Interpreting neural decoding models of working memory load using grouped model reliance

Simon Valentin[1,4]*, Maximilian Harkotte[2,4] Tzvetan Popov[3,4]

1 School of Informatics/University of Edinburgh/Edinburgh, United Kingdom **2** Department of Psychology/Eberhard Karls University Tübingen/Tübingen, Germany **3** Central Institute of Mental Health, Medical Faculty/University of Heidelberg, Mannheim, Germany **4** Department of Psychology/University of Konstanz/Konstanz, Germany

* simonvalentin@me.com

Abstract

The application of machine learning algorithms for decoding psychological constructs based on neural data is becoming increasingly popular. However, there is a need for methods that allow to interpret trained models, as a step towards bridging the gap between theory-driven cognitive neuroscience and data-driven decoding approaches. The present study demonstrates *grouped model reliance* as a model-agnostic permutation-based approach to this problem. Grouped model reliance indicates the extent to which a trained model relies on conceptually related groups of variables, such as frequency bands or regions of interest in electroencephalographic (EEG) data. Models were trained on within-participant single-trial EEG data from a Sternberg working memory task. Participants were asked to memorize a sequence of digits (0-9), varying randomly in length between one, four and seven digits, where EEG recordings for working memory load estimation were taken from a 3-second retention interval. Present results confirm previous findings, as both Random Forest and support vector machine models relied mostly on alpha-band activity across subjects. However, patterns in frequency and particularly topography varied considerably between individuals.

Informed by interpretations of decoding models, further analyses using cluster-based statistics identified inter-individual variability in neural correlates of working memory.

Author summary

Modern machine learning (ML) algorithms currently receive considerable attention for their predictive power in neural decoding applications. However, looking beyond predictive power, there is a need for interpretable methods, in order to answer questions such as which aspects of the input data a trained model relies upon to make predictions. In the present study, we demonstrate the use of grouped model-reliance as a generally applicable method for interpreting neural decoding models. We employed an experimental design in which a comparably small number of participants (10) completed a large number of trials (972) over multiple electroencephalography (EEG) recording sessions from a Sternberg working memory task. Trained decoding models consistently relied on alpha frequency activity, supporting existing research on the relationship between neural oscillations and working memory. However, results indicate large inter-individual variability in decoding models with respect to frequency and topography. This finding is also confirmed by additional analyses employing standard methods from cognitive neuroscience. In our view, there are two main implications of this study: First, model interpretability is critical to consolidate modern machine learning approaches to neural decoding and cognitive neuroscience theory. And second, present findings point to the importance of considering neural data on a single-subject level, instead of just a population average.

Introduction

The application of statistical algorithms to neural data is becoming an increasingly popular tool for explaining the link between biology and psychology [1, 2]. In particular, supervised learning algorithms are utilized to decode various psychological phenomena related, e.g. to perception, attention, and memory with promising success [3–6]. However, while these algorithms are optimized to provide accurate predictions, the interpretability of model parameters is often of secondary interest. The latter, however,

is particularly relevant in cognitive neuroscience, which tries to develop theoretical explanations for the link between psychological phenomena and their neural manifestation.

In order to provide a bridge between theory-driven cognitive neuroscience and data-driven decoding approaches, there is a need for methods that allow to interpret trained models. The present study demonstrates model reliance as a versatile and readily applicable approach to this problem. Here, models are trained to decode working memory load based on single-trial electroencephalographic (EEG) data, collected in multiple experimental sessions per participant. Consequently, model reliance is used to interpret models in terms of conceptually meaningful groups of features from a single-subject perspective.

For supervised learning problems, a broad distinction can be made between *useful* and *relevant* predictor variables [7–9]. While some predictor variables can be relevant for (i.e., related to) an outcome, they are not necessarily useful in building a predictive model, as they might contain redundant information. Univariate methods, such as simple linear regression can inform about the relevance of a predictor, but not necessarily about the usefulness when assessed together with other predictors in a multivariate model.

Usually, the importance of predictors in a multivariate model is assessed for individual predictors, such as partial regression coefficients in a linear regression model. However, as variables extracted from EEG or MEG data are often inter-correlated, questions about the usefulness of variables relate more to sets of related variables, rather than individual variables [10, 11]. For instance, when assessing the usefulness of certain topographical or spectral components for predicting a psychological phenomenon, neural activity may be shared across multiple brain regions or recording sites. The acquisition resolution of these components is usually on a more detailed level that is used for interpretation [12]. Hence, a method that assesses the usefulness of groups (or subsets) of variables describes their relation to psychological phenomena more appropriately. Ideally, such a method should be model-agnostic to enable comparisons of models trained using different learning algorithms.

A practical approach for assessing the usefulness of variables is given by permutation importance, used initially as a measure for the usefulness of variables in the Random

Forest algorithm [13]. The usefulness of a variable is quantified by the decrease in prediction performance when that predictor variable is randomly permuted, essentially destroying the association between that predictor variable and the outcome. An intuitive terminology for this idea for any learning algorithm and further theoretical development beyond the scope of the present study is given by *model reliance* [14]. Model reliance indicates the extent to which a model relies on specific variables in making predictions, i.e., how much performance decreases when permuting that variable. Crucially for the present problem, the method of permuting predictor variables can be adapted to permuting groups of conceptually (or statistically) related variables (such as frequency bands, as opposed to single frequencies) to measure their aggregate impact on predictive performance [15]. This is required as the importance of a group of variables (i.e. grouped model reliance) is not equivalent to the sum of individual model reliances, as demonstrated by Gregorutti et al. [15]. To emphasize the interpretation of this measure of variable importance as the decrease in accuracy of a trained model, the term model reliance is used. By design, this approach treats the model as a black-box, thereby making it a model agnostic method that can be used for any supervised learning algorithm.

Working memory is a widely studied psychological construct and refers to the temporary retention of information in the absence of sensory input, needed for a subsequent behavioral outcome. Neuronal oscillations are hypothesized to be involved in working memory by generating a temporal structure for the brain [16, 17]. Amplitude modulation of neuronal oscillations, in particular in the alpha frequency bands (8-14Hz), is a robust finding in psychophysiological research [18–20]. It is hypothesized that these power modulations aid the functional brain architecture during retention, protect against interference and thereby manifest in relevant behavioral outcomes, as measured by accuracy and reaction time [21, 22]. Thus, the scaling of alpha power with working memory load is considered as an essential neural manifestation of the psychological construct working memory. While alpha band activity has been established as a correlate of working memory across individuals, little information is available about how these correlates manifest on a single-subject basis and how they differ between individuals. Hence, employing single trial EEG analysis as well as assessing decoding models on a single-subject level can provide complementary information to that of

group-level statistics [23–26].

A widely used paradigm for studying working memory is the Sternberg task [27, 28]. Compared to other paradigms, this task has the advantage that the periods of encoding, retaining and recognizing stimuli are all separated in time [21, 22]. Subjects are first presented with a list of items, the number of which determines the working memory load. Following a retention interval of several seconds, a probe item is presented, and subjects indicate the membership of this item to the previously presented list.

In this study, using single-trial EEG data from a Sternberg task, Random Forest and support vector machine (SVM) models are trained on individual subjects to perform working memory load estimation and consequently interpreted using grouped model reliance. Informed by model interpretations, cluster-based statistics are employed to further probe the relationship between working memory and neural oscillations.

Materials and methods

Participants

Eleven subjects were recruited by advertisement at the local university ($M = 24.5$ years, $SD = 4.8$; 50% female) and reported no history of neurological and/or psychiatric disorders. One subject was excluded from the analysis, as data acquisition was interrupted during the second session of the experiment. All participants gave written informed consent in accordance with the Declaration of Helsinki prior to participation. The study was approved by the local ethics committee.

Stimulus material and procedure

Participants performed a Sternberg task [27] with alternating levels of difficulty (1, 4, or 7 items to be kept in memory) while the electroencephalogram (EEG) was recorded. The data of each participant were collected on three different sessions, which were, on average, $M = 4.35$ days ($SD = 2.72$) apart. Written informed consent was obtained from each subject prior to each session. One session comprised four practice trials and six main blocks, each consisting of 54 trials (lasting approximately nine minutes). In between blocks, participants were allowed to rest for a maximum of three minutes. Each

participant completed 324 trials per session, resulting in 972 trials in total. Participants were asked to memorize a sequence of digits (0–9), varying randomly in length between one, four or seven different digits. After an initial central fixation interval of 500 ms, the sequence of digits was presented serially. Each digit was presented for 1200 ms, followed by a blank screen for 1000 ms before the presentation of the next digit. After a 3000 ms retention interval (blank screen), a probe stimulus was presented in the center of the screen for 5000 ms. Participants were instructed to indicate whether the probe was part of the previously presented sequence. The right arrow key on a standard keyboard indicated 'yes' and the left arrow key 'no' answer. Participants' response was followed by positive or negative feedback for 500 ms. Finally, a blank screen was presented for 1000 ms, after which the next trial began. In order to balance the occurrence of each sequence length, there were nine positive trials (probe part of the study list) and nine negative trials (probe not in the sequence) for each sequence length within each block, presented in random order with respect to the sequence length.

Data acquisition

EEG was recorded with an ANT Neuro 128-electrode system (`www.ant-neuro.com`) with Ag/AgCl electrodes placed on a Waveguard cap with an equidistant hexagonal layout. Signals were sampled at 512 Hz, and electrode impedance was kept below 20 kOhm. The recording was DC and referenced to a common average reference.

Preprocessing

Data analysis was performed with the MATLAB based FieldTrip toolbox [29]. For each participant and channel, after demeaning and removing the linear trend across the session, independent component analysis (ICA) [30] was used to remove variance associated with eye blink and cardiac activity. Increased noise in the electrodes closest to the ears (LM, LE1, LE2, RM, RE1, RE2) in some participants led to their exclusion from all subsequent analyses. All trials per session and condition were included. Spectral analyses were conducted for each trial using a Fast Fourier Transformation (FFT) with a single Hanning taper for the retention interval of 3 sec. The predictor variables used for the classification model covered the frequency bands (delta 1–4Hz, theta 5–7Hz,

alpha 8–12Hz, beta 13–20Hz and low gamma 21–40Hz) in one Hz steps, for electrodes from nine regions of interest (ROI). Hence, there were a total of 4880 predictor variables (40 frequencies for each of 122 electrodes). ROI's were left/central/right occipital, left/central/right central and left/central/right frontal. The exact electrodes per group and a layout of their respective locations can be found in S1 Fig.

Decoding model training and evaluation

The Random Forest algorithm, a type of ensemble method, was used for all decoding analyses [13]. This algorithm was chosen for its ability to perform multiclass classification on a large number of possibly correlated and non-linearly related variables [13]. The number of trees in the forest was set to 5000 with all other hyperparameters set to default values. As an alternative model, further analyses employed an SVM model with a radial basis function (RBF) kernel with penalty term C set to 1 [31]. Classification accuracy was used as the performance metric for all models. Hence, no distinction was made between misclassifying a load 1 trial as load 4 or 7, for instance.

As the classification task comprises three balanced classes (load 1, 4 and 7), chance level accuracy corresponds to $33.\overline{33}\%$. All models were trained and tested using stratified and shuffled 10-fold cross-validation. Stratification ensures that the distribution of classes is the same for each fold of the cross-validation and can lead to more stable performance estimates than standard k-fold cross validation [8]. Simulation studies indicate the most robust performance for stratified k-fold cross-validation with k set to $k = 10$ [8], which is therefore employed in the present analyses. The reported accuracy and model reliance values correspond to the arithmetic means over all 10 folds of the cross-validation loop. Single-trial data from all 3 sessions and 6 blocks per session were pooled for each participant and used in the cross-validation procedure.

Model reliance

In order to make model reliance scores comparable across data sets, model reliance is defined as a ratio of the permuted error and the error on original predictors. As such, a more positive value indicates more model reliance on that predictor, whereas values

towards zero indicate less model reliance. In the case of groups of variables, this interpretation changes to the average reliance on the predictor variables in that group of variables. Values are normalized in order to make different sized groups of variables comparable [15]. This is achieved by dividing the model reliance score for a particular group of variables by the number of variables in that group.

As adapted from [15], X is a n by p matrix of observations[1] of predictor variables, respectively. $\mathbf{y}$ is a vector of outcomes of length n. f is a fitted model.

A group of variables (that is, columns) in X is indexed by a set J, where all $j \in J$ are $1 \leq j \leq p$. $\mathrm{ACC}_{\mathrm{bl}}$ refers to the baseline accuracy of f on X and $\mathbf{y}$, whereas $\mathrm{ACC}_{\mathrm{perm}_J}$ refers to the accuracy on the data after randomly permuting all variables indexed by J, essentially destroying the information encoded in those variables. The model reliance value, $\mathrm{MR}(X_J)$, for a group of variables, J, is given by

$$\mathrm{MR}(X_J) =_{def} \frac{1}{|J|}\left(\frac{1-\mathrm{ACC}_{\mathrm{perm}(J)}}{1-\mathrm{ACC}_{\mathrm{bl}}} - 1\right). \tag{1}$$

For every cross-validation fold, model reliance is averaged over 10 random permutations and subsequently averaged over all 10 cross-validation folds. [2]

Decoding analyses were implemented in Python, making use of the scikit-learn [34] module. Code for implementations of grouped model reliance in addition to Jupyter notebooks for all analyses will be made available under `https://github.com/simonvalentin/wmdecoding` after publication.

Non-parametric statistical testing with clusters

Effects of working memory load on neural data were probed by a cluster-based randomization approach [35]. This approach identifies clusters (in frequency and space) of activity on the basis of which the null hypothesis can be rejected, yet effectively addresses the multiple-comparison problem. The null hypothesis was that the trials per subject sampled from the three load conditions stems from the same distribution; thus

[1] When assessing MR for generalization performance, X and $\mathbf{y}$ should correspond to a test/validation set. However, in some scenarios, it might also be of interest to assess model reliance on the data a model was trained on.

[2] These values could be used to estimate parameters of dispersion and subsequently error bars and p-values for estimating the variability of model reliance estimates over the number of random permutations and across new model fits, using different random initialization and test sets (or folds) [32]. This, however, goes beyond the scope of this article, especially as estimating parameters of dispersion for k-fold cross-validation is nontrivial, given that the test sets are not independent of the training sets [33].

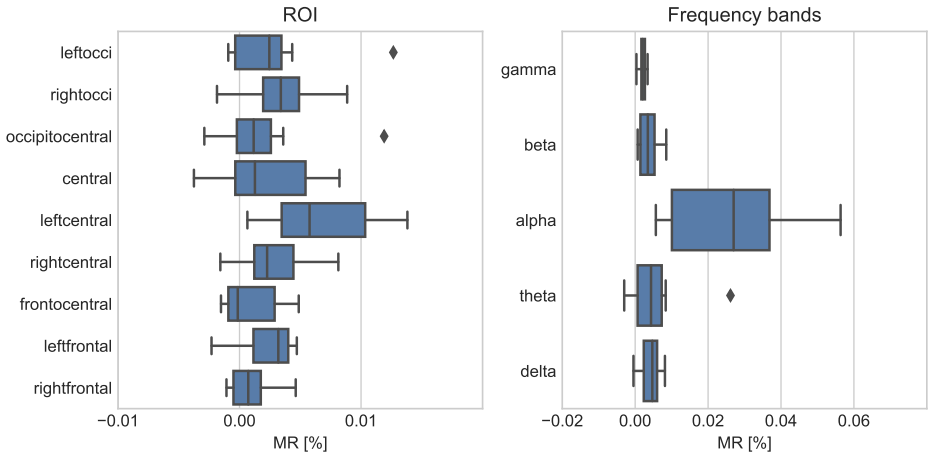

Fig 1. Box-Whisker plots of average model reliance (MR) per participant for different ROI's (left) and frequency bands (right).

the labels (i.e., load 1, load 4 and load 7) are exchangeable. Dependent samples F-tests were used as test statistics. Permutation of the labels was computed 1000 times resulting in a distribution of 1000 F-values. The original value of the test statistic was compared against this randomization distribution using an alpha level of 5%.

Results

Behavioral

In order to establish that the experimental manipulation yielded the expected behavioral effects, accuracy, as well as reaction times (RT) of participants, were recorded. In line with previous findings, both accuracy and reaction times decreased with an increase in working memory load (cf. Table 1).

Table 1. Means and standard deviations of reaction times and accuracies across subjects.

Load	**Reaction time [ms]**		**Accuracy [%]**	
	M	*SD*	*M*	*SD*
1	530.9	133.8	97.6	2.9
4	659.8	147.8	95.3	3.2
7	784.1	144.6	89.9	6.7

Values computed from averaged observations per participant.

Model reliance

Averaged across all subjects, within-subject classification accuracy was 48.51% ($SE = 1.25\%$) in a three-class classification task with a chance level accuracy of $33.\overline{33}\%$. As illustrated in Fig 1 (B, right) overall, trained models relied mostly on neural activity in the alpha frequency band. As described in the methods, model reliance is normalized according to the group size of predictor variables. Hence, large groups, such as the gamma band are penalized more than smaller groups, as, e.g. alpha. However, as reported in S2 Fig, even when no normalization for the group size is used, the interpretation that alpha-band activity is central for decoding performance holds. While model reliance for alpha band frequencies consistently emerges across subjects, there is variability in individual profiles, as presented in Fig 2.

In contrast, MR on scalp topographic features for classification accuracy is less decisive Fig 1(left). There is no group of predictors that clearly stands out as particularly important across subjects. Instead, scores show considerable intra-individual variability, as presented in Fig 2B.

Further, in order to assess whether this reflects only the reliance on these groups of features in a multivariate model, or also the relevance of the feature groups assessed separately, additional analyses were run. Here, a classifier was trained and tested only on separate feature groups (S3 Fig). This analysis further strengthened the interpretation that alpha band activity encodes information that is particularly relevant for decoding working memory load, while model reliance is broadly distributed across ROI.

In support of the notion of individual specificity and intra-individual variation, training and testing models across subjects yielded comparably poor generalization performance (accuracy 34.53%).

To explore whether fitting a different, but similarly performant model would result in comparable estimates of model reliance, additional analyses using an SVM algorithm were conducted. These analyses yielded similar results and interpretations, as illustrated in S4 Fig.

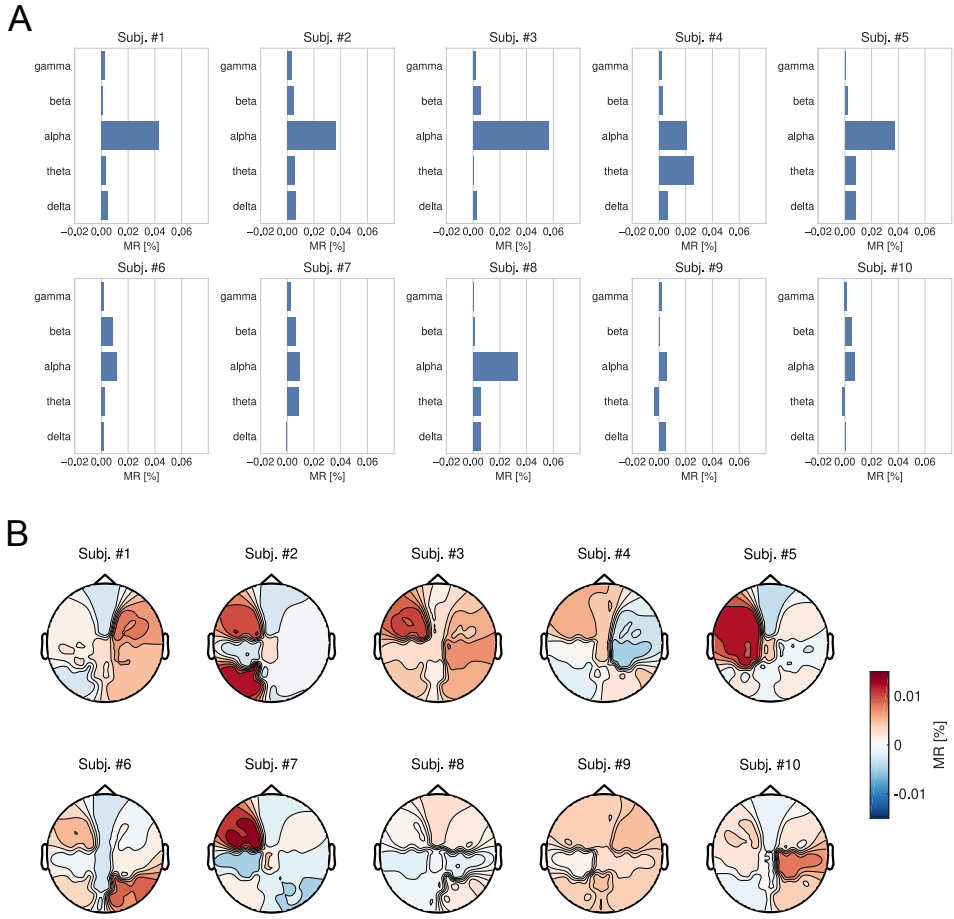

Fig 2. (A) Model reliances for each frequency band and individual participant. (B) Topography of model reliances for each individual participant.

Cluster-based inferential statistics

Model reliance scores imply that predictors from the alpha frequency group are of particular importance (being the most critical frequency band for all but one participant). Thus, cluster-based statistics were computed on the alpha-band for each participant. As shown in Fig 2A, the null hypothesis could be rejected for all but one individual. Descriptively, no clear topographic pattern could be identified across participants.

Power spectra were computed for those electrodes contained in clusters for which significant condition differences were found (Fig 3B)[3]. Crucially, some participants were

[3]As no cluster was found for Subj #9, power spectra were computed over electrodes selected from Subj #7. This individual was chosen due to the similar topographic representation of the main effect load.

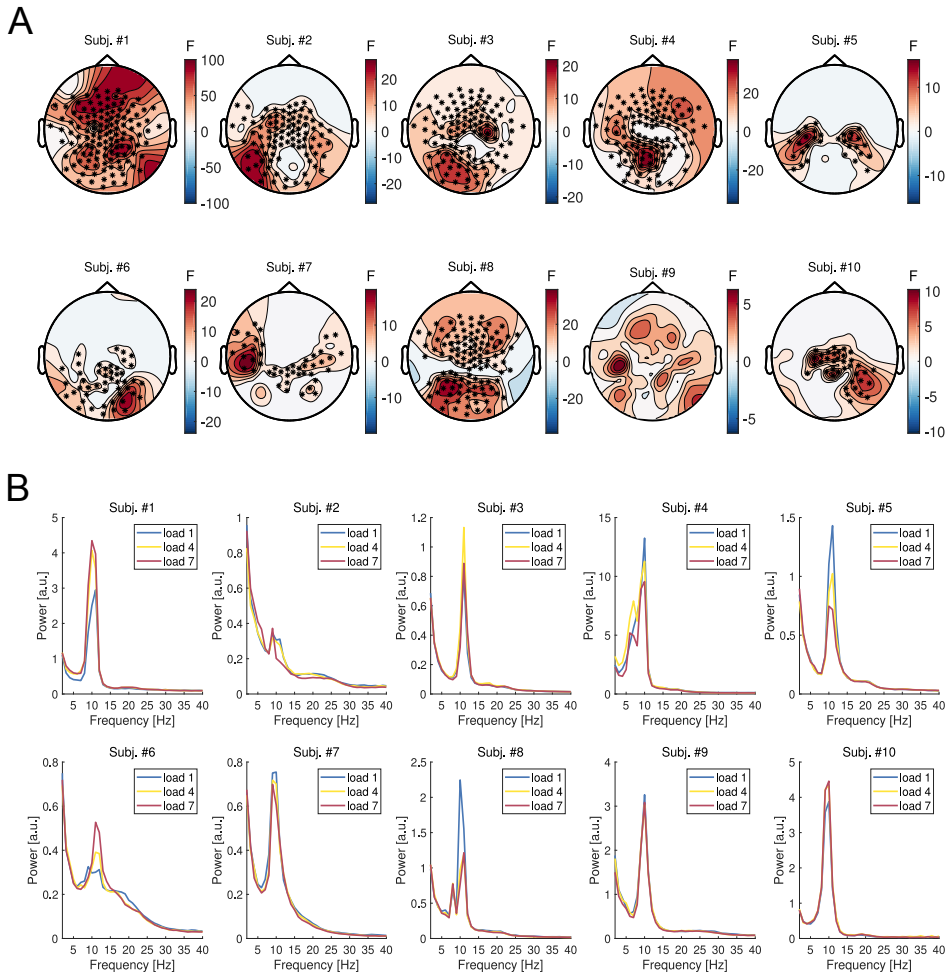

Fig 3. (A) Topography of the main effect "Load" illustrated for each individual participant. Warm colors indicate the spatial distribution of F-values. Asterisks denote electrodes corresponding to clusters on the basis of which the null hypothesis is rejected. (B) Power spectra averaged across the electrodes belonging to the corresponding clusters illustrated in A. Note that scales are plotted on an individual level, as condition differences within participants are of primary interest.

characterized by a positive relationship of alpha-band activity with increasing working memory load, yet others displayed a reverse ordering or very small to no difference.

Discussion

The primary aim of the present study was to interpret single-trial EEG decoding models of working memory load. Grouped model reliance was utilized to probe and interpret decoding models in frequency and space on a single-subject level. Decoding models performed with, on average, 48.51% ($SE = 1.25\%$) accuracy in a three-way

classification task. Model reliance scores suggest that for all participants, alpha band activity was particularly important for predicting working memory load. Alpha activity was the most critical frequency band for all but one participant (Subj. #4 for whom theta activity was most important). This is consistent with numerous studies reporting scaling of alpha amplitude with working memory load. [18, 19, 21, 36]. However, across participants, models do not appear to rely on some ROI more than on others. Instead, model reliance scores were found more distributed across different ROI than strongly clustered. Additional analyses employing an SVM model with an RBF-kernel yielded the same interpretations, supporting model reliance as a model-agnostic method.

Justified by model reliance scores, subsequent analyses focused on patterns of alpha band activity. Here, cluster-based permutation tests on a single subject level revealed inter-individual amplitude and topographic variation. For all but one subject (Subj. #9), there was a significant effect of WM load. The amplitude of alpha activity increased with load for some individuals (e.g., Subj. #1) and decreased in others (Subj. #5).

Training and testing models across subjects resulted in comparably poor generalization performance, indicating inter-subject variability on the physiological level. This is also in line with additional analysis using cluster-based permutation statistics, which did not reveal statistically significant condition differences across participants (see S5 Fig). Taken together, these findings indicate that while models relied most on alpha-band activity, the manifestation particularly in topography varies considerably between individuals.

A distinction can be made between an *activation-based* and *information-based* view on neural data analysis [2]. Briefly, the activation-based view focuses on patterns of de- and increases of activity (e.g., alpha power) and is typically utilized in cognitive neuroscience. The information-based view, on the other hand, is not restricted to activation but examines any change in the multivariate distribution of the data as information that can be used for making predictions. For an in-depth discussion of these ideas see [2].

Given that any information contained in the predictor variables may be used by the supervised learning algorithm to make predictions, preprocessing is needed to ensure that known confounding signals from EEG recordings do not distort model

interpretations. For instance, if prediction is the only objective, using ICA to remove ocular and cardiac artifacts would not be needed. The present study illustrates an example of employing an information-based method to provide insight into the different activity patterns underlying psychological phenomena. Hence, by probing decoding models on data that contains credible information about neural activity (with confounding signals such as blinks removed), information-based methods do not contradict the goal of interpreting neural phenomena, but can instead complement more common activation-based methods.

Given that there can be multiple, similarly performant solutions in high-dimensional data [37], model reliance, and hence interpretations, may also vary across models. In the present study, cross-validation and repeated random permutations were employed to obtain a representative value of what an "average" model relies upon, as was also confirmed by additional analyses employing an SVM model. Fisher et al. [14] propose Model-Class Reliance as a measure to obtain upper and lower bounds of the reliance on a particular feature for one model class. Future research may further explore this idea and grouping of predictors in relation to neuroelectric data and its associated challenges related to temporal, spatial and spectral structure in different machine learning models.

More broadly, the present results highlight the variability that exists in decoding models, even when assessing such an established construct as working memory in a relatively homogeneous university student sample. One aspect of heterogeneity relates to associating differences in behavior with neurophysiological differences (e.g., [38, 39]). However, in the present study, task performance was high in all subjects, while both model reliance, as well as cluster-based inferential statistics, indicate variability on the physiological level. One interpretation for these findings is that variability found in neurophysiological correlates does not arise from differences in cognitive abilities, but from differences of the physiological manifestation of working memory. Model reliance can provide insights into what predictors decoding models rely upon for a particular individual. For instance, the present data imply the importance of alpha activity across subjects, even though findings from cluster-based statistics reveal considerable variability and permit decision with respect to the dependent variable. Hence, rather than quantifying the response of a hypothetical "average individual," these findings support recent efforts aiming at quantifying heterogeneity [40, 41]. Given the data of a

larger sample of people, future research may look into clustering people whose models rely upon similar predictors and investigate whether systematic sources of this variability can be found.

Supporting information

S1 Fig. Definition of ROI in terms of electrodes. Note that electrodes LM, LE1, LE2, RM, RE1 and RE2 were excluded from the analyses.

S2 Fig. Model reliance without normalization for group sizes. Results reveal that alpha still emerges as the group of variables that trained models relied upon the most when not accounting for the size of the group of variables.

S3 Fig. Classification accuracy for training and testing on individual feature groups. Results highlight the difference between relevance and usefulness of predictor variables in multivariate models.

S4 Fig. Model reliance for SVM model. Model used an RBF kernel function, with penalty term C set to 1. Cross-validated classification accuracy was 48.21%.

S5 Fig Non-parametric statistical testing with clusters (A) Grand averaged power spectra over fronto-central (left) and occipital electrodes (right). Load conditions are denoted by color (red- load 1, yellow- load 4 and blue- load 7). (B) Topographic maps as a function of working memory load (x-axis) and frequency band (y-axis). Power spectra averaged across all participants for frontal (Subfigure A) and posterior/occipital electrodes (Subfigure B). Over frontal areas, there were two clear peaks in the theta (5–7HZ) and alpha (8–12Hz) frequency ranges, whereas posterior/occipital electrode sites were largely dominated by power increase in the alpha frequency range. Subfigure B illustrates the scalp-topographic distribution of activity as a function of working memory load (x-axis) and frequency bands (y-axis). In line with existing literature, theta activity displayed a maximum over mid-frontal electrodes whereas alpha activity was predominantly distributed over parieto-occipital electrodes. A main effect of working memory load was tested utilizing an independent F-statistic followed by a

post-hoc t-test evaluating simple condition differences. Neither test was able to reject the null hypothesis stating that the data stems from the same distribution.

Acknowledgments

We thank Brigitte Rockstroh for comments on the manuscript.

Author contributions

Conceptualization: S.V., M.H.

Formal analysis: S.V., T.P.

Funding acquisition: T.P.

Investigation: S.V., M.H.

Methodology: S.V., M.H.

Resources: T.P.

Software S.V.

Supervision: T.P.

Visualization: S.V., M.H., T.P.

Writing - original draft: S.V., M.H., T.P.

Writing - review & editing: S.V., M.H., T.P.

References

1. Bzdok D. Classical statistics and statistical learning in imaging neuroscience. Frontiers in Neuroscience. 2017;11(OCT):1–23. doi:10.3389/fnins.2017.00543.

2. Hebart MN, Baker CI. Deconstructing multivariate decoding for the study of brain function. NeuroImage. 2018;180(April 2017):4–18. doi:10.1016/j.neuroimage.2017.08.005.

3. Poldrack RA, Halchenko Y, Jos S. Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychological Science. 2009;20(11):1–16.

4. Poldrack RA. Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. Neuron. 2011;72(5):692–697. doi:10.1016/j.neuron.2011.11.001.

5. Tong F, Pratte M. Decoding Patterns of Human Brain Activity. Ssrn. 2012;doi:10.1146/annurev-psych-120710-100412.

6. Haxby JV, Connolly AC, Guntupalli JS. Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. Annual Review of Neuroscience. 2014;37(1):435–456. doi:10.1146/annurev-neuro-062012-170325.

7. Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence. 1997;97(1-2):245–271. doi:10.1016/S0004-3702(97)00063-5.

8. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997;97(1-2):273–324. doi:10.1016/S0004-3702(97)00043-X.

9. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research (JMLR). 2003;3(3):1157–1182. doi:10.1016/j.aca.2011.07.027.

10. McFarland DJ. Characterizing multivariate decoding models based on correlated EEG spectral features. Clinical Neurophysiology. 2013;124(7):1297–1302. doi:10.1097/MCA.0000000000000178.Endothelial.

11. da Silva FL. EEG: Origin and Measurement. In: Mulert C, Lemieux L, editors. EEG - fMRI: Physiological Basis, Technique, and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 19–38. Available from: `http://link.springer.com/10.1007/978-3-540-87919-0_2`.

12. Cohen MX. Where Does EEG Come From and What Does It Mean? Trends in Neurosciences. 2017;40(4):208–218. doi:10.1016/j.tins.2017.02.004.

13. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32. doi:10.1017/CBO9781107415324.004.

14. Fisher A, Rudin C, Dominici F. All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. arXiv preprint. 2018;.

15. Gregorutti B, Michel B, Saint-Pierre P. Grouped variable importance with random forests and application to multiple functional data analysis. Computational Statistics and Data Analysis. 2015;90:15–35. doi:10.1016/j.csda.2015.04.002.

16. Buzsáki G, Draguhn A. Neuronal olscillations in cortical networks. Science. 2004;304(5679):1926–1929. doi:10.1126/science.1099745.

17. Buzsáki G. Rhythms of the Brain. New York: Oxford University Press, Inc.; 2006.

18. Jensen O, Gelfand J, Kounios J, Lisman J. Oscillations in the Alpha Band (9-12 Hz) Increase with Memory Load during Retention in a Short-term Memory Task. Cerebral Cortex. 2002;12(8):877–882. doi:10.1093/cercor/12.8.877.

19. van Ede F. Mnemonic and attentional roles for states of attenuated alpha oscillations in perceptual working memory: A review. European Journal of Neuroscience. 2017;doi:10.1111/ejn.13759.

20. Kustermann T, Rockstroh B, Miller GA, Popov T. Neural network communication facilitates verbal working memory. Biological Psychology. 2018;136:119–126. doi:10.1016/j.biopsycho.2018.05.018.

21. Roux F, Uhlhaas PJ. Working memory and neural oscillations: Alpha-gamma versus theta-gamma codes for distinct WM information? Trends in Cognitive Sciences. 2014;18(1):16–25. doi:10.1016/j.tics.2013.10.010.

22. Esposito MD, Postle BR. The Cognitive Neuroscience of Working Memory. Annual Review of Psychology. 2015;66:115–142. doi:10.1146/annurev-psych-010814-015031.

23. Pernet CR, Sajda P, Rousselet GA. Single-Trial Analyses: Why Bother? Frontiers in Psychology. 2011;2:322. doi:10.3389/fpsyg.2011.00322.

24. Braga RM, Buckner RL, Braga RM, Buckner RL. Parallel Interdigitated Distributed Networks within the Individual Estimated by Intrinsic Functional Article Parallel Interdigitated Distributed Networks within the Individual Estimated by Intrinsic Functional Connectivity. Neuron. 2017;95(2):457–471. doi:10.1016/j.neuron.2017.06.038.

25. Bzdok D, Meyer-lindenberg A. Review Machine Learning for Precision Psychiatry : Opportunities and Challenges. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. 2018;3:223–230. doi:10.1016/j.bpsc.2017.11.007.

26. Smith PL, Little DR. Small is beautiful: In defense of the small-N design. Psychonomic Bulletin and Review. 2018;25:2083–2101. doi:10.3758/s13423-018-1451-8.

27. Sternberg S. High-speed scanning in human memory. Science, 153, 652–654. Science. 1966;153:652–654.

28. Sternberg S. Memory-Scanning: Mental Processes Revealed By Reaction-Time Experiments; 1969.

29. Oostenveld R, Fries P, Maris E, Schoffelen JM. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Computational Intelligence and Neuroscience. 2011;2011. doi:10.1155/2011/156869.

30. Jung TP, Makeig S, McKeown MJ, Bell AJ, Lee TW, Sejnowski TJ. Imaging brain dynamics using independent component analysis. Proceedings of the IEEE. 2001;89(7):1107–1122. doi:10.1109/5.939827.

31. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995;20(3):273–297. doi:10.1023/A:1022627411411.

32. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: A corrected feature importance measure. Bioinformatics. 2010;26(10):1340–1347. doi:10.1093/bioinformatics/btq134.

33. Bengio Y, Grandvalet Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. Journal of Machine Learning Research. 2004;5:1089–1105. doi:S0006291X03002249 [pii].

34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2012;12:2825–2830. doi:10.1007/s13398-014-0173-7.2.

35. Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods. 2007;164(1):177–190. doi:10.1016/j.jneumeth.2007.03.024.

36. Klimesch W. Alpha-band oscillations, attention, and controlled access to stored information. Trends in Cognitive Sciences. 2012;16(12):606–617. doi:10.1016/j.tics.2012.10.007.

37. Breiman L. Statistical Modeling: The Two Cultures. Statistical Science. 2001;16(3):199–231.

38. Dong S, Reder LM, Yao Y, Liu Y, Chen F. Individual differences in working memory capacity are reflected in different ERP and EEG patterns to task difficulty. Brain Research. 2015;1616:146–156. doi:10.1016/j.brainres.2015.05.003.

39. Maurer U, Brem S, Liechti M, Maurizio S, Michels L, Brandeis D. Frontal Midline Theta Reflects Individual Task Performance in a Working Memory Task. Brain Topography. 2014;28(1):127–134. doi:10.1007/s10548-014-0361-y.

40. Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. Biological Psychiatry. 2016;80(7):552–561. doi:10.1016/j.biopsych.2015.12.023.

41. Wolfers T, Doan NT, Kaufmann T, Alnæs D, Moberget T, Agartz I, et al. Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. JAMA Psychiatry. 2018;75(11):1146. doi:10.1001/jamapsychiatry.2018.2467.