NeuroImage 62 (2012) 177–188

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Evidence for neural encoding of Bayesian surprise in human somatosensation

Dirk Ostwald [a,b,*], Bernhard Spitzer [a], Matthias Guggenmos [a], Timo T. Schmidt [a], Stefan J. Kiebel [a,c], Felix Blankenburg [d]

[a] *Department of Neurology and Bernstein Center for Computational Neuroscience, Charité Universitätsmedizin Berlin, Germany*
[b] *School of Psychology, University of Birmingham, UK*
[c] *Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*
[d] *Dahlem Institute for Neuroimaging of Emotion, Freie Universität Berlin, Berlin, Germany*

ARTICLE INFO

Article history:
Accepted 26 April 2012
Available online 3 May 2012

Keywords:
Bayesian brain hypothesis
Somatosensory mismatch response
EEG single trial modeling
Computational neuroimaging

ABSTRACT

Accumulating empirical evidence suggests a role of Bayesian inference and learning for shaping neural responses in auditory and visual perception. However, its relevance for somatosensory processing is unclear. In the present study we test the hypothesis that cortical somatosensory processing exhibits dynamics that are consistent with Bayesian accounts of brain function. Specifically, we investigate the cortical encoding of Bayesian surprise, a recently proposed marker of Bayesian perceptual learning, using EEG data recorded from 15 subjects. Capitalizing on a somatosensory mismatch roving paradigm, we performed computational single-trial modeling of evoked somatosensory potentials for the entire peri-stimulus time period in source space. By means of Bayesian model selection, we find that, at 140 ms post-stimulus onset, secondary somatosensory cortex represents Bayesian surprise rather than stimulus change, which is the conventional marker of EEG mismatch responses. In contrast, at 250 ms, right inferior frontal cortex indexes stimulus change. Finally, at 360 ms, our analyses indicate additional perceptual learning attributable to medial cingulate cortex. In summary, the present study provides novel evidence for anatomical-temporal/functional segregation in human somatosensory processing that is consistent with the Bayesian brain hypothesis.

Introduction

The Bayesian brain hypothesis postulates that the brain uses probabilistic inference for perception and perceptual learning (Doya et al., 2007). These mechanisms can be implemented using Bayesian inference based on an internal generative model, which comprises a distribution over sensory data given an external cause (the sensory data likelihood) and a prior distribution over different causes (Friston, 2010; Knill and Pouget, 2004). Perception is modeled as the process of computing a posterior distribution over causes using the generative model and sensory input, while perceptual learning is explained as the updating of the brain's representation of the prior distribution based on the inferred posterior distribution over causes (Friston, 2003; Kersten et al., 2004).

It has been suggested that these Bayesian mechanisms are encoded by neuronal populations whose responses to novel sensory input are interpreted as dynamics induced by the violation of prior expectations (Mumford, 1992; Rao and Ballard, 1999; Strange et al., 2005). Typical neurobiological markers of this violation are EEG novelty responses such as the auditory mismatch negativity (aMMN) (Näätänen et al., 2011), the P300 (Polich, 2007), or the dishabituation of laser evoked-potentials (Mouraux and Iannetti, 2008; Wang et al., 2010). In the framework of the Bayesian brain hypothesis, one way to formally quantify the novelty of sensory input is Bayesian surprise, a recently proposed information theoretic quantity (Baldi and Itti, 2010; Itti and Baldi, 2009). Bayesian surprise quantifies the effect sensory input has on the internal generative model as the divergence between the encoded prior and posterior distribution over causes. Representing Bayesian surprise may enable an observer like the brain to efficiently and dynamically encode the statistical (ir)regularities of its environment.

While empirical evidence suggests a role of Bayesian perceptual learning for shaping neural responses in auditory and visual perception (Garrido et al., 2009a, 2009b, 2009c; Harrison et al., 2007; Rao and Ballard, 1999), its relevance for somatosensory processing is unclear. Here, we address the hypothesis that somatosensory processing, as assessed with somatosensory mismatch responses (sMMRs) in EEG, exhibits dynamics that are consistent with Bayesian theories of perceptual learning and specifically, the encoding of Bayesian surprise.

Although less studied than the aMMN, a number of investigations have previously described novelty or mismatch responses for somatosensory evoked potentials (SEPs) (Näätänen, 2009). Consistently, a fronto-parietal negative shift between 100 and 200 ms contralateral to the side of stimulation has been observed for unexpected stimuli (Akatsuka et al., 2007b; Kekoni et al., 1997; Kida et al., 2004;

* Corresponding author at: Bernstein Center for Computational Neuroscience, Charité Universitätsmedizin Berlin, Philippstr. 13 Haus 6, 10115 Berlin, Germany. Fax: +49 30 2093 6771.
E-mail addresses: dirk.ostwald@bccn-berlin.de, dirk.ostwald@mpib-berlin.mpg.de (D. Ostwald).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2012.04.050

Restuccia et al., 2007; Shinozaki et al., 1998; Spackman et al., 2007, 2010), while some studies additionally reported earlier mismatch responses 60 to 90 ms after stimulus onset (Akatsuka et al., 2007a, 2007b; Götz et al., 2011). Further, in analogy to other sensory modalities, somatosensory oddball stimuli that capture the observer's attention elicit a parietal positive response at about 300 ms post-stimulus, commonly referred to as P300 (Restuccia et al., 2009; Tarkka et al., 1996). While the precise neural generation mechanism of the sMMR remains to be established, it provides a unique experimental tool to investigate perceptual learning in the somatosensory system from a Bayesian perspective.

We investigate our hypothesis by capitalizing on recently developed model-based analyses of single-trial EEG or fMRI activity (Friston and Dolan, 2010; Mars et al., 2008, 2010). To derive single-trial estimates of Bayesian surprise we employ a sequential Bayesian stimulus probability learning algorithm and, to account for the assumption that the brain uses finite time-windows to dynamically update its generative model, we employ a forgetting mechanism in the learning scheme (Harrison et al., 2011; Kiebel et al., 2008a). Using EEG data from 15 subjects, we perform computational modeling of evoked somatosensory potentials on the source level and for the entire peri-stimulus time period. By means of Bayesian model selection, we identify both cortical substrates and critical time windows of ongoing Bayesian surprise encoding in the cortical-temporal hierarchy of somatosensory processing.

Materials and methods

Participants

Fifteen healthy volunteers (21–31 years, six females) participated in the experiment after providing written informed consent. The study was approved by the Ethical Committee of the Charité University Hospital Berlin and corresponded to the Human Subjects Guidelines of the Declaration of Helsinki.

Stimuli

Electrical stimuli of 0.2 ms duration were delivered to the left median nerve via adhesive electrodes attached to the wrist. Two intensity levels (low/high stimulus amplitude) were adjusted on an individual subject basis to account for subject specific sensory thresholds. The low stimulus intensity (mean 4.0 ± 1.6 STD mA) was determined to be close to detection threshold but clearly noticeable for every stimulus replication. The high stimulus intensity (6.0 ± 2.2 STD mA) was chosen to be markedly distinguishable from the low stimulus intensity, but not painful and below the motor threshold.

Experimental procedure

Upon familiarization with the experimental stimulation, participants underwent nine to ten experimental runs of an oddball-like roving paradigm (Baldeweg et al., 2004): stimuli were delivered in consecutive trains of alternating stimulus intensity with a constant inter-stimulus interval of 650 ms (Fig. 1). In contrast to classical oddball paradigms, which comprise the repeated presentation of standard stimuli occasionally interrupted by the presentation of physically different deviant stimuli (Näätänen et al., 1978), in roving paradigms stimuli with different physical properties can take on the role of both deviant (oddball) and standard stimulus. By averaging over deviant and standard potentials evoked by physically different stimuli, this allows to discount differential responses to the physical stimulus per se in observed mismatch effects.

The length of each stimulus train was chosen at random from the set {2,4,8,16}, using equal probabilities. The participants were instructed to count the number of stimulus trains per experimental run, i.e., to attend to the changes from low to high and high to low amplitude. Thereby, the applied paradigm differed from classical mismatch tasks in so far that participants directed their attention to the stimuli. To render the counting task nontrivial, the number of stimulus trains in each run was sampled at random from a normal distribution with expectation 72 and standard deviation 5. Consequently, ca. 72 stimulus trains corresponded to roughly 500 stimuli delivered per run, and to about 5000 stimuli per subject in total. As the first stimulus in a stimulus train (high or low stimulus amplitude) is, by definition, a deviant, approximately 720 deviant responses were recorded per subject. After each experimental run, the subjects reported the number of experimental trains and were informed about the correct outcome.

EEG recording and pre-processing

EEG data were recorded using a 64-channel active electrode system at a sampling rate of 2048 Hz (ActiveTwo, BioSemi), with electrodes placed in an elastic cap according to the extended 10–20 system. Individual electrode locations were registered with respect to three fiducial markers (left and right preauricular points and nasion) using an electrode positioning system (Zebris Medical) to improve subsequent source space analyses. All further data preprocessing steps were performed using Statistical Parametric Mapping (SPM8) (Litvak et al., 2011). Specifically, the data were downsampled to a sampling rate of 512 Hz, referenced against average reference, band-pass filtered (1 to 40 Hz) and corrected for eyemovements using a topological confound approach originally developed by Berg and Scherg (1994) and implemented in SPM8 (Litvak et al., 2007). The data were epoched using a peri-stimulus time interval of −100 ms to 600 ms. Trials containing amplitudes larger than 150 μV were excluded from further analysis. SEPs for experimental conditions of interest were computed using standard averaging and were averaged across subjects to yield grand mean SEPs. The experimental conditions of interest were 1) the somatosensory evoked potential averaged over all stimuli, abbreviated ‘SEP’, 2) the response to deviant stimuli, averaged across high and low intensity stimuli, abbreviated ‘Deviant’, 3) the response to stimuli immediately preceding deviant stimuli, averaged across high and low intensity stimuli, abbreviated ‘Standard’. It should be noted that this nomenclature differs

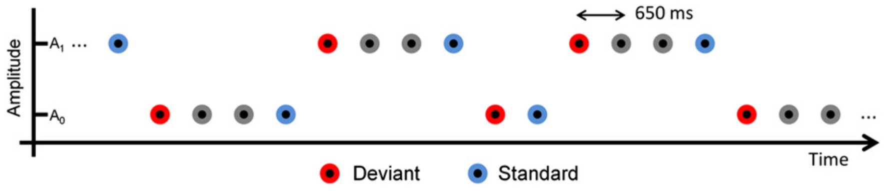

Fig. 1. Experimental paradigm. Electrical stimuli of two amplitudes, high (A_1) and low (A_0), were delivered to the median nerve with an inter-stimulus interval of 650 ms. Trains of identical stimuli, i.e. of either high or low amplitude, comprised 2, 4, 8 or 16 stimuli. For clarity only the cases of trains of 2 and 4 stimuli are shown in the figure. The first stimulus in each train of identical stimuli was labeled a deviant. To compare deviant (red) and standard (blue) responses based on the same number of trials, only those stimuli immediately preceding a deviant stimulus were labeled standard. This experimental paradigm is an adaptation of a previously established roving paradigm for the somatosensory domain (Baldeweg et al., 2004).

from standard mismatch negativity paradigms, which usually define as ‘Standard’ stimuli all non-deviant stimuli. However, defining the ‘Standard’ condition in this manner allows estimating both ‘Deviant’ and ‘Standard’ conditions from the same number of trials. Further, by collapsing the ‘Deviant’ and ‘Standard’ conditions over both stimulus amplitudes allows discounting differential responses to the physical properties of the stimulus per se.

Anatomical network identification

As we aimed to perform the model-based trial-by-trial analyses based on anatomically specific substrates (rather than on the electrode level), a combination of previous findings on the generators of the SEP as well as source reconstructions of the present data set was used to identify a set of six brain regions involved in the generation of the trial-by-trial EEG response. The anatomical localization of these sources enabled us to project the trial-by-trial electrode space EEG data onto a set of oriented equivalent current dipoles (ECDs) placed at corresponding locations as described below. Previous research established that the cortical SEP comprises the succession of (at least) three EEG components reflecting different processing stages of somatosensory information (Niedermeyer and Silva, 2004; Thees et al., 2003): 1) the parietal N20 reflecting a dipolar generator in primary somatosensory cortex (S1) situated in the posterior bank of the rolandic fissure, 2) the fronto-central P45/N60 complex of unclear origin, and 3) the parieto-temporal N140 peaking in the 60–160 ms latency range, presumably reflecting additional stimulus processing in secondary somatosensory cortex (S2). The localization of contralateral S1 and bilateral S2 for the current data set is hence obligatory. Moreover, experimental evidence suggests the involvement of frontal regions in the generation of mismatch responses, e.g., for the aMMN (Näätänen et al., 2011; Rinne et al., 2000, 2005, 2006; Tse and Penney, 2008). We thus were also interested in monitoring trial-by-trial activity in frontal cortex, the exact anatomical location being derived from source reconstruction of the deviant response of the present data set. Finally, cingulate cortex has been repeatedly implicated in the generation of oddball as well as P300 responses across a wide range of experimental paradigms (Linden, 2005; Thees et al., 2003), motivating the inclusion of a cingulate source.

In summary, strong anatomical–functional evidence suggests the inclusion of contralateral primary and bilateral secondary somatosensory, bilateral inferior frontal, and cingulate cortex in a six-dipole model of single-trial EEG responses for the current paradigm. To identify the precise anatomical substrates in the current study in a data-informed manner, we employed a two-stage procedure to obtain MNI coordinates and moment vectors for a multiple dipole model: 1) Using grand mean event-related potentials (ERPs) a distributed source localization of the regions suggested by the literature was performed to obtain MNI coordinates, and 2) the moments of ECDs located at these MNI coordinates were fitted, again based on grand mean ERPs. The distributed source localization and dipole orientation fitting procedures are detailed below. Both the condition-specific grand mean ERPs and time windows used in these analyses were chosen to be consistent with previous findings on the neural generators of somatosensory and mismatch ERPs as discussed in the section above and are documented in detail in Tables 1 and 2. The final set of six location and six moment vectors was subsequently used to project the data of each trial, for each subject, to anatomical brain space.

Statistical distributed source localization

At the first stage, the sources of the evoked EEG activity were determined using the distributed source reconstruction algorithm as implemented in SPM8. A forward model was constructed for each subject using a 8196 vertex template cortical mesh co-registered at the individual electrode positions via three fiducial markers. The lead field for the forward model was generated using the three-shell BEM EEG head model as provided by SPM8. Source estimates were computed on the canonical cortical mesh using multiple sparse priors (Friston et al., 2008) under group constraints (Litvak and Friston, 2008). Source power increases were statistically analyzed at the group level using one-sample t-tests. For display purposes statistical parametric maps were thresholded at $p<0.001$ (uncorrected), and random field theory was used to control for family-wise error in source space (Kiebel and Friston, 2004; Worsley, 1994). Finally, the SPM Anatomy toolbox was employed to establish cytoarchitectonic references (Eickhoff et al., 2005). This procedure enabled the reliable identification of four of the six sources (Table 1). The exact procedure used for obtaining these four sources and the remaining two sources to form a six-dipole model is described below under results.

Dipole fitting and timecourse extraction

At the second stage, the six obtained MNI source locations were used to fit ECDs to project the evoked electrode data into source space. To this end, grand mean group evoked potentials were subjected to the Variational Bayes — Equivalent Current Dipole (VB-ECD) algorithm implemented in SPM8 (Kiebel et al., 2008b). For each source, either a single dipole or a symmetric dipole pair was chosen with tight location priors centered on the sets of coordinates obtained from the distributed source analysis, unless otherwise noted. The moment parameters of the respective dipolar sources were then optimized using VB-ECD, and the posterior moment expectations normalized to a Euclidean norm of 1. The dipole specific SEPs, prior locations and moment expectations, as well as the peri-stimulus time-points analyzed are listed in Table 2. Finally, the trial-by-trial event-related electrode space data were projected onto the set of fixed and oriented ECDs using SPM8's spm_eeg_extract_waveforms.m function (Litvak et al., 2011).

Functional model of evoked source activity

To relate single-trial source activity to Bayesian principles of perceptual learning, we computed Bayesian surprise for each single trial using a sequential (online) Bayesian learning algorithm of stimulus probabilities (Bishop, 2007) (pp. 68–78). Briefly, the model assumes that the brain implements a trial-by-trial Bayesian parameter learning scheme starting from an uninformative prior and computes Bayesian surprise as the divergence between the parameter prior and posterior probability density functions (PDF) at the single-trial level. Moreover, variants of this model assume that the brain only incorporates observed trials which lie in a variable time-window of the close past into its estimation of the parameter PDF, where the length of the time-window is governed by an exponential forgetting (i.e. relative down-weighting) of stimulus observations in the distant past.

Formally, the model employed here assumes that the probability of observing a low ($S=0$) or high ($S=1$) intensity stimulus on the n-th trial is described by a Bernoulli distribution with expectation $\mu\in[0,1]$

$$p(S)=\mu^{S}(1-\mu)^{1-S}. \tag{1}$$

Here μ is the probability of observing a stimulus of high intensity on the n-th trial. To model the initial uncertainty about the parameter μ the model assumes a uniform beta prior distribution over μ, which, on each trial, is sequentially updated according to the observed data likelihood to form a posterior distribution over μ. The posterior distribution over μ after observing l stimuli of low intensity and m stimuli of high intensity, i.e. after $l+m$ trials in total, is given by

$$p(\mu|m,l)=\frac{\Gamma(m+l)}{\Gamma(m)\Gamma(l)}\mu^{m}(1-\mu)^{l} \tag{2}$$

Table 1
Distributed source reconstruction statistics.

Label	ERP	PST window	p-Cluster (FWE)	Z-value	p-Peak (FWE)	MNI coordinates			Cytotechtonic reference
S1	SEP	18–25 ms	<0.001	5.87	<0.001	48	−30	50	Right postcentral gyrus
				5.78	<0.001	36	−30	64	Area 3b (90% [40–100%])
				5.76	<0.001	44	−32	60	Area 1 (50% [40–80%])
						42	**−31**	**58**	Area 4a (20% [0–20%])
rS2	SEP	30–160 ms	0.001	4.14	0.074	**62**	**−34**	**12**	Right superior temporal gyrus
									IPC (PFcm): 30% [10–40%]
									IPC (PF): 20% [0–50%]
									(Activation extending into OP1 (60% [20–60%]))
lS2			0.117	3.92	0.151	**−60**	**−44**	**12**	Left superior temporal gyrus
rIFG	Deviant	340–360 ms	0.001	3.91	0.165	48	16	12	Right inferior frontal gyrus
				3.82	0.214	36	24	8	(pars triangularis)
				3.81	0.220	40	30	2	Area 45 40% [20–50%]
						41	**23**	**7**	

Column 1: Based on previous findings in the literature, the group distributed source reconstruction method implemented in SPM8 (Litvak and Friston, 2008) was used to determine the MNI coordinates of four sources of interest (S1, rS2, lS2, rIFG). Column 2: The statistical evaluation was based on reconstructing the source activity of the 'SEP' and the 'Deviant' waveform. Column 3: Peri-stimulus times of interest. These were derived from the grand mean, see Fig. 3 and text. Column 4: Corrected p-values obtained by using one-sample t-tests and family-wise error (FWE) correction at the cluster level (Worsley, 1994). Columns 5–7: Up to three peak MNI coordinates more than 8 mm apart, their corresponding Z and p-values at the voxel level. For multiple peak clusters, the arithmetic mean of the peak coordinates was used as the source MNI coordinate, set in bold in column 7. These coordinates were entered into SPM8's Anatomy toolbox (Eickhoff et al., 2005) to obtain an anatomical and probabilistic cytotechtonic reference, as reported in column 8.

where $\Gamma : \mathbb{R} \rightarrow \mathbb{R}$ denotes the gamma function. Corresponding to the sequential Bayesian learning approach, this distribution then acts as the prior distribution for inference on the $(l+m+1)$-th trial.

In order to implement a forgetting kinetic, instead of using the accumulative stimulus counts l_n and m_n on the $l_n + m_n = n$-th trial, the model employed here weights the stimulus counts with an exponential function over trials, such that

$$l_{w_n} = \sum_{i=0}^{n} \exp\left(-\frac{1}{\tau}(n-i)\right) l_i \tag{3}$$

and

$$m_{w_n} = \sum_{i=0}^{n} \exp\left(-\frac{1}{\tau}(n-i)\right) m_i \tag{4}$$

yield the weighted stimulus counts m_{w_n}, l_{w_n} at the n-th trial and $\tau > 0$ denotes the time constant of the forgetting kinetic. This results in a modulated posterior distribution on the n-th trial/prior distribution on the $(n+1)$-th trial given by

$$p\left(\mu | m_{w_n}, l_{w_n}\right) = \frac{\Gamma\left(m_{w_n} + l_{w_n}\right)}{\Gamma\left(m_{w_n}\right)\Gamma\left(l_{w_n}\right)} \mu^{m_{w_n}} (1-\mu)^{l_{w_n}}. \tag{5}$$

With respect to the Bayesian brain hypothesis, the product of the sensory data likelihood (1) and prior distribution (5) form the internal generative model over the external cause μ at the $(n+1)$-th trial, the formation of the posterior distribution (5) corresponds to perception on the n-th trial, and the iterative exchange of prior and posterior distributions based on weighted stimulus counts corresponds to perceptual learning with forgetting.

Finally, the model quantifies its degree of perceptual learning on the n-th trial as Bayesian surprise, i.e. the Kullback–Leibler divergence

Table 2
Equivalent current dipoles.

Label	ERP	PST point	Location prior exp.			Moment prior exp.			Location posterior exp.			Moment posterior exp.		
S1	SEP	21 ms	42	−31	58	0	0	0	42	−31	58	−0.0037	0.5743	0.8186
			(Informative)			(Uninformative)								
rS2	SEP	140 ms	62	−34	12	0	0	0	62	−34	12	−0.8890	0.4335	0.1766
			(Informative symmetric pair)			(Uninformative)								
lS2			−62	−34	12	0	0	0	−62	−34	12	0.8279	0.5174	0.2162
			(Informative symmetric pair)			(Uninformative)								
rIFG	Deviant	351 ms	41	23	7	0	0	0	41	23	7	0.0832	0.0184	0.9964
			(Informative symmetric pair)			(Uninformative)								
lIFG			−41	23	7	0	0	0	−41	23	7	0.0886	0.0184	0.9959
			(Informative symmetric pair)			(Uninformative)								
MC	Deviant–standard	351 ms	0	0	0	0	0	0	−1	−21	36	0.0253	0.3975	0.9172
			(Uninformative)			(Uninformative)								

After MNI coordinates for the anatomical regions of interest had been determined (Table 1), the normalized ECD moments were obtained by using SPM8's VB-ECD method with informative prior location expectation and uninformative prior moment expectation (Kiebel et al., 2008b). Column 1: Sources of interest with medial cingulate cortex (MC). Columns 2 and 3: The event-related potentials (ERP) and peri-stimulus time points (PST point) analyzed for each source. For the bilateral S2 and the IFG sources, a coupled symmetric ECD pair was fitted, where the location prior expectation was set to the MNI coordinates of the right hemisphere sources and their homologue coordinates. Columns 6 and 7: The posterior means of the location and moments (normalized). For the MC source, a single ECD model with uninformative location and moment prior was used, i.e. the location of the MC source is based solely on the VB-ECD solution.

between the prior and posterior distribution over μ at trial n (Baldi and Itti, 2010; Cover and Thomas, 1991; Itti and Baldi, 2009).

$$\begin{aligned}\text{Bayesian Surprise} &:= \mathrm{KL}\left(p\left(\mu|m_{w_{n-1}}, l_{w_{n-1}}\right)||p\left(\mu|m_{w_n}, l_{w_n}\right)\right)\\ &= \int p\left(\mu|m_{w_{n-1}}, l_{w_{n-1}}\right)\ln\left(\frac{p\left(\mu|m_{w_{n-1}}, l_{w_{n-1}}\right)}{p\left(\mu|m_{w_n}, l_{w_n}\right)}\right)d\mu \end{aligned} \tag{6}$$

Due to the use of conjugate priors, the Kullback–Leibler divergence can be evaluated analytically as a function of the parameters $m_{w_{n-1}}, l_{w_{n-1}}$ and m_{w_n}, l_{w_n} which significantly simplifies the integration over μ (Penny, 2001).

We used this approach to generate subject- and session-specific trial-by-trial Bayesian surprise sequences using the output of Eq. (6) for each single trial. These sequences were used as regressor/predictor variables in the trial-by-trial EEG data analysis, see next section. To implement varying degrees of forgetting, we varied the time-window of temporal stimulus integration from long to short, (Eqs. (3) and (4)), yielding a set of five Bayesian surprise models with different time constants τ, abbreviated BS0, BS1, BS2, BS3 and BS4, where we chose $\tau_0 = \infty, \tau_1 = 8, \tau_2 = 4, \tau_3 = 2.\bar{6}$ and $\tau_4 = 2$. This choice of time constants was motivated by sampling a wide range of possible temporal integration windows, while simultaneously maximizing the differences between the resulting regressors. Note that larger time-constants τ correspond to longer time windows. We show an example for one of the Bayesian surprise regressors ($\tau_2 = 4$) in Fig. 2. In Table 3, we translate the five time-constants to the weighting of the trial history.

Additionally, we generated three control regression models that implement less complex but widely used conventional hypotheses about the functional origin of trial-by-trial source amplitude variations: 1) a regressor indexing deviant stimuli with 1's and standard stimuli with 0's ('stimulus change model', model SC), 2) a parametric model implementing a linear relationship between the expression of the evoked source activity and the number of standards preceding a deviant stimulus ('linearly modulated stimulus change model', model LIN), and 3) a constant null regressor comprising a vector of 1's (model M0). Model M0 was used as a baseline model to compare all other models against. Importantly, note that model SC is the standard model for the analysis of (auditory) mismatch negativity studies. For the roving paradigm, model LIN has also been used in modified form in Baldeweg et al. (2004).

Functional model evaluation using parametric empirical Bayes

Each of the functional models provides a single, stimulus sequence-specific regressor. We used a parametric empirical Bayes (PEB) approach as implemented in SPM8's spm_PEB.m function (Friston et al., 2002a, 2002b, 2007) to fit the models and to compute the corresponding Bayesian model evidences for subsequent model selection (Gelman et al., 1995; Penny, 2012). The Bayesian model evidence framework enables the formal statistical comparison of computational models by accounting for the accuracy-complexity trade-off in explaining experimental data and constitutes a well-established approach in statistics (Hoeting et al., 1999), computational psychology (Pitt and Myung, 2002), and neuroimaging (Woolrich, in press).

To this end, the single subject, single session, single peri-stimulus time bin data for $n \in \mathbb{N}$ trials was modeled according to the two-level linear model

$$p(\bar{y}|\lambda) = N(\bar{y}; \bar{X}\bar{\theta}, C(\lambda) + C_\theta) \tag{7}$$

i.e., the probability distribution of the augmented data $\bar{y} \in \mathbb{R}^{n+2}$ was assumed to be multivariate normal with expectation $\mu = \bar{X}\bar{\theta} \in \mathbb{R}^{n+2}$ and parameterized covariance $\Sigma = (C(\lambda) + C_\theta) \in \mathbb{R}^{n+2 \times n+2}, (C(\lambda) + C_\theta) \succ 0$, where $\lambda \in \mathbb{R}^2$ refers to the covariance constraint weighting coefficients, the free parameters of the hierarchical linear model. Following the notation in Friston et al. (2002a, 2002b, 2007), the augmented data is given by

$$\bar{y} = \begin{pmatrix} y \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^{n+2} \tag{8}$$

where $y \in \mathbb{R}^n$ refers to the single subject, single session (comprising $n \in \mathbb{N}$ trials), single peri-stimulus time bin data, which, following standard approaches to multiple linear regression, was normalized to a mean of zero and a variance of 1 (z-score normalization). The two additional zeros encode the expectation of the second level error and the prior of the second level parameters. The augmented design matrix is given by

$$\bar{X} = \begin{pmatrix} X^{(1)} & X^{(1)}X^{(2)} \\ & I_2 \end{pmatrix} \in \mathbb{R}^{n \times 2} \tag{9}$$

where $X^{(1)} \in \mathbb{R}^{n \times 1}$ denotes the BS0–BS4/SC/LIN/M0-model specific regressor normalized to a mean of zero and a l_2-norm of 1 (except for the regressor of model M0, which was not normalized), $X^{(2)} = 0 \in \mathbb{R}$ is the second level design matrix allowing single level Bayesian inference with priors on the parameters, and $I_2 \in \mathbb{R}^{2 \times 2}$ is the identity matrix.

Further,

$$\bar{\theta} = \begin{pmatrix} \varepsilon^{(2)} \\ \theta^{(2)} \end{pmatrix} \in \mathbb{R}^2 \tag{10}$$

is a vector of latent variables corresponding to the second level error and linear parameter obtained by substitution of the hierarchical form of the model

$$y = X^{(1)}\theta^{(1)} + \varepsilon^{(1)} \tag{11}$$

$$\theta^{(1)} = X^{(2)}\theta^{(2)} + \varepsilon^{(2)}. \tag{12}$$

The parameterized covariance of the model $C(\lambda)$ is given by

$$C(\lambda) = \textstyle\sum_{i=1}^{2} \lambda_i Q_i \tag{13}$$

where

$$Q_1 = \begin{pmatrix} I_n & 0 \\ 0 & 0_2 \end{pmatrix} \in \mathbb{R}^{n+2 \times n+2} \tag{14}$$

with the identity matrix $I_n \in \mathbb{R}^{n \times n}$ and the zero matrix $0_2 \in \mathbb{R}^{2 \times 2}$ embeds the independence assumption over trials (justified by the separation of neighboring trials by 650 ms) and

$$Q_2 = \begin{pmatrix} 0_{nn} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{n+2 \times n+2} \tag{15}$$

with the zero matrix $0_n \in \mathbb{R}^{n \times n}$ embeds the second level covariance constraint. Finally, an uninformative prior for the second level parameters was specified by setting

$$C_\theta = \begin{pmatrix} 0_{n+1} & 0 \\ 0 & \exp(32) \end{pmatrix} \in \mathbb{R}^{n+2 \times n+2}. \tag{16}$$

Upon specification of the hierarchical linear model for each single-trial regressor, the model parameters $\lambda \in \mathbb{R}^2$ were estimated using an EM algorithm for maximum likelihood estimation and the model log-evidence was approximated using the variational free energy (Friston et al., 2007; Garrido et al., 2007, 2009a). For single subjects, the model

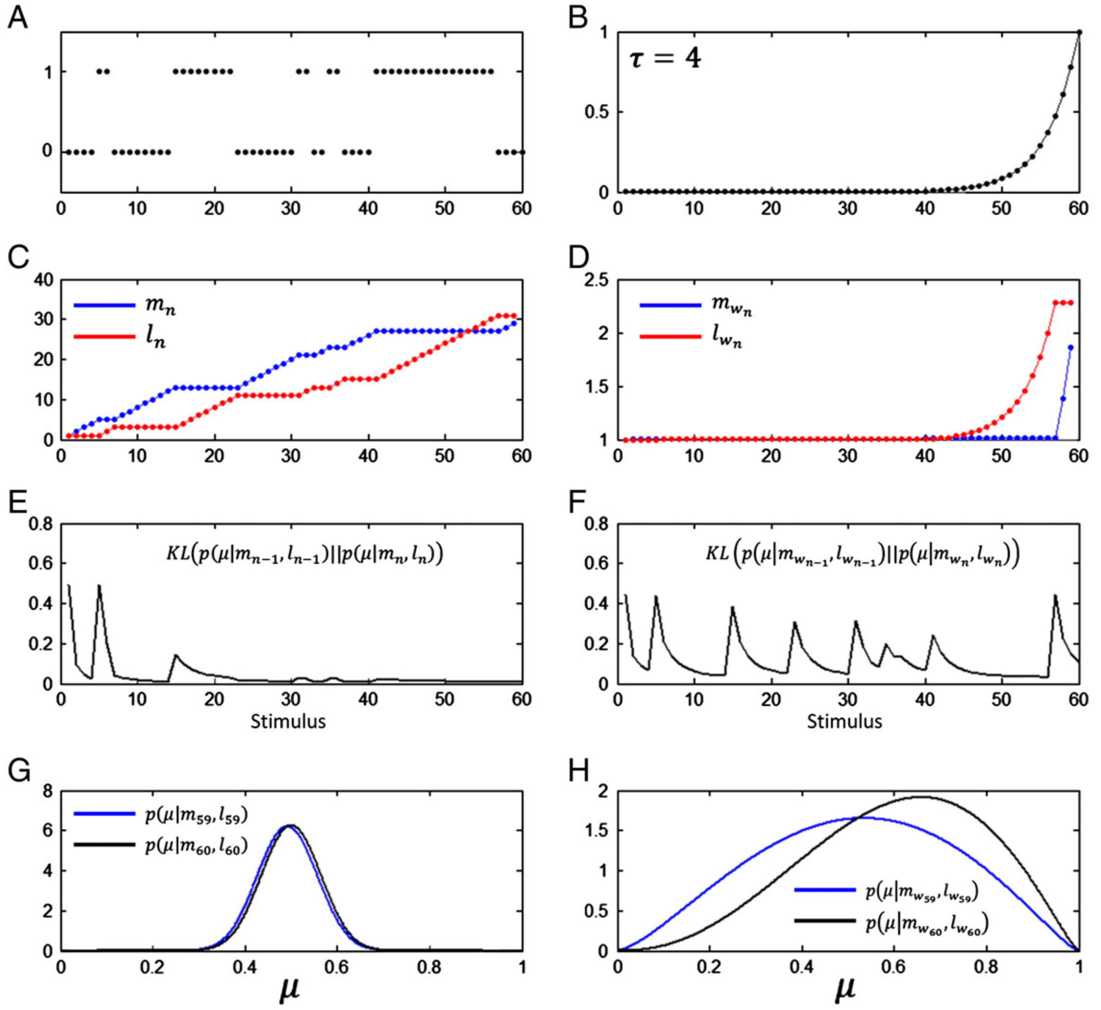

Fig. 2. Computational model. A) Typical stimulus sequence of 60 stimuli alternating between two stimulus amplitudes $A_0 = 0$ and $A_1 = 1$. B) Stimulus-specific weights implemented a forgetting kinetic with a time constant $\tau = 4$ for the 60th stimulus. C) Increase of the beta distribution parameters m_n and l_n over trials, implementing Bayesian learning up to stimulus 60 for no trial weighting. D) Result of applying the weighting function of B) to the temporal evolution of m_n and l_n, resulting in the weighted parameters m_{w_n} and l_{w_n}. E) Illustration of the Bayesian surprise regressor without forgetting or infinite time of integration constant $\tau = 0$ for the stimulus sequence depicted in A). The predicted surprise for this stimulus sequence is large at the beginning and for the first switch of stimulus amplitudes, but close to zero for the remaining amplitude switches. F) Bayesian surprise predictor obtained for the stimulus sequence depicted in A), but under application of the forgetting kinetic shown in B). Note that the amount of Bayesian surprise decreases with the number of stimuli in an identical train of stimuli and increases with the number of preceding stimuli of the opposite amplitude. G) Prior and posterior probability density functions over the parameter μ for the 60th trial of the stimulus sequence shown in A) without forgetting. H) Prior and posterior probability density functions over the parameter μ for the 60th trial of the stimulus sequence shown in A) under the forgetting kinetic depicted in B).

log-evidences across experimental runs were averaged to obtain an estimate of the subject-specific model log-evidence for each model. Following Garrido et al. (2007, 2009a), the single-subject model log-evidences were summed over subjects to yield the group log-evidence for each model, as a function of dipole and peri-stimulus time-bin. Additionally, for a subset of time windows, the group log-evidences were averaged over time and the pairwise differences (i.e. log Bayes factors) between models were plotted. The ensuing difference maps are thresholded at a group log-evidence difference of 3, indicating strong evidence of a particular model, compared to another model (Penny et al., 2004).

Table 3
Interpretation of the Bayesian surprise model time constants τ_0 to τ_4.

BS model	τ	63% (s/stim)	99% (s/stim)
BS0	∞	∞	∞
BS1	8	5.2/8	26.0/40
BS2	4	2.6/4	13.0/20
BS3	2.6	1.7/6	8.6/13.3
BS4	2	1.3/2	6.5/10

Column 1: The five Bayesian surprise models BS0–BS4 were derived from the same set of governing Eqs. (1)–(6), but with varying values of the time constant τ in Eqs. (3) and (4) listed in Column 2. Columns 3 and 4 list the time in seconds and the number of stimuli (ISI 650 ms) corresponding to a 63% and 99% down-weighting of past observations for each of the models/time constants.

Results

Event-related potentials

Grand mean event-related potentials and electrode space results for the somatosensory mismatch response are shown in Fig. 3. Inspection of the grand mean SEP at the channel level, obtained by averaging over all experimental trials, confirmed the presence of well-established SEP components (N20, N45/P60, N140, Fig. 3A). Fig. 3B depicts the 'Deviant–Standard' difference waveform, i.e. the difference between averages evoked by deviant and their immediately preceding standard stimuli (averaged over both stimulus intensities). Fig. 3C depicts the grand mean SEP and the 'Deviant–Standard' waveforms averaged over electrodes over contralateral (C4, C6, CP4, CP5) and ipsilateral (C3, C5, CP3, CP5) somatosensory cortices, showing the expected pattern of stronger contralateral responses. Across all electrodes (Fig. 3B), early sMMR effects were observed around 40 and 85 ms post-stimulus, whereas more pronounced effects were observed at 140, 200, and 350 ms post-stimulus. The corresponding topographies of these differences across electrodes are shown in the upper row of Fig. 3D. The two early effects display fronto-central and parieto-central negativities. The sMMR at 140 ms exhibits a bilateral centro-parietal topography with a stronger contralateral negativity, the sMMR at 200 ms a fronto-contralateral negativity and the sMMR at 350 ms a fronto-central positivity. To establish the reliability of these effects across subjects, we performed paired two-sample t-tests between 'Deviant' and 'Standard' responses in the time windows indicated by the gray bars in Fig. 3B. For the electrodes exhibiting the most pronounced effects as shown in the lower row of Fig. 3D, these differences were significant ($p_{FWE} < 0.05$, Bonferroni corrected for the number of electrodes) for the effects observed at 40, 140, and 350 ms post-stimulus (Table 4). The T-value plots across post-stimulus time (Fig. 3D, middle row) indicate a relatively low between-subject variability in the expression of the observed effects. In summary, a reliable sMMR was recorded using the current experimental paradigm.

Distributed source localization and ECD orientation fitting

Figs. 4A–C summarizes the statistical analysis of the distributed source reconstruction results and Fig. 4D shows the set of oriented ECD sources used as basis for the trial-by-trial analyses (cf. Tables 1 and 2 for details).

To localize sources, we selected time windows of interest based on the peak times of the most prominent deflections in both the 'SEP' and the 'Deviant–Standard' difference response (Figs. 3A, B). To localize S1, the SEP was reconstructed in a time-window of 18–25 ms (i.e. around the N20 effect, Fig. 3A) after stimulus onset, resulting in the expected activation pattern of contralateral S1 (Fig. 4A, Table 1). Likewise, reconstruction of the SEP in a time-window of 130–160 ms (i.e., around the N140 effect, Fig. 3A) after stimulus onset resulted in bilateral activation of posterior S2 (Fig. 4B, Table 1). Right inferior frontal gyrus activity is typically implicated in the response to the deviant and was therefore located by reconstructing the 'Deviant' response in a time window 340–360 ms (as identified from the difference response, Fig. 3B) after stimulus onset (Fig. 4C, Table 1). In accordance with previous studies on the aMMN this source was mirrored for the left hemisphere to derive a symmetric frontal source pair (Rinne et al., 2000, 2005). The distributed source localization analysis did not reveal a significant activation of cingulate cortex for the 'Deviant' condition. Hence, the VB-ECD method with uninformative prior location was used to spatially localize this source at the time point of maximal expression identified from the 'Deviant–Standard' difference response, i.e., at 351 ms, Fig. 3B. The deficiency of the distributed source localization analysis to detect activation of cingulate cortex

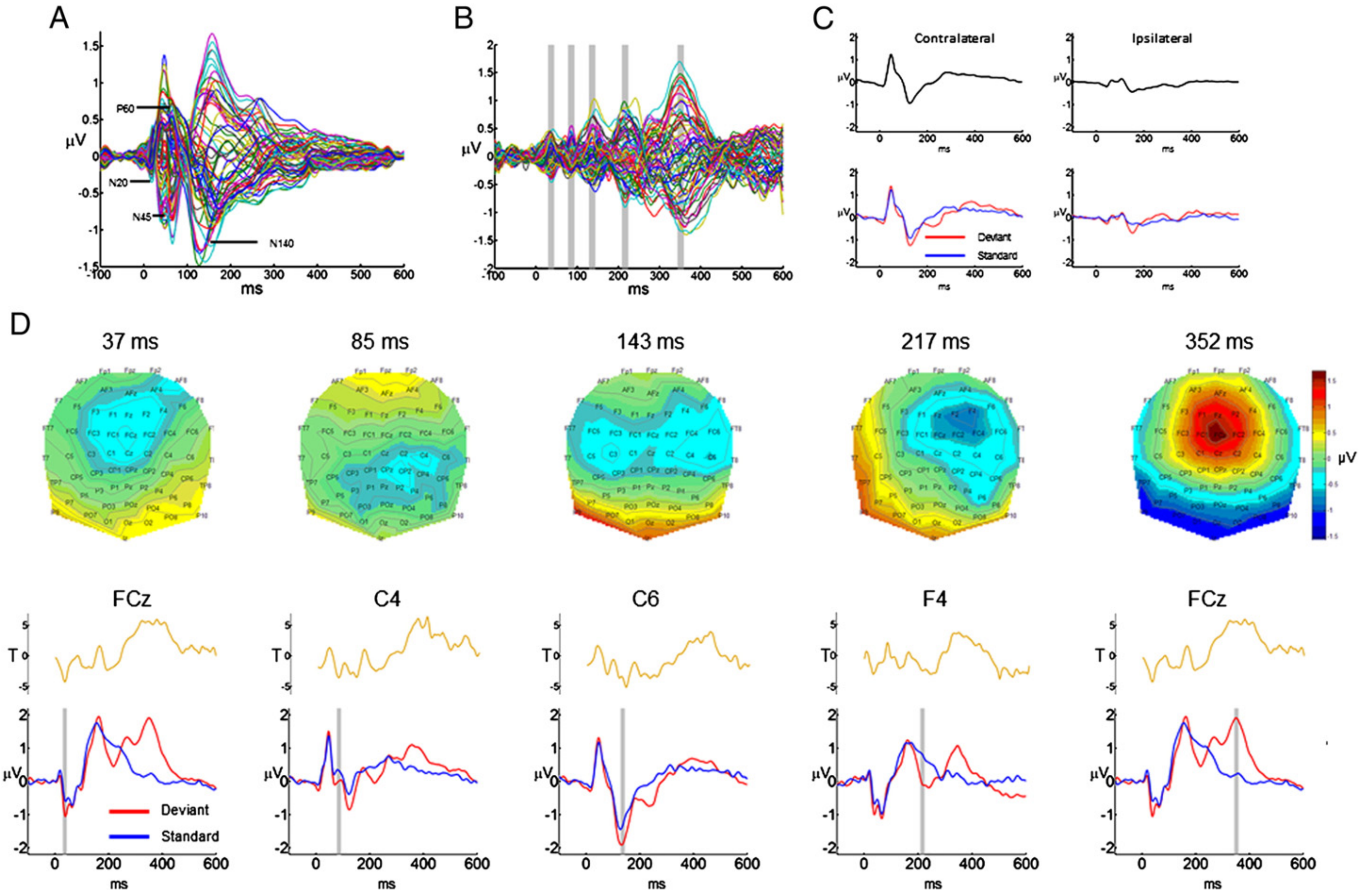

Fig. 3. Event related potentials. A) Grand mean SEP across all stimuli and subjects, for all electrodes. The classical SEP peaks (N20, N45/P60 and N140) are labeled. B) Grand mean 'Deviant–Standard' difference waveform, for all electrodes. The largest differences between deviant and standard SEPs are observed in time windows around 140, 200 and 350 ms post-stimulus, while smaller differences are found around 40 and 85 ms post-stimulus. In conjunction, the difference waveforms up to 200 ms are here referred to as sMMR, while the difference around 350 ms reflects the P300. C) Grand mean SEP (upper row) and grand mean 'Deviant' and 'Standard' waveforms (lower row) averaged over contralateral (C4, C6, CP4, CP6, left panels) and ipsilateral (C3, C5, CP3, CP5, right panels) electrodes above somatosensory cortices. D) Upper row: topographies for the 'Deviant–Standard' difference waveform for all electrodes at the time points indicated by the gray bars in panel B. Middle row: post-stimulus T-values across time for the contrast 'Deviant–Standard' for the electrodes which express the largest 'Deviant–Standard' difference effects for time points 40, 85, 140, 200, and 350 ms. Lower row: the peri-stimulus waveforms for the 'Deviant' (red) and 'Standard' (blue) conditions are shown for the selected electrodes.

may be most likely due to the known bias of source imaging methods to superficial sources (Fuchs et al., 1999; Michel et al., 2004) which is alleviated, but not abolished, by the multiple sparse prior approach used in the current study (Friston et al., 2008). As discussed above, previous findings in the literature as well as the observed potential topography for deviant responses in the current study speak for the inclusion of a cingulate source in the ECD model. As a caveat, based on the approach taken here, we cannot rule out the possibility that activity in cortical areas other than cingulate cortex is captured by the mediate cingulate source. Hence, the topographical interpretation of our modeling results for this source exhibits some uncertainty. On the other hand, the interpretation of the temporal–functional specificity of the modeling results for this source is unaffected by any spatial misattribution of scalp EEG activity.

After MNI coordinates were derived for each of the sources, the VB-ECD method with informative location priors was used to determine the normalized moments for ECDs located at the respective MNI coordinates as documented in Table 2.

Plausibility of the anatomical source model

To investigate whether the data reduction furnished by projecting the electrode space data onto the six ECDs of fixed orientation is sensible, the activity time-courses of the six ECDs were computed for the grand mean SEP (Fig. 5A). This plot indicates a neurobiologically plausible spatiotemporal activity pattern: The S1 dipole captures most of the early N20 activity, while for the N140 activity the S2 and IFG sources contribute most. Late components >200 ms post-stimulus are captured by a mixture of sources, but in particular by the MC source.

To investigate to which degree the data reduction onto a limited set of basis vectors was able to capture the grand mean SEP activity, the channel percent variance explained (PVE) was computed over time (Fig. 5B). Performing the analysis for hierarchical subversions of the six dipole model comprising 1) only S1, 2) S1 and bilateral S2, 3) S1, bilateral S2, bilateral IFG and 4) the complete six ECD model revealed that the PVE over time was largest for the complete model, in particular between 120 and 300 ms.

In summary, based on both the results of previous studies and the present analysis, we concluded that the six ECD model forms an appropriate anatomical basis for the evaluation of trial-by-trial EEG data.

Model-based trial-by-trial analyses

Having established a reliable recording of the sMMR in electrode space and a plausible anatomical source reconstruction, we studied the functional specialization of each of the six sources. To this end, the peri-stimulus trial-by-trial electrode data were projected onto the identified set of ECDs, each peri-stimulus time-point's trial-by-trial ECD activity was modeled using seven different functional models (BS0-BS4, SC and LIN), and for each model, the group model log-evidence was determined using parametric empirical Bayes.

Table 4
Statistical evaluation of the evoked 'Deviant–Standard' difference waveform.

Electrode label	Time window	T-value (df)	p-value
FCz	29–44 ms	T(14) = −3.96	p = 0.001 (sig.)
C4	78–93 ms	T(14) = −2.86	p = 0.012 (n.s.)
C6	129–145 ms	T(14) = −4.33	p = 0.001 (sig.)
F4	209–225 ms	T(14) = −2.36	p = 0.033 (n.s.)
FCz	344–359 ms	T(14) = 5.55	p<0.001 (sig.)

Table 4 reports the results of a series of paired two-sample t-tests for statistical significance of the difference between 'Deviant' and 'Standard' waveforms in the time-windows expressing the most pronounced potentials across subjects (see Fig. 3B). Columns 1 and 2: For each time-window, the average potential at the electrode expressing the maximum potential at the group level was computed for each subject and condition (standard/deviant) and subjected to a paired two-sample t-test. Columns 3 and 4: Statistical significance (sig.) was established using Bonferroni correction for multiple testing over electrodes at a level of $p_{FWE}<0.05$.

In Fig. 6A, each ECD-specific panel depicts the group model log-evidences for the seven models, relative to the constant model M0, over peri-stimulus time. As indicated by the prominent peaks in the log model-evidence maps in three post-stimulus time-windows (around 140, 250, and 360 ms, dotted rectangles) for three different anatomical source ECDs (rS2, rIFG, and MC), Bayesian surprise encoding exhibits a high degree of anatomical–temporal specificity. In the following, we provide a detailed account of this key result by considering the model comparisons for each time-window and ECD source in turn.

In the time window around 140 ms post-stimulus, single-trial variability of contra-lateral secondary somatosensory cortex is best explained by Bayesian surprise models with forgetting (Fig. 6A, rS2, models BS1–BS4). Pair-wise comparison of the time-averaged model log-evidences (Fig. 6B, rS2, 109–171 ms) shows that for this ECD source and time-window, these models explain the data better than the Bayesian surprise model without forgetting (BS0) and both conventional models (stimulus change/linearly modulated stimulus change, SC and LIN, respectively). Next, in the time window around 250 ms post-stimulus, the highest model log-evidences are observed for right inferior frontal cortex (Fig. 6A, rIFG). Here, the Bayesian surprise models with forgetting (BS2–BS4) and the conventional stimulus change (SC) model explain the data best. Specifically, the SC model performs better than all models except the BS4 model for the time-integrated model log-evidence (Fig. 6B, rIFG, 210–291 ms) and at 254 ms supervenes the BS4 model in its ability to explain the data (model log-evidence difference SC − BS4: 3.5). Finally, in a time window around 360 ms post-stimulus, the largest model log-evidences are observed for mediate cingulate cortex (Fig. 6A, MC). Here, Bayesian surprise models with forgetting (BS2–BS4) perform best over an extended period of time, with model BS4 explaining the observed data better than all other models except BS3 (Fig. 6B, panel MC, 310–416 ms). In contrast to the results for secondary somatosensory cortex at 140 ms, the superiority of the Bayesian surprise models over the stimulus change model is less pronounced (model log-evidence difference for rS2 at 140 ms BS4 − SC: 15.5 vs. model log-difference for MC at 365 ms BS4 − SC: 6.9).

Besides these main log-evidence peaks for the rS2, IFG and MC sources, Fig. 6B indicates that also for sources S1 and lS2, the Bayesian surprise models provide better accounts of the data than the conventional/BS0 models during early processing (Fig. 6B, 109–171 ms). For intermediate processing (210–291 ms), for S1 only, the Bayesian surprise models BS3/BS4 are better than the conventional/BS0 models. The log-evidence map for the lIFG source shows only a limited degree of temporal variation and no clear superiority for any of the models. Finally, over all sources and time-windows, model BS0 finds the least support by the data.

Discussion

In the current study we have shown that EEG markers of central somatosensory processing exhibit dynamics that are consistent with the Bayesian brain hypothesis of perceptual learning. Using a model-based approach for the analysis of trial-by-trial EEG source activity, we were able to demonstrate spatiotemporal specific encoding of Bayesian surprise as expressed by the present sequence processing models (Eqs. (1) to (6)). Moreover, we could show that the resulting models are, for some critical processing stages, better in explaining cortical source activity than conventional models that are typically used to analyze mismatch negativity studies.

Specifically, we find that a lower level source (rS2), at an early processing stage (140 ms), is more prominently involved in the representation of Bayesian surprise than in the representation of modulated stimulus change as assessed by conventional models. Critically, in the

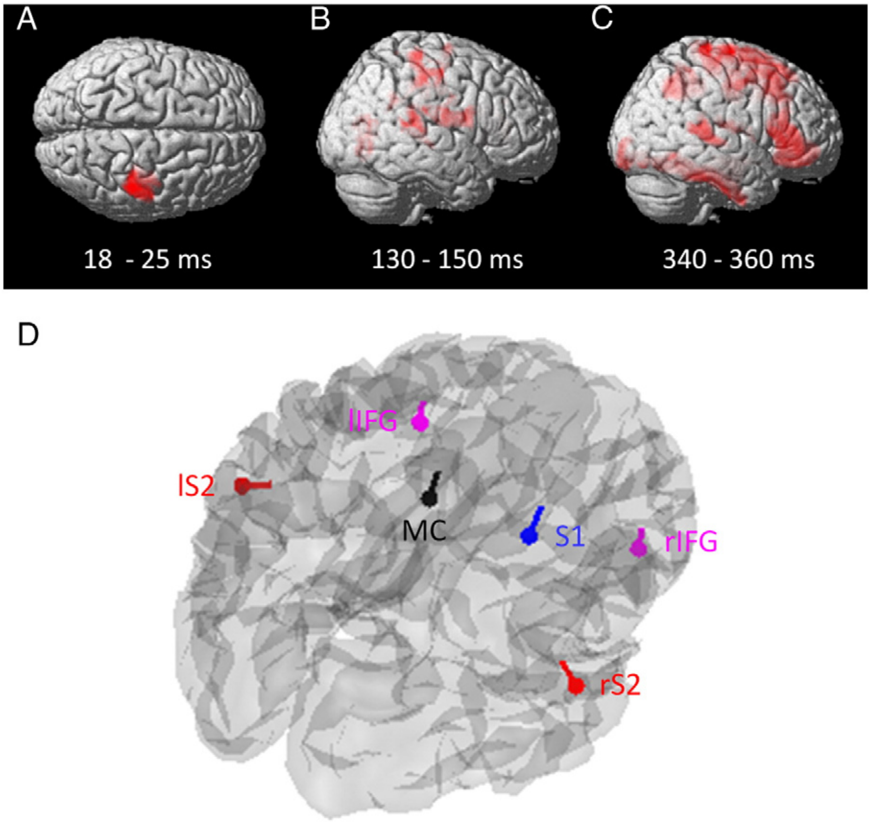

Fig. 4. Distributed source reconstruction and ECD fitting results. A–C) Statistical group distributed source reconstruction results for time-windows of 18–25 ms, 130–150 post-stimulus of the 'SEP' (A and B) and 340–360 ms post-stimulus of the 'Deviant' ERP. The p-value maps displayed were thresholded at $p<0.001$ (uncorrected) and overlaid onto SPM8's standard single subject brain. D) The complete six source ECD model with normalized moments obtained by VB-ECD overlaid on the standard MNI cortical mesh provided by SPM8 (S1: right primary sensory cortex, rS2: right secondary somatosensory cortex, lS2: left secondary somatosensory cortex, rIFG: right inferior frontal gyrus, lIFG left inferior frontal gyrus, MC: medial cingulate cortex).

present study, Bayesian surprise acts as a marker of perceptual learning by signaling the adjustment of the brain's internal generative model. Our finding is hence in line with previous imaging and electrophysiological studies that implicate secondary somatosensory cortex in fast perceptual learning (Pleger et al., 2003; Romo et al., 2003). In contrast, for a subsequent stimulus processing stage around 250 ms, a high level frontal source (rIFG) is more strongly involved in the representation of stimulus change as compared to Bayesian surprise. This finding is in accordance with the "salience network"- theory, which implicates higher level frontal/insular cortex in the bottom-up detection of salient events and switching between other large-scale networks to facilitate access to attention and working memory resources (Menon and Uddin, 2010; Vinod, 2011). Finally, the model analyses indicate additional perceptual learning attributable to the cingulate cortex (MC) at a later processing step (around 360 ms). This result lends support to recent suggestions that perceptual learning manifests itself at several temporal stages of the EEG response (Hamamé et al., 2011; Song et al., 2005), where later stages are thought to predominantly reflect learning-induced modulations of attention (Gilbert et al., 2001). The reduced superiority of Bayesian surprise over stimulus change for this late processing step might thus be indicative of an interaction between perceptual learning and saliency detection. Taken together, the present study provides novel evidence for spatiotemporal/functional segregation in human somatosensory processing. More specifically, early-processing/low-level-cortical stages (S1, contralateral S2) may implement passive short-term Bayesian perceptual learning, downstream intermediate-processing/high-level-cortical stages (rIFG) index active stimulus engagement, and late-processing/high-level-cortical stages (MC) may reflect learning-induced updating of top-down attentional control mechanisms. Importantly, without computational modeling of single trial EEG data at the source level, we would not have found the evidence for these anatomic-temporal functional differences in the somatosensory system.

In contrast to standard Bayesian online-learning schemes (Bishop, 2007), the present study shows that the cortical learning signal is not likely to arise from a Bayesian learner that incorporates all previously observed stimuli of an experimental session with equal weight (model BS0). Rather, we find strong evidence for temporally-adaptive Bayesian learners (BS1–BS4) that give higher weight to temporally close, rather than distant, observations. Specifically, as implied by the generally low explanatory power of model BS0 and by the lower explanatory power of models BS1/BS2 compared to models BS3/BS4 (for a number of ECDs and time windows, see Fig. 6A), the time window of stimulus integration in the current experimental paradigm is probably shorter than 30 s and closer to the 5–10 s range. As shown in the fourth column of Table 3, models BS3 and BS4 exhibit almost complete suppression of stimuli more than 8.6 s and 6.5 s in the past, respectively. A comparison with the timing parameters of the experimental paradigm (inter-stimulus interval 0.65 s, average length of identical stimulation ~5 s, maximal length of identical stimulation ~10 s) suggests that the somatosensory system may use an optimized integration window for the average temporal statistics of the stimulation sequence.

A number of studies have previously addressed the encoding of surprise in the auditory and visual domain using model-based trial-by-trial analyses of evoked EEG and fMRI data (Harrison et al., 2006, 2011; Mars et al., 2008; Strange et al., 2005). Besides providing an extension to the somatosensory domain, the present study makes the following novel contributions: first, the only previous EEG study that used a trial-by-trial model-based analysis of surprise encoding

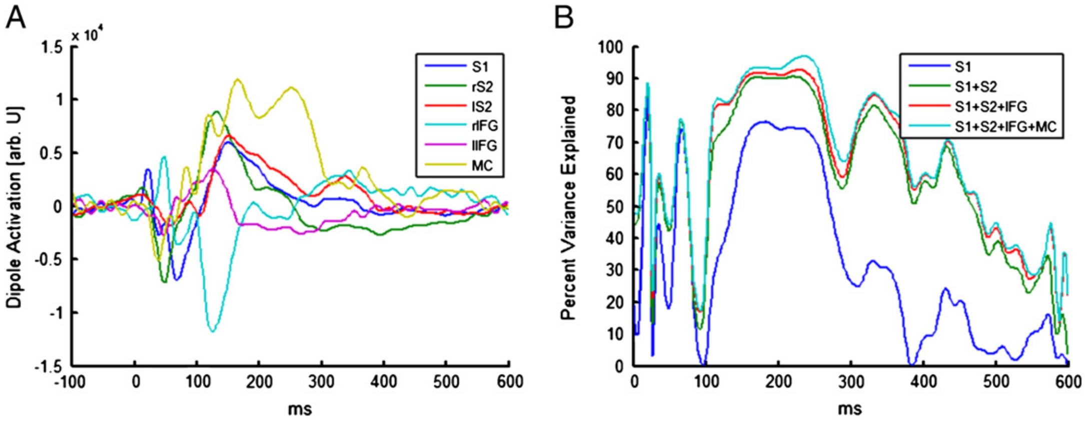

Fig. 5. Plausibility of the six source ECD model. A) ECD activity waveforms obtained by projecting the grand mean SEP onto the oriented ECDs. B) The channel percent variance explained (PVE) for the complete six ECD model (S1 + S2 + IFG + MC) and three hierarchical subversions of this model comprising only the S1 ECD (S1), the S1 and bilateral S2 ECDs (S1 + S2), and the S1, bilateral S2, and bilateral IFG ECDs (S1 + S2 + IFG + MC).

similar to the one employed here focused on a single spatiotemporal feature, namely the P300 amplitude at electrode Pz (Mars et al., 2008). The present study goes beyond this approach by explicitly analyzing all peri-stimulus time bin features of cortical source activity, thereby allowing for a comprehensive analysis of the spatiotemporal activity in brain space. Second, with respect to the model-based regressors, most previous studies used the negative log of the current stimulus probability estimates as measure of surprise (Harrison et al., 2006; Mars et al., 2008; Strange et al., 2005). While we are not claiming that Bayesian surprise is necessarily a better measure of surprise in perceptual learning schemes, it has the benefit of representing the degree to which the internal model is updated on a given trial, rather than the improbability of a stimulus under the current state of the model. This has the advantage that it reflects updating of the internal model over all possible states of external causes, rather than being conditioned on the point estimate of a single cause. In this sense, Bayesian surprise allows the observer to economically evaluate a broad range of possible external scenarios weighted by their inferred uncertainty, rather than relying on a possibly unreliable point estimate of the external cause for future predictions. Third, most previous studies (with the notable exception of Harrison et al., 2011) did not explicitly address the question of the temporal scale of stimulus integration, as provided by the exponential forgetting kinetic used here. We found clear evidence that somatosensory processing represents the stimulus sequence at a specific time-scale which may be related to the temporal statistics of the input

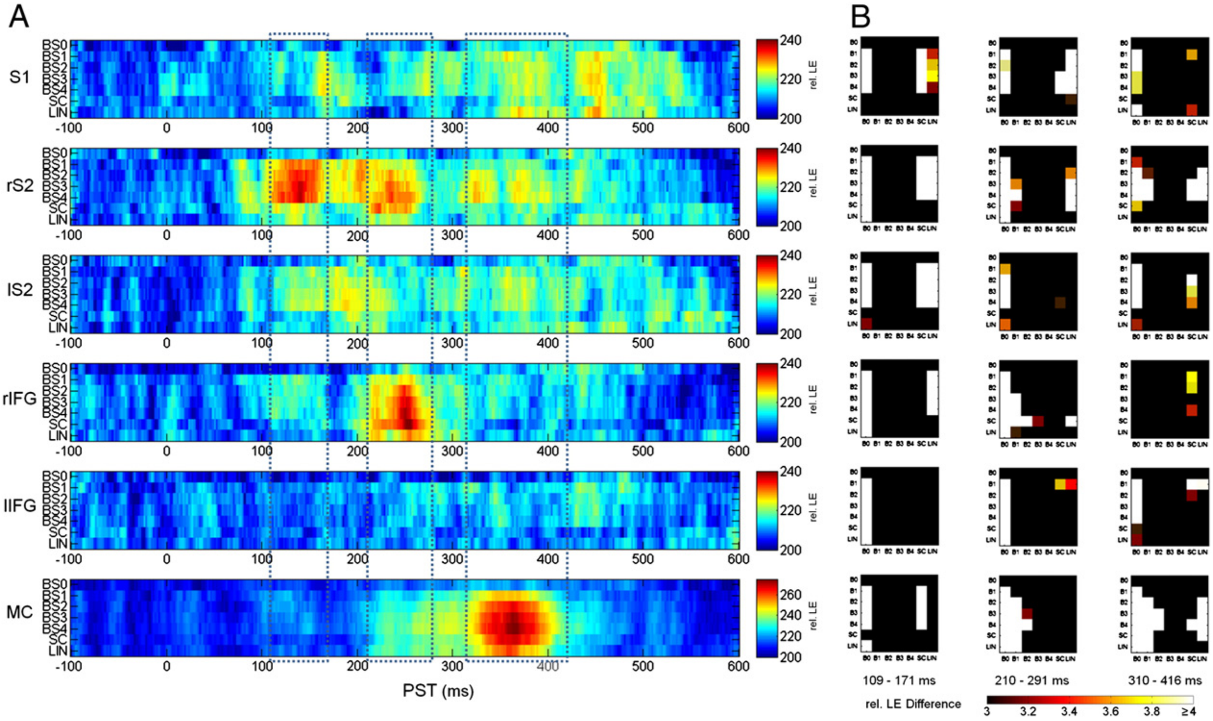

Fig. 6. Computational modeling results. A) Relative group model log evidences for all ECD sources and peri-stimulus time (PST) −100 to 600 ms (S1: right primary sensory cortex, rS2: right secondary somatosensory cortex, lS2: left secondary somatosensory cortex, rIFG: right inferior frontal gyrus, lIFG: left inferior frontal gyrus, MC: medial cingulate cortex). Each row within each ECD panel depicts the model log-evidence over peri-stimulus time for a specific model relative to the constant null model M0 (BS0–BS4: Bayesian surprise models with different time constants for the exponential forgetting constant, see Table 3. SC: stimulus change model, LIN: linear model). The dotted rectangles indicate time-windows of interest further evaluated in panel B. B) Pair-wise model log evidence comparisons for the three time-windows identified in A (109–171 ms, 210–291 ms, 310–416 ms). The squares of each panel color code the difference in model log evidence between the model indicated as row and the model indicated as column. For example, the color in the square at location (B1, SC) denotes the group model log evidence $\ln(y|\mathrm{B1}) - \ln(y|\mathrm{SC})$.

sequences. Fourth, and probably most importantly, mismatch responses in EEG have traditionally been assumed to be elicited by stimulus change (Näätänen et al., 2011), whereas more recent accounts have explicitly addressed the theoretical notion that mismatch responses may be evidence of internal model adjustments in response to unexpected input (Garrido et al., 2009a; Winkler et al., 2009). The computational modeling approach employed here enables us to formally define and statistically test this notion using single trial EEG data.

We conclude the discussion with some considerations of the methodological approach employed: First, the current model-based trial-by-trial EEG analyses are conditioned on the selection of the anatomical location and orientation parameters of the ECDs used for data set projection. While this technique allows us to selectively monitor 'virtual' neural activity in specified brain areas, the data reduction entailed by this procedure might provide different results when using a different source model. Although we have taken great care to appropriately motivate the selection of sources based on both previous literature as well as on source reconstruction results of the present data set, we cannot rule out that there is some other source model that is more plausible than the one employed. Second, the underlying assumption of the present Bayesian surprise model is that the generative model used by the brain samples each stimulus identically and independently. This assumption enabled us to employ a simple update rule and is reasonably plausible given that we explicitly model sequences of stimuli. Third, fitting separate predictor functions derived from the Bayesian surprise model with different time constants τ of 'forgetting' allowed us to obtain a rough estimate of the optimal forgetting constant for a given ECD and time-window. This approach may be replaced by treating the time constant τ as a free parameter of a nonlinear optimization problem. In such a framework, it would be possible to evaluate a continuous parameter space, possibly in a Bayesian fashion, and to directly obtain source-specific estimates of the temporal dynamics of perceptual learning.

Conclusion

In summary, our current study indicates that the dynamics of single-trial somatosensory EEG responses can be explained by a formal model of Bayesian perceptual learning. Specifically, we have shown that in a somatosensory mismatch paradigm Bayesian surprise signals are encoded by multiple cortical regions of the somatosensory network in a temporally specific manner and found that Bayesian surprise signals can provide a better explanation for source-reconstructed single-trial EEG signals than conventional models typically employed for mismatch negativity studies.

Acknowledgments

This work was supported by the BMBF Bernstein II initiative (Förderkennzeichen: 01GQ1001C).

References

Akatsuka, K., Wasaka, T., Nakata, H., Kida, T., Hoshiyama, M., Tamura, Y., Kakigi, R., 2007a. Objective examination for two-point stimulation using a somatosensory oddball paradigm: an MEG study. Clin. Neurophysiol. 118, 403–411.

Akatsuka, K., Wasaka, T., Nakata, H., Kida, T., Kakigi, R., 2007b. The effect of stimulus probability on the somatosensory mismatch field. Exp. Brain Res. 181, 607–614.

Baldeweg, T., Klugman, A., Gruzelier, J., Hirsch, S.R., 2004. Mismatch negativity potentials and cognitive impairment in schizophrenia. Schizophr. Res. 69, 203–217.

Baldi, P., Itti, L., 2010. Of bits and wows: a Bayesian theory of surprise with applications to attention. Neural Netw. 23, 649–666.

Berg, P., Scherg, M., 1994. A multiple source approach to the correction of eye artifacts. Electroencephalogr. Clin. Neurophysiol. 90, 229–241.

Bishop CM., 2007. Pattern Recognition and Machine Learning. First ed. 2006. Corr. 2nd printing. ed. Springer.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory, Ninety-ninth ed. Wiley-Interscience.

Doya, K., Ishii, S., Pouget, A., 2007. Bayesian Brain: Probabilistic Approaches to Neural Coding, first ed. MIT Press.

Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., Zilles, K., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25, 1325–1335.

Friston, K., 2003. Learning and inference in the brain. Neural Netw. 16, 1325–1352.

Friston, K.J., 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11, 127–138.

Friston, K.J., Dolan, R.J., 2010. Computational and dynamic models in neuroimaging. Neuroimage 52, 752–765.

Friston, K.J., Glaser, D.E., Henson, R.N.A., Kiebel, S., Phillips, C., Ashburner, J., 2002a. Classical and Bayesian inference in neuroimaging: applications. Neuroimage 16, 484–512.

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002b. Classical and Bayesian inference in neuroimaging: theory. Neuroimage 16, 465–483.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. Neuroimage 34, 220–234.

Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., Henson, R., Flandin, G., Mattout, J., 2008. Multiple sparse priors for the M/EEG inverse problem. Neuroimage 39, 1104–1120.

Fuchs, M., Wagner, M., Kohler, T., Wischmann, H.A., 1999. Linear and nonlinear current density reconstructions. J. Clin. Neurophysiol. 16 (3), 267–295.

Garrido, M.I., Kilner, J.M., Kiebel, S.J., Stephan, K.E., Friston, K.J., 2007. Dynamic causal modelling of evoked potentials: a reproducibility study. Neuroimage 36, 571–580.

Garrido, M.I., Kilner, J.M., Kiebel, S.J., Friston, K.J., 2009a. Dynamic causal modeling of the response to frequency deviants. J. Neurophysiol. 101, 2620–2631.

Garrido, M.I., Kilner, J.M., Kiebel, S.J., Stephan, K.E., Baldeweg, T., Friston, K.J., 2009b. Repetition suppression and plasticity in the human brain. Neuroimage 48, 269–279.

Garrido, M.I., Kilner, J.M., Stephan, K.E., Friston, K.J., 2009c. The mismatch negativity: a review of underlying mechanisms. Clin. Neurophysiol. 120, 453–463.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman and Hall, Boca Raton.

Gilbert, C.D., Sigman, M., Crist, R.E., 2001. The neural basis of perceptual learning. Neuron 31, 681–697.

Götz, T., Huonker, R., Miltner, W.H.R., Witte, O.W., Dettner, K., Weiss, T., 2011. Task requirements change signal strength of the primary somatosensory M50: oddball vs. one-back tasks. Psychophysiology 48, 569–577.

Hamamé, C.M., Cosmelli, D., Henriquez, R., Aboitiz, F., 2011. Neural mechanisms of human perceptual learning: electrophysiological evidence for a two-stage process.

Harrison, L.M., Duggins, A., Friston, K.J., 2006. Encoding uncertainty in the hippocampus. Neural Netw. 19, 535–546.

Harrison, L.M., Stephan, K.E., Rees, G., Friston, K.J., 2007. Extra-classical receptive field effects measured in striate cortex with fMRI. Neuroimage 34, 1199–1208.

Harrison, L.M., Bestmann, S., Rosa, M.J., Penny, W., Green, G.G.R., 2011. Time scales of representation in the human brain: weighing past information to predict future events. Front. Hum. Neurosci. 5, 37.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14, 382–417.

Itti, L., Baldi, P., 2009. Bayesian surprise attracts human attention. Vision Res. 49, 1295–1306.

Kekoni, J., Hämäläinen, H., Saarinen, M., Gröhn, J., Reinikainen, K., Lehtokoski, A., Näätänen, R., 1997. Rate effect and mismatch responses in the somatosensory system: ERP-recordings in humans. Biol. Psychol. 46, 125–142.

Kersten, D., Mamassian, P., Yuille, A., 2004. Object perception as Bayesian inference. Annu. Rev. Psychol. 55, 271–304.

Kida, T., Nishihira, Y., Wasaka, T., Nakata, H., Sakamoto, M., 2004. Passive enhancement of the somatosensory P100 and N140 in an active attention task using deviant alone condition. Clin. Neurophysiol. 115, 871–879.

Kiebel, S.J., Friston, K.J., 2004. Statistical parametric mapping for event-related potentials: I. Generic considerations. Neuroimage 22, 492–502.

Kiebel, S.J., Daunizeau, J., Friston, K.J., 2008a. A hierarchy of time-scales and the brain. PLoS Comput. Biol. 4, e1000209.

Kiebel, S.J., Daunizeau, J., Phillips, C., Friston, K.J., 2008b. Variational Bayesian inversion of the equivalent current dipole model in EEG/MEG. Neuroimage 39, 728–741.

Knill, D.C., Pouget, A., 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 27, 712–719.

Linden, D.E.J., 2005. The P300: where in the brain is it produced and what does it tell us? Neuroscientist 11, 563–576.

Litvak, V., Friston, K., 2008. Electromagnetic source reconstruction for group studies. Neuroimage 42, 1490–1498.

Litvak, V., Komssi, S., Scherg, M., Hoechstetter, K., Classen, J., Zaaroor, M., Pratt, H., Kahkonen, S., 2007. Artifact correction and source analysis of early electroencephalographic responses evoked by transcranial magnetic stimulation over primary motor cortex. Neuroimage 37, 56–70.

Litvak, V., Mattout, J., Kiebel, S., Phillips, C., Henson, R., Kilner, J., Barnes, G., Oostenveld, R., Daunizeau, J., Flandin, G., Penny, W., Friston, K., 2011. EEG and MEG data analysis in SPM8. Comput. Intell. Neurosci. 2011, 852961.

Mars, R.B., Debener, S., Gladwin, T.E., Harrison, L.M., Haggard, P., Rothwell, J.C., Bestmann, S., 2008. Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. J. Neurosci. 28, 12539–12545.

Mars, R.B., Shea, N.J., Kolling, N., Rushworth, M.F.S., 2010. Model-based analyses: promises, pitfalls, and example applications to the study of cognitive control. Q J Exp Psychol (Hove) 1–16.

Menon, V., Uddin, L.Q., 2010. Saliency, switching, attention and control: a network model of insula function. Brain Struct. Funct. 214, 655–667.

Michel, C.M., Murray, M.M., Lantz, G., Gonzales, S., Spinelli, L., de Peralte, R.G., 2004. EEG source imaging. Clin. Neurophysiol. 115, 2195–2222.

Mouraux, A., Iannetti, G.D., 2008. A review of the evidence against the "first come first served" hypothesis Comment on Truini et al. [Pain 2007; 131:343–7] Pain 136, 219–225.

Mumford, D., 1992. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol. Cybern. 66, 241–251.

Näätänen, R., 2009. Somatosensory mismatch negativity: a new clinical tool for developmental neurological research? Dev. Med. Child Neurol. 51, 930–931.

Näätänen, R., Gaillard, A.W., Mäntysalo, S., 1978. Early selective-attention effect on evoked potential reinterpreted. Acta Psychol. (Amst) 42, 313–329.

Näätänen, R., Kujala, T., Winkler, I., 2011. Auditory processing that leads to conscious perception: a unique window to central auditory processing opened by the mismatch negativity and related responses. Psychophysiology 48, 4–22.

Niedermeyer, E., Silva, F.L.D., 2004. Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, Fifth ed. Lippincott Raven.

Penny, W.D., 2012. Comparing dynamic causal models using AIC, BIC and free energy. Neuroimage 59, 319–330.

Penny, W.D., 2001. Kullback-Leibler Divergences of Normal, Gamma, Dirichlet and Wishart Densities (Technical Report). Wellcome Department of Cognitive Neurology.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. Neuroimage 22, 1157–1172.

Pitt, M.A., Myung, I.J., 2002. When a good fit can be bad. Trends Cogn. Sci. 6, 421–425.

Pleger, B., Foerster, A.F., Ragert, P., Dinse, H.R., Schwenkreis, P., Malin, J.P., Nicolas, V., Tegenthoff, M., 2003. Functional imaging of perceptual learning in human primary and secondary somatosensory cortex. Neuron 40, 643–653.

Polich, J., 2007. Updating P300: an integrative theory of P3a and P3b. Clin. Neurophysiol. 118, 2128–2148.

Rao, R.P., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79–87.

Restuccia, D., Della Marca, G., Valeriani, M., Leggio, M.G., Molinari, M., 2007. Cerebellar damage impairs detection of somatosensory input changes. A somatosensory mismatch-negativity study. Brain 130, 276–287.

Restuccia, D., Zanini, S., Cazzagon, M., Del Piero, I., Martucci, L., Della Marca, G., 2009. Somatosensory mismatch negativity in healthy children. Dev. Med. Child Neurol. 51, 991–998.

Rinne, T., Alho, K., Ilmoniemi, R.J., Virtanen, J., Näätänen, R., 2000. Separate time behaviors of the temporal and frontal mismatch negativity sources. Neuroimage 12, 14–19.

Rinne, T., Degerman, A., Alho, K., 2005. Superior temporal and inferior frontal cortices are activated by infrequent sound duration decrements: an fMRI study. Neuroimage 26, 66–72.

Rinne, T., Särkkä, A., Degerman, A., Schröger, E., Alho, K., 2006. Two separate mechanisms underlie auditory change detection and involuntary control of attention. Brain Res. 1077, 135–143.

Romo, R., Hernández, A., Zainos, A., Salinas, E., 2003. Correlated neuronal discharges that increase coding efficiency during perceptual discrimination. Neuron 38, 649–657.

Shinozaki, N., Yabe, H., Sutoh, T., Hiruma, T., Kaneko, S., 1998. Somatosensory automatic responses to deviant stimuli. Brain Res. Cogn. Brain Res. 7, 165–171.

Song, Y., Ding, Y., Fan, S., Qu, Z., Xu, L., Lu, C., Peng, D., 2005. Neural substrates of visual perceptual learning of simple and complex stimuli. Clin. Neurophysiol. 116, 632–639.

Spackman, L.A., Boyd, S.G., Towell, A., 2007. Effects of stimulus frequency and duration on somatosensory discrimination responses. Exp. Brain Res. 177, 21–30.

Spackman, L.A., Towell, A., Boyd, S.G., 2010. Somatosensory discrimination: an intracranial event-related potential study of children with refractory epilepsy. Brain Res. 1310, 68–76.

Strange, B.A., Duggins, A., Penny, W., Dolan, R.J., Friston, K.J., 2005. Information theory, novelty and hippocampal responses: unpredicted or unpredictable? Neural Netw. 18, 225–230.

Tarkka, I.M., Micheloyannis, S., Stokić, D.S., 1996. Generators for human P300 elicited by somatosensory stimuli using multiple dipole source analysis. Neuroscience 75, 275–287.

Thees, S., Blankenburg, F., Taskin, B., Curio, G., Villringer, A., 2003. Dipole source localization and fMRI of simultaneously recorded data applied to somatosensory categorization. Neuroimage 18, 707–719.

Tse, C.-Y., Penney, T.B., 2008. On the functional role of temporal and frontal cortex activation in passive detection of auditory deviance. Neuroimage 41, 1462–1470.

Vinod, M., 2011. Large-scale brain networks and psychopathology: a unifying triple network model. Trends Cogn. Sci. 15, 483–506.

Wang, A.L., Mouraux, A., Liang, M., Iannetti, G.D., 2010. Stimulus novelty, and not neural refractoriness, explains the repetition suppression of laser-evoked potentials. J. Neurophysiol. 104, 2116–2124.

Winkler, I., Denham, S.L., Nelken, I., 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn. Sci. 13, 532–540.

Woolrich MW., in press. Bayesian inference in FMRI. Neuroimage. DOI: 10.1016/j.neuroimage.2011.10.047.

Worsley, K.J., 1994. Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields. Adv. Appl. Probab. 26, 13–42.