#OHBM2021
@JoramSoch

Can behavioral responses be decoded indirectly?

Abstract #2307
Poster #2307

Decoding behavioral responses from fMRI without learning behavioral responses from fMRI

Joram Soch[1,2,9,•], John-Dylan Haynes[1–8]

Bernstein Center for Computational Neuroscience Berlin

[1] Bernstein Center for Computational Neuroscience, [2] Berlin Center for Advanced Neuroimaging, [3] Berlin School of Mind and Brain, Berlin, Germany
[4] Clinic for Neurology, Charité – Universitätsmedizin Berlin, Germany, [5] Department of Psychology, Humboldt-Universität zu Berlin, Germany
[6] EXC NeuroCure, Charité Berlin, [7] EXC Science of Intelligence, TU Berlin, [8] CRC Volition and Cognitive Control, TU Dresden, Germany
[9] German Center for Neurodegenerative Diseases, Göttingen, Germany, • BCCN Berlin, Philippstraße 13, Haus 6, 10115 Berlin, Germany;
e-mail address: **joram.soch@bccn-berlin.de**.

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN

Introduction

The data acquired during a functional magnetic resonance imaging (fMRI) experiment can usually be categorized into experimental design **X** (e.g. experimental conditions, modulator variables), measured signals **Y** (i.e. BOLD signals in several voxels) and behavioral data **Z** (e.g. button presses, stimulus ratings).

In multivariate pattern analysis (MVPA) of fMRI data [1,2], behavioral data are typically decoded by training an algorithm to predict the recorded responses **Z** based on the measured data **Y** (direct response decoding, dRD).

Here we show that this can be achieved without constructing an explicit mapping from fMRI signals to behavioral responses. In fact, behavioral data can also be decoded when first reconstructing the experimental design **X** from measured data **Y** and then predicting behavioral responses **Z** from the reconstructed design **X** (indirect response decoding, iRD).

Theory

Here, we compared three decoding algorithms (see Figure 1):

(1) *Stimulus-based response decoding* (sbRD): First, behavioral data are directly predicted from the experimental design, $Z = g(X)$, e.g. using a logistic regression model predicting discrete button presses from continuous stimulus variables.

$$X \rightarrow Z$$

(2) *Direct response decoding* (dRD): Second, behavioral responses are directly predicted from measured signals, $Z = h(Y)$, e.g. by training a classifier to distinguish behavioral choices based on trial-wise fMRI response amplitudes.

$$Y \rightarrow Z$$

(3) *Indirect response decoding* (iRD): Here, the experimental design is decoded from measured signals, $X = f^{-1}(Y)$, e.g. by training a classifier to distinguish experimental conditions based on fMRI responses. Then, the estimated behavioral model from (1) is used to indirectly predict behavioral responses from the experimental design reconstructed from fMRI signals: $Z = g(X) = g(f^{-1}(Y))$.

$$Y \rightarrow X \rightarrow Z$$

All these models are estimated in a cross-validated fashion, using leave-one-out cross-validation over fMRI recording sessions [3]. Behavioral models (1) were estimated using linear regression or logistic regression, depending on the type of experimental design variables and behavioral response variables considered (see Table 1). Neurophysiological mappings from fMRI data to behavioral choices (2) or experimental conditions (3) were established using whole-brain support vector machines (SVM) for classification or regression, depending on whether categorical or parametric variables were decoded (see Table 1).

Experiment

We analyzed the entire data set ($N = 108$ subjects) from the Neuroimaging Analysis and Replication and Prediction Study (NARPS) [4,5]. In this experiment (see Figure 2), subjects were offered a mixed gamble in each trial, with certain amounts of money to win or lose (= experimental design X), and then indicated favorability of the bet (= behavioral responses Z) using a four-point Likert scale [6]. In our investigations, we tested several analysis strategies:

Ana	Design X	Signals Y	Behavior Z	Model X → Z
1	high/low gain high/low loss	whole-brain fMRI signals	accept/reject	conditional probabilities
2	gain value loss value	whole-brain fMRI signals	accept/reject	logistic regression
3	high/low gain high/low loss	whole-brain fMRI signals	favorability rating	linear regression
4	gain value loss value	whole-brain fMRI signals	favorability rating	linear regression

Table 1. *Analyses used for indirect response decoding.* Both, experimental design variables and subjects' behavioral responses, can be regarded as categorical or parametric.

Methods: *Stimulus-based, direct and indirect response decoding.*

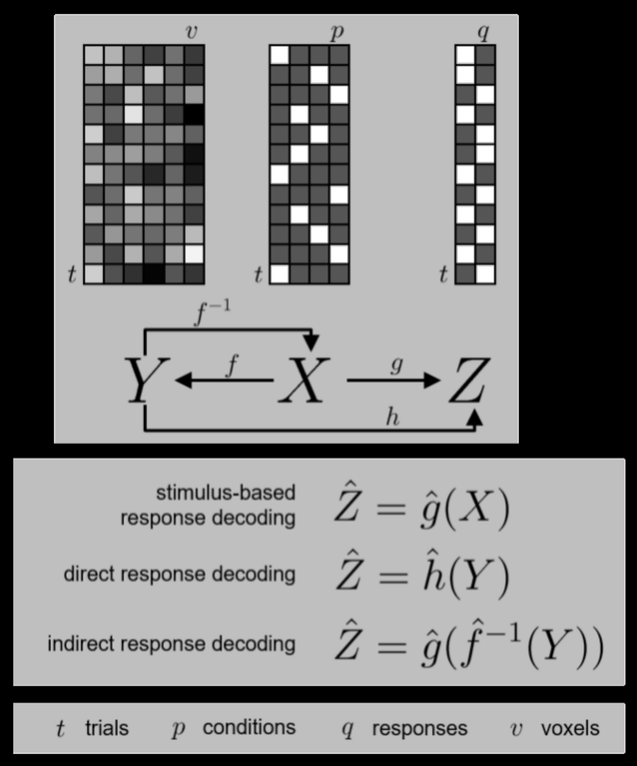

Figure 1. *Theory behind neurobehavioral decoding.* The analyses operate on the trial-by-voxel matrix **Y** (e.g. fMRI signals in a searchlight, region of interest or the whole brain), the trial-by-condition matrix **X** (i.e. experimental design) and the trial-by-response matrix **Z** (i.e. behavioral responses). The theory assumes a neurophysiological model f (from X to Y, inverse f^{-1}), a purely behavioral model g (from X to Z) as well as a mapping for direct response decoding h (from Y to Z).

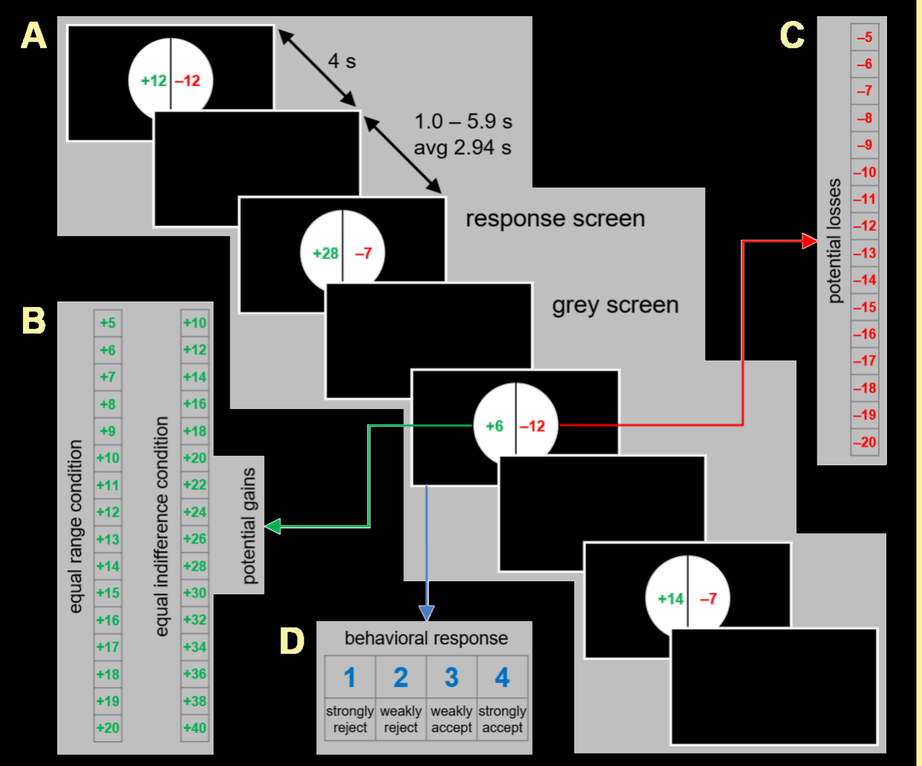

Figure 2. *Experimental design of mixed gambling task.* **(A)** In each trial, a certain amount of money to win (green) and a certain amount of money to loose (red) were displayed. **(B)** Potential gains, in the equal range (left) and equal indifference (right) condition, a between-subject factor. **(C)** Potential losses, in both experimental conditions. Each combination of potential gain and potential loss was presented exactly once to each subject, resulting in $16 \times 16 = 256$ trials. **(D)** In each trial, subjects made their response on a four-point scale.

Results: *Decoding accuracies for different decoding analyses.*

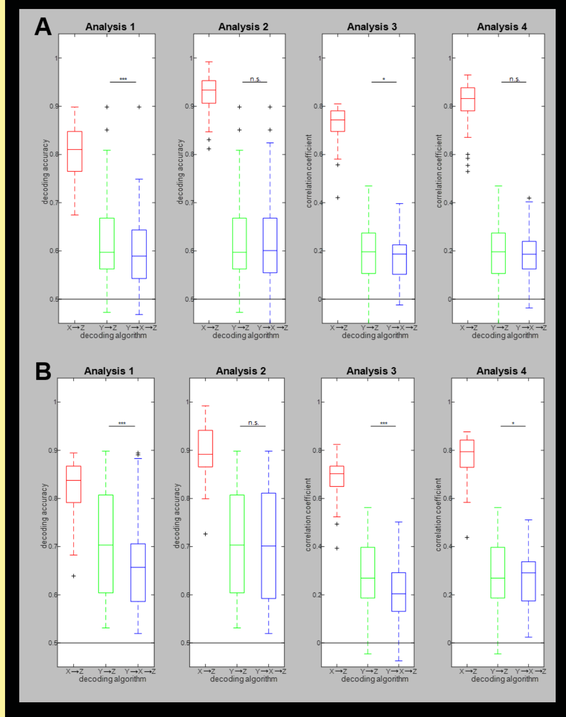

Figure 3. *Decoding accuracies as a function of decoding method.* Performances of three decoding algorithms (red: sbRD; green: dRD; blue: iRD) are visualized using box plots for four analysis types (see Table 1), separately for the **(A)** equal range and the **(B)** equal indifference condition. dRD and iRD are tested against each other using a two-tailed paired t-test (n.s. = not significant; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$).

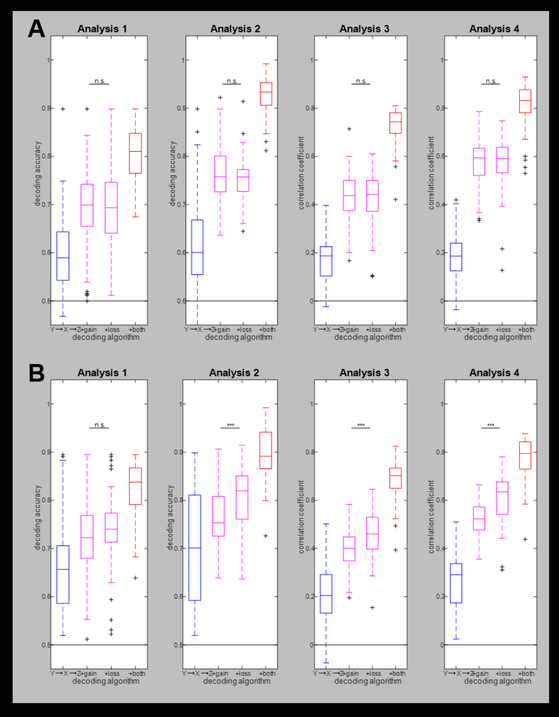

Figure 4. *Decoding accuracies as a function of information supply.* Performances of indirect response decoding for **(A)** equal range and **(B)** equal indifference, depending on whether both experimental dimensions were reconstructed from the data (blue), whether one of the experimental dimensions was not reconstructed, but supplied to the behavioral model (magenta) or whether both experimental dimensions were known (red).

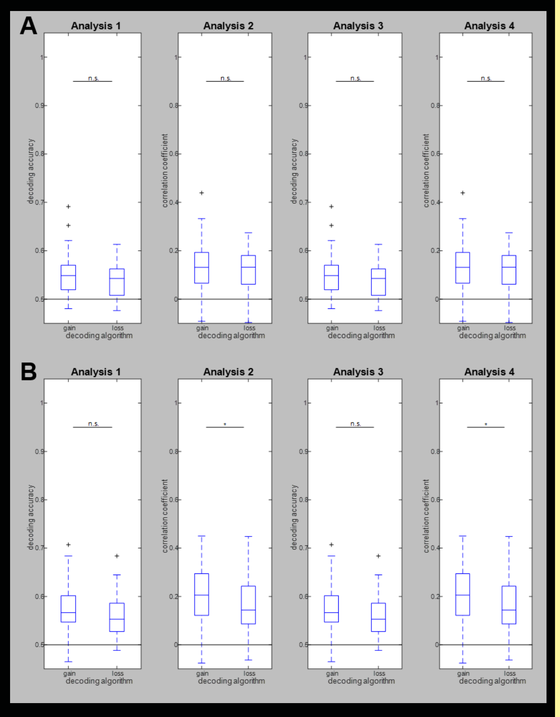

Figure 5. *Decoding accuracies for experimental design variables.* Performances of reconstructing the experimental design from measured fMRI signals. The layout follows the one of Figure 4 and gives results for four analysis types (see Table 1), separately for the **(A)** equal range and the **(B)** equal indifference condition.

Discussion

In this proof-of-concept study, we have demonstrated that behavioral responses can be decoded without training on neurophysiological data measured during behavioral responses (*direct response decoding*, dRD), but rather indirectly by taking a detour via the experimental design (*indirect response decoding*, iRD). This is particularly interesting, because dRD is commonly seen as a sanity check, the decoding accuracy of which should not be exceeded by other analyses. It is also worth noting that in our example, just one response dimension (favorability), but two design dimensions (gain and loss) had to be decoded. We hypothesize that decoding the design from the data acts as a feature reduction mechanism which helps iRD predicting behavior using the psychologically most meaningful factors. In the future, we want to investigate the performance of iRD when the mapping from experimental conditions is more deterministic [7,8] or completely random [9,10].

References

(1) Haynes JD & Rees G. 2006. Decoding mental states from brain activity in humans. Nature Reviews Neuroscience. 7(7):523-534. doi:10.1038/nrn1931
(2) Haynes JD. 2015. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron. 87(2):257-270. doi:10.1016/j.neuron.2015.05.025
(3) Soch J et al. 2020. Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding. NeuroImage. 209:116449. doi:10.1016/j.neuroimage.2019.116449
(4) Botvinik-Nezer R et al. 2019. fMRI data of mixed gambles from the Neuroimaging Analysis Replication and Prediction Study. Sci Data. 6(1):106. doi:10.1038/s41597-019-0113-7
(5) Botvinik-Nezer R et al. 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature. 582(7810):84-88. doi:10.1038/s41586-020-2314-9
(6) Likert R. 1932. A technique for the measurement of attitudes. Archives of Psychology. 22 140:55-55.
(7) Reverberi C et al. 2012. Compositionality of Rule Representations in Human Prefrontal Cortex. Cereb Cortex. 22(6):1237-1246. doi:10.1093/cercor/bhr200
(8) Hebart MN et al. 2012. Human visual and parietal cortex encode visual choices independent of motor plans. NeuroImage. 63(3):1393-1403. doi:10.1016/j.neuroimage.2012.08.027
(9) Soon CS et al. 2008. Unconscious determinants of free decisions in the human brain. Nature Neuroscience. 11(5):543-545. doi:10.1038/nn.2112
(10) Schultze-Kraft M et al. 2016. The point of no return in vetoing self-initiated movements. PNAS. 113(4):1080-1085. doi:10.1073/pnas.1513569112