, sometimes with multiple people. In every image one person will be marked with a red square. While viewing the image we ask you to focus on the emotion of that person in three different ways.

With some images we ask you to focus on HOW this person expresses his or her emotion. Here we ask you to identify expressions in the face or body that are informative about the emotional state that the person is experiencing.

With other images we ask you to focus on WHAT this person may feel in his or her body. Here we ask you to identify sensations, such as a change in heart rate, breathing or other internal feeling, that the person might feel in this situation.

With other images we ask you to focus on WHY this person experiences an emotion. Here we ask you to identify a specific reason or cause that explains why the person feels what he or she feels.

Every image will be presented for six seconds. During this period we ask you to silently focus on HOW this person expresses emotion, WHAT this person feels in his/her body, and WHY this person feels an emotion.

Before you will enter the scanner we will practice. I will show you three images and will ask you to perform each of the three instructions out loud.

It is important to note that there are no correct or incorrect answers, it is about how you interpret the image. For the success of the study it is very important that you apply the HOW, WHAT or WHY instruction for each image. Please do not skip any images and try to apply each instruction with the same motivation. It is also important to treat every image separately, although it is possible that you have similar interpretations for different images.

The three instructions are combined with the images in blocks. In every block you will see five images with the same instruction. Each block will start with a cue that tells you what to focus on in that block.

Each image is combined with all three instructions, so you will see the same image multiple times. In between images you will sometimes see a black screen for a longer period of time.

Do you have any questions?

Full instruction for the self-focused emotion imagery task

Translated from Dutch; task presented second.

In this study we are interested in how the brain responds when people imagine different aspects of emotion. In the scanner you will see sentences that describe aspects of emotional experience. We ask you to try to imagine the content of each sentence as rich and detailed as possible.

Some sentences describe actions and expressions. We ask you to imagine that you are performing this action or expression. Other sentences describe sensations or feelings that you can have *inside* your body. We ask you to imagine that you are experiencing this sensation or feeling. Other sentences describe

emotional situations. We ask you to imagine that you are experiencing this specific situation.

We ask you to always imagine that YOU have the experience. Thus, it is about imagining an action or expression of your body, a sensation inside your body, or a situation that you are part of.

I will give some examples now.

For each sentence you have six seconds to imagine the content. All sentences will be presented twice. In between sentences you will sometimes see a black screen for a longer period of time. For this experiment to succeed it is important that you imagine each sentence with the same motivation, even if you have seen the sentence before. Please do not skip sentences.

Do you have any questions?

A.3 Behavioral results

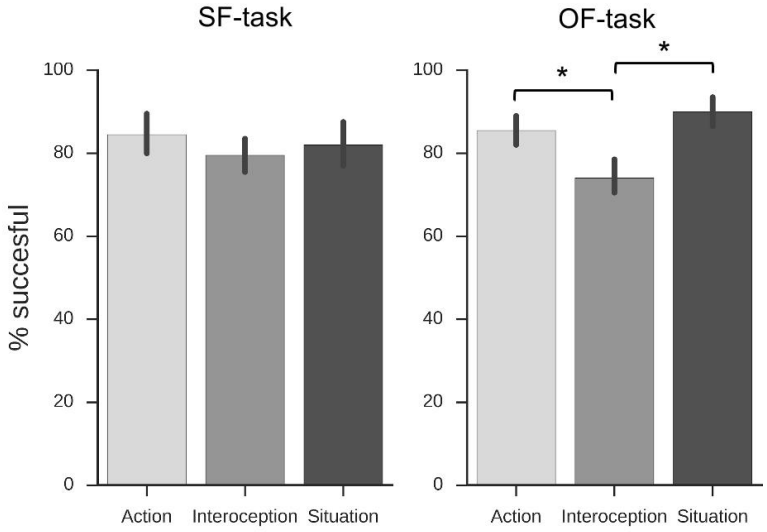


FIGURE A.1 Mean percentage of trials successfully executed for the SF-task (left panel) and OF-task (right panel). Error bars indicate 95% confidence intervals. A one-way ANOVA of the success-rates of the SF-task (left-panel) indicated no significant overall differences, $F(2, 17) = 1.03$, $p = 0.38$. In the OF-task (right panel) however, a one-way ANOVA indicated that success-rates differed significantly between classes, $F(2, 17) = 17.74$, $p < 0.001$. Follow-up pairwise comparisons (Bonferroni corrected, two tailed) revealed that interoception-trials ($M = 74.00$, $SE = 2.10$) were significantly less successful ($p < 0.001$) than both action-trials ($M = 85.50$, $SE = 1.85$) and situation trials ($M = 90.00$, $SE = 1.92$).

A.4 Optimization results

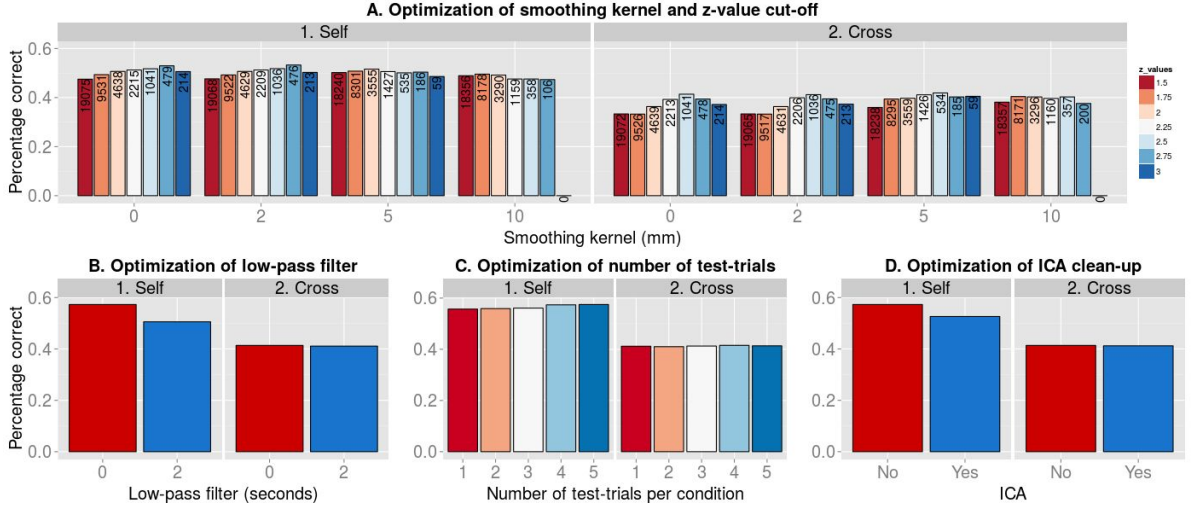


FIGURE A.2 Results of the parameter-optimization procedure. Reported scores reflect the classification scores averaged over subjects and classes (i.e. the diagonal of the confusion matrix). All optimization analyses were iterated 5000 times. **A**) Classification results for different smoothing kernels (0, 2, 5, and 10 mm) and z-value threshold for differentiation scores during feature selection (see MVPA pipeline section in the main text for a description of the particular feature selection method we employed). Numbers reflect the average number of voxels selected across iterations. **B**) Classification results of using a low-pass filter (2 seconds) or not. **C**) Classification results for different numbers of test-trials per class (1 to 5). **D**) Classification results when preprocessing the data with Independent Component Analysis (ICA) or not.

TABLE A.2 Parameters assessed in the optimization set

Parameter	Options	Final choice
Smoothing kernel	0 mm, 2 mm, 5 mm, 10 mm	5 mm
Feature selection threshold	1.5, 1.75, 2, 2.25, 2.5, 2.75, 3	2.3
Number of test-trials	1, 2, 3, 4, 5	4
Low-pass filter	2 seconds vs. none	None
ICA denoising	ICA vs. no ICA	No ICA

Note: The first set of parameters we evaluated in the optimization-set were different smoothing factors and feature selection thresholds (see MVPA pipeline section in the main text). On average, across the self- and cross-analysis, a 5 mm smoothing kernel yielded the best results in combination with a feature selection threshold of 2.25, which we rounded up to 2.3 as this number represents a normalized (z-transformed) score, which corresponds to the top 1% scores within a normal distribution. Next, the difference between using a low-pass (of 2 seconds, i.e. 1 TR) versus none was assessed, establishing no low-pass filter as the optimal choice. Next, different numbers of test-trials (1 to 5) per class per iteration were assessed. Four trials yielded the best results. Lastly, the effect of "cleaning" the data with an independent component analysis was examined (FSL: MELODIC and FIX; Salimi-Khorshidi et al., 2014). Not performing ICA yielded the best results. These parameters – 5 mm smoothing kernel, 2.3 feature selection thresholded, no low-pass filter, and four test-trials per iteration – were subsequently used in the analysis of the validation set.

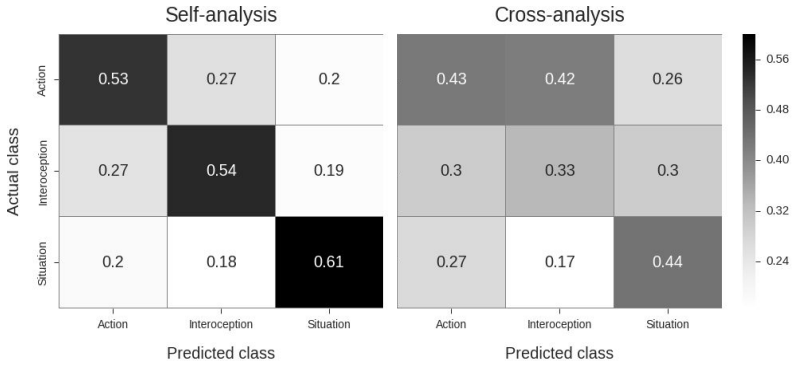


FIGURE A.3 Confusion matrices displaying precision-values yielded by the classification analysis of the optimization dataset with the final set of parameters. Because no permutation statistics were calculated for the optimization set, significance was calculated using a one-sample independent *t*-test against chance-level classification (i.e. 0.333) for each cell in the diagonal of the confusion matrices. Here, all t-statistics use a degrees of freedom of 12 (i.e. 13 subjects - 1) and are evaluated against a significance level of 0.05, Bonferroni-corrected. For the diagonal of the self-analysis confusion matrix, all values were significantly above chance-level, all $p < 0.0001$. For the diagonal of the cross-analysis confusion matrix, both the action (43% correct) and situation (44% correct) classes scored significantly above chance, $p = 0.014$ and $p = 0.0007$ respectively. Interoception was classified at chance level, $p = 0.99$, which stands in contrast with the results in the validation-set.

A.5 Bagging procedure

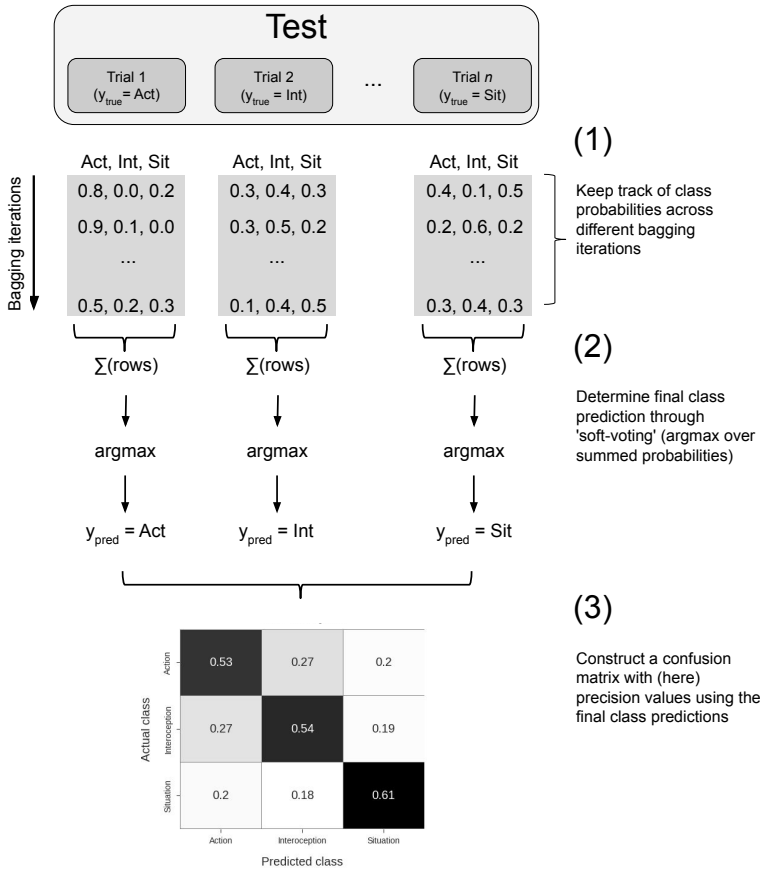


FIGURE A.4 Schematic overview of the bagging procedure. Class probabilities across different bagging iterations are summed and the class with the maximum probability determines each trial's final predicted class, which are subsequently summarized in a confusion matrix on which final recall/precision scores are calculated.

A.6 Precision vs. recall

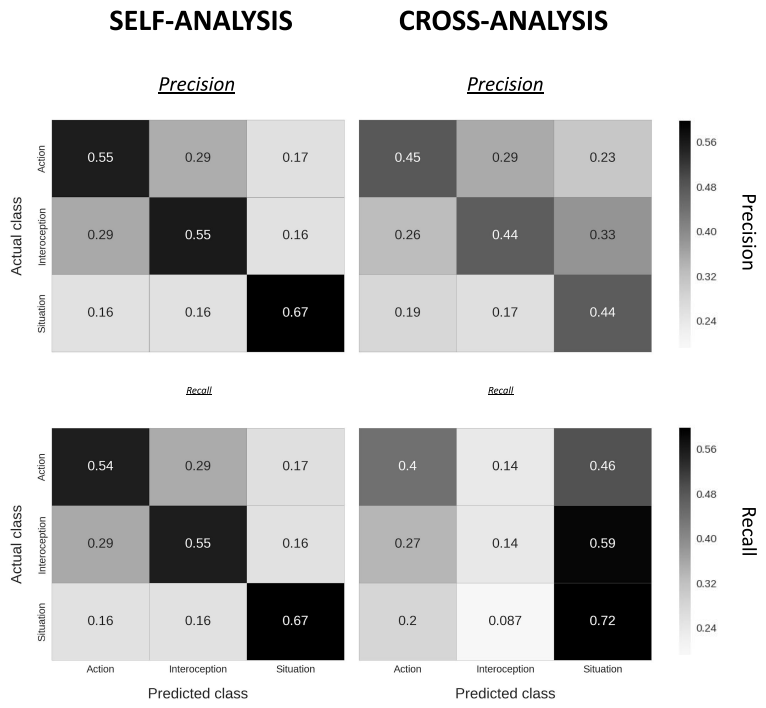


FIGURE A.5 A comparison between precision and recall confusion matrices of the self- and cross-analysis of the validation dataset. Precision refers to the amount of true positive predictions of a given class relative to all predictions for that class. Recall refers to the amount of true positive predictions of a given class relative to the total number of samples in that class. In the self-analysis, all classes were decoded significantly above chance for both precision and recall (all $p < 0.001$). In the cross-analysis, all classes were decoded significantly above chance for precision (all $p < 0.001$); for recall both action and situation were decoded significantly above chance ($p = 0.0013$ and $p < 0.001$, respectively), while interoception was decoded below chance. All p-values were calculated using a permutation test with 1300 permutations (as described in the Methods section in the main text). When comparing precision and recall scores for both analyses, precision and recall showed very little differences in the self-analysis, while the cross-analysis shows a clear difference between metrics, especially for interoception and situation. For the interoception class, the relatively high precision score (44%) compared to its recall score (14%) suggests that trials are very infrequently classified as interoception, but when they are, it is (relatively) often correct. For the situation class, the relatively high recall score (72 %) compared to its precision score (44%) suggests that situation is strongly over-classified, which is especially clear in the lower-right

A.7 Self vs. other classification

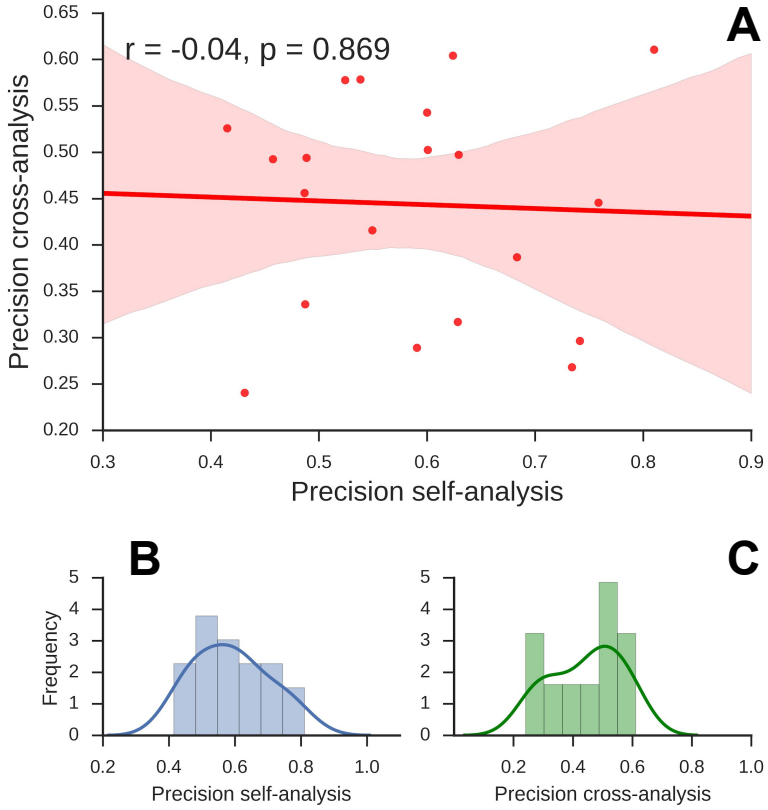


FIGURE A.6 Relation between self- and cross-analysis scores across subjects and their respective distributions. Note that the scores here represent the average of the class-specific precision scores. **A**) There is no significant correlation between precision-scores on the self-analysis and the corresponding scores on the cross-analysis, r = -0.04, p = .86, implying that classification scores in the self-analysis is not predictive of scores in the cross-analysis. **B**) The distribution of precision-scores in the self-analysis, appearing to be normally distributed. **C**) The distribution of precision-scores in the cross-analysis, on the other hand, appears to be bimodal, with one group of subjects having scores around chance level (0.333) while another group of subjects clearly scores above chance level (see individual scores and follow-up analyses in (ref:fig-shared-states-S4).

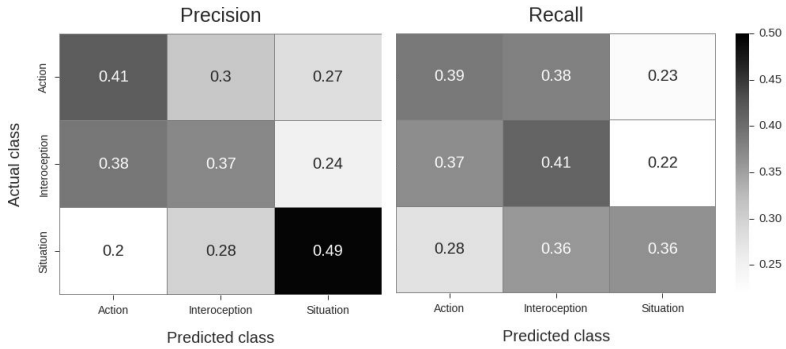


FIGURE A.7 Confusion matrices with precision (left matrix) and recall (right matrix) estimates of the other-to-self decoding analysis. The MVPA-pipeline used was exactly the same as for the (self-to-other) cross-analysis in the main text. *P*-values corresponding to the classification scores were calculated using a permutation analysis with 1000 permutations of the other-to-self analysis with randomly shuffled class-labels. Similar to the self-to-other analysis, the precision-scores for all classes in the other-to-self analysis were significant, $p(\text{action}) < 0.001$, $p(\text{interoception}) = 0.008$, $p(\text{situation}) < 0.001$. For recall, classification scores for action and interoception were significant (both $p < 0.001$), but not significant for situation ($p = 0.062$). The discrepancy between the self-to-other and other-to-self decoding analyses can be explained by two factors. First, the other-to-self classifier was trained on fewer samples (i.e. 90 trials) than the self-to-other classifier (which was trained on 120 trials), which may cause a substantial drop in power. Second, the preprocessing pipeline and MVPA hyperparameters were optimized based on the self-analysis and self-to-other cross-analysis. Given the vast differences between the nature of the self- and other-data, these optimal preprocessing and MVPA hyperparameters for the original analyses may not cross-validate well to the other-to-self decoding analysis.

A.8 Condition-average results

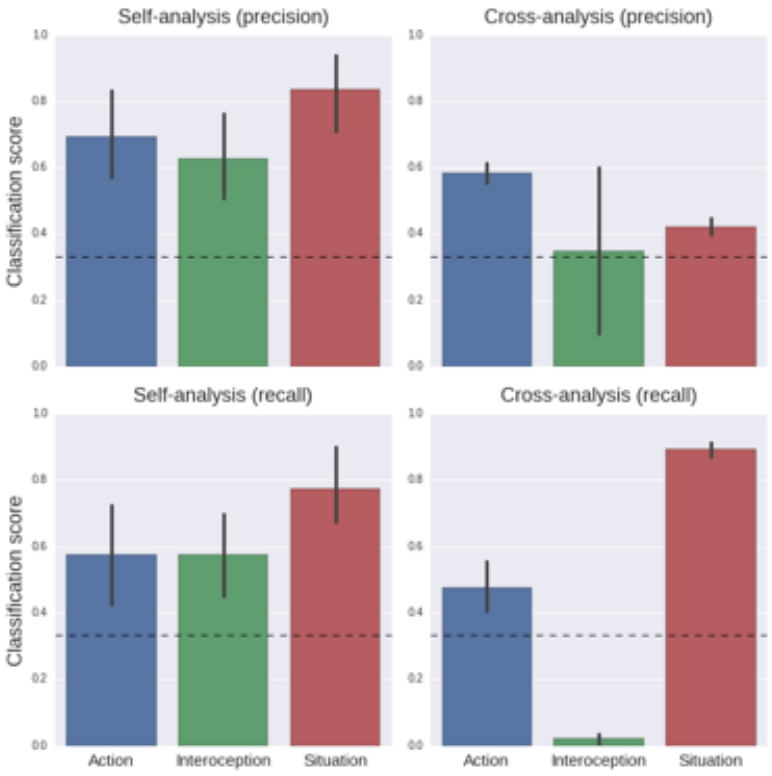


FIGURE A.8 Results of MVPA analyses using condition-average voxel patterns across subjects instead of single-trial patterns within subjects. Here, patterns are estimated in a GLM in which each condition, as opposed to each trial, is modeled with a single regressor, from which whole-brain t-value patterns were extracted. In this condition-average multi-voxel pattern analysis, condition-average patterns across subjects were used as samples. The condition-average patterns were extracted from the univariate first-level contrasts. In total, this yielded 120 samples for the self-data (3 conditions x 2 runs x 20 participants) and 60 samples for the other-data (3 conditions * 20 participants). For these analyses, the same hyperparameters were used as the original analyses reported in the main text, except with regard to the cross-validation and bagging procedure. Here, we used (stratified) 10-fold cross-validation without bagging. The upper panels show precision scores (per class) for the self- and (self-to-other) cross-analysis; the lower panels show results from the same analyses but expressed in recall-estimates (error bars indicate 95% confidence intervals). Apart from interoception in the cross-analysis (both precision and recall), all scores were significant ($p < 0.001$) in a permutation test with 1000 permutations. These results largely replicate our findings as reported in the main text. This

A.9 Individual subject scores

TABLE A.3 Mean general classification scores per subject for the self- and cross-analysis on the validation-set only.

Subject nr.	Self-analysis precision	Cross-analysis precision	Session	Part of optimization-set?
1	0.758	0.445	2	y
2	0.487	0.336	2	y
3	0.629	0.316	1	y
4	0.524	0.577	2	y
5	0.457	0.492	1	y
6	0.741	0.296	2	y
7	0.600	0.542	1	y
8	0.431	0.240	2	y
9	0.629	0.497	1	y
10	0.734	0.268	2	y
11	0.683	0.386	1	y
12	0.415	0.525	2	y
13	0.623	0.604	1	y
14	0.810	0.610	1	n
15	0.538	0.578	1	n
16	0.486	0.455	1	n
17	0.549	0.415	1	n
18	0.488	0.494	1	n
19	0.590	0.289	1	n
20	0.600	0.502	1	n

Note: Supplementary Table 3 suggests individual variability in the extent to which neural resources are shared between self- and other-focused processes. In the SF-task all subjects demonstrated a mean classification score well above .33 (i.e., score associated with chance). When generalizing the SF-classifier to the OF-task, however, the classification scores appear to be bimodally distributed (see Supplementary Figure 5C). As can be seen in Table 3, some subjects demonstrated a relatively high mean classification score (i.e., > .45), whereas other subjects demonstrated a classification score at chance level or lower. Note that there is no significant difference between the OF classification scores for subjects who participated in the experiment for the first or second time ("Session" column in table; *t*(18) = 1.73, p = 0.10), nor for subjects who were or were not part of the optimization-set ("Part of optimization-set?" column in table; *t*(18) = -.95, p = 0.35), suggesting that inclusion in the optimization-set or session ordering is not a confound in the analyses. Regarding individual variability in self-other neural overlap, it is important to note that in the field of embodied cognition, there is increasing attention for the idea that simulation is both individually and contextually dynamic (Oosterwijk & Barrett, 2014; Winkielman, Niedenthal, Wielgosz & Kavanagh, 2015; see also Barrett, 2009). To better distinguish between meaningful individual variation and variation due to other factors (e.g., random noise), future research should test a priori formulated hypotheses about how and when individual variation is expected to occur.

A.10 Brain region importance

TABLE A.4 Most important voxels in terms of their average weight across iterations and subjects.

Brain region	k	Max	Mean	Std
Frontal pole	1827	5.05	2.35	0.52
Occipital pole	1714	5.15	2.45	0.56
Supramarginal gyrus anterior	1573	7.48	2.84	0.91
Lateral occipital cortex superior	1060	4.52	2.18	0.39
Lateral occipital cortex inferior	923	4.73	2.36	0.49
Angular gyrus	856	4.52	2.24	0.40
Supramarginal gyrus posterior	806	4.49	2.29	0.45
Middle temporal gyrus temporo-occipital	798	4.00	2.33	0.48
Temporal pole	711	4.38	2.37	0.54
Precentral gyrus	568	3.54	2.14	0.31

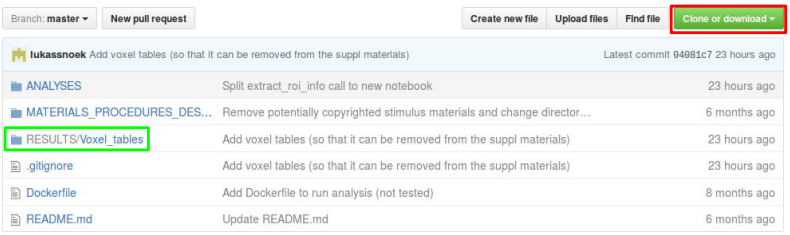
Superior temporal gyrus posterior	549	3.64	2.27	0.41
Superior frontal gyrus	510	3.83	2.18	0.38
Postcentral gyrus	489	4.61	2.43	0.60
Inferior frontal gyrus pars-triangularis	488	4.22	2.35	0.50
Inferior frontal gyrus parsoper-cularis	441	3.54	2.14	0.31
Middle temporal gyrus posterior	417	5.68	2.34	0.52
Occipital fusiform	400	4.28	2.14	0.37
Middle temporal gyrus anterior	398	5.68	2.58	0.76
Middle frontal gyrus	300	3.01	2.06	0.25
Precuneus	282	3.34	2.14	0.31

Note: Brain regions were extracted from the Harvard-Oxford (bilateral) Cortical atlas. A minimum threshold for the probabilistic masks of 20 was chosen to minimize overlap between adjacent masks while maximizing coverage of the entire brain. The column *k* represents the absolute number of above-threshold voxels in the masks. The columns *Max*, *Mean*, and *Std* represent the maximum, mean, and standard deviation from the *t*-values included in the masks. Note that the *t*-values, corresponding to the mean weight across subjects normalized by the standard error of the weights across subjects (after correcting for a positive bias when taking the absolute of the weights), were thresholded at a minimum of 1.75, referring to a *p*-value of 0.05 of a one-sided *t*-test against zero with 19 degrees of freedom (i.e. *n* – 1). Note that this *t*-value map was not corrected for multiple comparisons, and is intended to visualize which regions in the brain were generally involved in our sample of subjects. The *X*, *Y*, and *Z* columns represent the MNI152 (2mm) coordinates of the peak (i.e. max) *t*-value for each listed brain region.

A.11 General note about tables with voxel-coordinates

In order to keep the Supplementary materials concise and orderly, we chose not to include the actual tables with the peak coordinates of all significant clusters of the univariate analyses of the self- and other-task data (as these would amount to 25 pages). These tables, however, can be downloaded (as simple tab-separated-value files) from the study's Github respository[1]. The voxel tables are listed under: `SharedStates/RESULTS/Voxel_tables/*.tsv` (see green box in image below), and can be downloaded by cloning the remote Github repository locally or downloading the ZIP-file from the website (see red box in image below).

[1] https://github.com/lukassnoek/SharedStates


Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
lukassnoek Add voxel tables (so that it can be removed from the suppl materials)
Latest commit 04081c7 23 hours ago
ANALYSES
Split extract_roi_info call to new notebook
23 hours ago
MATERIALS_PROCEDURES_DES...
Remove potentially copyrighted stimulus materials and change director...
6 months ago
RESULTS/Voxel_tables
Add voxel tables (so that it can be removed from the suppl materials)
23 hours ago
.gitignore
Add voxel tables (so that it can be removed from the suppl materials)
23 hours ago
Dockerfile
Add Dockerfile to run analysis (not tested)
8 months ago
README.md
Update README.md
6 months ago

Appendix B

Supplement to Chapter 3

The following supplementary methods describe the methods for the additional analyses done related to controlling for confounds in decoding analyses of (simulated) fMRI data and the validation of using linear confound models for brain size. All code for these simulations, analyses, and results for the fMRI-related sections can be found in `functional_MRI_simulation.ipynb` notebook. The code for the validation of linear confound models can be found in the notebook `empirical_analysis_gender_classification.ipynb`. Both notebooks are stored in the project's Github repository (https://github.com/lukassnoek/MVCA).

B.1 Supplementary methods

Functional MRI simulation

Rationale

The simulations of fMRI data as described here are meant to test the efficacy of our proposed methods for confound control (CVCR) and those proposed by others ["Control for confounds during pattern estimation"; Woolgar et al. (2014)) when applied to fMRI data instead of structural MRI data, as we did in the main text. One reason to suspect differences in how these methods behave between structural and functional MRI data is that

we need to estimate the feature patterns (X) from time series for fMRI data while the feature patterns in structural MRI data are readily available. Moreover, samples in pattern analyses of fMRI data are often correlated (due to temporal autocorrelation in the fMRI signal) while it is reasonable to believe that structural MRI yields independent samples (often individual subjects).

Generation of the data (X, y, C)

In these simulations, we generated fMRI time series data across a grid of "voxels" (K) which may or may not activate in trials from different conditions. Additionally, we allow for the additive influence of a confounding variable with a prespecified correlation to the target variable (corresponding to the "additive" model from Woolgar et al., 2014). In short, we generate voxel signal (s) as a function of both true effects of the trials from different conditions (β_X), the effect of the confound (β_C), and autocorrelated noise (ε):

$$s = X\beta_X + C\beta_C + \varepsilon, \varepsilon \sim \mathcal{N}(0, V) \tag{B.1}$$

where V specifies the covariance matrix of a signal autocorrelated as described by to an AR(1) process (we use $\varphi_1 = 0.5$). Note that, here, X and C refer to time-varying (HRF-convolved) predictors instead of, as in in the main text, arrays of features (voxels) across different samples. In this simulation, we evaluate two types of MVPA approaches. In the first approach, which we call "trial-wise decoding", an activity pattern is estimated for each trial separately using the least-squares all technique (LSA; Abdulrahman & Henson, 2016). In LSA, each trial gets its own regressor in a first-level GLM. This approach is often used when there is only a single fMRI run available. In the second approach, which we call "run-wise decoding", an activity pattern is estimated for each condition separately. Now, suppose one acquires

a single run with two conditions and twenty trials per condition. In the trial-wise decoding approach, one would estimate the patterns of 40 trials (twenty for condition 1 and twenty for condition 2). Alternatively, suppose that one acquires ten runs with two conditions and again twenty trials per run. In the run-wise decoding approach, one would estimate in total twenty patterns (ten for condition 1 and ten for condition 2).

Formally, for I trials across P conditions, X is ultimately of shape T (timepoints) $\times (I \times P + 1)$ in the trial-wise decoding approach and $T \times (P+1)$ in the run-wise decoding approach (the $+1$ refers to the addition of an intercept). In our trial-wise simulations, we simulate 40 trials (I) across 2 conditions (P). In our run-wise simulations, we simulate 40 trials across 2 conditions in 10 runs. Note that the length of the fMRI run is automatically adjusted to the number of trials (I), conditions (P), trial duration, and interstimulus interval (ISI); increasing any of these parameters will increase the length of the run. We use a trial duration of 1 second and a jittered ISI between 4 and 5 seconds (mean = 4.5 seconds).

The initial (non-HRF-convolved) confound (C, with shape $N \times 1$) is generated with a prespecified correlation to the target (y) by adding noise (ε) drawn from a standard normal distribution multiplied by a scaling factor (γ):

$$C = y + \gamma\varepsilon \tag{B.2}$$

This process starts with a very small scaling factor (γ). If the correlation between the target and the confound is too high, the scaling factor is increased and the process is repeated, which is iterated until the desired correlation has been found. The confound is then scaled from 0 to 1 using min-max scaling. After scaling, similar to the single-trial regressors (X_j), the confound

(C) is also convolved with an HRF (the SPM default), representing a regressor which is parametrically modulated by the value of the confound for each trial (which could represent, for example, reaction time). The confound, C, now represents a time-varying array of shape $T \times 1$. This process is identical for the trial-wise and run-wise decoding approaches. However, when evaluating the efficacy of confound regression (as explained in the next section) in the context of run-wise decoding, we used the means per condition of the confound instead of the trial-wise confound values.

The simulation draws the true activation parameters for the trials from different conditions ($y \in \{0, 1\}$ for $P = 2$) from a normal distribution with a specified mean and standard deviation. To generate null data (i.e., without any true difference across, e.g., two conditions), we generate the true parameters as follows:

$$\beta_{X(y=p)} \sim \mathcal{N}(\mu, \sigma) \tag{B.3}$$

where, in the case of null data, μ represents the same mean for all conditions $p = 0, \ldots, P-1$. The weight of the confound (β_C) is also drawn from a normal distribution with a prespecified mean (μ_C) and standard deviation (σ_C):

$$\beta_C \sim \mathcal{N}(\mu_C, \sigma_C) \tag{B.4}$$

The weights for both X and C are drawn independently for the K voxels (we use $10 \times 10 = 100$ voxels for all our simulations). However, to simulate spatial autocorrelation in our grid of artificial voxels, we smooth all T 2D "volumes" (of shape $\sqrt{K} \times \sqrt{K}$) separately with a 2D Gaussian smoothing kernel with a prespecified standard deviation. For our simulations, we use a standard deviation of 1 for the kernel.

Estimating activity patterns from the data

After generating the signals (s) of shape $T \times K$ (in which the 2D voxel dimension has been flattened to a single dimension), the "activity" parameters ($\hat{\beta}_X$) for the trials (for trial-wise decoding) or conditions (for run-wise decoding) are estimated across all K voxels. We estimate these parameters using a generalized least squares (GLS) model on y using the design matrix X and covariance matrix V:

$$\hat{\beta}_X = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{B.5}$$

where $\hat{\beta}_X$ is of shape $N \times K$ (where N refers to the amount of trials in trial-wise decoding and the amount of conditions in run-wise decoding). Note that C is not part of the design matrix X, here. This would amount to the "control for confounds during pattern estimation" discussed in Supplementary Methods section "Controlling for confounds during pattern estimation". Before entering these activity estimates ($\hat{\beta}_X$) in our decoding pipeline, we divide these them by their standard error to generate t-values (as advised in Misaki et al., 2010):

$$t_{\hat{\beta}_X} = \frac{\hat{\beta}_X}{\sqrt{\sigma^2 \mathrm{diag}(X^T V^{-1} X)^{-1}}} \tag{B.6}$$

where σ^2 is the sum-of-squares error divided by the degrees of freedom ($T - N - 1$).

To summarize, in all of our simulations, we keep the following "experimental" parameters constant: we simulate data from two conditions ($y \in \{0, 1\}$, which we refer to as "condition 0" and "condition 1"), each condition has 40 trials, trial duration is 1 second, ISIs are jittered between 4 and 5 seconds (mean = 4.5), noise is autocorrelated according to an AR(1) process (with $\varphi_1 =$

0.5), and the data is spatially smoothed using a Gaussian filter with a standard deviation of 2 across a 10×10 grid of voxels. For run-wise decoding, we simulate 10 runs.

Testing confound regression on simulated fMRI data

In this simulation, we aim to test whether confound regression is able to remove the influence of confounds in fMRI data for both trial-wise and run-wise decoding analyses. Similar to the analyses reported in the main article, we contrast WDCR and CVCR, expecting that WDCR leads to similar negative bias while CVCR leads to similar (nearly) unbiased results. We use the same pipeline as in the other simulations, which consists of a normalization step to ensure that each feature has a mean of zero and a unit standard deviation, and a support vector classifier with a linear kernel (regularization parameter $C = 1$). The decoding pipeline uses a 10-fold stratified cross-validation scheme. For the run-wise decoding analyses, this is equivalent to a leave-one-run-out cross-validation scheme. Model performance in this simulation is reported as accuracy (as there is no class imbalance). The simulation was repeated 10 times for robustness.

Specifically, we evaluate and report model performance after confound regression both in the trial-wise decoding and run-wise decoding context and for different strengths of the correlation between the target and the confound ($r_{Cy} \in \{0, 0.1, \ldots, 0.9, 1\}$). Because arguably the temporal autocorrelation of fMRI data is the most prominent difference between fMRI and structural MRI data, and thus might impact decoding analyses differently (Gilron et al., 2016), we additionally test CVCR on data that differs in the degree of autocorrelation. We manipulate autocorrelation by temporally smoothing the signals

(y) before fitting the first-level GLM using a Gaussian filter with increasing widths ($\sigma_{\text{filter}} = \{0, 1, \ldots, 5\}$), yielding data with increasing autocorrelation. We used a grid of 4×4 voxels for this simulation to reduce computation time.

Controlling for confounds during pattern estimation

As discussed in the main text, one way to potentially control for confounds in fMRI data is to remove their influence when estimating the activity patterns in an initial first-level model (Woolgar et al., 2014). In this section of the Supplementary methods, we tested the efficacy of this method on simulated fMRI data in both trial-wise decoding and run-wise decoding contexts. Note that the original article on this method (Woolgar et al., 2014) performed run-wise decoding.

Specifically, we tested whether confounds can be controlled by adding the (HRF-convolved) confound to the design matrix X in the first-level pattern estimation procedure. According to @[Woolgar2014-jb], when assuming that the confound has a truly additive effect, adding the confound to the design matrix will yield (single-trial) pattern estimates ($\hat{\beta}_X$X) that only capture unique variance (i.e., no variance related to the confound). In our simulations, we tested two versions of this method. In one version, which we call the "default" version, the confound is added to the design matrix and the (GLS) model is estimated on using the design-matrix including both the single-trial (for trial-wise decoding) or condition regressors (for run-wise decoding) and the confound regressor. In the other version, which we call the "aggressive" version (reflecting the same terminology as the fMRI-denoising method "ICA-AROMA"; Pruim et al., 2015), the confound regressor is first regressed out of the signal ($s_{\text{corr}} = s - C\hat{\beta}_C$) before fitting the regular first-level model using

the design matrix (X) without the confound regressor. The reason for testing these two methods is because it is unclear from the original Woolgar et al. articles (Woolgar et al., 2011, 2014) which version was used and whether the two versions yield different results.

In our simulation, we varied the correlation between the (non-HRF-convolved) confound and the target (here, $y \in \{0, 1\}$). Importantly, we generated data without any effect, i.e., the parameters of the trials from the two different conditions ($\beta_{X|y=0}$ and $\beta_{X|y=1}$) were (independently) drawn from the same normal distribution with a mean of 1 and standard deviation of 0.5. Similar to the other simulations, we use patterns of t-values in our decoding pipeline (unless explicitly stated otherwise). For both trial-wise and run-wise decoding, we report results from the default and aggressive procedure. All analyses are documented in the aforementioned notebook containing the fMRI simulations.

Linear vs nonlinear confound models: predicting VBM and TBSS data based on brain size

In the main text, we used linear models to regress out variance associated with brain size from VBM and TBSS data. Here, we test whether a linear model of the relation between brain size and VBM and TBSS data is suitable, and whether possibly a non-linear model should have been preferred. We do this by performing a model comparison between linear, quadratic, and cubic regression models. All analyses and results can be found in the `brain_data_vs_brainsize.ipynb` notebook from the Github repository associated with this article.

For all analyses in this section, brain size with and without second and third degree polynomials were used as independent variables. As target variables, we created four voxel sets: first, we selected 500 random voxels from the VBM and TBSS data

(voxel set A). Second, to inspect how large the *misfit* of a linear model could be, we selected as target variables the 500 voxels which have the highest quadratic (voxel set B) and cubic (voxel set C) correlation with brain size. Finally we select the 500 voxels with highest linear correlation with brain size (voxel set D), to inspect how large the misfit of a polynomial model is in these voxels.

We applied a 10-fold cross-validation pipeline which consisted of scaling features to mean 0 and standard deviation 1, and fitting an ordinary least squares regression model. Explained variance (R^2) was used as a metric of model performance. The pipeline was repeated 50 times with random shuffling of samples.

For each voxel, we calculated the difference between model performance of linear and polynomial models. Negative differences ($\Delta R^2_{\text{linear}-\text{polynomial}} < 0$) indicate that the polynomial model has higher cross-validated R^2 than a linear model, and thus, that a linear confound regression model would leave variance arguably associated with the confound in the target voxel. We plot the distributions of $\Delta R^2_{\text{linear}-\text{polynomial}}$ to inspect for how many voxels this is the case, and for how many voxels linear models perform better.

B.2 Supplementary results

The following supplementary results describe the results from the supplementary analyses related to controlling for confounds in decoding analyses of (simulated) fMRI data.

Testing confound regression on simulated fMRI data

Here, we evaluated the efficacy of confound regression (both WDCR and CVCR) on simulated fMRI data in both trial-wise

and run-wise decoding analyses across different strengths of the correlation between the target and the confound. Similar to the results reported in the main article, we find that WDCR yields consistent below chance accuracy in both the trial-wise and run-wise decoding analyses (Supplementary Figure B.1, upper panels) and that CVCR yields (nearly) unbiased results for both trial-wise and run-wise decoding (Supplementary Figure B.1, lower panels).

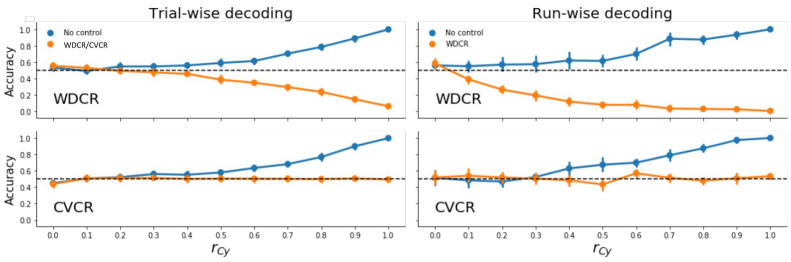


FIGURE B.1 Model performance when using WDCR (upper panels) and CVCR (lower panels) to remove the influence of confounds in simulated fMRI data across different correlations between the confound and the target (r_{Cy}). Error-bars reflect the 95% CI across iterations.

Moreover, CVCR effectively controls for confounds on fMRI data with varying amounts of autocorrelation for both trial-wise and run-wise decoding analyses, as is shown in Supplementary Figure B.2.

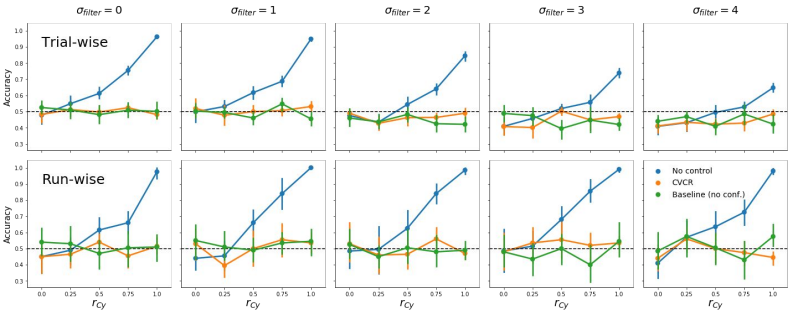


FIGURE B.2 Model performance using CVCR versus no control and baseline (data with no confound) for different levels of autocorrelation (after smoothing with a Gaussian filter with an increasing standard deviation, σ_{filter}) for trial-wise and run-wise decoding. Note that for trial-wise decoding, high autocorrelation leads to below chance-accuracy for CVCR, but this is also present in the baseline data, which suggests that high autocorrelation in general leads to negative bias (at least in our simulation).

Controlling for confounds during pattern estimation

Here, we tested the efficacy of controlling for confound during pattern estimation (as proposed by Woolgar et al., 2014). Similar to the previous Supplementary analyses, we evaluated this method's efficacy in both trial-wise and run-wise decoding analyses. We furthermore evaluated both the "default" (add the confound to the first-level design matrix) and "aggressive" (regress the confound from the signal before fitting the first-level model) approaches.

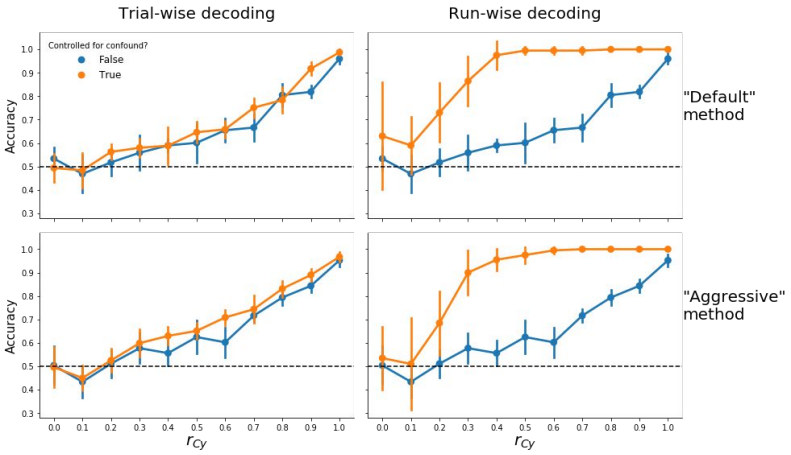


FIGURE B.3 Model performance when controlling for confounds during pattern estimation using the "default" (upper panels) and "aggressive" (lower panels) versions for both trial-wise (left panels) and run-wise decoding (right panels). Note that, in these analyses, patterns of t-values from the first-level model are used as features.

As can be seen in Supplementary Figure B.3, this method fails to control for confounds for all variants that are tested (trial-wise vs. run-wise decoding, "default" vs. "aggressive|). Below, we provide a potential explanation of this bias. We argue that the mechanism underlying this bias is different in trial-wise decoding than in run-wise decoding. We will first focus on the trial-wise decoding analyses.

Explanation for bias in trial-wise decoding analyses

To supplement this explanation, in Supplementary Figure B.4 we visualized the distribution of parameter estimates from the first-level model, $\hat{\beta}_X$ across the two conditions after controlling for the confound during pattern estimation using the "aggressive" version in trial-wise decoding (but the graphs are similar when plotting the data from the "default" version; graphs for

run-wise decoding analyses are, however, different, which will be discussed later).

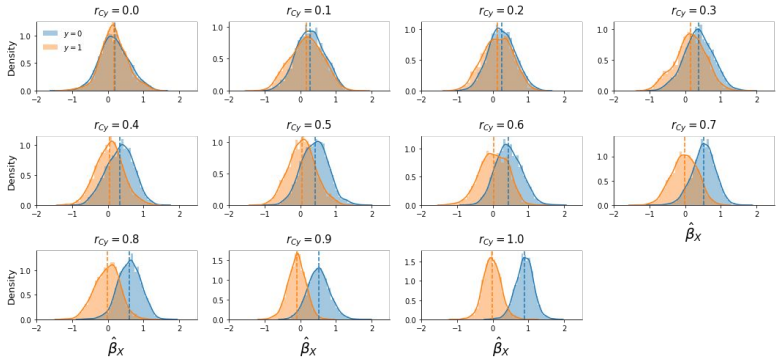


FIGURE B.4 Distribution of first-level parameter estimates, $\hat{\beta}_X$, for the two conditions (condition 0 in blue, condition 1 in orange) across different correlations between the target and the confound (r_{Cy}), with the colored dashed lines indicating the mean feature value for each condition.

When inspecting Supplementary Figure B.4, recall that the data were generated with a confound that was positively correlated with the target. Given that the target variable represents the two different conditions ($y \in \{0, 1\}$), the existence of a positive correlation between the target and the confound implies that the confound increases the activation of the voxel in trials from condition 0. The confound's effect on the voxel increases with higher correlations between the target and the confound. For example, suppose trials from condition 0 and condition 1 both truly activate the voxel with 1 unit ($\beta_{X|y=0} = 1$ and $\beta_{X|y=1} = 1$; Supplementary Figure B.5, upper panel), and that the confound is perfectly correlated with the target ($r_{Cy} = 1$) and thus activates the voxel additionally with 1 unit ($\beta_C = 1$). In this case, without confound control, one would expect the voxel to be activated in response to trials from condition 1 with a magnitude of 2

($\hat{\beta}_{X|y=1} \approx \beta_{X|y=1} + \beta_C$; Supplementary Figure B.5, middle panel). If one in fact would control for the confound by regressing out the confound from the signal (i.e., the "aggressive" approach), one would completely remove both the true effect ($\beta_{X|y=1}$) and the confound effect (β_C), driving the estimated activation parameter for condition 1 towards 0 ($\hat{\beta}_{X|y=1} \approx 0$). The activation parameter for condition 0 is unaffected by removing the confound, as they are uncorrelated, and will thus be estimated correctly ($\hat{\beta}_{X|y=0} = 1$; see Supplementary Figure B.5, lower panel). In this way, controlling for the confound created an artificial "effect": trials from condition 0 seem to activate the voxel more than trials from condition 1 ($\hat{\beta}_{X|y=0} > \hat{\beta}_{X|y=1}$). We believe that this phenomenon underlies the positive bias when controlling for confounds during pattern estimation in trial-wise decoding analyses.[1]

[1] One could argue that this issue only poses a problem when the true parameters are non-zero ($\beta_{X|y=0} = \beta_{X|y=1} \neq 0$) and when the true parameters are in fact all zero ($\beta_{X|y=0} = \beta_{X|y=1} = 0$), there would be no positive bias. This is indeed the case, but we note that the true parameters are never known in empirical anlayses, so we nonetheless advise against using this method.

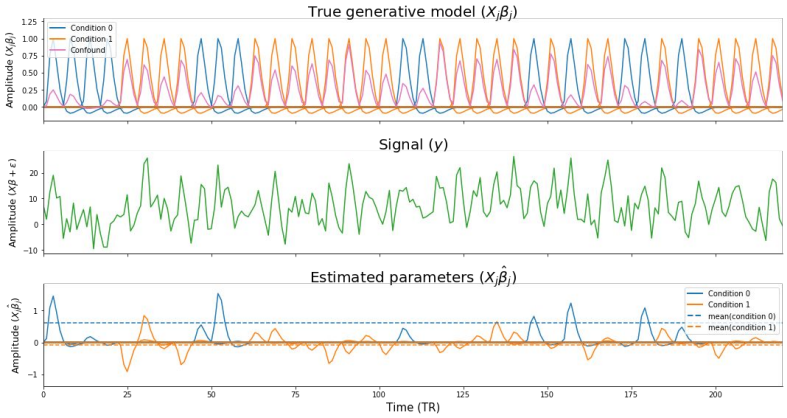


FIGURE B.5 Visualization of the issue underlying positive bias arising when controlling for confounds during pattern estimation. The upper panel ("true generative model") shows the individual single-trial regressors for the different conditions, scaled by their true weight (here, $\beta_{X|y=0} = \beta_{X|y=1} = 1$) and the confound (here, $r_{Cy} = 0.9$). The middle panel ("signal") shows the signal resulting from the generative model (including noise, ε). The lower panel ("estimated parameters") shows the estimated model parameters for the different single-trial regressors. The dashed lines represent the average estimated parameter per condition, which shows that the estimated parameters of the condition that is correlated with the confound are driven towards zero.

Explanation for bias in run-wise decoding analyses

We believe that the cause of bias in run-wise decoding analyses after controlling for confounds during pattern estimation is different than the cause of bias in trial-wise decoding analyses. Upon further inspection of the results of this simulation, we found that in the specific case of run-wise decoding with the "default" approach (i.e., including the confound in the first-level model instead of regressing the confound out of the signal before fitting the first-level model), there is no bias when using patterns of parameter estimates (X) instead of patterns of t-values ($t(X)$; Supplementary Figure B.6, upper row, left panel). Indeed, when

visualizing the distributions of the feature values (Supplementary Figure B.6, lower row), using the "raw" parameter estimates ($\hat{\beta}_X$, left column) or t-values (right column), it is clear that the bias only arises when using t-values. In fact, this bias in t-values is caused by unequal variance (Supplementary Figure B.6, middle panel) of the parameter estimates. The cause of the increased variance for condition 1, here, is due to the fact that a positive correlation between the confound and target (r_{Cy}) results in a relatively higher correlation between the regressor of condition 1 and the confound regressor compared to the correlation between the regressor of condition 0 and the confound regressor. (Note that if the correlation would be negative, e.g., $r_{Cy} = -0.9$, then the reverse would be true.) This issue of classifiers picking up differences in parameter variance in the process of estimating patterns for MVPA has been termed "variance decoding", which is described in detail in Görgen et al. (2017) and Hebart & Baker (2017).

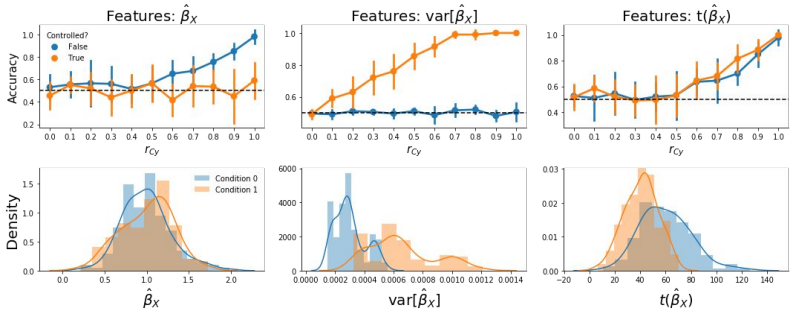


FIGURE B.6 Visualization of model performance and feature distributions based on patterns of "raw" parameter estimates ($\hat{\beta}_X$), variance of parameter estimates (var($\hat{\beta}_X$)), or *t*-values ($t(\hat{\beta}_X)$) after controlling for confounds. The upper row shows the average accuracy across folds across different values of the correlation between the confound and the target (r_{Cy}) for the different types of features. Note that the middle panel shows that "variance decoding" only occurs when controlling for confounds, as model performance is at chance when using patterns of variance estimates (the blue line in the middle panel). The lower row represents the distributions of feature values for the three different statistics when $r_{Cy} = 0.9$.

To summarize, we found that controlling for confound during pattern estimation leads to positive bias in all cases except in run-wise decoding using the "default" approach. We believe that cross-validated confound regression (CVCR) is nevertheless preferable to this method because it both controls for confounds effectively and allows the use of *t*-values (or other statistics based on parameter estimates, like multivariate noise-normalized parameter estimates), which has been shown to be more sensitive than using "raw" parameter estimates in MVPA (Guggenmos et al., 2018; Misaki et al., 2010; Walther et al., 2016).

Linear vs. nonlinear confound models: predicting VBM and TBSS intensity using brain size

Supplementary Figure B.7 (top left panel) shows the distributions of difference in cross-validated R^2 (Linear - Polynomial) for the VBM data with 500 randomly selected voxels (voxel set A). A linear model performs (slightly) better (positive R^2) than a quadratic model for 86.6% of these voxels (mean $\Delta R^2_{\text{linear}-\text{quadratic}} = 0.009$, $SD = 0.006$), and better than a cubic model for 90.8% of the voxels (mean $\Delta R^2_{\text{linear}-\text{cubic}} = 0.019$, $SD = 0.027$). Note that it can be expected that polynomial and cubic models perform better in a minority of the voxels simply due to random noise in the data (since we compare 500 R^2-values), even if the "true" underlying relation is linear. To visualize the quality of fit of these models fit, brain size is plotted against VBM voxel intensity for a randomly selected voxel from this set in the bottom left panel. Lines are regression lines for the linear, quadratic, and cubic models. Supporting the use of a linear model, there is no clear deviation from bivariate normality.

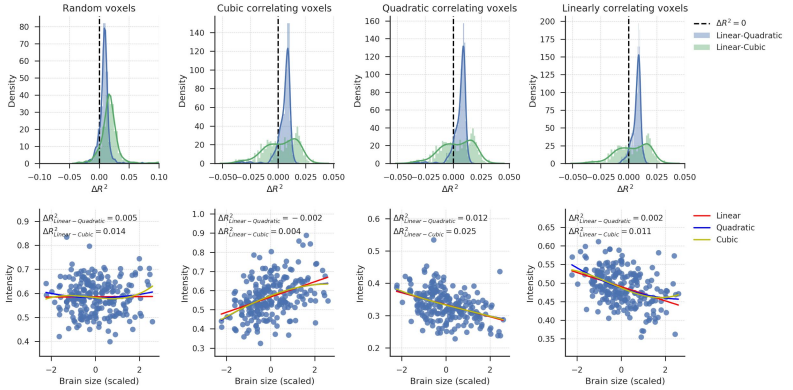


FIGURE B.7 Top row: R^2 distributions for the four voxel sets of the VBM data. Density estimates are obtained by kernel density estimation with a Gaussian kernel and Scott's rule (Scott, 1979) for bandwidth selection. Bottom row: scatter plots of the relation between brain size (scaled to mean 0 and SD 1) and voxel intensity from randomly selected voxels from each voxel sets. These panels are included to visualize the quality of model fits and to inspect whether there are no obvious misfits, i.e., whether the models miss patterns in the data. This is not the case — all models seem to fit the distributions reasonably well.

To explore further how well linear models perform in voxels where we expect polynomial models to perform best, we plotted the R^2 distributions for the 500 voxels with the highest overall quadratic correlation with brain size (voxel set B; second column). These are the voxels where a quadratic model would remove most variance (if used in a confound regression procedure). Also for these voxels, a linear model performs equally well as or better than a quadratic model ($R^2_{\text{linear}-\text{quadratic}} > 0$ for 89.4% of the voxels, mean $\Delta R^2_{\text{linear}-\text{quadratic}} = 0.006$, $SD = 0.007$) and a cubic model ($\Delta R^2_{\text{linear}-\text{cubic}} > 0$ for 57.8% of these voxels, mean $\Delta R^2_{\text{linear}-\text{cubic}} = 0.003$, $SD = 0.015$). For a randomly selected voxel from this set, a scatter plot is included in the bottom panel to visualize how well the regression models fit. The plot indicates no obvious misfit of any of the models.

The same analysis was also performed for the 500 voxels that have the highest overall cubic correlation with brain size (voxel set C; third column). The histograms look similar to the histograms in the second column because voxel sets B and C consisted of largely the same voxels. It should not thus not be surprising that linear models again perform well compared to quadratic models ($\Delta R^2_{\text{linear}-\text{quadratic}} > 0$ for 90.4% of the voxels, mean $\Delta R^2_{\text{linear}-\text{quadratic}} = 0.006$, $SD = 0.006$), and compared to cubic models ($\Delta R^2_{\text{linear}-\text{cubic}} > 0$ for 57.6% of these voxels, mean $\Delta R^2_{\text{linear}-\text{cubic}} = 0.003$, $SD = 0.014$). The bottom panel shows a randomly selected voxel from this voxel set, which again shows no obvious misfit.

Finally, we inspect the 500 voxels with the highest overall linear correlation with brain size (voxel set D, fourth column). Again, these turned out the be partly the same voxels as in set B and C. Therefore, linear models perform again equally well as or better than quadratic models ($\Delta R^2_{\text{linear}-\text{quadratic}} > 0$ for 94.2% of the voxels, mean $\Delta R^2_{\text{linear}-\text{quadratic}} = 0.007$, $SD = 0.004$) and cubic models ($\Delta R^2_{\text{linear}-\text{quadratic}} > 0$ for 59.2% of the voxels, mean $\Delta R^2_{\text{linear}-\text{quadratic}} = 0.004$, $SD = 0.014$). The bottom panel show a randomly selected voxel from this voxel set, and indicates that all models capture the structure in the data.

Together, these results seem to imply that for all voxel sets, linear models perform mostly equally well as or better than polynomial models. Yet, it is interesting to inspect the "worst case" voxels; that is, to inspect how large the maximal misfit of a linear model is. Therefore, in Supplementary Figure B.8, we plot the relation between brain size and VBM intensity for the voxel with most negative $\Delta R^2_{\text{linear}-\text{cubic}}$ from voxel set A (left panel) and voxel set B. For comparison, we also plot the relation between brain size and the voxel where a linear model performs better than a cubic model (selected from voxel set B). For the selected voxels, $\Delta R^2_{\text{linear}-\text{cubic}}$ values are -0.036 (left panel), -0.039(middle panel)

and 0.032. Especially in the middle and right panel, the difference in fit between the linear and cubic model is mostly apparent the tails of the brain size distribution, where the model fit is based on least observations. For most brain sizes, both models make similar predictions about voxel intensity.

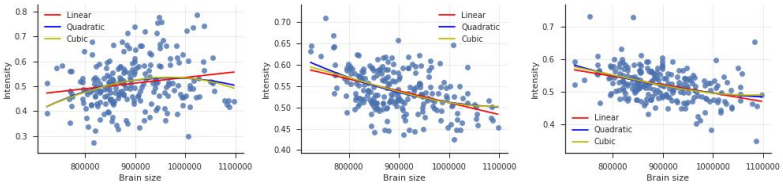


FIGURE B.8 Visualisation of the relation between brain size and VBM intensity for three voxels. The left two voxels have most negative $\Delta R^2_{\text{linear}-\text{cubic}}$ (i.e., the cubic model performs maximally better than the linear model) in voxel sets A and B, respectively. The voxel plotted in the right panel has the most positive $\Delta R^2_{\text{linear}-\text{cubic}}$ in voxel set B.

We repeated the same analyses for the TBSS data, and summarize the results in Supplementary Figure B.9. Since the results are qualitatively the same as for the VBM data, and lead to the same conclusions, we do not discuss them in detail. Those interested can find additional details in the notebook of this simulation.

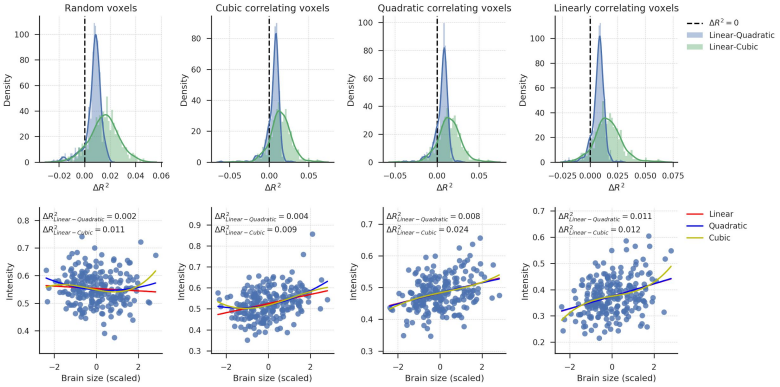


FIGURE B.9 Top row: R^2 distributions for the four voxel sets of the TBSS data. Density estimates are obtained by kernel density estimation with a Gaussian kernel and Scott's rule (Scott, 1979) for bandwidth selection. Bottom row: scatter plots of the relation between brain size (scaled to mean 0 and SD 1) and voxel intensity from randomly selected voxels from each voxel sets. These panels are included to visualize the quality of model fits and to inspect whether there are no obvious misfits, i.e., whether the models miss patterns in the data. This is not the case — all models seem to fit the distributions well.

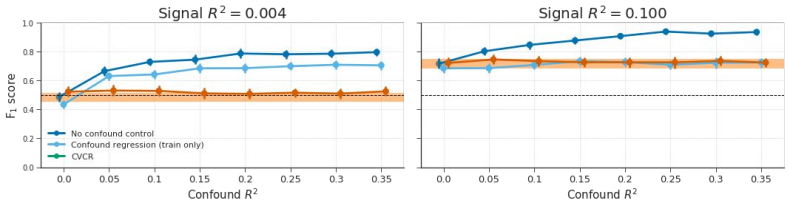


FIGURE B.10 Model performance of fully cross-validated confound regression (CVCR) versus confound regression on the train-set only ("train only") on simulated data without any experimental effect (signal $R^2 = 0.004$; left graph) and with some experimental effect (signal $R^2 = 0.1$; right graph) for different values of confound R^2 (cf. Figure 3.8 in the main text). The orange line represents the average performance (± 1 SD) when confound R^2 = 0, which serves as a "reference performance" for when there is no confounded signal in the data. For both graphs, the correlation between the target and the confound, r_{yC}, is fixed at 0.65. The reason for testing this version of confound regression (i.e., on the train-set only) is because it reduces the computation time substantially compared to fully cross-validated confound regression (as it does not have to compute $X_{\text{test}} = X_{\text{test}} - C_{\text{test}}\hat{\beta}_C$). However, this method seems to yield substantial bias when there is (almost) no signal (left graph), but intriguingly not when there is true signal (right graph).

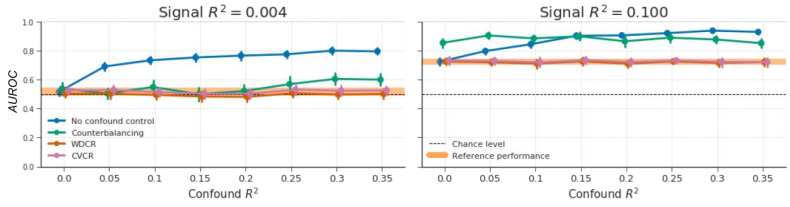


FIGURE B.11 Model performance of the different evaluated methods for confound control but using the AUC-ROC metric to measure model performance instead of F_1 score, as this latter metric has been criticized because it neglects false negatives (Powers, 2011). The results are highly similar to results obtained when using the F_1 score (cf. Figure 3.8 in the main text).

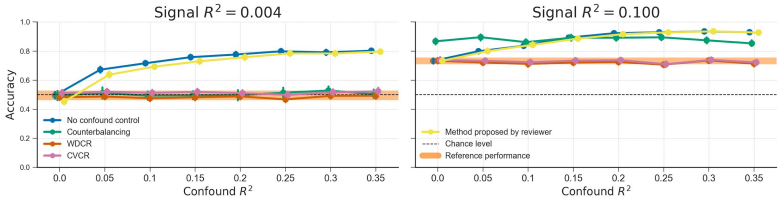


FIGURE B.12 Model performance of the different evaluated methods for confound control, including a method proposed by a reviewer. This method entails training the decoding model on data including the confound as a predictor (i.e., an implementation of the "Include confound in model" method), but setting the confound values to their mean in the test set. The rationale is that the decoding model cannot profit from the confound in the test set. However, contrary to expectations, this method performs similarly to not controlling for confounds.

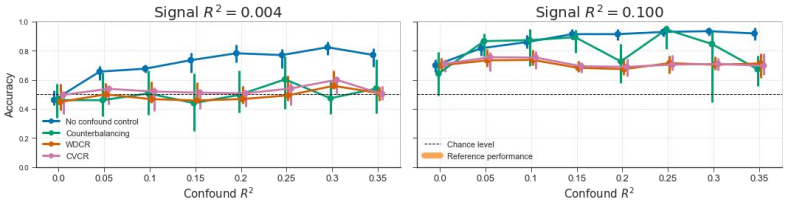


FIGURE B.13 Reproduction of Figure 3.8 from the main text ("generic simulation" results), but with the random subsampling procedure instead of the targeted subsampling procedure (from only a single iteration due to time constraints). This procedure attempts to find a subsample of the data of a given size without a correlation between target and confound for 10.000 tries. If such a subsample cannot be found, the subsample size is decreased by 1, after which again 10.000 attempts are made to find a good subsample with the new size. The results from counterbalancing, here, are qualitatively similar to the results when using the "targeted subsampling" method (cf. Figure 3.8 in the main text), albeit much slower.

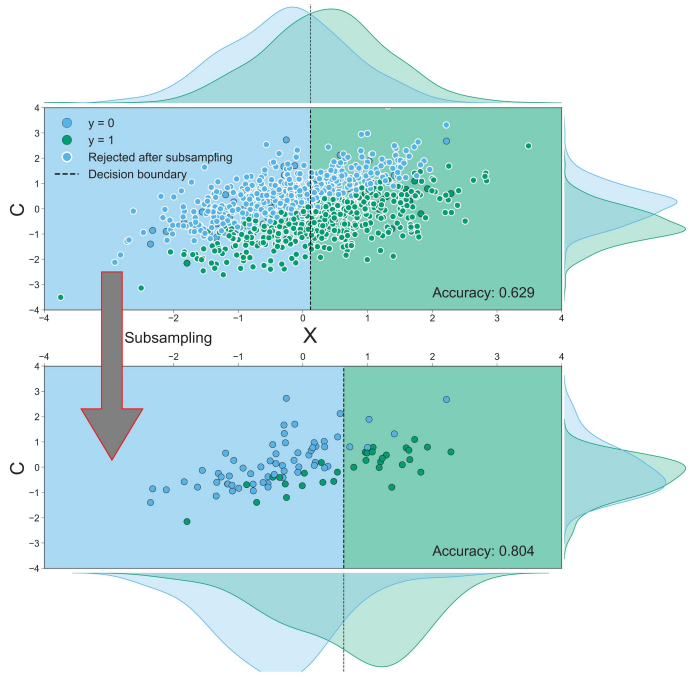


FIGURE B.14 Reproduction of Figure 3.10 from the main text, but with the random subsampling procedure instead of the targeted subsampling procedure. This procedure attempts to find a subsample of the data of a given size without a correlation between target and confound for 10.000 tries. If such a subsample cannot be found, the subsample size is decreased by 1, after which again 10.000 attempts are made to find a good subsample with the new size. The plot shows that also random subsampling can induce a positive bias, even with extreme power loss (90% smaller sample).

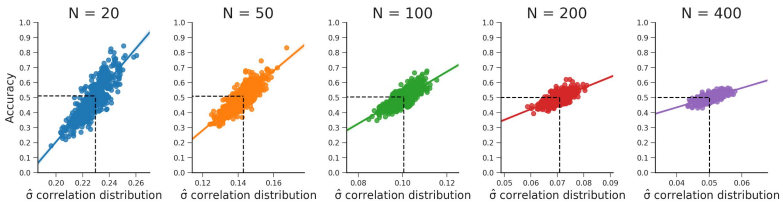


FIGURE B.15 These plots show that the relationship between the standard deviation of the empirical feature-target correlation distribution, $sd(r_{yX})$, and accuracy holds for different samples sizes (i.e., values for N). Note that the predicted accuracy based on the standard deviation expected from the sampling distribution is at 0.5 for every plot. The data were generated in the same manner as reported in the WDCR follow-up section.

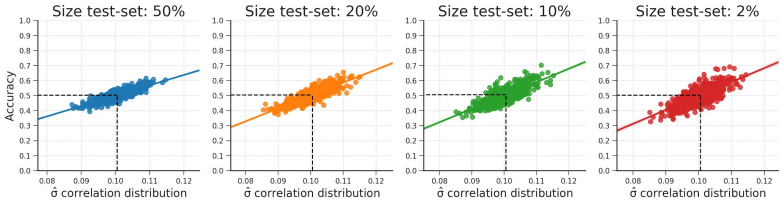


FIGURE B.16 These plots show that the relationship between the standard deviation of the empirical feature-target correlation distribution and accuracy also holds for sizes of the test-set (replicating results from H. Jamalabadi et al., 2016). Note that the predicted accuracy is again at 0.5 for every plot. The data were generated in the same manner as reported in the WDCR follow-up section.

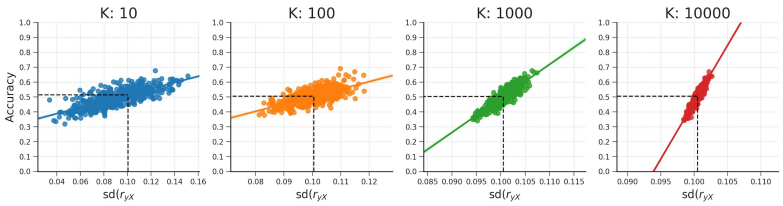


FIGURE B.17 These plots show that the relationship between the standard deviation of the empirical feature-target correlation distribution, $sd(r_{yX})$, and accuracy also holds for different numbers of features (K). Note that the predicted accuracy based on $sd(r_{yX})$ is approximately at 0.5 for every plot. The data were generated in the same manner as reported in the WDCR follow-up section.

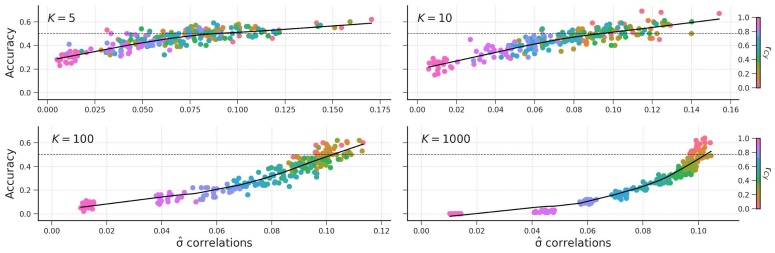


FIGURE B.18 The relation of the standard deviation of the correlation distribution and accuracy for different values of K.

Appendix C

Supplement to Chapter 5

Appendix D

Supplement to Chapter 6

Appendix E

Supplement to Chapter 7

Appendix F

Data, code and materials

Bibliography

Abdulkadir, A., Ronneberger, O., Tabrizi, S. J., & Klöppel, S. (2014). Reduction of confounding effects with voxel-wise gaussian process regression in structural MRI. *2014 International Workshop on Pattern Recognition in Neuroimaging*, 1–4.

Abdulrahman, H., & Henson, R. N. (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.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.*, *8*, 14.

Alizadeh, S., Jamalabadi, H., Schönauer, M., Leibold, C., & Gais, S. (2017). Decoding cognitive concepts from neuroimaging data using multivariate pattern analysis. *Neuroimage*, *159*, 449–458.

Allefeld, C., Görgen, K., & Haynes, J.-D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *Neuroimage*, *141*, 378–392.

Allefeld, C., & Haynes, J.-D. (2014). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated manova. *Neuroimage*, *89*, 345–357.

Anderson, M. L. (2016). Précis of after phrenology: Neural reuse and the interactive brain. *Behavioral and Brain Sciences*, *39*.

Andrews-Hanna, J. R., Smallwood, J., & Spreng, R. N. (2014). The default network and self-generated thought: Compo-

nent processes, dynamic control, and clinical relevance. *Annals of the New York Academy of Sciences*, *1316*(1), 29.

Bangalore, S. S., Prasad, K. M. R., Montrose, D. M., Goradia, D. D., Diwadkar, V. A., & Keshavan, M. S. (2008). Cannabis use and brain structural alterations in first episode schizophrenia—a region of interest, voxel based morphometric study. *Schizophr. Res.*, *99*(1), 1–6.

Barnes, J., Ridgway, G. R., Bartlett, J., Henley, S. M. D., Lehmann, M., Hobbs, N., Clarkson, M. J., MacManus, D. G., Ourselin, S., & Fox, N. C. (2010). Head size, age and gender adjustment in MRI studies: A necessary nuisance? *Neuroimage*, *53*(4), 1244–1255.

Barrett, L. F. (2012). Emotions are real. *Emotion*, *12*(3), 413.

Barrett, L. F., & Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: Towards an integrative functional architecture of the brain. *Current Opinion in Neurobiology*, *23*(3), 361–372.

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, *16*(7), 419–429.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1281–1289.

Bastiaansen, J. A., Thioux, M., & Keysers, C. (2009). Evidence for mirror systems in emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2391–2404.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Brodtmann, A., Puce, A., Darby, D., & Donnan, G. (2009). Regional fMRI brain activation does correlate with global brain volume. *Brain Research*, *1259*, 17–25.

Brosch, T., Bar-David, E., & Phelps, E. A. (2013). Implicit race bias decreases the similarity of neural representations of black and white faces. *Psychological Science*, *24*(2), 160–166.

Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. *Front. Neurosci.*, *11*, 543.

Carlson, T. A., & Wardle, S. G. (2015). Sensible decoding. *Neuroimage*, *110*, 217–218.

Carr, L., Iacoboni, M., Dubeau, M.-C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences*, *100*(9), 5497–5502.

Chekroud, A. M., Ward, E. J., Rosenberg, M. D., & Holmes, A. J. (2016). Patterns in the human brain mosaic discriminate males from females. *Proc. Natl. Acad. Sci. U. S. A.*, *113*(14), E1968.

Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C., Initiative, A. D. N., & others. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*, *60*(1), 59–70.

Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P., & Singer, T. (2016). Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. *Nature Communications*, *7*(1), 1–12.

Craddock, R. C., Holtzheimer, P. E., 3rd, Hu, X. P., & Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magn. Reson. Med.*, *62*(6), 1619–1628.

Craig, A. D., & Craig, A. (2009). How do you feel–now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10*(1).

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., & Alzheimer's Disease Neuroimaging Initiative. (2011). Automatic classification of patients with alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *Neuroimage*, *56*(2), 766–781.

Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., & Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *Neuroimage*, *97*, 271–283.

Decety, J. (2011). Dissecting the neural mechanisms mediating empathy. *Emotion Review*, *3*(1), 92–108.

Del Giudice, M., Lippa, R. A., Puts, D. A., Bailey, D. H., Bailey, J. M., & Schmitt, D. P. (2016). Joel et al.'s method systematically fails to detect large, consistent sex differences. *Proc. Natl. Acad. Sci. U. S. A.*, *113*(14), E1965.

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self-and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *24*(8), 1742–1752.

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput. Biol.*, *13*(4), e1005508.

Dixon, L. (1999). Dual diagnosis of substance abuse in schizophrenia: Prevalence and impact on outcomes. *Schizophr. Res.*, *35 Suppl*, S93–100.

Douaud, G., Smith, S., Jenkinson, M., Behrens, T., Johansen-Berg, H., Vickers, J., James, S., Voets, N., Watkins, K., Matthews, P. M., & James, A. (2007). Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain*, *130*(Pt 9), 2375–2386.

Dubois, J., Galdi, P., Han, Y., Paul, L. K., & Adolphs, R. (2018). Resting-state functional brain connectivity best predicts the personality dimension of openness to experience. *Personality Neuroscience*, *1*.

Dukart, J., Schroeter, M. L., Mueller, K., Initiative, A. D. N., & Others. (2011). Age correction in dementia–matching to a healthy brain. *PLoS One*, *6*(7), e22193.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185.

Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of emotional information in voice-sensitive cortices. *Current Biology*, *19*(12), 1028–1033.

Etzel, J. A., Valchev, N., & Keysers, C. (2011). The impact of certain methodological choices on multivariate analysis of fMRI data with support vector machines. *Neuroimage*, *54*(2), 1159–1167.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, *8*(9), 396–403.

Gelder, B. de, Van den Stock, J., Meeren, H. K., Sinke, C. B., Kret, M. E., & Tamietto, M. (2010). Standing up for the body. Recent progress in uncovering the networks involved in the perception of bodies and bodily expressions. *Neuroscience & Biobehavioral Reviews*, *34*(4), 513–527.

Gilbert, S. J., Swencionis, J. K., & Amodio, D. M. (2012). Evaluative vs. Trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia*, *50*(14), 3600–3611.

Gilron, R., Rosenblatt, J. D., & Mukamel, R. (2016). Addressing the" problem" of temporal correlations in mvpa analysis. *2016 International Workshop on Pattern Recognition in Neuroimaging (Prni)*, 1–4.

Gilron, R., Rosenblatt, J., Koyejo, O., Poldrack, R. A., & Mukamel, R. (2017). What's in a pattern? Examining the

type of signal multivariate analysis uncovers at the group level. *Neuroimage*, *146*, 113–120.

Glezerman, M. (2016). Yes, there is a female and a male brain: Morphology versus functionality. *Proceedings of the National Academy of Sciences*, *113*(14), E1971–E1971.

Goldstein, J. M., Seidman, L. J., Horton, N. J., Makris, N., Kennedy, D. N., Caviness, V. S., Jr, Faraone, S. V., & Tsuang, M. T. (2001). Normal sexual dimorphism of the adult human brain assessed by in vivo magnetic resonance imaging. *Cereb. Cortex*, *11*(6), 490–497.

Good, C. D., Johnsrude, I., Ashburner, J., Henson, R. N., Friston, K. J., & Frackowiak, R. S. (2001). Cerebral asymmetry and the effects of sex and handedness on brain structure: A voxel-based morphometric analysis of 465 normal adult human brains. *Neuroimage*, *14*(3), 685–700.

Görgen, K., Hebart, M. N., Allefeld, C., & Haynes, J.-D. (2017). The same analysis approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. *Neuroimage*.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife*, *7*.

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *Neuroimage*, *173*, 434–447.

Gur, R. C., Turetsky, B. I., Matsui, M., Yan, M., Bilker, W., Hughett, P., & Gur, R. E. (1999). Sex differences in brain gray and white matter in healthy young adults: Correlations with cognitive performance. *J. Neurosci.*, *19*(10), 4065–4072.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389–422.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, *87*, 96–110.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *Neuroimage*, *62*(2), 852–855.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron*, *87*(2), 257–270.

Hebart, M. N., & Baker, C. I. (2017). Deconstructing multivariate decoding for the study of brain function. *Neuroimage*.

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *Elife*, *7*.

Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C., & Gais, S. (2016). Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. *Human Brain Mapping*, *37*(5), 1842–1855.

Jamalabadi, H., Alizadeh, S., Schönauer - Human brain …, M., & 2016. (2016). Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. *Wiley Online Library*.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782–790.

Jimura, K., & Poldrack, R. A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, *50*(4), 544–552.

Joel, D., & Fausto-Sterling, A. (2016). Beyond sex differences: New approaches for thinking about variation in brain struc-

ture and function. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *371*(1688), 20150451.

Johnstone, T., Ores Walsh, K. S., Greischar, L. L., Alexander, A. L., Fox, A. S., Davidson, R. J., & Oakes, T. R. (2006). Motion correction and the use of motion covariates in multiple-subject fMRI analysis. *Hum. Brain Mapp.*, *27*(10), 779–788.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Kendall, M. G., & Stuart, A. (1973). Functional and structural relationship. *The Advanced Theory of Statistics*, *2*, 399–343.

Keysers, C., & Gazzola, V. (2014). Dissociating the ability and propensity for empathy. *Trends in Cognitive Sciences*, *18*(4), 163–166.

Kostro, D., Abdulkadir, A., Durr, A., Roos, R., Leavitt, B. R., Johnson, H., Cash, D., Tabrizi, S. J., Scahill, R. I., Ronneberger, O., Klöppel, S., & Track-HD Investigators. (2014). Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. *Neuroimage*, *98*, 405–415.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009a). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009b). Circular analysis in systems neuroscience: The dangers of double dipping. *Nat. Neurosci.*, *12*(5), 535–540.

Krishnan, A., Woo, C.-W., Chang, L. J., Ruzic, L., Gu, X., López-Solà, M., Jackson, P. L., Pujol, J., Fan, J., & Wager, T. D. (2016). Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *Elife*, *5*, e15166.

Kveraga, K., Boshyan, J., Adams Jr, R. B., Mote, J., Betz, N., Ward, N., Hadjikhani, N., Bar, M., & Barrett, L. F. (2015).

If it bleeds, it leads: Separating threat from mere negativity. *Social Cognitive and Affective Neuroscience*, *10*(1), 28–35.

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage*, *54*(3), 2492–2502.

Lamm, C., & Majdandžić, J. (2015). The role of shared neural activations, mirror neurons, and morality in empathy–a critical comment. *Neuroscience Research*, *90*, 15–24.

Lang, P. J. (2005). International affective picture system (iaps): Affective ratings of pictures and instruction manual. *Technical Report*.

Lang, P. J., Bradley, M. M., Cuthbert, B. N., & others. (1997). International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, *1*, 39–58.

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding attended information in short-term memory: An EEG study. *J. Cogn. Neurosci.*, *25*(1), 127–142.

Legrand, D., & Ruby, P. (2009). What is self-specific? Theoretical investigation and critical review of neuroimaging results. *Psychological Review*, *116*(1), 252.

Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitations. *Psychological Bulletin*, *137*(5), 834.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *The Behavioral and Brain Sciences*, *35*(3), 121.

Long, B., Yu, C. P., & Konkle, T. (2017). A mid-level organization of the ventral stream. *bioRxiv*.

Lüders, E., Steinmetz, H., & Jäncke, L. (2002). Brain size and grey matter volume in the healthy human brain. *Neuroreport*, *13*(17), 2371–2374.

McGrath, J., Saha, S., Chant, D., & Welham, J. (2008). Schizophrenia: A concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.*, *30*, 67–76.

Medford, N., & Critchley, H. D. (2010). Conjoint activity of anterior insular and anterior cingulate cortex: Awareness and response. *Brain Structure and Function*, *214*(5-6), 535–549.

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage*, *53*(1), 103–118.

Mischkowski, D., Crocker, J., & Way, B. M. (2016). From painkiller to empathy killer: Acetaminophen (paracetamol) reduces empathy for pain. *Social Cognitive and Affective Neuroscience*, *11*(9), 1345–1353.

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *Neuroimage*, *103*, 130–138.

Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends Cogn. Sci.*, *19*(10), 551–554.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, *56*(2), 400–410.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006a). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006b). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.*, *10*(9), 424–430.

O'Brien, L. M., Ziegler, D. A., Deutsch, C. K., Frazier, J. A., Herbert, M. R., & Locascio, J. J. (2011). Statistical adjustments for brain size in volumetric neuroimaging studies: Some practical implications in methods. *Psychiatry Res.*, *193*(2), 113–122.

Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *J. Mach. Learn. Res.*, *11*(Jun), 1833–1863.

Oosterwijk, S., & Barrett, L. F. (2014). Embodiment in the construction of emotion experience and emotion understanding. *Routledge Handbook of Embodied Cognition. New York: Routledge*, 250–260.

Oosterwijk, S., Lindquist, K. A., Anderson, E., Dautoff, R., Moriguchi, Y., & Barrett, L. F. (2012). States of mind: Emotions, body feelings, and thoughts share distributed neural networks. *NeuroImage*, *62*(3), 2110–2128.

Oosterwijk, S., Mackey, S., Wilson-Mendenhall, C., Winkielman, P., & Paulus, M. P. (2015). Concepts in context: Processing mental state concepts with internal or external focus involves different neural systems. *Social Neuroscience*, *10*(3), 294–307.

Parkinson, C., Liu, S., & Wheatley, T. (2014). A common cortical metric for spatial, temporal, and social distance. *Journal of Neuroscience*, *34*(5), 1979–1987.

Parra, L. C., Spence, C. D., Gerson, A. D., & Sajda, P. (2005). Recipes for the linear analysis of EEG. *Neuroimage*, *28*(2), 326–341.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R.,

Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, *12*(Oct), 2825–2830.

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience*, *30*(30), 10127–10134.

Popov, V., Ostarek, M., & Tenison, C. (2018). Practices and pitfalls in inferring neural representations. *NeuroImage*, *174*, 340–351.

Powers, D. M. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv Preprint arXiv:2010.16061*.

Pruim, R. H., Mennes, M., Rooij, D. van, Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). ICA-aroma: A robust ica-based strategy for removing motion artifacts from fMRI data. *Neuroimage*, *112*, 267–277.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–360.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press.

Ramakrishnan, K., Scholte, H. S., Groen, I. I. A., Smeulders, A. W. M., & Ghebreab, S. (2014). Visual dictionaries as intermediate features in the human brain. *Front. Comput. Neurosci.*, *8*, 168.

Rao, A., Monteiro, J. M., Mourao-Miranda, J., & Alzheimer's Disease Initiative. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *Neuroimage*, *150*, 23–49.

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2017). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *Br. J. Philos. Sci.*

Rosenblatt, J. D. (2016). Multivariate revisit to "sex beyond the genitalia". *Proc. Natl. Acad. Sci. U. S. A.*, *113*(14), E1966–7.

Rütgen, M., Seidel, E.-M., Silani, G., Riečansky, I., Hummer, A., Windischberger, C., Petrovic, P., & Lamm, C. (2015). Placebo analgesia and its opioidergic regulation suggest that empathy for pain is grounded in self pain. *Proceedings of the National Academy of Sciences*, *112*(41), E5638–E5646.

Sabatinelli, D., Fortune, E. E., Li, Q., Siddiqui, A., Krafft, C., Oliver, W. T., Beck, S., & Jeffries, J. (2011). Emotional perception: Meta-analyses of face and natural scene processing. *Neuroimage*, *54*(3), 2524–2533.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), 605–610.

Sedgwick, P. (2013). Analysing case-control studies: Adjusting for confounding. *BMJ: British Medical Journal*, *346*.

Sepehrband, F., Lynch, K. M., Cabeen, R. P., Gonzalez-Zacarias, C., Zhao, L., D'Arcy, M., Kesselman, C., Herting, M. M., Dinov, I. D., Toga, A. W., & Clark, K. A. (2018). Neuroanatomical morphometric characterization of sex differences in youth using statistical learning. *Neuroimage*, *172*, 217–227.

Singer, T. (2012). The past, present and future of social neuroscience: A european perspective. *Neuroimage*, *61*(2), 437–449.

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., Ciccarelli, O., Cader, M. Z., Matthews, P. M., & Behrens, T. E. J. (2006). Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *Neuroimage*, *31*(4), 1487–1505.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and struc-

tural MR image analysis and implementation as FSL. *Neuroimage*, *23 Suppl 1*, S208–19.

Smith, S. M., & Nichols, T. E. (2018). Statistical challenges in "big data" human neuroimaging. *Neuron*, *97*(2), 263–268.

Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience*, *21*(3), 489–510.

Spunt, R. P., & Lieberman, M. D. (2012). An integrative model of the neural systems supporting the comprehension of observed emotional behavior. *Neuroimage*, *59*(3), 3050–3059.

Stelzer, J., Buschmann, T., Lohmann, G., Margulies, D. S., Trampel, R., & Turner, R. (2014). Prioritizing spatial accuracy in high-resolution fMRI data using multivariate feature weight mapping. *Frontiers in Neuroscience*, *8*, 66.

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *Neuroimage*, *77*, 157–165.

Uddin, L. Q., Iacoboni, M., Lange, C., & Keenan, J. P. (2007). The self and social cognition: The role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, *11*(4), 153–157.

Van Haren, N. E., Cahn, W., Hulshoff Pol, H. E., & Kahn, R. S. (2013). Confounders of excessive brain volume loss in schizophrenia. *Neurosci. Biobehav. Rev.*, *37*(10 Pt 1), 2418–2423.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *Neuroimage*, *48*(3), 564–584.

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, *180*, 68–77.

Waarde, J. A. van, Scholte, H. S., Oudheusden, L. J. B. van, Verwey, B., Denys, D., & Wingen, G. A. van. (2014). A func-

tional MRI marker may predict the outcome of electroconvulsive therapy in severe and treatment-resistant depression. *Mol. Psychiatry*, *20*, 609.

Wacholder, S., Silverman, D. T., McLaughlin, J. K., & Mandel, J. S. (1992). Selection of controls in case-control studies. III. Design options. *Am. J. Epidemiol.*, *135*(9), 1042–1050.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, *137*, 188–200.

Waytz, A., & Mitchell, J. P. (2011). Two mechanisms for simulating other minds: Dissociations between mirroring and self-projection. *Current Directions in Psychological Science*, *20*(3), 197–200.

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., & Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *Neuroimage*, *110*, 48–59.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS One*, *11*(3), e0152719.

Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K., & Barsalou, L. W. (2011). Grounding emotion in situated conceptualization. *Neuropsychologia*, *49*(5), 1105–1127.

Woolgar, A., Golland, P., & Bode, S. (2014). Coping with confounds in multivoxel pattern analysis: What should we do about reaction time differences? A comment on todd, nystrom & cohen 2013. *Neuroimage*, *98*, 506–512.

Woolgar, A., Thompson, R., Bor, D., & Duncan, J. (2011). Multivoxel coding of stimuli, rules, and responses in human frontoparietal cortex. *Neuroimage*, *56*(2), 744–752.

Wurm, M. F., & Lingnau, A. (2015). Decoding actions at different levels of abstraction. *Journal of Neuroscience*, *35*(20), 7727–7735.

Yu-Feng, Z., Yong, H., Chao-Zhe, Z., Qing-Jiu, C., Man-Qiu, S., Meng, L., Li-Xia, T., Tian-Zi, J., & Yu-Feng, W. (2007).

Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain and Development*, *29*(2), 83–91.

Zaki, J., & Ochsner, K. N. (2012). The neuroscience of empathy: Progress, pitfalls and promise. *Nature Neuroscience*, *15*(5), 675–680.

Zaki, J., Wager, T. D., Singer, T., Keysers, C., & Gazzola, V. (2016). The anatomy of suffering: Understanding the relationship between nociceptive and empathic pain. *Trends in Cognitive Sciences*, *20*(4), 249–259.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging*, *20*(1), 45–57.

Contributions to the chapters

List of other publications

van der Maas, H.L.J., **Snoek, L.**, & Stevenson, C. (2021). How much intelligence is there in artificial intelligence? A 2020 update.

Hoogeveen, S., **Snoek, L.**, & Van Elk, M. (2020). Religious belief and cognitive conflict sensitivity: A preregistered fMRI study. *Cortex, 129*, 247–265.

van Elk, M., & **Snoek, L.** (2020). The relationship between individual differences in gray matter volume and religiosity and mystical experiences: A preregistered voxel-based morphometry study. *European Journal of Neuroscience, 51*(3), 850–865.

Van Mourik, T., **Snoek, L.**, Knapen, T., & Norris, D. G. (2018). Porcupine: a visual pipeline tool for neuroimaging analysis. *PLoS computational biology, 14*(5), e1006064.

Nederlandse samenvatting (Summary in Dutch)

Replace this with the Dutch title of your thesis

The summary goes here.

Acknowledgments

This section is optional, but theses typically include acknowledgments (*dankwoord* in Dutch) at the end. You may want to mix languages to thank people in their native tongue (though most Dutch speakers write it entirely in Dutch). But the standard language of the thesis template is English. You can switch temporarily by wrapping the text in language tags like so: `[Your Dutch text here]{lang=nl}`. This is important for things like hyphenation to work properly.