NeuroImage 196 (2019) 302–317

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Trial-by-trial surprise-decoding model for visual and auditory binary oddball tasks

Alireza Modirshanechi, Mohammad Mahdi Kiani, Hamid Aghajan *

Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran

ARTICLE INFO

Keywords:
Neural decoding
Bayesian brain
Surprise coding
Machine learning
Information therory
EEG single trial analysis

ABSTRACT

Having to survive in a continuously changing environment has driven the human brain to actively predict the future state of its surroundings. Oddball tasks are specific types of experiments in which this nature of the human brain is studied. Detailed mathematical models have been constructed to explain the brain's perception in these tasks. These models consider a subject as an ideal observer who abstracts a hypothesis from the previous stimuli, and estimates its hyper-parameters - in order to make the next prediction. The corresponding prediction error is assumed to manifest the subjective surprise of the brain. While the approach of earlier works to this problem has been to suggest an encoding model, we investigated the reverse model: if the stimuli's surprise is assumed as the cause of the observer's surprise, it must be possible to decode the surprise of each stimulus, for every single subject, given only their neural responses, i.e. to tell how unexpected a specific stimulus has been for them. Employing machine learning tools, we developed a surprise decoding model for binary oddball tasks. We constructed our model using the ideal observer proposed by Meyniel et al. in 2016, and applied it to three datasets, one with visual, one with auditory, and one with both visual and auditory stimuli. We demonstrated that our decoding model performs very well for both of the sensory modalities with or without the presence of the subject's motor response.

1. Introduction

To survive in an environment which is continuously changing over time, humans need to be able to predict the future state of their surroundings at all times. Various mathematical theories have been developed to explain the mechanism with which our brains predict the future, at least in very basic situations (Friston, 2010, 2005, 2009; Friston and Kiebel, 2009; Knill and Pouget, 2004). The core assumption of these theories is that humans behave as near-optimal observers, wherein optimally may be derived from a Bayesian, information theoretic, or other points of view (Friston, 2010; Knill and Pouget, 2004). Despite the use of essentially similar mathematical approaches in considering the brain's prediction procedure as an optimization problem, there are lots of differences in applying these models to real-world specific situations, clarifying the relations between mathematical variables and experimental observations.

Among the large number of experimental tasks to investigate evidences of the predictive nature of the human brain, one of the most famous experiments is the oddball paradigm task, i.e. the task during which subjects are given a sequence of random stimuli, composed of a mix of frequent and rare types (Garrido et al., 2009; Huettel et al., 2002; Huettel and McCarthy, 2004; Imada et al., 1993; Kolossa et al., 2013; Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Näätänen, 2000; Ostwald et al., 2012; Rubin et al., 2016; Squires et al., 1976). Various evidences based on behavioral data as well as brain signals support the hypothesis of the predictive brain: it continually predicts the next stimulus given the previous ones. In addition, different landmarks represented by EEG signals, such as P300 (Kolossa et al., 2013; Luck, 2004; Mars et al., 2008; Meyniel et al., 2016; Sur and Sinha, 2009) and Mismatch Negativity (MMN) (Garrido et al., 2009; Lieder et al., 2013; Luck, 2004; Meyniel et al., 2016; Näätänen, 2000; Sur and Sinha, 2009; Symonds et al., 2017), as well as some landmarks represented by behavioral data, like the delay of subjects in responding to target stimuli (Huettel et al., 2002; Meyniel et al., 2016), are reported to be correlated with the error of the brain's prediction, or the subjective surprise of each stimulus, which can be defined in a few different mathematical forms (Baldi, 2002; Barto et al., 2013; Faraji et al., 2018; Itti and Baldi, 2009; Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Ostwald et al., 2012).

Several detailed mathematical models have been proposed to explain

* Corresponding author.
E-mail addresses: alireza.modirshanechi@epfl.ch (A. Modirshanechi), mahdikiani@ee.sharif.edu (M.M. Kiani), hamid.aghajan@ugent.be (H. Aghajan).

https://doi.org/10.1016/j.neuroimage.2019.04.028
Received 31 July 2018; Received in revised form 26 February 2019; Accepted 8 April 2019
Available online 11 April 2019

the experimental observations in oddball tasks and the way subjective (stimulus) surprise is encoded in the mentioned landmarks (Kolossa et al., 2013; Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Ostwald et al., 2012; Rubin et al., 2016; Squires et al., 1976). All of these models suggest a system whose responses to a sequence of stimuli has a significant statistical relation with the subject's responses to the same sequence (Kolossa et al., 2013; Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Ostwald et al., 2012; Squires et al., 1976). Although, roughly in all of the models, the response of the simulated encoder is considered as the surprise of the stimuli, the subject's response does not have a unified definition. Rather, in each model, one of the different traditionally accepted behavioral or biological landmarks is considered as the subject's response (Kolossa et al., 2013; Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Rubin et al., 2016; Squires et al., 1976). Among these models, an exception is the work of Ostwald et al. in 2012 (Ostwald et al., 2012) which searches for one time point of EEG signals which is most related to their model's responses. Pursuing the approach of more general theories about the predictive nature of the brain, most of these papers consider the subject as a near-ideal observer, or a predictive machine. The observer abstracts a hypothesis, or a rule from the previous stimuli, and actively estimates, or tunes its hyper-parameters. By doing so, the observer makes a prediction about the future stimulus. While various mathematical definitions of the subjective surprise have been proposed in different papers, a general, non-mathematical definition of surprise is often given as the level of being unexpected for the presented stimulus. The proposed models also use various assumptions about the brain's abstraction, and the way it forms the hypothesis and estimates the hyper-parameters. However, the more important issue is the way these works evaluate their models and assumptions, and compare them with each other. In almost all related literature, at least one of the following approaches or assumptions has been employed, resulting in reduced generality of the evaluation procedure.

1 Evaluating their work on averaged data rather than doing a trial-by-trial analysis (Meyniel et al., 2016; Squires et al., 1976): It is obvious that an *on-average* analysis is not necessarily valid for all various situations; an extreme example is the weakness of traditional Event Related Potential (ERP) analyses in describing the sequential effect in oddball tasks, i.e. the brain's response is a function of not only the occurrence probability of the stimuli, but also their order of occurrence (Luck, 2004; Meyniel et al., 2016). It should be mentioned that even in the work of Kolossa et al. in 2013 (Kolossa et al., 2013), evaluation is performed by averaging over sequences. Furthermore, in the work of Lieder et al. in 2013 (Lieder et al., 2013), evaluation is based on the analysis of only deviant stimuli, due to their objective to study the underlying mechanism of MMN generation.
2 Being concerned only with the statistically significant relationship between the model's and the brain's responses, rather than the predictive power of the model: This approach only states a statistical fact about the brain's perception, and cannot propose a model whose response can be assumed to be identical to the brain's response on a trial-by-trial basis (Rubin et al., 2016). It is noteworthy that this approach has been usually employed with the goal of choosing one model from a set of models (Mars et al., 2008; Ostwald et al., 2012), usually in order to find better evidences for some psychological hypotheses (Lieder et al., 2013), or explaining the encoding procedure of surprise (Kolossa et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Ostwald et al., 2012).
3 Investigating the statistical relationship between only "one" landmark represented by the EEG signal and the model's output (Kolossa et al., 2013; Lieder et al., 2013; Mars et al., 2008; Ostwald et al., 2012; Squires et al., 1976): This approach does not consider the possibility that the subjective surprise is coded as a multi-dimensional function of different landmarks, or in the shape of the waveform.

In order to suggest a more reliable evaluation method to investigate the relation between experimental data and theoretical models, which avoids the mentioned approaches or assumptions, in this paper, we solve the reverse problem: If the surprise calculated by a near-ideal observer is identical with the brain's subjective surprise, we must be able to decode the surprise of "each stimulus" (i.e. in single trial level, and not limited to one type of stimuli), and for "every specific subject", given only their "neural responses", i.e. tell how unexpected a specific stimulus is for them. Therefore, by employing machine learning tools, we propose a surprise decoding model for binary oddball tasks. Our model is constructed based on the assumption that the brain's perception about the sequences is governed by one of the systems (near-ideal observers) which has been developed by the recent work of Meyniel et al. in 2016 (Meyniel et al., 2016) – because of its relatively high descriptive power outperforming earlier works, its simplicity, and its independence from the sensory domain. Then, we apply our model to two open access datasets, one with visual and auditory stimulus modalities developed by (Walz et al., 2015, 2014, 2013) for our main analysis, and one with visual stimuli developed by (Robbins et al., 2018) for benchmarking and for demonstrating that the presence of asymmetric behavior data in the first dataset does not create a bias in our decoding results. By doing so, we evaluate our decoding model, and investigate the relation of its precision with the sensory domain. Considering the results of some of the earlier works such as (van Dinteren et al., 2014; Tsolaki et al., 2015), which reported a significant effect for the age and gender of a subject on the latency and amplitude of his or her EEG P300 and MMN components, we also investigate the relation of our decoding model with the age and the gender of subjects. We also explore the most informative temporal features of the EEG signal about the sensory input's surprise, and compare it with earlier findings.

2. Materials and methods

Summary of methods is depicted in Fig. 1.

2.1. Experiment

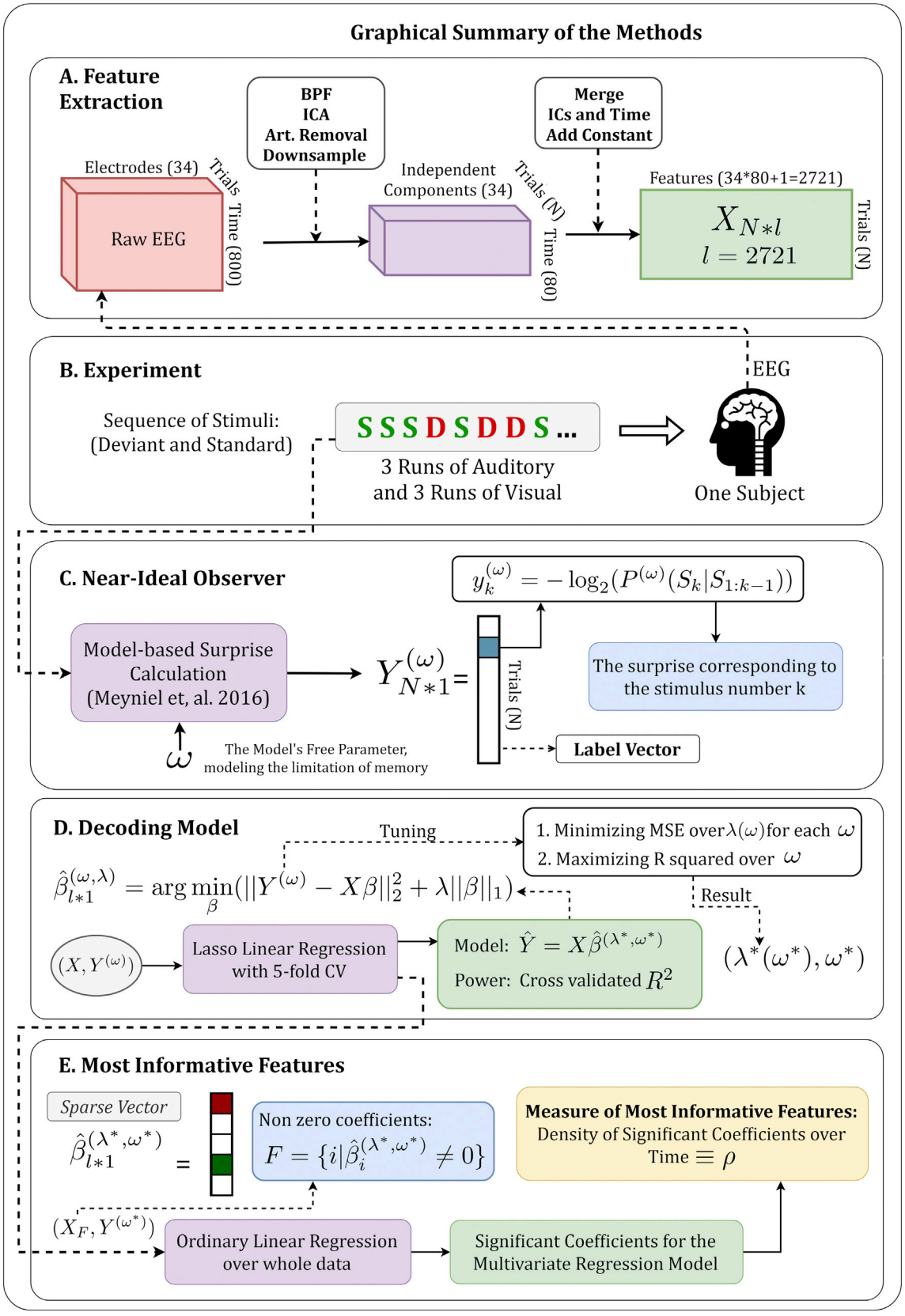
2.1.1. Main dataset

The dataset used in this work for the main analysis is an open access dataset which was originally collected and published by Walz et al. (Walz et al., 2013, 2015, 2014). Methods for data acquisition, and the description of experimental tasks are precisely described in (Walz et al., 2013). However, they are briefly mentioned here for reference.

Seventeen adult subjects (six females, eleven males; age mean: 27.7 years, age range: 20–40 years) participated in two separate auditory and visual oddball tasks. Each task consisted of three runs of a random sequence of stimuli - 135 stimuli per run. The deviant's (rare stimulus) occurrence probability for both of the tasks was 0.2. The first two stimuli of each sequence were fixed to be standards. Subjects were asked to use a button press to respond to the deviant/target stimuli (error rate: $0.8 \pm 0.6\%$ for the auditory and $0.6 \pm 0.5\%$ for the visual task, range: 0.3%–2.4%). Duration of each stimulus was 200 ms, and they were presented uniformly every 2–3 s. The standard and deviant stimuli for the auditory task were, respectively, a 390 Hz pure tone, and a broadband "laser gun" sound. For the visual task, a large red circle (4.45deg visual angles) and a small green circle (1.15deg visual angles) on isoluminant gray background were used as the standard and deviant.

EEG and fMRI data were simultaneously recorded during the task, as well as the button press responses. EEG signals were sampled at a rate of 1000 samples per second, and were re-referenced to the 34-channel electrode space. They were preprocessed by the original provider to remove gradient artifacts. The re-referenced version of the EEG signals, used in the original provider's study, was used in this work.

The experiment was in nature very similar to what were used in the other available datasets (Kolossa et al., 2013; Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Ostwald et al., 2012). However, there is an advantage for our study in using the dataset of this experiment. Each



(caption on next page)

Fig. 1. Graphical Summary of the Methods. The whole process was done separately for each subject. The details in the first two subfigures are related to only the main dataset, and not the benchmarking one. **A. Feature Extraction:** A band-pass filter was applied to the EEG signals which were recorded during the visual or auditory tasks. The trials which were affected by artifacts were removed based on a specific threshold. The trials were also transformed from the electrode space to the Independent Components space - which was constructed by applying the InfoMax algorithm (Bell and Sejnowski, 1995) using functions supplied by the EEGLAB open source toolbox (Delorme and Makeig, 2004). Finally, trials were downsampled and reshaped in order to make the final feature matrix. **B. Experiment:** The dataset which was used in our work had been collected and published in (Walz et al., 2013, 2015, 2014). 17 subjects had participated in three runs of an auditory as well as a visual binary oddball task. EEG signals and fMRI had been recorded during the experiments, but only the EEG signals were used in our work. **C. Near-Ideal Observer:** By applying the transition probability model, which had been developed by (Meyniel et al., 2016), on the sequences of the stimuli of experimental tasks, the surprise value of each single trial was calculated. Assuming that the subjects acted like the near-ideal observer, the calculated surprises were considered as labels for the decoding model. **D. Decoding Model:** Lasso linear regression was employed to make the decoding model (Hastie et al., 2001). The free parameters of lasso (the weight of norm-l1 regularization) and the transition probability model (the rate for the leakage of memory) were tuned respectively by minimizing the mean square error and maximizing the R-squared value. Decoding power of the model was calculated using cross-validation. **E. Most Informative Features:** The features corresponding to non-zero coefficients of the decoding model were used to train a multivariate ordinary linear regression model over the entire data – without cross-validation. The ordinary linear regression model was used to make it simpler to statistically test the significance of the selected features – by also taking into account the correlations of different features. Therefore, the significant coefficients were chosen by applying the Benjamini and Hochberg algorithm to control False Discovery Rate in large-scale simultaneous hypothesis testing (Efron, 2010). Finally, to compare the level of being informative for different time slots, the time density of significant features, ρ, was calculated.

subject participated in both visual and auditory tasks in this dataset. This allows us to examine the brain's oddball responses in two modalities, and compare the performance of our decoding model on these modalities.

In contrast to some other works (Kolossa et al., 2013; Mars et al., 2008; Ostwald et al., 2012), there is an asymmetry in the experiment design of the mentioned dataset: subjects responded to the deviant stimuli by pressing a key, and did nothing in response to the standard ones. Therefore, one might argue that there may be a correlation between the brain's response to the deviant stimuli and the subject's decision to perform a motor response, and that this correlation could enhance the decoding performance of our model. Such an argument could then mean that the decoding result may not be informative about the brain's perception of surprise, and might be decoding the motor response of the subject. To investigate this issue, our decoding model was also applied to another open access dataset in which subjects responded to both types of stimuli by pressing keys. The symmetric response to both standard and deviant stimuli by the subject would in this case allow for the assessment of the decoding performance in spite of any such correlations.

2.1.2. Benchmarking dataset

The benchmarking dataset is an open access dataset which was recently published in (Robbins et al., 2018). The dataset was originally collected for the study reported in (David Hairston et al., 2014), but it has since been used as test data in several other technical studies, such as (Bigdely-Shamlo et al., 2015; Su et al., 2018).

Eighteen adult subjects participated in a visual oddball task. No information about the age and gender of the participants was included in the dataset. The task consisted of three blocks of random sequences, each of which contained approximately 89 stimuli. The original provider has reported that the deviant's occurrence probability was equal to 1/7; however, based on our analyses it is not same for all subjects, and its average is roughly 1/8. Subjects were asked to respond to the standard and deviant stimuli separately by different press buttons (error rate: 3.8 $\pm$ 2.4%, range: 0.4%–11.3%). The dataset provider did not include information on whether the subjects responded to stimuli with only one finger, with two fingers of the same hand, or with two fingers of different hands – there is also the possibility that there was no specific protocol. Duration of each stimulus was 150 ms, and they were presented with the frequency of 0.5 $\pm$ 0.1 Hz. The standard stimulus was an image of a US soldier, and the deviant stimulus was an image of enemy combatant. Images were presented with size of 152*375 pixels on a Dell P2410 monitor whose distance from subjects was approximately 70 cm.

EEG signals and button press responses were recorded during the task. EEG signals were sampled at a rate of 512 samples per second in a 64-channel electrode space. They were preprocessed by the original provider for removal of line noise, and detection as well as interpolation of bad channels (Bigdely-Shamlo et al., 2015); the ICA transformation matrix was also calculated for each subject (exactly the same process which is used in this work on the main dataset and described in section 2.2.1), and the ICs corresponding to artifacts were specified (Winkler et al., 2011). The cleaned version of EEG signals was used in this work.

In the following sections, the focus of our presentation will be on the processes we applied to the main dataset. As mentioned above, similar processes have also been applied to the benchmark data by the producers of the dataset.

2.2. EEG preprocessing and feature extraction

Graphical summary of this part is depicted in Fig. 1A.

2.2.1. Independent component analysis (ICA)

With the aim of blindly separating task-dependent brain sources from the irrelevant sources, such as eye artefacts, the technique of independent component analysis was exploited (Jung et al., 2001, 2000). The InfoMax ICA algorithm (Bell and Sejnowski, 1995) was applied to EEG signals separately for each subject and each modality. This process was done by using the functions supplied by the EEGLAB open source toolbox (Delorme and Makeig, 2004). Finally, 34 transformation matrices (2 modalities multiplied by 17 subjects) were computed. By using these matrices, the signals were transformed from the electrode space to the independent component (IC) space. Since EEG signals had been re-referenced to 34 electrodes by the developer (Walz et al., 2013), the dimension of both spaces were equal to 34.

This transformation, in addition to the automatic feature selection method which was used in our work, and is described in the following parts, eliminated the redundant relations between electrodes, and chose the best features for the decoding model. Furthermore, by doing so, the artefacts were blindly separated and removed – this procedure was conducted based on the earlier methods of ICA-based artefact removal (Jung et al., 2001, 2000), but with an automatic approach.

It should be mentioned that in this work, ICA is considered only as a technique in the process of feature extraction, and the output sources are not interpreted as any meaningful EEG sources.

2.2.2. Feature matrix generation

A digital bandpass filter was applied to the transformed EEG signals – bandwidth = 0.5–38 Hz, Kaiser FIR BPF (length = 501, beta = 3) (Oppenheim and Schafer, 1975). Then, trials were extracted by segmenting the signals from the onset of each stimulus to 600 ms after that. The baseline of each trial was corrected using the 200 ms time interval preceding the onset of the corresponding stimulus. Then, the trials whose maximum amplitudes in any of the electrodes were more than 250μV were excluded – the number of remaining trials is defined as N, which is around 360 for each subject. Finally, the remaining trials were downsampled by 10, which means that their sampling rate was converted from 1000 Hz to 100 Hz. In addition, and in order to evaluate the performance

of our feature selection method, the baseline, i.e. a 200 ms window preceding the onset of the stimuli, was also considered to be part of the features. Therefore, for each specific stimulus, there were 80 time points (20 points for baseline and 60 points for post-stimulus signal) for each of the 34 independent components (ICs). Concatenating these time points with each other, there were 2720 (80 time points multiplied by 34 ICs) features for each single trial (the dimension of this primary feature space is called l which is equal to $2720 + 1 = 2721$; the 1 is for the constant regressor). As a result, for each subject and each modality, the feature matrix, X_{N*l}, is defined as an N by l matrix with rows corresponding to the trials and columns corresponding to the features – see Fig. 1A.

2.3. Near-ideal observer

It has traditionally been assumed that the brain's responses to a binary sequence of stimuli are the same as the responses of a near-ideal observer to the same sequence (Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Ostwald et al., 2012). Therefore, in our study, a model developed by Meyniel et al. in 2016 (Meyniel et al., 2016) was considered to represent this near ideal observer. This model was named "transition probability model" with "leaky integration inference style" in (Meyniel et al., 2016). Detailed formulation and experimental evaluations are precisely described in (Meyniel et al., 2016). However, a brief summary of the model is mentioned here for reference – graphical summary can be found in Fig. 2.

2.3.1. Transition probability model

The model is constructed based on this assumption: the brain hypothesizes that the generation of stimuli in a binary sequence is governed by a "Markovian" generative process, i.e. the conditional probability distribution of each stimulus given past observations is a function of only the previous stimulus rather than all of the earlier stimuli (Meyniel et al., 2016). In other words, the brain abstracts a pattern from the observations, which can be compressed to only two parameters: the occurrence probability of a deviant stimulus (D) after a standard (S), which is defined as $\theta_{D|S}$, and the reverse, which is denoted by $\theta_{S|D}$ (Fig. 2A). Based on this assumption, in order to make a prediction about the future stimulus, the brain needs to make an estimate of these two parameters.

2.3.2. Leaky integration inference style

It is assumed that the brain's inference style is based on the Bayes' rule. Furthermore, the other assumption is that the brain uses previous stimuli to make a prediction about the future ones, and actively adds new presented stimuli to its set of observations – a phenomenon called "active

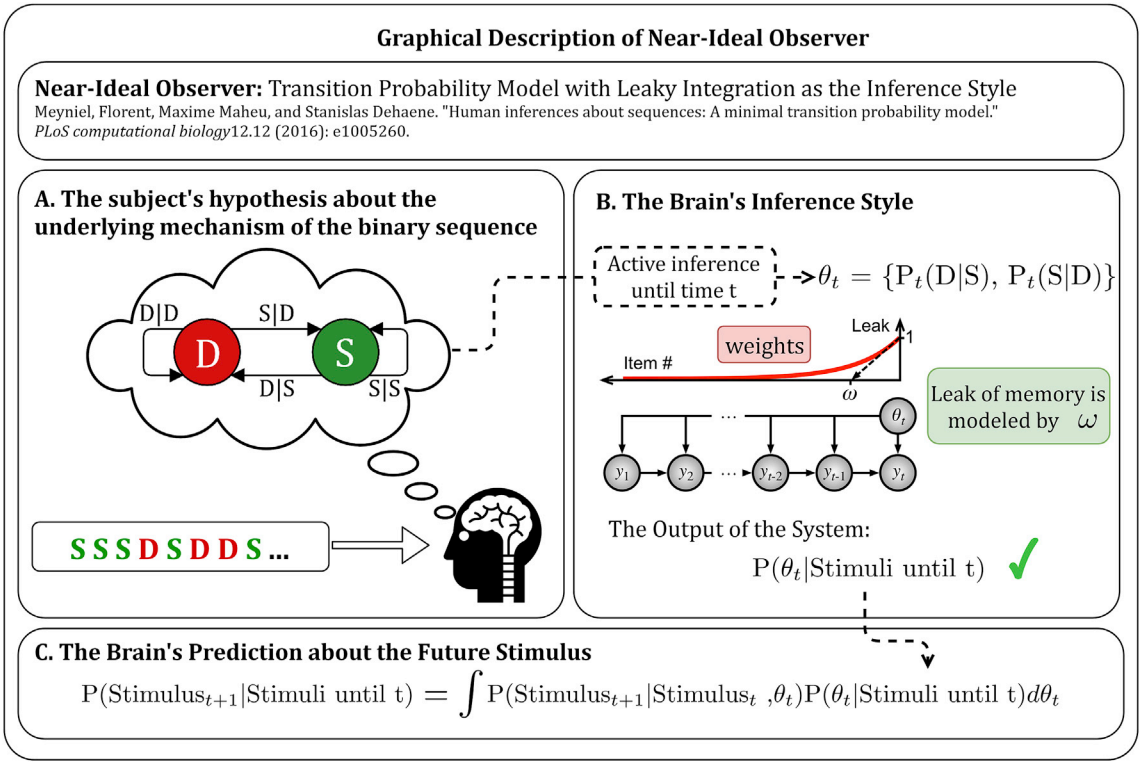


Fig. 2. Graphical Description of Near-Ideal Observer. The decoding model was constructed based on the assumption that the brain's perception about the sequences was governed by one of the systems (near-ideal observers) which had been developed by the recent work of Meyniel et al. in 2016 (Meyniel et al., 2016) – because of its relatively high descriptive power in outperforming earlier works, its simplicity, and its independence from the sensory domain. In fact, Meyniel et al. discussed four different types of inference for their model, but only the one which they called "leaky integration inference" was used for developing our decoding model – this inference model is described more precisely in part B of figure. **A. Subjective Hypothesis about the Underlying Mechanisms of the Sequence:** It was assumed that the brain considers that the sequence of stimuli is generated by a "Markovian" generative process, i.e. only the previous stimulus is informative about the upcoming one. Based on this consideration, the brain continuously predicts the future stimulus given the previous observations. **B. The Brain's Inference Style:** To predict the future stimulus, the brain needs to have an estimation about the hyper-parameters of the assumed "Markovian" generative process. Therefore, it was assumed that the brain uses active Bayesian inference, i.e. at each time point, the brain predicts the future stimulus, and when the stimulus is presented, the brain uses it to update its estimation of the hyper-parameters. However, since the human memory is limited, the ideal Bayesian estimation cannot be implemented in the brain. On that account, it is assumed that the brain considers an exponential set of weights for the past stimuli: considering the recent stimuli as more important than the older ones. This limitation of memory is controlled by a free parameter, ω. The output of the inference model is a conditional probability distribution for the hyper-parameters of the model, given the past observations. (The curve showing weights and the graph of observations are adapted from (Meyniel et al., 2016).) **C. The Brain's Prediction about the Future Stimulus:** It is assumed that the brain predicts the future stimulus by calculating its conditional probability distribution given the past observations. To this end, it should use the conditional probability distribution of the hyper-parameters as the result of the inference model.

inference" (Meyniel et al., 2016).

If the brain's memory was infinite, it would be reasonable to assume that the brain perfectly uses all of its observations for estimating the model's parameters. Assuming that the brain's memory is infinite, using the Bayes' rule, and the Markovian assumption, the likelihood function of parameters can be written as below:

$$P(S_{1:t}|\theta) = P(S_1|\theta)\prod_{i=2}^{t} P(S_i|S_{i-1},\theta) \tag{1}$$

where S_i is the stimulus number i, $S_{1:t}$ is the set $\{S_1,\ldots,S_t\}$, and $\theta = [\theta_{S|D}, \theta_{D|S}]$ is the vector consisting of the parameters of the Markovian process. Considering the uniform distribution for the first stimulus, Eq. (1) can be written as:

$$P(S_{1:t}|\theta) = \frac{1}{2}\left(\theta_{S|D}^{N_{S|D}}\left(1-\theta_{S|D}\right)^{N_{D|D}}\right)\left(\theta_{D|S}^{N_{D|S}}\left(1-\theta_{D|S}\right)^{N_{S|S}}\right) \tag{2}$$

where $N_{S|D}$ is the number of transitions from a deviant stimulus to a standard; $N_{D|D}$, $N_{D|S}$, and $N_{S|S}$ are also defined in a similar manner. It is assumed that the prior probability distribution of the parameter θ is a uniform distribution. Considering this assumption, and using Eq. (2), after some simple mathematical manipulations, it can be shown that the *a posteriori* probability distribution function of the parameters equals the product of two Beta distributions:

$$P(\theta|S_{1:t}) = Beta\left(\theta_{S|D}|N_{S|D}+1,\ N_{D|D}+1\right)Beta\left(\theta_{D|S}|N_{D|S}+1,\ N_{S|S}+1\right) \tag{3}$$

Finally, the Bayesian prediction about the next stimulus can be derived using Eq. (3) as mentioned below:

$$P(S_{t+1}|S_{1:t}) = \int P(S_{t+1}|\theta,S_t)P(\theta|S_{1:t})d\theta \tag{4}$$

Then, the corresponding surprise of each stimulus is produced by the following equation. This is the value that our model is designed to decode based on the EEG signals:

$$y_{t+1} = -\log_2 P(S_{t+1}|S_{1:t}) \tag{5}$$

This style of inference is called "Perfect Integration" in (Meyniel et al., 2016). However, since the brain has a finite memory, Eq. (3) should be modified for its estimation model. Therefore, Meyniel et al. considered the weighted $N_{S|D}$ instead of the absolute $N_{S|D}$ – the same assumption is also applied to $N_{D|D}$, $N_{D|S}$, and $N_{S|S}$. Values of $N_{S|D}$, $N_{D|D}$, $N_{D|S}$, and $N_{S|S}$ are computed by applying an exponential weight of $\exp\left(-\frac{k}{\omega}\right)$ for the k^{th} stimulus preceding the current one, where ω is a free parameter which is used to control the decay of memory over time. This variation from the "Prefect Integration" inference is called "Leaky Integration" by Meyniel et al., (2016). It should be noted that Eq. (3) is applied for both of the mentioned inference models, but with different values for the parameters $N_{S|D}$, $N_{D|D}$, $N_{D|S}$, and $N_{S|S}$.

2.3.3. Reasons for choosing transition probability model

In the Introduction section, it was mentioned that the model of Meyniel et al. was used in this work because of its relatively high descriptive power in outperforming earlier works, its simplicity, and its independence from the sensory domain. Other options for observer model were models which had been proposed by Mars et al. in 2008 (Mars et al., 2008), Ostwald et al. in 2012 (Ostwald et al., 2012), Kolossa et al. in 2013 (Kolossa et al., 2013), and Lieder et al. in 2013 (Lieder et al., 2013). Based on the following reasons, we believe that the best model for our study is the Transition Probability Model.

First, using the transition probabilities instead of the probability of occurrence as the parameters which the brain uses for prediction makes the model of Meyniel et al. more general and more descriptive than the ones developed by (Mars et al., 2008; Ostwald et al., 2012) – this is discussed in detail in (Meyniel et al., 2016), but as an instance, estimating the occurrence probability cannot explain the brain's perception about a continuously alternating sequence of binary stimuli. Second, analyzing experimental data, Meyniel et al. demonstrated that their model outperformed the one proposed by (Kolossa et al., 2013) – considering Bayesian Information Criteria for model selection, they demonstrated that their model is not only more descriptive, but also simpler (Meyniel et al., 2016). Third, in contrast to the model proposed by (Lieder et al., 2013), which is specifically designed for a special auditory oddball task with single tone stimuli, the Transition Probability Model is independent from the modality of stimuli, and can be used for both of the auditory and visual tasks.

2.4. Label generation (surprise calculation)

Graphical summary of this part is depicted in Fig. 1C. Subjects were considered to be near-ideal observers. Furthermore, similar to the earlier works (Mars et al., 2008; Ostwald et al., 2012), the different experimental runs for each subject were assumed to be separate and independent from each other; in other words, the prior belief of each subject at the beginning of each experimental run was considered to be a uniform probability distribution on types of stimuli. Then, by applying the Transition Probability Model on the sequences of stimuli, the subjective surprise of every single stimulus was calculated as a function of ω (the model's free parameter). As a result, for each subject and each modality, the label vector $Y_{N*1}^{(\omega)}$ is defined as an N by 1 vector whose entries are equal to the model-based surprise of the stimuli (y_t – which is defined in Eq. (5)). The process of tuning ω for each subject and each modality is described in section 2.5.2.

2.5. Decoding model

Graphical summary of this part is depicted in Fig. 1D.

2.5.1. Lasso Linear Regression

As it was mentioned earlier in section 2.2.2, the dimension of the feature space is equal to $l = 2721$, but the number of observations (trials), N, is around 360. Considering this, and the fact that all of the time points are merged together, whether they are related to the task or not, it is necessary to regularize the model to avoid over-fitting. Therefore, our decoding model was constructed by using Lasso Linear Regression, which is a regularized version of the ordinary Linear Regression (Hastie et al., 2001). The procedure of applying this regression model can be expressed as below:

$$\widehat{\beta}_{l*1}^{(\lambda,\omega)} = \underset{\beta\in\mathbb{R}^l}{\operatorname{argmin}}\left(\left|Y_{N*1}^{(\omega)} - X_{N*l}\beta_{l*1}\right|_2^2 + \lambda|\beta_{l*1}|_1\right) \tag{6}$$

where $\widehat{\beta}_{l*1}^{(\lambda,\omega)}$ is the regression coefficient vector, and λ is the regularization parameter. Using this notation, the Mean Square Error and R-squared value are defined as functions of λ and ω:

$$MSE(\lambda,\omega) \triangleq \frac{\left|Y_{N*1}^{(\omega)} - X_{N*l}\widehat{\beta}_{l*1}^{(\lambda,\omega)}\right|_2^2}{N} \tag{7}$$

$$R^2(\lambda,\omega) \triangleq 1 - \frac{MSE(\lambda,\omega)}{\sigma^2_{Y_{N*1}^{(\omega)}}} \tag{8}$$

2.5.2. Parameter tuning

For each subject and each modality, the best λ was found as a function of ω by minimizing the $MSE(\lambda,\omega)$ over λ, employing a 5-fold cross-validation (folds were constructed by partitioning shuffled data samples, no matter from which run of the experiment they were taken); the tuned parameter is defined as $\lambda^*(\omega)$. However, since the label vector by

itself is a function of ω (see section 2.4), minimizing $MSE(\lambda^*(\omega),\omega)$ over ω does not necessarily lead to the best regression model. As a matter of fact, a trivial solution of the MSE optimization problem can be found by limiting ω to zero. In this case, since the surprise calculation procedure is memory-less, surprise of each stimulus is roughly equal to 1 bit. As a consequence, by estimating only a constant amount of surprise for all stimuli (independent from the corresponding features), it is possible to achieve approximately zero MSE, which is obviously not the desirable solution for a decoding model. Therefore, ω was tuned by maximizing $R^2(\lambda^*(\omega),\omega)$, which is a normalized measure for the goodness of regression models – the process is identical to minimizing the normalized MSE, the second term in Eq. (8). By doing so, the best coefficient vector, $\widehat{\beta}_{l*1}^{(\lambda^*,\omega^*)}$, was chosen.

2.5.3. Decoded surprise and decoding power

The decoded surprise vector, $\widehat{Y}_{N*1}$, was calculated by multiplying the feature matrix by the coefficient vector, which is formulized in Eq. (9).

$$\widehat{Y}_{N*1} = X_{N*l}\widehat{\beta}_{l*1}^{(\lambda^*,\omega^*)} \tag{9}$$

To measure the precision of the decoding model, the 5-fold cross-validated R-squared, defined in Eq. (10), was considered as its power (folds were constructed as described in the previous section).

$$R^2 \triangleq 1 - \frac{\mathrm{E}_{CV}\left[MSE\left(\lambda^*,\omega^*\right)\right]}{\sigma^2_{Y_{N*1}^{(\omega^*)}}} \tag{10}$$

where $\mathrm{E}_{CV}[MSE(\lambda^*,\omega^*)]$ is the average of MSE over folds of cross-validation. Finally, by applying this model on each subject, and for each modality, we had 34 (17 subjects multiplied by 2 modalities) decoding models; the decoding power was calculated separately for each of them.

2.5.4. Null distribution of decoding power

Consider the null hypothesis as the following: subjective surprise (Y) is independent from EEG signals (X). Given the condition that this hypothesis is true, R-squared values corresponding to applying the decoding model on the different permutations of label vector, when the feature matrix is fixed, are coming from the same distributions. In other words, when X and Y are independent from each other, order of assigning the components of Y to the rows of X is not important. Therefore, for each subject and each task, we considered 100 random permutations of the label vector (the number of all possible permutations > 10300). Then, to make an estimation of the chance level for the decoding power for each of them, the R-squared value was computed. Finally, the null distribution for each task was estimated by combining the samples calculated for each individual subject ($17*100 = 1700$ samples).

2.6. Most informative temporal features

Graphical summary of this part is depicted in Fig. 1E. First, a subset of features corresponding to the non-zero coefficients of $\widehat{\beta}_{l*1}^{(\lambda^*,\omega^*)}$ was considered for further analysis. Then, multivariate ordinary Linear Regression was applied on the selected set of features to statistically verify whether each coefficient is significant. By comparing the time density of these significant coefficients over different time slots, most informative temporal features were specified.

2.6.1. Selected features

Considering the fact that the coefficient vector of Lasso Linear Regression is sparse, i.e. most of its entries are zero (Hastie et al., 2001), the set of features corresponding to the non-zero coefficients were selected as the most predictive and descriptive features – this is mathematically defined in Eq. (11):

$$F = \left\{ i \middle| i \in \{1,2,\ldots,l\}, \widehat{\beta}_i^{(\lambda^*,\omega^*)} \neq 0 \right\} \tag{11}$$

For each subject and each modality, the selected feature matrix, $(X_F)_{N*|F|}$, is defined as an N by $|F|$ (size of the set F) matrix whose rows correspond to trials and whose columns are a subset of the columns of the original feature matrix, X_{N*l}.

2.6.2. Significant coefficients

As a matter of fact, there are some very recent works, such as (Barber and Candes, 2016; Weinstein et al., 2017), which propose novel methods of finding the significant coefficients of Lasso Regression. However, these methods are extremely more complicated than the traditional method which is used to find the significant coefficients of Ordinary Linear Regression (Hastie et al., 2001). Therefore, to find the significant features for surprise decoding, an ordinary linear regression was trained on the space of selected features, formulized in Eq. (12).

$$\widehat{\beta}^*_{|F|*1} = \underset{\beta\in\mathbb{R}^{|F|}}{\arg\min} \left| Y_{N*1}^{(\omega^*)} - (X_F)_{N*|F|}\beta_{|F|*1} \right|_2^2 \tag{12}$$

Applying the traditional method for testing the hypothesis of whether β_i^* is zero or not, as described in (Hastie et al., 2001), one p-value as well as one t-statistics was calculated for each entry of the coefficient vector. Then, using the algorithm of Benjamini and Hochberg for controlling the False Discovery Rate (FDR) in large-scale simultaneous hypothesis testing (Efron, 2010), the significant coefficients were found (the FDR was controlled to be less than 0.1).

It is worthy to mention that in these types of problems, p-values are calculated by considering the effect of correlations between different features. This approach is fundamentally different from testing the effect of each individual feature on the decoding procedure – similar to what was done in (Ostwald et al., 2012). To make this difference clear, consider a hypothetical example of investigating the correlations between the values of various time-points of a set of realizations of a discrete Markovian process. We know that there is a significant correlation between the value of an observation and the value of each of the observations preceding it. However, once we have the most recent one, the value of each time-point by definition is independent from the older observations. In other words, although there are several informative observation, the most recent one is also the most informative one. This phenomenon can be observed by multivariate, and not univariate analysis. In this paper, based on the same idea, we search for the "most informative" temporal features and not for all of the informative ones.

2.6.3. Density of significant coefficients

The measure of being informative for a time interval, for example from t_1 to t_2 was defined as the time density of significant coefficients corresponding to this interval. This time density function was separately calculated for each IC, each subject, and each modality. The mathematical definition is stated in Eq. (13).

$$\rho(t_1,t_2) \triangleq \frac{\#C_S(t_1,t_2)}{t_2 - t_1} \tag{13}$$

where $\#C_S(t_1,t_2)$ is the number of significant entries of the coefficient vector, $\widehat{\beta}^*_{|F|*1}$, corresponding to a specific subject, IC, and modality, and in the desired time interval. In other words, by applying Eq. (13) to all of ICs and subjects for each time interval, and for each modality, there are 578 (17 subject multiplied by 34 ICs) time densities. After excluding the ICs which did not have any significant coefficients, the averages of time densities for different time intervals were compared using t-test, considering each IC of each subject as one sample.

2.6.4. Desired time intervals

To investigate the temporal properties of surprise encoding, the

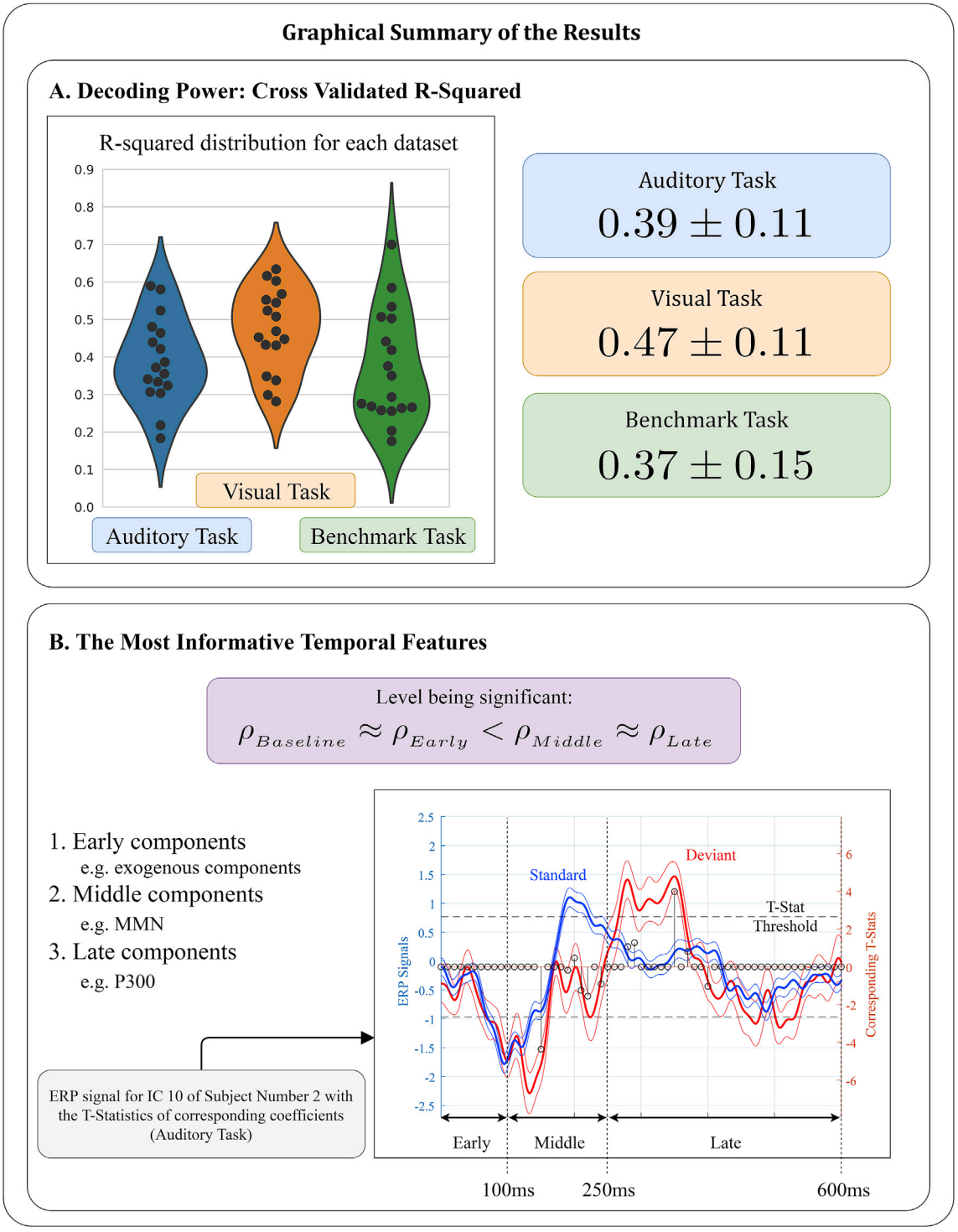


Fig. 3. Graphical Summary of the Results. The decoding model was constructed separately for each subject, and for the benchmarking dataset (visual) and the main dataset (auditory and visual tasks). **A. Decoding Power, Cross Validated R-Squared:** The distribution of R-squared values of the decoding models are shown in the form of clouds of dots for all of the three tasks. Each sample point is related to one subject. **B. The Most Informative Temporal Features:** Time intervals corresponding to the Early, Middle, and Late components are specified on an ERP waveform of a subject; the t-statistics of the coefficients corresponding to each time point are also depicted – the threshold which was calculated by the algorithm of Benjamini and Hochberg is also specified in the figure. Consider that the t-statistics were calculated completely independently from the ERP waveforms – by a multivariate regression model; these are depicted on the same figure to show the consistency of our method with the earlier works. According to the result of applying *t*-test to compare the densities of significant coefficients over time, ρ, the Early components of the EEG evoked potentials are not different from the Baseline for the decoding procedure, and are significantly less informative than the Middle and Late components. Also the effect of ρ_{Late} and ρ_{Middle} on the performance of the model is the same.

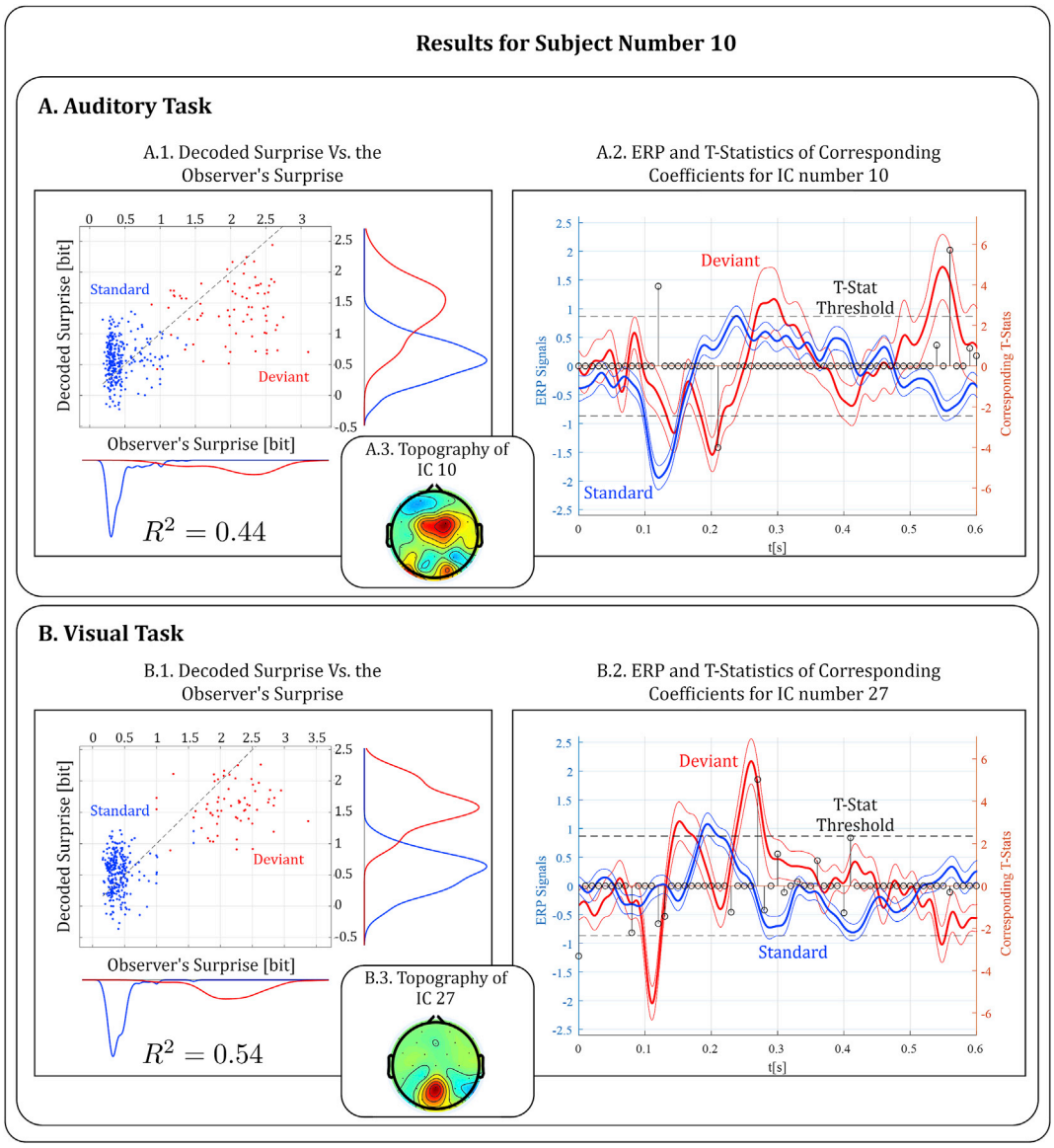


Fig. 4. Results for Subject Number 10. A. Auditory Task: 1. The decoded surprise value is depicted versus the surprise value which was calculated by the ideal observer model used as the label. Decoded values are calculated by 5-fold cross-validation, i.e. each time point is estimated by 4/5 data which does not contain this point. The marginal distribution over each dimension is also depicted separately for standards and deviants. 2. The waveform of Event Related Potentials (ERP) of the Independent Component (IC) which had the greatest t-statistics in absolute value is shown. 3. The brain topography map of the IC for which the ERP is depicted is shown. Red color corresponds to the positive values and blue color corresponds to the negative ones. In order to clarify the possible ambiguities, it should be mentioned that the t-statistics are derived using the multivariate linear regression analysis, described in section 2.6.2, and not by comparing the ERP waveforms of standard and deviant stimuli. These values are depicted on the ERP signals only to show the consistency of our method with the earlier results. **B. Visual Task:** The figures are similar to what is mentioned for part A, but for the visual tasks.

200 ms pre-stimulus as well as 600 ms post-stimulus time intervals of evoked potentials were partitioned to three parts:

1. -200 ms to 0 (Baseline): the time interval which causally cannot be informative about the surprise of stimuli. This interval was considered for analysis of the chance level of the time density of significant coefficients.
2. 0–100 ms post stimulus (Early Components): the time interval which mainly contains exogenous components, which are mostly related to physical parameters of the stimuli (Sur and Sinha, 2009)
3. 100 ms–250 ms post stimulus (Middle Components): the time interval which was segmented due to the fact that the Mismatch Negativity (MMN) component, which is assumed to be correlated with subjective surprise, usually happens in this interval (Garrido et al., 2009; Lieder et al., 2013; Näätänen, 2000)
4. 250 ms–600 ms post stimulus (Late Components): the time interval in which the most important component is P300, which is traditionally considered as an index for subjective surprise (Kolossa et al., 2013; Luck, 2004; Mars et al., 2008; Sur and Sinha, 2009)

Using the defined time density, these time intervals were compared with each other. The average of time densities over the ICs and subjects are defined respectively as $\rho_{Baseline}$ ρ_{Early}, ρ_{Middle}, and ρ_{Late}.

3. Results

The graphical summary of the results is depicted in Fig. 3. The details of applying the model to a specific subject is also shown in Fig. 4, which consists of three charts: A.1 & B.1: Graphs showing the decoded surprise versus the surprise calculated by the ideal observer model; A.2 & B.2:

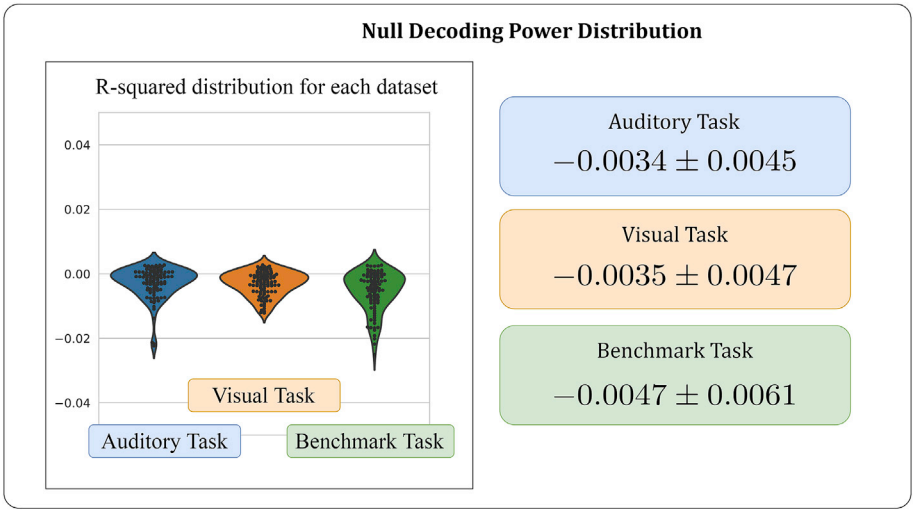


Fig. 5. Null Decoding Power Distribution. The null distribution of R-squared values of the decoding models are shown in the form of clouds of dots for all of the three tasks. The distributions were estimated by repeating the decoding models for different permutations of the label vector, while the feature matrix was fixed. To have a more informative, less crowded figure, only 100 samples are shown in each cloud, but the means and standard deviations are calculated by the 1700 samples of the main dataset's visual and auditory tasks, and the 1800 samples of the benchmarking dataset's visual task.

Waveforms of Event Related Potentials (ERP) for the Independent Component (IC) which had the greatest t-statistics in absolute value; A.3 & B.3: Brain topography maps of the mentioned ICs. These results are described in detail in the following sections.

3.1. Decoding power

The distributions of R-Squared values, which are considered as the decoding power, are shown in Fig. 3A. The results demonstrate that the model could decode the subjective surprise of every single stimulus, for each subject, and for the auditory, visual, and benchmarking tasks with high precision. To compare the results with chance level, the null distributions of the decoding power are shown in Fig. 5 (the means and standard deviations are reported in Table 1).

Comparing the result for the benchmarking task and the main task (more details in section 3.2), demonstrates that the asymmetry between the standards and deviants in the first dataset, i.e. subjects responded only to the deviant stimuli, does not have a significant effect on the decoding power of our model. Therefore, the reasonable performance of our model even in the case that the subjects are responding to both types of stimuli can be considered as an evidence against the possibility that the motor cortex activities enhance our model's performance. This issue is discussed in more detail in Appendix A. Also, the results of applying our decoding model to a third dataset in which the subjects did not have to press buttons are provided and analyzed in Appendix A.

3.2. Effect of modality

The p-values of the *t*-tests for comparing the averages of the decoding power of the three examined tasks are reported in Table 2. In addition, for each of these tests, the Bayes factor is reported using the approach proposed by (Rouder et al., 2009): instead of the traditional p-value, the ratio of marginal likelihoods is computed by assuming a specific prior distribution for the model's parameters. This approach makes it possible not only to reject the null hypothesis, but also to accept it. In our work, we employed the online platform developed by (Rouder et al., 2009) – considering scaled JZS prior with $r = 1$. Traditionally, Bayes factors of 1/3 and 1/10 (3 and 10) are respectively considered as sufficient and strong evidences for rejecting (or accepting) the null hypothesis.

According to the results, we can accept the indifference between the auditory task of the main dataset and the visual task of the benchmark dataset. However, we cannot make a strong statement about the relation of the visual task of the main dataset with the other tasks. This is because according to the p-value, the decoding model performed slightly better in this task than in the two others, but based on the Bayes factor, there is no sign for rejecting or accepting the null hypothesis – which means that in this case, observing the significant difference from the point of view of traditional hypothesis testing may be due to its bias toward accepting the alternative hypothesis.

The correlation between the decoding powers for the auditory and the visual tasks of the main dataset was also calculated (since the sample population of the benchmarking dataset was different from the visual and auditory tasks of the main dataset, it was impossible to analyze the correlation between the decoding powers for the benchmarking dataset and the main dataset). According to the results, we could neither reject nor accept the null hypothesis, i.e. we cannot make any strong statement about the significant correlation between the decoding power of the auditory and the visual tasks – Table 2. For this part, the Bayes factor was calculated using the approach proposed by (Liang et al., 2008; Rouder and Morey, 2012) and by employing the online platform developed by (Rouder et al., 2009).

Table 1
Statistics of decoding power for the visual, auditory, and benchmarking tasks – chance level statistics are computed using 1700 samples for the auditory and visual tasks of the main dataset, and 1800 samples for the visual task of the benchmarking dataset.

	Results	*Chance Level*
Main Dataset: Auditory Task	0.39 ± 0.11	-0.0034 ± 0.0045
Main Dataset: Visual Task	0.47 ± 0.11	-0.0035 ± 0.0047
Benchmarking Dataset: Visual Task	0.37 ± 0.15	-0.0047 ± 0.0061

Table 2
The p-values and Bayes factors of the statistical tests investigating significant difference in averages of and significant correlation between decoding powers.

	T-Test		Correlation Test	
	P-Value	Bayes Factor	P-Value	Bayes Factor
Auditory vs. Visual (Main Dataset)	0.020	1/2.70	0.254	1.49/1
Visual (Main Dataset) vs. Benchmarking Dataset	0.026	1/2.30	–	–
Auditory (Main Dataset) vs. Benchmarking Dataset	0.676	3.76/1	–	–

3.3. Age and gender dependency

The p-values as well as Bayes factors of the t-tests comparing the averages of the decoding power corresponding to male subjects and

Table 3
The p-values and Bayes factors of the statistical tests investigating significant difference in averages of decoding power for male and female subjects, in addition to the p-values and Bayes factors of the statistical tests investigating significant correlation between decoding power and the age of subjects.

	T-Test for Gender		Correlation Test for Age	
	P-Value	Bayes Factor	P-Value	Bayes Factor
Auditory	0.347	1.11/1	0.809	2.35/1
Visual	0.487	2.41/1	0.835	2.34/1

female subjects are reported in Table 3. We did not observe any significant gender-dependent difference. However, based on the Bayes factors, we could not accept the indifference between the male and female subjects. Furthermore, the correlation between the different measures of decoding power and the age of subjects was tested – results are also reported in Table 3. We did not observe any significant correlation between the age and the model's performance either. While extremely large p-values suggested that the model's performance may be independent from the age of subjects, based on the corresponding Bayes factors we could not accept that the age and decoding power are uncorrelated – which means that in this case, the lack of significant dependency from the point of view of traditional hypothesis testing cannot necessarily be considered as an indifference.

3.4. Standards versus deviants

The conditional probability distributions of the surprise values of the standard as well as the deviant stimuli, for one of the subjects, are shown in Fig. 4A1 and Fig. 4B1. Obviously, the average surprise value of the deviants was more than that of the standards; this was due to the difference between the global occurrence probabilities of standards and deviants, and is considered as the "global effect" in (Meyniel et al., 2016). However, there were also some standard stimuli which had more surprise values than some of the deviant stimuli; this phenomenon, which is considered as "local effect" by (Meyniel et al., 2016) is related not only to the absolute amount of occurrence probabilities of the stimuli, but also on their order of presentation in the sequence, e.g. the occurrence of a deviant after a long sequence of consecutive deviants have significantly less surprise than the occurrence of a standard in the same place.

The decoded surprise values versus the observer's surprise values for the same subject are also shown in the same figure. As it is illustrated in these plots, the decoding model worked for the low-surprise deviants and the high-surprise standards as accurately as for the other stimuli. In other words, the model could reasonably well decode the surprise value of every single stimulus, independent from its physical features. As a matter of fact, this characteristic of our model is its most significant distinction with respect to the existing classification models for separating standards from deviants (Blankertz et al., 2011). In other words, our decoding model makes an estimation about the perception of the brain about a stimulus, instead of the type of that stimulus. This variation may be employed in BCI (Brain-Computer Interface) systems like the P300 Speller (Guger et al., 2009) to improve their performance. However, to have stronger evidences for this result, one should also analyze the performance of our decoding model for the tasks in which the occurrence probabilities of standards and deviants are comparable.

3.5. Most informative temporal features

Time intervals corresponding to Baseline and the Early, Middle, and Late components are specified on an ERP waveform of a subject in Fig. 3B. The t-statistics of the coefficients corresponding to each time point are also depicted in the figure – the threshold which was calculated by the algorithm of Benjamini and Hochberg is also specified in the figure. Similar waveforms, but for another subject, are shown in Fig. 4A2 and Fig. 4B2.

Table 4
The p-values and Bayes factors of the statistical tests investigating significant difference between averages of time densities of significant coefficients in four time intervals of Baseline and Early, Middle, and Late components.

	Auditory Task		Visual Task	
	P-Value	Bayes Factor	P-Value	Bayes Factor
$\rho_{Baseline}$ vs. ρ_{Early}	0.94	19.5/1	0.220	9.60/1
$\rho_{Baseline}$ vs. ρ_{Middle}	3.25e-6	1/2.5e3	2.98e-7	1/2.4e4
$\rho_{Baseline}$ vs. ρ_{Late}	4.73e-6	1/1.8e3	3.95e-8	1/1.7e5
ρ_{Early} vs. ρ_{Middle}	4.70e-5	1/2.0e2	1.80e-9	1/3.4e6
ρ_{Early} vs. ρ_{Late}	2.54e-4	1/41	1.76e-11	1/3.1e8
ρ_{Middle} vs. ρ_{Late}	0.149	6.92/1	0.137	6.74/1

In order to clarify the possible ambiguities, it should be mentioned that the t-statistics are derived using the multivariate linear regression analysis, described in section 2.6.2, and not by comparing the ERP waveforms of standard and deviant stimuli. These values are depicted on the ERP signals only to show the consistency of our method with the earlier results.

3.5.1. Significant coefficients

Since the Lasso Linear Regression was used for the decoding model (section 2.5), and as it can be seen in the mentioned figures, most of the coefficients were equal to zero, which means that the coefficient vector was a sparse vector. Comparing the t-statistics with the corresponding time points of the ERP signal, it can be seen that the feature selection method blindly chose the features which were correspondent to the time points in which there was a large difference between the standard and the deviant ERPs. Further to this, the method blindly removed the redundancy between the features, i.e. in most cases, only one feature was selected from a set of informative but correlated features, for example the ones in the neighborhood of each extremum of the ERP - like the Markovian process conceptual example in section 2.6.2).

The result of the feature selection procedure demonstrates that our method is consistent with but more general and powerful than the methods reported in earlier works, which mostly considered the amplitude of one extremum point as a landmark for defining subjective surprise (Lieder et al., 2013; Mars et al., 2008; Ostwald et al., 2012). In our work, the surprise value was decoded by a linear combination of the extracted feature amplitudes. Another distinctive feature of our model is that the weights of the selected time points are tuned automatically as part of the proposed procedure. This feature allows us to find more complex patterns which are related to the surprise of the stimuli; for example, as it can be implied by Fig. 3B, it seems that the weighted average of the P300 and N100 amplitudes is used for surprise decoding. In fact, this result is the most important advantage of our decoding approach over the existing encoding approaches, in which it is not obvious which features of the EEG signal could be pointed to as landmarks of surprise. However, as more complex models are also more difficult to analyze, the main limitation of our decoding approach compared to simple encoding models is that its intermediate results may be hard to interpret.

3.5.2. Early, middle, and late components

The p-values as well as the Bayes factors of the *t*-tests investigating the significant differences between the averages of time densities of significant coefficients ($\equiv \rho$), in four time intervals of Baseline and Early, Middle, and Late components are reported in Table 4.

According to Table 4, for both the auditory and visual tasks, ρ_{Early} was significantly less than ρ_{Middle} and ρ_{Late}. In addition, based on the corresponding Bayes factor, the null hypothesis for the invariance of ρ_{Early} and $\rho_{Baseline}$ is accepted, which means that the Early components are completely non-informative about the surprise of stimuli. This result is consistent with, and even can be considered as a stronger evidence for the earlier works which reported that the Early components of ERP signal are

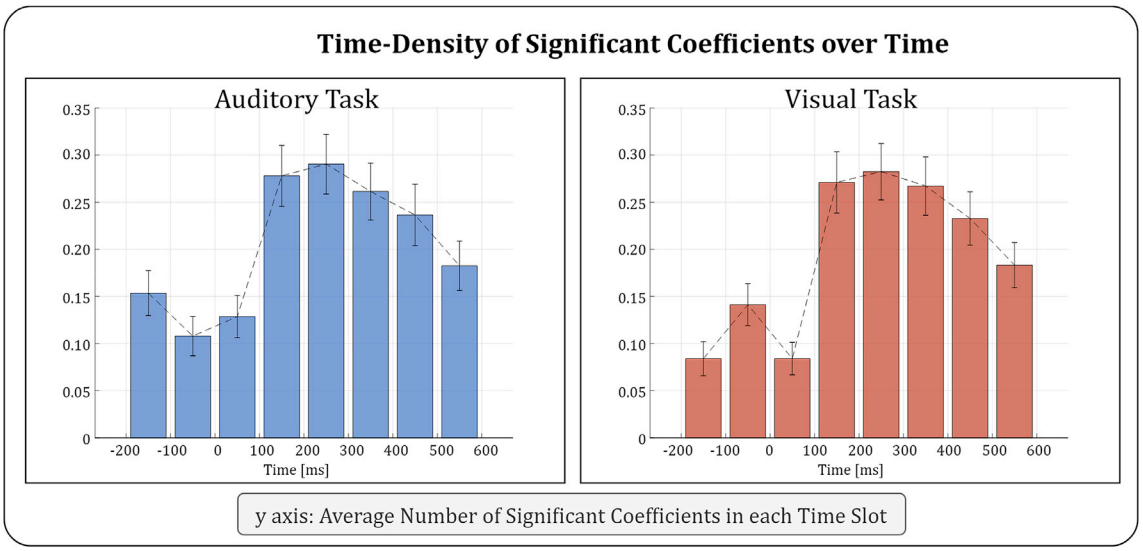


Fig. 6. Time Density of Significant Coefficients over Time (for 100 ms intervals). The defined time densities of the significant coefficients, ρ, for the eight 100 ms intervals were calculated. The average of the time density for each time interval as well as the standard deviation of the average (standard deviation of samples divided by the square root of number of samples) are shown separately for the auditory and visual tasks.

mostly functions of exogenous features of stimuli rather than the cognitive, or endogenous features (Luck, 2004; Sur and Sinha, 2009).

The results of comparing the Late and the Middle components demonstrated that both of them are significantly more informative than chance level, $\rho_{Baseline}$. In addition, and based on the calculated Bayes factors, the null hypothesis of invariance between ρ_{Middle} and ρ_{Late} is accepted.

In addition to the analysis of the mentioned four time intervals, we segmented the 600 ms intervals succeeding and 200 ms preceding the onset of the stimuli to eight 100 ms-intervals. Averages of the number of significant coefficients for each of these time intervals, for both of the modalities, are shown in Fig. 6. According to these plots and Table 4, the components between 100 ms and 400 ms are slightly more informative than the other components.

4. Discussion

In the current study, we developed a trial-by-trial surprise-decoding model for binary oddball tasks, and applied it to two open access datasets. We demonstrated that it is possible to decode the subjective surprise of each stimulus (i.e. in single trial level, and for both the standard and deviant stimuli), for every single subject, given only his or her EEG signals – for both the auditory and visual sensory modalities. Further to this, by analyzing the performance of our model, we did not observe any dependencies between the age and the gender of the subjects with the model's decoding powers; we also did not observe any significant difference between the performances of our model for the two different modalities. However, employing Bayesian approaches for hypothesis testing, we also could not accept the invariance in the mentioned situations. In the following subsections, we discuss the results of our work from different perspectives, and finally suggest some ideas for further studies.

4.1. An evidence for the Bayesian brain hypothesis

Our model related the output of the near-ideal observer which was governed by the Transition Probability Model (proposed by Meyniel et al., (2016)) with the neural responses of a subject. Therefore, the most primary interpretation of the results is that we can consider our model's decent performance as a new experimental evidence for the earlier theoretical works. Specifically, our work is a more general and stronger verification for the Transition Probability Model, in particular, and the Bayesian brain hypothesis in general (Knill and Pouget, 2004) – the reasons are described below.

As we mentioned detail in the introduction, there are three common approaches to work with experimental data, at least one of which has been employed by any of the earlier works to evaluate their theoretical models. We claimed that these approaches reduced the level of generalization of the evaluation procedure. To find a more reliable, general, and strong relationship between the experimental data and the theoretical models, in our study we avoided to follow the mentioned approaches, and developed our model in a manner to achieve the following characteristics:

1. Decoding the subjective surprise of each specific trial in the sequence of stimuli rather than considering only some on-averaged situations:
   - Trial-by-trial analysis allowed us to take into account the fluctuations which might be removed by averaging over subsets of stimuli. In other words, our model investigated the dynamics of the brain's responses to the sequence of stimuli. This approach is more general than analyzing the averaged form of the brain's responses, which only can be true for stationary situations – as is practiced in (Meyniel et al., 2016; Squires et al., 1976).
   - In contrast to the work of Lieder et al. in 2013 (Lieder et al., 2013) in which the evaluation was based on the analysis of only deviant stimuli, our work investigated the surprise value of every single trial – no matter if they were deviants or standards. Since the work of Lieder et al. was focused on the underlying mechanisms of MMN generation, it only considered the unexpected stimuli, i.e. the ones which change a repeating sequence, which obviously imposed a limitation on the generality of their results.
   - Our results demonstrated that although the EEG signals in the non-averaged form are noisy and difficult to interpret, evoked potentials are sufficient for decoding the subjective surprise of every single stimulus – by employing complex machine learning tools and analyzing signals in a multi-dimensional space. As a result, averaging over sequences employed in the work of Kolossa et al. in 2013 (Kolossa et al., 2013), which only used the P300 value, was not necessary. This is a significant advantage for extending the applicability of oddball test studies as averaging over the samples often results in a limited number of data points for being used in the subsequent analysis. However, it should be mentioned that a

disadvantage of this complex method compared to the earlier simple encoding models is that the interpretation of the meaning of features in the high-dimensional space can be much more difficult.

2. Constructing a decoding model instead of investigating the surprise encoding procedure:
   - Although there are lots of similarities between the theories for the encoding procedures and the decoding models, most earlier works were only concerned with the way surprise is encoded in neural responses (Kolossa et al., 2013; Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Ostwald et al., 2012). The focus of those works was usually on finding a significant statistical relationship between the output of their model and a landmark of surprise represented in EEG signals (Kolossa et al., 2013; Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Ostwald et al., 2012) rather than finding a reversible function relating the neural responses to the subjective surprise. Our results demonstrated that we are able to precisely decode the output of the Transition Probability Model, in particular, by looking only at EEG signals. In other words, complementing the earlier findings, our results proposed that not only can we observe a significant relationship between the neural responses and the near-ideal observer's output, but we are also able to precisely decode this output by using neural responses.
   - One may also attempt to address the second shortcoming of earlier works (being concerned only with the statistically significant relationship between the model's and the brain's responses, rather than the predictive power of the model) by still employing an encoding model and adopting a regression-type analysis rather than a factorial-type analysis. Such a likely approach may then seem to challenge our motivation for using a decoding model instead of an encoding model and render it an unnecessarily more complex method. We believe that while from the decoding perspective, in these types of model-based approaches, the specific feature of the stimuli which the model is to decode (in this work the subjective surprise of each stimulus) is well-defined, in the encoding analysis there is not a unified definition for the feature of the brain's responses in which the properties of the stimuli can be encoded. Therefore, usually in the encoding models, some very simple features of EEG signals or behavioral data are considered as the objective (Kolossa et al., 2013; Lieder et al., 2013; Meyniel et al., 2016; Ostwald et al., 2012; Squires et al., 1976). We admit that there is a tradeoff between how easy the model is to interpret and how informative it can be about the experimental phenomena. As it is more clarified in the following item, using the current approach allows us to find nontrivial complex patterns which are more informative about the surprise of the stimuli, but indeed more difficult to interpret.
3. Using multi-dimensional analysis instead of considering only one traditionally accepted landmark:
   - Multi-dimensional analysis of EEG signals allowed us to find the patterns which are much more informative than the amplitude of only one time point, or one famous component. This fact was not considered in almost all of the earlier studies like (Kolossa et al., 2013; Lieder et al., 2013; Mars et al., 2008; Ostwald et al., 2012). The only exception is the work of Squires et al. in 1976 (Squires et al., 1976). However, they also considered only three traditionally accepted landmarks (N200, P300, and large slow rate) as the features of evoked potentials.
   - Our automatic methods for feature selection, which blindly found the most informative features, allowed us to not be biased by the findings of the earlier works. It is conceptually similar to what was done in the study of Ostwald et al. in 2012 (Ostwald et al., 2012), but in a high-dimensional space, and by using a linear combination of the amplitudes of different time points of EEG signals. The most important feature of our model in this respect was that the weight of each single time point was fine-tuned automatically – while maximizing the decoding power.

Furthermore, we applied our model on 17 subjects who participated separately in an auditory as well as a visual oddball task – and as well on a completely different dataset with a visual task for benchmarking. As far as we know, our work is the first work which investigated the brain's responses of a specific group of subjects for two different modalities from the Bayesian perspective. The high precision of our model for both of this modalities, in addition to the three variations in the approach of data analysis, suggest that the results of our decoding model is a much more general and a stronger verification for the Transition Probability Model and the Bayesian brain hypothesis.

4.2. Perception decoding versus stimulus decoding

In contrast to the classifiers whose aim is to discriminate between the standard stimuli and the deviant ones (Blankertz et al., 2011), our model focuses on the subject's perception about the stimuli rather than the types of stimuli. As a matter of fact, the output of our model for each stimulus is a continuous number (regression model) which represents the subjective surprise corresponding to that stimulus; this is fundamentally different from the models whose output is a categorical variable (classification model) representing the type of the stimulus.

The results demonstrated that it is possible to reasonably decode the surprise of every single stimulus, independent from its physical features. Due to the dynamics in the sequence of stimuli, there can be some standard samples which may have more subjective surprise than some deviant samples. This phenomenon, which is called “local effect” by (Meyniel et al., 2016), suggests that the brain's perception of standards and deviants is not a discrete and fixed perception. In other words, by assigning an occurrence probability to each type of stimulus at each time, the type of stimulus which is thought to be deviant can be changed over time, i.e. there can be low-surprise deviants as well as high-surprise standards. The results demonstrated our decoding model could also decode the surprise of the low-surprise deviants and the high-surprise standards as precisely as for the other stimuli.

The difference between the decoding model which we proposed and the mentioned classification models suggests that our approach may be useful for improving the performance of BCI systems, like the P300 Speller (Guger et al., 2009).

In contrast to some previous works, such as (Mars et al., 2008; Meyniel et al., 2016; Squires et al., 1976), some other works, such as (Bonala et al., 2008) claimed that the local effect is coded in the likelihood of P300 generation and not in its amplitude. In this case, our results can be considered as an evidence in favor of the former and against the latter. In other words, since the likelihood of P300 generation is a statistical property for trials, it cannot be decoded in the single-trial analysis. Based on our results, even if the results of the mentioned previous works about the likelihood of the P300 generation and the surprise of stimuli are considered to be true, one cannot conclude that it is the only landmark of the surprise in the brain's response.

4.3. P300 versus MMN

The results of our blind search for the most informative features of evoked potentials were consistent with some of the earlier findings. We demonstrated that, for surprise decoding, the Early components of evoked potentials (0–100 ms) are not informative about the surprise of the stimuli at all; in this case, they are similar to baseline signals, which causally should not contain any information about the stimuli. We also observed that both the Middle (100 ms–250 ms) and the Late (250 ms–600 ms) components are significantly more informative than chance level – for both of the modalities. This is roughly equivalent to what earlier works have reported about these components: That the Early components are mostly functions of exogenous features of stimuli rather than the cognitive, or endogenous features (Luck, 2004; Sur and Sinha, 2009).

However, the more important part of our study was to examine

whether the Middle components are more important than or as important as the Late components in the decoding model. As it is depicted in Fig. 6, the most informative features of evoked potentials are within the interval between 100 ms and 400 ms. In addition, the hypothesis of indifference between the Late components and the Middle ones was accepted by employing Bayesian hypothesis testing approach. The results are in contrast with the general expectation that the Late components of evoked potentials, which contain P300 (Kolossa et al., 2013; Luck, 2004; Mars et al., 2008; Sur and Sinha, 2009), should be the most informative components.

Although there are debates about the underlying mechanisms of MMN generation (Garrido et al., 2009; Lieder et al., 2013), looking from the perspective of Predictive Coding, MMN is correlated with the brain's prediction error (Garrido et al., 2009) – which in our work could be considered as the surprise. Considering the fact that MMN usually happens between 100 ms and 250 ms (Garrido et al., 2009; Lieder et al., 2013; Näätänen, 2000), our results suggest that the components near MMN are as important as P300 for investigating the surprise coding mechanism of the brain.

4.4. Future studies

One of the fundamental limitations of our model is that it can be used only for the "binary" oddball tasks, and not the tasks with more than two types of stimuli. This limitation is inherited from the Transition Probability Model (proposed by Meyniel et al., (2016) in 2016) which was employed in our work as the model for the near-ideal observer. The challenge of extending this model to be applied to the tasks with more than two types of stimuli is the following: When there are more than two types of stimuli, in addition to the probabilities of transitions between all the types, the sensory distances of every two types are needed for modeling the brain's response. This necessary information is not considered in the work of Mars et al. in 2008 (Mars et al., 2008), which used 4 types of visual stimuli for its oddball tasks. To further clarify this point, consider a task in which there are three monotone auditory types of stimuli, 1 kHz, 1.2 kHz, and 2 kHz. The transition between the 1 kHz stimulus and the 1.2 kHz one cannot be considered the same as the transition between the 1 kHz stimulus and the 2 kHz one. As a matter of fact, extending the near-ideal observer model would make it dependent on the sensory domain – because while it should be possible to define a sensory distance for any given case, the challenge will be to propose a general model which can be used for any sensory domain. The only work that we know which investigates this issue is the work of Lieder et al. in 2013 (Lieder et al., 2013). However, although their method is much more complicated than similar works (Mars et al., 2008; Meyniel et al., 2016; Ostwald et al., 2012), it is concerned only with one very special task with monotone auditory stimuli, and could not be applied to other tasks even with a reasonable amount of modifications.

Another limitation of our work, as well as almost all of the earlier studies, is the rather simple design of the experiments. Specifically, in the datasets that we used, the sequences of stimuli were generated as white Bernoulli noise, i.e. each trial was independent from the previous ones, and it only could have two possible values. Similar tasks were also used in earlier works like (Kolossa et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Squires et al., 1976). Even in the works which have more complicated tasks like (Lieder et al., 2013; Ostwald et al., 2012), the tasks were identical for all of the runs and subjects. As far as we know, although there are a few studies that use different approaches for generating sequences of stimuli (like (Imada et al., 1993)), there are no model-based analysis reported on the effect of the procedure of sequence generation in oddball tasks. For further study, the level of complexity, and the dependency of the stimuli on each other can be investigated.

Several earlier studies have demonstrated that there are differences in the brain's responses of healthy people compared to the brain responses of Parkinson's (Cavanagh et al., 2018), Alzheimer's (Golob and Starr, 2000; Polich and Corey-Bloom, 2005), or Schizophrenia (Baldeweg et al., 2004) patients during oddball tasks. One of the ideas for future studies is to investigate the effects of these kinds of abnormalities on the performance of our decoding model, specifically for medical diagnosis, and to compare that with the performance of traditional methods of analysis. Furthermore, and as we mentioned earlier in sections 4.2, our decoding model may be also adapted to be used for improving BCI systems.

Conflicts of interest

None.

Acknowledgment

The authors wish to thank J. M. Walz and collaborators for their publicly available experimental data, which was originally collected for their study reported in (Walz et al., 2013). The authors wish to thank K. Robbins and collaborators for their publicly available experimental data, which was published as the data paper (Robbins et al., 2018). The authors also wish to thank A. Kübler and collaborators at the University of Würzburg, Germany, for their publicly available dataset on www.bnci-horizon-2020.eu/database/data-sets.

Appendix A. The Effect of Motor Response

One can raise a concern about the possibility that the asymmetric motor response to the deviant stimuli in the main dataset could be a source of additional discriminating features in the EEG signal, which might boost the performance of the decoder in our model. Similarly, when the benchmarking dataset is considered, in which the subject responds to both stimuli types by pressing different buttons, the confound effect by the motor response could still be a source of concern in the performance of the surprise decoder.

In order to address this problem, we use two approaches: 1. Employing the method of proof by contradiction, through examining two direct consequences of the possibility of the confound effect, and 2. Applying our decoding model on another dataset with no motor response effect.

A.1. Proof by Contradiction

If the motor response is to significantly boost the performance of our model, the corresponding increase in the decoding power of our model should be a function of both: 1) the occurrence probability of deviants, P_D, and 2) the complexity of the difference between the motor responses corresponding to different stimuli. The dependency of the decoding power to these two factors is explained below:

1) For small P_D (when deviants are much more surprising than standards), a small increase in P_D should cause an increase in the decoding power – since by increasing P_D, the number of surprising stimuli with a different motor response also increases.
2) When the difference between two motor responses is more complex, it is also more complicated to distinguish between them by decoding models. As a result, an increase in the level of complexity of the difference between two motor responses should cause a decrease in the decoding power.

We consider two situations which enable us to investigate these two consequences of the mentioned possibility of a confound effect:

1) For the main dataset P_D is equal to 0.2, and for the benchmarking dataset P_Ds are concentrated around $\frac{1}{8} = 0.125$. In addition, in order to decode motor responses, for the main dataset (with asymmetric motor response: button press only for the deviants), it is sufficient to discriminate between movement and no movement, while for the benchmarking dataset (with symmetric motor response: button presses for both stimuli types), it is necessary to discriminate between two different types of movement – which means that the difference in the motor responses for the benchmarking dataset is more complex than this difference for the main dataset.

Considering both of these differences between the main and the benchmarking dataset, if there is a confound effect between the surprise and motor responses which boosted our decoder's performance, then we should observe a significantly better performance by applying our model to the main dataset. This is because the main dataset enjoys both a slightly higher value for P_D and a lower complexity level, which both point to a higher decoding power for it. However, as it is mentioned in section 3.2, Table 1, Table 2, and Fig. 3, we were able to accept the null hypothesis when comparing the performance of the auditory task of the main dataset and the visual task of the benchmarking dataset. We were also not able to reject the null hypothesis when comparing the visual task of the main dataset and the visual task of the benchmarking dataset. These mean that the decoding powers of our model were not noticeably different between the two datasets.

2) For the benchmarking dataset, P_D is not the same for all subjects – it is in the range of $\frac{1}{8.2}$ to $\frac{1}{7.7}$. Therefore, in this case, it is possible to investigate the relation between the decoding power of our model for each subject, R^2, and P_D (the ratio of the number of deviants to the number of all stimuli) corresponding to that subject. As we mentioned earlier, if the confound effect between the surprise and motor responses could have boosted our decoder's performance, there should also be a significant positive correlation between R^2 and P_D. Using correlation test, we observed P-Value = 0.53 and Bayes-Factor = 2.1, which means that although we can neither accept nor reject the null hypothesis, our evidences are in favor of accepting it – for the details of hypothesis testing approach, look at section 3.2. This means that an increase in the value of P_D (between subjects) did not increase the decoding power of our model.

These two observations demonstrate that our results are not consistent with the possibility that motor responses are significantly boosting our decoder's performance. Therefore, we can indirectly conclude that the confound effect between the surprise and motor responses is not a critical issue for our model.

A.2. A Third Dataset with no Motor Response

The dataset used for this part is openly accessible on the website www.bnci-horizon-2020.eu/database/data-sets with the name “Auditory oddball during hypnosis (005–2014)” (Anon, n.d.).

Two adult subjects (one male and one female, both right-handed) participated in four separate auditory oddball tasks. Each task consisted of 420 standard and 60 deviant stimuli. The standard and deviant stimuli were respectively short complex high (440 + 880+1760 Hz) and low (247 + 494+988 Hz) tones – intensity was 70 dB. The duration of each stimulus was 50 ms, and stimulus-onset asynchrony was 900 ms. During the first two tasks, subjects were conscious, and during the last two tasks, they were hypnotized. In our analysis, only the first two tasks have been used. In one of the first two tasks, the stimuli were presented in the passive condition (i.e. subjects were asked to only listen to a sequence of tones), and in the other, the stimuli were presented in the active condition (i.e. subjects were asked to count the occurrences of the deviant stimuli).

EEG (27 channels) and EOG (horizontal and vertical) signals were recorded with the frequency of 512 Hz. For the processing of the EEG signals, the procedure which is explained in Section 2.2 was performed. In addition, the trials for which the amplitude of one of the EOG channels was more than 200 μV were excluded. The decoding model was applied to the data of each subject separately for the active and passive conditions.

The results are reported in Table A1. The average R-squared value for both subjects and both conditions is equal to 0.38 with a standard deviation of 0.16. Similar to what was explained in Section 3, the null distribution of R-squared was also computed: The average is equal to −0.0039, and the standard deviation is equal to 0.0062. Therefore, all of the four decoding powers are significantly more than the average of the null distribution.

In addition, we did not observe any significant difference between the results for the third dataset and the ones which we previously examined – the results are reported in Table A2. This can be because of the very small number of samples for the third dataset, but considering the Bayes Factors reported in Table A2, the evidence for accepting the indifference between the results for the third dataset on the one hand, and the auditory task of the main dataset as well as the visual task of the benchmarking dataset on the other hand, is very close to be significant. The small P-Value and Bayes Factor for testing the difference between our results for the third dataset and our results for the visual task of the main dataset can be considered similar to our results reported in Section 3.2: We cannot make a strong statement about the relation of the visual task of the main dataset with the other tasks.

The fact that our decoding model performed decently in the absence of the motor response, and for both the active and passive conditions, is an evidence against the possibility of a confound effect between the surprise and motor responses in our decoding model.

Table A.1
Decoding power (R-squared values computed by 5-fold cross validation) corresponding to the analysis of the third dataset – for both subjects and in both active (i.e. counting the deviant stimuli) and passive (i.e. only listening to the stimuli) conditions. The standard deviations are calculated over cross validation folds.

	Passive Condition	Active Condition (Counting)
Subject 1	0.26 ± 0.08	0.54 ± 0.03
Subject 2	0.49 ± 0.06	0.22 ± 0.08
Mean and Standard Deviation	0.38 ± 0.16	
Mean and Standard Deviation for the Null distribution	-0.0039 ± 0.0062	

Table A.2
The statistics of testing the differences between the average decoding power for the third dataset and the other datasets in which subjects had to press buttons.

	Auditory (Main Dataset)	Visual (Main Dataset)	Benchmarking Dataset
P-Value	0.84	0.16	0.95
Bayes Factor	2.71/1	1.25/1	2.76/1

References

Anon. n.d. "BNCI Horizon 2020." Retrieved (www.bnci-horizon-2020.eu/database/data-sets).

Baldeweg, Torsten, Klugman, Anthony, Gruzelier, John, Hirsch, Steven R., 2004. Mismatch negativity potentials and cognitive impairment in Schizophrenia. Schizophr. Res. 69 (2–3), 203–217.

Baldi, Pierre, 2002. A Computational Theory of Surprise. Ppp. 1–25. In: Information, Coding and Mathematics. Springer US, Boston, MA.

Barber, Rina Foygel, Candes, Emmanuel J., 2016. A Knockoff Filter for High-Dimensional Selective Inference. ArXiv Preprint ArXiv:1602.03574.

Barto, Andrew, Mirolli, Marco, Baldassarre, Gianluca, 2013. Novelty or surprise? Front. Psychol. 4 (DEC), 1–15.

Bell, Anthony J., Sejnowski, Terrence J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7 (6), 1129–1159.

Bigdely-Shamlo, Nima, Mullen, Tim, Kothe, Christian, Su, Kyung-Min, Robbins, Kay A., 2015. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. Front. Neuroinf. 9, 16.

Blankertz, Benjamin, Lemm, Steven, Treder, Matthias, Haufe, Stefan, Klaus-Robert Müller, 2011. Single-trial analysis and classification of ERP components — a tutorial. Neuroimage 56 (2), 814–825.

Bonala, Bharat, Boutros, Nashaat N., Jansen, Ben H., 2008. Target probability affects the likelihood that a P300 will Be generated in response to a target stimulus, but not its amplitude. Psychophysiology 45 (1) , 93–99.

Cavanagh, James F., Kumar, Praveen, Mueller, Andrea A., Richardson, Sarah Pirio, Mueen, Abdullah, 2018. Diminished EEG habituation to novel events effectively classifies Parkinson's patients. Clin. Neurophysiol. 129 (2), 409–418.

David Hairston, W., et al., 2014. Usability of four commercially-oriented EEG systems. J. Neural Eng. 11 (4), 046018.

Delorme, Arnaud, Makeig, Scott, 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134 (1), 9–21.

Efron, Bradley, 2010. Large-Scale Inference. Cambridge University Press, Cambridge.

Faraji, Mohammadjavad, Preuschoff, Kerstin, Gerstner, Wulfram, 2018. Balancing new against old information: the role of puzzlement surprise in learning. Neural Comput. 30 (1), 34–83.

Friston, Karl, 2005. A theory of cortical responses. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 360 (1456), 815–836.

Friston, Karl, 2009. The free-energy principle: a rough guide to the brain? Trends Cognit. Sci. 13 (7), 293–301.

Friston, Karl, 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11 (2), 127–138.

Friston, K., Kiebel, S., 2009. Predictive coding under the free-energy principle. Phil. Trans. Biol. Sci. 364 (1521), 1211–1221.

Garrido, Marta I., Kilner, James M., Stephan, Klaas E., Friston, Karl J., 2009. The Mismatch negativity: a review of underlying mechanisms. Clin. Neurophysiol. 120 (3), 453–463.

Golob, E., Starr, A., 2000. Effects of stimulus sequence on event-related potentials and reaction time during target detection in Alzheimer's disease. Clin. Neurophysiol. 111 (8), 1438–1449.

Guger, Christoph, et al., 2009. "How many people are able to control a P300-based brain–computer Interface (BCI)? Neurosci. Lett. 462 (1), 94–98.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, 2001. The Elements of Statistical Learning. Springer series in statistics, New York.

Huettel, Scott A., McCarthy, Gregory, 2004. What is odd in the oddball task?: prefrontal cortex is activated by dynamic changes in response strategy. Neuropsychologia 42 (3), 379–386.

Huettel, Scott A., Mack, Peter B., McCarthy, Gregory, 2002. Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. Nat. Neurosci. 5 (5), 485–490.

Imada, T., Hari, R., Loveless, N., McEvoy, L., Sams, M., 1993. Determinants of the auditory Mismatch response. Electroencephalogr. Clin. Neurophysiol. 87 (3), 144–153.

Itti, Laurent, Baldi, Pierre, 2009. Bayesian surprise attracts human attention. Vis. Res. 49 (10), 1295–1306.

Jung, T.P., et al., 2000. Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. Clin. Neurophysiol.: Official Journal of the International Federation of Clinical Neurophysiology 111 (10), 1745–1758.

Jung, T.P., et al., 2001. Analysis and visualization of single-trial event-related potentials. Hum. Brain Mapp. 14 (3), 166–185.

Knill, David C., Pouget, Alexandre, 2004. The bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 27 (12), 712–719.

Kolossa, Antonio, Fingscheidt, Tim, Wessel, Karl, Kopp, Bruno, 2013. A model-based approach to trial-by-trial P300 amplitude fluctuations. Front. Hum. Neurosci. 6.

Liang, Feng, Paulo, Rui, Molina, German, Clyde, Merlise A., Berger, Jim O., 2008. Mixtures of g priors for bayesian variable selection. J. Am. Stat. Assoc. 103 (481), 410–423.

Lieder, Falk, Jean, Daunizeau, Garrido, Marta I., Friston, Karl J., Stephan, Klaas E., 2013. "Modelling trial-by-trial changes in the Mismatch negativity" edited by O. Sporns. PLoS Comput. Biol. 9 (2), e1002911.

Luck, Steven J., 2004. An Introduction to the Event-Related Potential Technique. MIT press.

Mars, Rogier B., et al., 2008. Trial-by-Trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. J. Neurosci. 28 (47), 12539–12545.

Meyniel, Florent, Maheu, Maxime, Dehaene, Stanislas, 2016. "Human inferences about sequences: a minimal transition probability model" edited by S. J. Gershman. PLoS Comput. Biol. 12 (12), e1005260.

Näätänen, Risto, 2000. Mismatch negativity (MMN): perspectives for application. Int. J. Psychophysiol. 37 (1), 3–10.

Oppenheim, Alan V., Schafer, Ronald W., 1975. Digital Signal Processing. Prentice Hall.

Ostwald, Dirk, et al., 2012. Evidence for neural encoding of bayesian surprise in human somatosensation. Neuroimage 62 (1), 177–188.

Polich, John, Corey-Bloom, Jody, 2005. Alzheimers disease and P300: review and evaluation of task and modality. Curr. Alzheimer Res. 2 (5), 515–525.

Robbins, Kay, Su, Kyung min, David Hairston, W., 2018. An 18-subject EEG data collection using a visual-oddball task, designed for benchmarking algorithms and headset performance comparisons. Data in Brief 16, 227–230.

Rouder, Jeffrey N., Morey, Richard D., 2012. Default Bayes factors for model selection in regression. Multivariate Behav. Res. 47 (6), 877–903.

Rouder, Jeffrey N., Speckman, Paul L., Sun, Dongchu, Morey, Richard D., Iverson, Geoffrey, 2009. Bayesian t tests for accepting and rejecting the null hypothesis. Psychon. Bull. Rev. 16 (2), 225–237.

Rubin, Jonathan, Ulanovsky, Nachum, Nelken, Israel, Tishby, Naftali, 2016. "The representation of prediction error in auditory cortex" edited by F. E. Theunissen. PLoS Comput. Biol. 12 (8), e1005058.

Squires, K., Wickens, C., Squires, N., Emanuel, Donchin, 1976. The effect of stimulus sequence on the waveform of the cortical event-related potential. Science 193 (4258), 1142–1146.

Su, Kyung-min, David Hairston, W., Robbins, Kay, 2018. EEG-annotate: automated identification and labeling of events in continuous signals with applications to EEG. J. Neurosci. Methods 293, 359–374.

Sur, Shravani, Sinha, V.K., 2009. Event-related potential: an overview. Ind. Psychiatry J. 18 (1), 70.

Symonds, Renée M., et al., 2017. Distinguishing neural adaptation and predictive coding hypotheses in auditory change detection. Brain Topogr. 30 (1), 136–148.

Tsolaki, Anthoula, Kosmidou, Vasiliki, Hadjileontiadis, Leontios, Kompatsiaris, Ioannis (Yiannis), Tsolaki, Magda, 2015. Brain source localization of MMN, P300 and N400: aging and gender differences. Brain Res. 1603, 32–49.

van Dinteren, Rik, Arns, Martijn, Marijtje, L., Jongsma, A., Kessels, Roy.P.C., 2014. "P300 development across the lifespan: a systematic review and meta-analysis" edited by F. Di Russo. PLoS One 9 (2), e87347.

Walz, J.M., et al., 2013. Simultaneous EEG-FMRI reveals temporal evolution of coupling between supramodal cortical attention networks and the brainstem. J. Neurosci. 33 (49), 19212–19222.

Walz, Jennifer M., et al., 2014. Simultaneous EEG–fMRI reveals a temporal cascade of task-related and default-mode activations during a simple target detection task. Neuroimage 102 (P1), 229–239.

Walz, Jennifer M., et al., 2015. Prestimulus EEG alpha oscillations modulate task-related FMRI BOLD responses to auditory stimuli. Neuroimage 113, 153–163.

Weinstein, Asaf, Barber, Rina, Candes, Emmanuel, 2017. A Power and Prediction Analysis for Knockoffs with Lasso Statistics. 1712.06465.

Winkler, Irene, Haufe, Stefan, Tangermann, Michael, 2011. Automatic classification of artifactual ICA-components for artifact removal in EEG signals. Behav. Brain Funct. 7 (1), 30.