Brain signatures of surprise in EEG and MEG data

Zahra Mousavi¹, Mohammad Mahdi Kiani¹, Hamid Aghajan^{1*}, ^{#a}

¹Department of Electrical Engineering, Sharif University of Technology, Tehran,

Iran

* Corresponding Author

Email: aghajan@ee.sharif.edu (HA)

^{#a} Current address of the corresponding author: Department of Electrical

Engineering, Sharif University of Technology, Tehran, Iran

9 **Abstract**

10 The brain is constantly anticipating the future of sensory inputs based on past
 11 experiences. When new sensory data is different from predictions shaped by recent
 12 trends, neural signals are generated to report this surprise. Existing models for
 13 quantifying surprise are based on an ideal observer assumption operating under one
 14 of the three definitions of surprise set forth as the Shannon, Bayesian, and
 15 Confidence-corrected surprise. In this paper, we analyze both visual and auditory
 16 EEG and auditory MEG signals recorded during oddball tasks to examine which
 17 temporal components in these signals are sufficient to decode the brain's surprise
 18 based on each of these three definitions. We found that for both recording systems
 19 the Shannon surprise is always significantly better decoded than the Bayesian
 20 surprise regardless of the sensory modality and the selected temporal features used
 21 for decoding.

22 **Keywords:** Shannon surprise, Bayesian surprise, confidence-corrected surprise, R-
 23 squared of regression model, ERP temporal components, generative distribution,
 24 decoding power, middle temporal components, late temporal components, mismatch
 25 negativity (MMN), P300, oddball task

26 **Author summary**

27 A regression model is proposed for decoding the level of the brain's surprise in
 28 response to sensory sequences using selected temporal components of recorded EEG

29 and MEG data. Three surprise quantification definitions (Shannon, Bayesian, and
30 Confidence-corrected surprise) are compared in offering decoding power. Four
31 different regimes for selecting temporal samples of EEG and MEG data are used to
32 evaluate which part of the recorded data may contain signatures that represent the
33 brain's surprise in terms of offering a high decoding power. We found that both the
34 middle and late components of the EEG response offer strong decoding power for
35 surprise while the early components are significantly weaker in decoding surprise. In
36 the MEG response, we found that the middle components have the highest decoding
37 power while the late components offer moderate decoding powers. When using a
38 single temporal sample for decoding surprise, samples of the middle segment possess
39 the highest decoding power. Shannon surprise is always better decoded than the
40 other definitions of surprise for all the four temporal feature selection regimes.
41 Similar superiority for Shannon surprise is observed for the EEG and MEG data
42 across the entire range of temporal sample regimes used in our analysis.

43 **Introduction**

44 The predictive coding framework [1] states that the brain is constantly predicting its
45 next sensory input. Past inputs are used by the brain to form prior knowledge in the
46 Bayesian brain model [2, 3]. In fact, the results of brain functions such as perception,
47 eliminating ambiguity, attention, and decision making are dependent on the way the
48 current sensory input and the knowledge gained from previous experiences are
49 combined in the hierarchical inference model of the brain [1] (for review see [4]).

50 An input different from what the brain has predicted will be surprising in that it
51 generates a form of response measurable by brain imaging techniques. This surprise
52 has been quantified in the literature as a variable representing the error in predicting
53 the future stimuli based on the expectation of the near-optimal observer [5-15].
54 Studying the characteristics of surprise plays an important role in understanding how
55 learning happens in the brain. In this context, surprise is often represented by a
56 parameter that the brain attempts to minimize during the process of learning [14, 16-
57 18]. A remarkable observation is that the unpredictability of the stimulus, which
58 leads to a high value of surprise, elicits the attention of the observer [6, 7], and
59 produces distinct brain responses [9, 12, 19].

60 The Shannon surprise [10] has been widely used as a measure for quantifying
61 surprise based on the likelihood of the data [9, 11-13, 15, 19, 20]. Faraji et al. (2018)
62 introduce the notion of “puzzlement” surprise for when the brain encounters an input
63 not matching the model it has created for the world, and consider the Shannon
64 surprise as a quantification of this surprise. They also introduce an alternative
65 quantification of the puzzlement surprise, named the Confidence-corrected surprise,
66 by comparing the posterior distribution of the input with that of a naïve observer
67 (who bases his model on the most recent input and a uniform prior). To the scope of
68 our knowledge, it is not yet examined whether this new definition better corroborates
69 experimental data than the Shannon surprise does.

70 After the brain has utilized the new input to update its model of the world, an
71 “enlightenment” surprise ensues [14]. The Bayesian surprise, based on comparing
72 the estimated generative distribution of the received stimuli before and after the
73 arrival of each input, was introduced by Baldi (2002) and has been used thereafter by
74 many researchers [9, 12, 18, 67]. The Bayesian surprise can be regarded as a model
75 for describing the “enlightenment” surprise [14].

76 Bayesian surprise, also called shifts in beliefs or model updating, is believed to map
77 to a completely different region of the human brain from Shannon surprise. Several
78 recent reports based on fMRI (Functional Magnetic Resonance Imaging) data posit
79 that Shannon and Bayesian surprises modulate different brain regions [21-23]. In
80 other words, Shannon and Bayesian surprise definitions model distinct cognitive
81 processes in the brain, with Shannon surprise reflecting the unlikeliness of the input,
82 and Bayesian surprise representing model updating.

83 In the context of Event-Related Potential (ERP) experiments, temporal components
84 such as the Mismatch Negativity (MMN) and P300 have been widely believed to
85 manifest underlying surprise events in the brain, and hence, reflect the brain’s error
86 in anticipating the next sensory input [24-29]. These surprise-related components
87 have been often described by a stimulus-referenced information-theoretical surprise
88 in experiments in which subjects are presented with a sequence of stimuli randomly
89 selected from a finite set [9, 12, 13]. Abnormal values in these components have also
90 been proposed as biomarkers for some cognitive disorders such as Schizophrenia and

11 Alzheimer's disease [30-33], reflecting their importance not only in understanding
12 the behavior of the normal brain in handling surprise, but also in the detection of a
13 number of brain disorders. Similar surprise-related temporal components have been
14 reported in the studies of MEG (Magnetoencephalography) signals [34-36].
15 An experiment specifically designed to account for the different brain reactions to
16 rare and frequent stimuli is the oddball task, in which a randomly composed
17 sequence of standard and deviant stimuli is presented to a subject [27, 28, 37].
18 Previous surprise modeling studies mainly base their conclusions on a single
19 component extracted from the brain signals, with the MMN [27] or the P300 [12, 24,
20 25] or both [29] serving as the main such target components. While such single
21 component analysis simplifies the ensuing effort to develop an encoder or a decoder
22 for the brain surprise, it ignores the contribution of other temporal components
23 corresponding to different post-stimulus latencies. Recent studies have developed
24 models using the entire temporal signals, assuming that the entire epoch of the
25 response is modulated by the statistical properties of the input sequence [15, 38].
26 In this paper, we set out to study which of the three mentioned information-theoretic
27 quantifications of surprise, reflecting distinct concepts of unlikelyness (Shannon) or
28 unexpectedness (Confidence-corrected) of the input, and belief updating (Bayesian),
29 each happening in different brain regions, fits better to empirical EEG data of visual
30 or auditory modalities and auditory MEG data to model the brain's response in an
31 oddball experiment. The derivation of these three information-theoretic surprise

112 values are based on an assumption that the observer tries to learn the probability
113 transition matrix of stimuli .This question falls within the scope of the Bayesian
114 brain theory by seeking out how the brain creates and updates a probabilistic model
115 for the incoming sequence of stimuli as a precursor for predicting future sensory
116 inputs. The paper formulates its definition of surprise assuming two transition
117 probability values between the standard and deviant stimuli to describe the
118 generative distribution of the stimuli sequence [13].

119 The study is applied to an experimental oddball dataset with separate runs of visual
120 and auditory tasks recorded by a multi-channel EEG system. Also, we have repeated
121 the study on the recorded MEG of another auditory oddball task. The recorded data
122 is used to train the trial-by-trial decoder of Modirshanechi et al. (2019) as a regressor
123 to measure the surprise of each stimulus. Surprise labels for training the decoder are
124 generated from the sequence of stimuli. Three separate decoder systems are hence
125 trained with labels generated by each of the three definitions of surprise, and the
126 decoding powers of these decoders are compared.

127 For each definition of surprise and each sensory modality, samples of recorded data
128 in each epoch are selected in four different regimes in order to also examine the
129 relative decoding power attained by using different temporal segments:

- 130 • Entire epoch: First, the entire recorded epoch is used as input to train the decoder.
- 131 • Samples: Second, the value of a single time instance sampled from every epoch is
132 used as the decoder input. The decoding power of the set of decoders, each trained

133 based on a specific time instance, determines the relative importance of single
134 temporal EEG/MEG samples in estimating the stimuli surprise.

135 • Intervals: Third, the decoder inputs are selected as aggregates of temporal samples
136 from the start of the epoch to the target time instance. This allows the decoder to
137 utilize the dependency among the temporal EEG/MEG samples. In this regime, the
138 earliest time instance in the epoch for which the decoding power surpasses a fixed
139 confidence level is subsequently detected.

140 • Segments: Finally, in order to compare the middle and late segments of the
141 EEG/MEG record which contain the well-known points of mismatch negativity
142 (MMN) and P300, respectively, two different decoders are trained with each of
143 these segments as input.

144 The paper sets forth comparative results for all the mentioned surprise decoders, and
145 elaborates on the relative efficacy of the underlying definitions of surprise as well as
146 the importance of the different time components in decoding the surprise of the
147 stimuli.

148 **Results**

149 The feature vectors extracted from the EEG/MEG data for each subject are used as
150 input samples to train a surprise decoder. The decoder uses the result of applying a
151 definition of surprise to the input stimuli sequence as the label during the training
152 phase. The decoder we use for this analysis was introduced by Modirshanechi et al.

(2019), and we modified its input features as well as the surprise labels to fit our analysis with three surprise definitions and four temporal feature selection regimes as described earlier. We provide a brief description of the decoder's operation to better motivate our results. More details about the decoder can be found in [15].

The decoder mainly consists of two modules of linear regression. A Lasso linear regression method takes the feature matrix $S_{N \times p}$ (extracted from the EEG data according to one of the 4 temporal sample selection regimes) as its input, and the label vector $Y_{N \times 1}$ (calculated from the input stimuli sequence according to one of the 3 three definitions of surprise) as its labels. The Lasso regressor aims to minimize the reconstruction error while observing an added sparsity term, and is used to eliminate the input features which might be irrelevant to reconstructing the surprise, and hence help avoid overfitting. To evaluate the trained model with cross-validation, we used the R-squared measure. Noticing that the decoding power is a function of the integration coefficient w (the parameter defining the coefficients of the window of integration; see Materials and Methods), we have reported the maximum decoding power across all w s for each regression by employing the best integration window coefficients. After the removal of features with zero coefficients by the Lasso regressor, the remaining features were presumed effective in describing the surprise.

Tables 1-3 and Figs 1-8 summarize our results on both the EEG and MEG data, which are described in detail here.

173 **Table 1. Decoding power and chance level for the three definitions of surprise for the temporal**
174 **regimes of *Entire epoch* and the three *Early*, *Middle*, and *Late* segments.**

	EEG Visual Task			EEG Auditory Task			MEG Auditory Task		
Decoding Power	Shannon	Confidence-Corrected	Bayesian	Shannon	Confidence-Corrected	Bayesian	Shannon	Confidence - Corrected	Bayesian
Entire epoch	0.482	0.331	0.264	0.375	0.278	0.198	0.134	0.070	0.033
Early	0.017	0.006	0.002	0.026	0.027	0.005	0.027	0.009	0.009
Middle	0.348	0.233	0.195	0.306	0.229	0.167	0.141	0.095	0.046
Late	0.414	0.287	0.227	0.340	0.248	0.185	0.088	0.056	0.031
Chance Level									
Entire epoch	-0.0038 ± 0.0046	-0.0038 ± 0.0049	-0.0038 ± 0.0050	-0.0030 ± 0.0044	-0.0033 ± 0.0043	-0.0033 ± 0.0043	-0.0031 ± 0.0050	-0.0031 ± 0.0051	-0.0031 ± 0.0050
Early	-0.0038 ± 0.0047	-0.0038 ± 0.0049	-0.0035 ± 0.0048	-0.0031 ± 0.0042	-0.0033 ± 0.0045	-0.0032 ± 0.0042	-0.0031 ± 0.0050	-0.0031 ± 0.0051	-0.0031 ± 0.0051
Middle	-0.0038 ± 0.0046	-0.0038 ± 0.0047	-0.0038 ± 0.0047	-0.0028 ± 0.0063	-0.0033 ± 0.0042	-0.0031 ± 0.0043	-0.0031 ± 0.0050	-0.0031 ± 0.0051	-0.0031 ± 0.0051
Late	-0.0038 ± 0.0046	-0.0038 ± 0.0049	-0.0038 ± 0.0047	-0.0030 ± 0.0041	-0.0033 ± 0.0043	-0.0032 ± 0.0042	-0.0031 ± 0.0050	-0.0031 ± 0.0051	-0.0004 ± 0.0069

175

176 **Table 2. P-values of the t-test for comparing three temporal segments of *Early*, *Middle*, and *Late*.**

<i>EEG Visual Task</i>	Middle vs. Late	Early vs. Middle	Early vs. Late
<i>Bayesian</i>	0.33	8.77E-10	9.32E-14
<i>Shannon</i>	0.17	1.89E-11	1.05E-16
<i>Confidence- Corrected</i>	0.17	1.46E-09	1.79E-14
<i>EEG Auditory Task</i>			
<i>Bayesian</i>	0.55	5.23E-14	4.83E-08
<i>Shannon</i>	0.41	1.24E-13	8.11E-11

<i>Confidence- Corrected</i>	0.55	1.01E-10	6.59E-08
MEG Auditory Task			
<i>Bayesian</i>	0.44	0.005	0.16
<i>Shannon</i>	0.17	0.0006	0.02
<i>Confidence- Corrected</i>	0.16	0.0006	0.006

177 Highlights indicate p-values < 0.05.

178 **Table 3. P-values of the t-test of comparing three surprise quantifications.**

EEG Visual Task	Shannon vs. Bayesian	Bayesian vs. Confidence-Corrected	Shannon vs. Confidence-Corrected
<i>Entire epoch</i>	4.61E-07	0.056	1.15E-4
<i>Middle</i>	4.10E-04	0.335	0.006
<i>Late</i>	6.10E-07	0.061	1.7E-4
EEG Auditory Task			
<i>Entire epoch</i>	2.24E-06	0.008	0.008
<i>Middle</i>	3.56E-06	0.005	0.008
<i>Late</i>	5.9E-04	0.094	0.048
MEG Auditory Task			
<i>Entire epoch</i>	0.002	0.108	0.051
<i>Middle</i>	0.004	0.058	0.211
<i>Late</i>	0.048	0.238	0.265

179 Highlights indicate p-values < 0.05.

180

181 **Fig 1. Surprise decoding powers of the three surprise quantifications for every single temporal**
182 **Sample in EEG auditory and visual tasks.** Dashed lines indicate selected temporal segments (colors
183 should be used in printing).

Fig 2. Surprise decoding powers of the three surprise quantifications for the temporal *Interval* [-200ms, t] in EEG auditory and visual tasks. Dashed lines indicate selected temporal segments (colors should be used in printing).

Fig 3. MEG auditory task results. a) Surprise decoding powers of the three surprise quantifications for every single temporal *Sample*. **b)** Surprise decoding powers of the three surprise quantifications for the temporal *Interval* [-200ms, t] (colors should be used in printing).

196 **Fig 4. P-values of comparing decoding powers of three surprise quantifications using temporal**
197 ***Intervals [-200ms, t] in EEG auditory and visual tasks. Colors should be used in printing.***

Fig 5. P-values of comparing decoding powers of three surprise quantifications of using each single temporal *Sample* in EEG auditory and visual tasks. Colors should be used in printing.

Fig 6. P-values of comparing decoding powers of three surprise quantifications. a) *Sample* and b) *Intervals* regimes in MEG auditory task (colors should be used in printing).

208 **Fig 7. The decoding power of different values of w over time for different definitions of surprise,**
 209 **Intervals or Samples regimes and two sensory modalities in EEG data.** Colors denote the
 210 decoding power and should be used when printing.

212 **Fig 8. The decoding power of different values of w over time for different definitions of surprise,**
213 **Intervals or Samples regimes and two sensory modalities in MEG data. Colors denote the**
214 **decoding power and should be used when printing.**

217 **Decoding power**

218 Here we describe the decoding powers of the three quantifications of surprise when
 219 employed to supply labels for training surprise decoders. In temporal regimes based
 220 on the *Entire epoch* and each of the three *Segments* (of Early, Middle, and Late),
 221 there is only one decoder per subject. Hence, the R^2 (R-squared) values that are to be
 222 reported will be single numbers for each case averaged on the subjects, which are
 223 reported in Table 1. For the temporal regimes of *Interval* and *Samples*, since the
 224 target time is swept from the start of the epoch to the current sample time t , for each
 225 value of t there will be a decoder trained, for which an R^2 value will be reported.
 226 These R^2 values for these two temporal sampling regimes are illustrated as curves
 227 versus time t and are plotted in Figs 1 to 4.

228 *Entire Epoch and Segments regimes*: The R^2 values using the entire samples, as well
 229 as the segments named Early, Middle and Late time intervals are presented in Table
 230 1 for both auditory and visual tasks of EEG and the auditory MEG task, and for the
 231 three surprise quantifications. The chance level is calculated and shown for every
 232 condition to have a criterion for assessing the R^2 values as decoding powers.
 233 Comparing the decoding powers achieved in each task and for each definition of
 234 surprise with the corresponding chance level, we observe that for the Early segment,
 235 the R^2 values and the chance levels are close to each other and hence, the hypothesis
 236 of independence between the surprise values (with any surprise definition) and the
 237 Early segments is not violated. Unlike this low performance of the Early segment in

describing the theoretical surprise values, it can be observed in Table 1 that the Middle and Late segments demonstrate significant values of decoding power.

Samples regime: The decoding power of the regression model, whose input feature is a single time sample at a target time t of the response in different ICA components, is illustrated in Fig 1 for the EEG and in Fig 3.a for the MEG data, for different values of $t \in [-200\text{ms}, 600\text{ms}]$. Due to the employment of only one time sample in each epoch for decoding the trial's surprise, it is understandable to have relatively lower R^2 levels. Interestingly, in Fig 1, two major local maxima are observed in each curve. This means that for all definitions of surprise and for both the auditory and visual task data, one can identify one time sample in the Middle segment and one in the Late segment each having a peak decoding power. These time samples with peak R^2 values can serve as statistically meaningful placeholders for the well-known N200 and P300 components of the ERP terminology, respectively. The curves of Fig 3.a appear different from those of Fig 1 in that the second local maximum points in the MEG corresponding to the Late segments are smaller than the peaks of the Middle segment, whereas the maximum points occurring in the Late segments of the EEG analysis in Fig 1 had large values.

Interval regime: Figs 2 and 3.b illustrate the decoding powers of decoders trained using an interval of temporal samples in the range of $[-200\text{ms}, t]$ for different values of t . This regime is expected to reveal at which time enough evidence has been accumulated from the response for achieving a confident decoding performance. As

259 expected, all the curves in this regime are approximately monotonically increasing
 260 with time as more data is used by the decoder with increasing t . In each curve, the R^2
 261 value stays close to zero until around 100ms, when there is a considerable jump in
 262 the decoding power. This increase occurs in the temporal range which we called the
 263 Middle segment in our *Segments regime*. The decoding power continues to increase
 264 gradually afterwards. This behavior is observed for all three definitions of surprise
 265 and for both stimuli modalities in the EEG analysis. However, the curves of the
 266 MEG and EEG analyses have a difference in that the MEG curves appear more flat
 267 after the first rise. This can be explained by the lower decoding power of the Late
 268 samples in decoding surprise, which was discussed in the *Samples regime*.

269 The time point at which the jump in the decoding power occurs in the *Interval*
 270 *regime* is expected to be located on the temporal axis at an earlier time than the point
 271 in the *Samples regime* with the highest decoding power. This is because the *Interval*
 272 *regime* tends to accumulate evidence about the surprise from the beginning of the
 273 epoch to the present target time, and hence feeds its surprise regressor with a larger
 274 amount of information compared to the decoder in *samples regime*, which only uses
 275 a single point in time to decide whether a surprise has occurred. Nevertheless, the
 276 fact that the *Samples regime* is able to identify one single time point in the Middle
 277 segment which can determine whether a surprise has occurred with significance (for
 278 any of the three definitions of surprise) is itself a remarkable observation in our
 279 study. The gradual increase in the decoding power of the *Interval regime* eventually

280 matches that of the *Entire epoch regime* as more time points are recruited to
281 accumulate evidence for the occurrence of surprise (compare the maximum decoding
282 power of Figs 6 and 7.b with the first row of Table 1).

283 **Significance of temporal features**

284 Based on the calculated R^2 values achieved by the different temporal samples in
285 decoding surprise, here we discuss the relative significance of these samples or
286 intervals in revealing whether a surprise has occurred.

287 Table 2 shows the p-values of the t-test for comparing the decoding powers of the
288 three segments of *Early*, *Middle*, and *Late*, employing data from the different
289 subjects as the statistical samples.

290 For EEG experiments, it can be seen from this Table that there is a statistically
291 significant distinction between the decoding powers of the *Early* and *Middle*
292 segments, and that of the *Early* and *Late* segments (p-values < 0.05). However, we
293 did not observe any significant difference in the R^2 value of the *Middle* and *Late*
294 segments.

295 For MEG experiments, the Table indicates that there is a statistically significant
296 distinction between the decoding powers of the *Early* and *Middle* segments (p-values
297 < 0.05). However, we did not observe any significant difference in the R^2 value of
298 the *Middle* and *Late* segments and the R^2 value of the *Early* and *Late* segments. This

299 is because Late components in MEG are not as powerful as the ones in EEG data for
300 decoding surprise.

301 As indicated in Fig 1, for the *Samples regime* which examines the significance of
302 each sample time point in decoding surprise, two main local maxima occur, one in
303 each of the *Middle* and *Late* segments. The first peak always has a higher power in
304 decoding surprise, and its power is noticeably higher in decoding the Shannon
305 surprise. The global maximum of the decoding power curves across time always
306 occurs in the *Middle* segment, reflecting the relative importance of components like
307 N200 in decoding surprise. For the MEG data, Fig 3.a illustrates power of the Middle
308 components in decoding surprise.

309 In addition, in the monotonically ascending R^2 curves of Figs 2 and 3.b, which
310 illustrate the decoding powers of *Intervals* (i.e. using all temporal samples in a range
311 from the beginning of the epoch to the present target time), a considerable increase in
312 decoding power occurs (only) in the Middle segment, and the curves asymptotically
313 saturate afterwards. This is true for all quantifications of surprise. In fact, these
314 curves indicate that there is a range of temporal samples in the Middle segment
315 which possess a critical role in describing surprise, to the extent that adding samples
316 from later times (e.g. points in the Late segment) does not noticeably enhance the
317 surprise decoding power. In other words, recruiting only the temporal samples in the
318 [100ms, 250ms] interval seems to suffice for describing the theoretical surprise in
319 both the puzzlement and enlightenment senses. To the extent of our knowledge, this

320 observation has not been reported earlier, and furthermore, our analysis is the first to
321 quantify this sufficiency through rigorous statistical analysis and for the different
322 quantifications of surprise.

323 **Comparison of surprise quantifications**

324 Here we compare the relative powers of the three quantifications of surprise which
325 were employed to create labels for training decoders.

326 When the *Entire epoch*, or each of the *Segments of Middle* or *Late* are used as
327 temporal segments, our results (Table 1) show a clear advantage for the Shannon
328 surprise as the most decodable surprise in terms of the model's average R^2 value,
329 while the Bayesian surprise is the least decodable surprise. The results for the *Early*
330 *segment* are not used in this comparison as the decoding powers are close to the
331 chance level, reflecting the very low amount of useful information in this segment
332 for decoding surprise.

333 Furthermore, for the *Entire epoch*, or the *Middle* or *Late* temporal segments, in the
334 EEG data, the superiority of the Shannon surprise over the other two definitions of
335 surprise is the only significant difference (with p-value less than 0.05) (Table 2). In
336 other words, the statistical difference between the Bayesian and the Confidence-
337 corrected surprises is not significant, while a significant difference exists between
338 the R^2 values of the Shannon surprise and the Bayesian surprise, and those of the
339 Shannon surprise and the Confidence-corrected surprise (Table 3). For a similar

340 temporal segmentation of the epoch for the MEG analysis, the only significant
341 difference occurs between the Shannon surprise and the Bayesian surprise, in favor
342 of the Shannon surprise.

343 In the case of the *Samples* temporal point selection regime, the Shannon surprise is
344 significantly better decoded than the Bayesian surprise in both visual and auditory
345 tasks of the EEG analysis and in the MEG analysis, while none of the other
346 differences in the R^2 value is significant (Figs 5 and 6.a).

347 Finally, in the *Intervals* regime of the EEG analysis in which temporal sample points
348 are added progressively, all of the two-way comparisons between the decoding
349 powers of the three surprise definitions are nearly significant (Fig 4), and adding
350 every temporal component does not change the order of R^2 for these three surprise
351 values in the visual task. The only exception is in the visual task in that the
352 distinction between R^2 values of the Bayesian and Confidence-corrected surprise is
353 not statistically significant (Fig 4).

354 In Fig 6.b, for the *Intervals* regime of the MEG analysis, the only significant
355 difference belongs to comparing the Shannon and Bayesian surprise values.

356 **The effect of integration coefficient**

357 In this part, we will examine the effect of the integration coefficient w on the
358 decoding power of the three quantifications of surprise for the two regimes of
359 *Intervals* and *Samples*.

360 In Figs 7 and 8 the decoding powers of the designed decoder are plotted for different
361 integration coefficients in the range of [1,100]. The decoding power is shown by
362 colors defined on the right.

363 Two local maxima are again observed for the *Samples* regime in the Middle and Late
364 components. In the *Intervals* regime, the decoding power constantly increases in time
365 as expected, and hence the color becomes bolder.

366 Two different behaviors can be observed for the three surprise quantifications. For
367 the Shannon and Confidence-corrected surprise values, when w is not small a
368 relatively high decoding power is observed. However, for the Bayesian surprise w
369 needs to be relatively small in order to obtain high decoding powers.

370 According to the Bayesian definition of surprise, after the brain has observed enough
371 stimuli to correctly learn the generative distribution, there will not be a high amount
372 of surprise in response to any type of stimuli because at that time the difference
373 between the two distributions involved in the calculation of the KL divergence
374 diminishes. In addition, for larger integration coefficients, estimates of the
375 underlying parameter θ will be closer to its true value. This improvement enables the
376 Shannon and Confidence-corrected surprises to produce more accurate predictions
377 for the brain surprise

378 In addition, in the *Samples* regime of the EEG analysis, the best integration
379 coefficient is not much dependent on time. In other words, the best w is not much
380 different for the middle and late components (Fig 7).

381 Discussion

382 In this study, we aim to examine how the surprise of an ideal observer is reflected in
383 the temporal data recorded by EEG and MEG systems, and which surprise model can
384 be best described by these temporal measurement samples. On the basis of the
385 Bayesian brain assumption, an ideal observer is described as an observer who
386 attempts to estimate the generative distribution of the incoming stimuli. The
387 estimated distribution is updated after receiving each stimulus. The distribution is
388 parameterized in binary oddball experiments using a transition probability matrix
389 (e.g. as in [13]).

390 Based on the estimated probability of the stimulus, we have calculated three kinds of
391 surprise values in the present work to predict how surprising receiving each stimulus
392 could be. The three quantifications of surprise employed in our study are: Shannon,
393 Bayesian and Confidence-corrected surprise values.

394 After calculating the surprise values based on the incoming stimuli sequence, we
395 employed a regression model with a subset of the temporal ERP components as its
396 input features, and one of the surprise values as its labels, to estimate the surprise of
397 the brain's response. We utilized the decoding model of Modirshanechi, 2019 to

398 extract the surprise of the response from the input feature vector. There is a
399 difference in the decoder we employed and the one introduced by Modirshanechi,
400 2019. We have not used the second ordinary regression block in [15] to acquire the
401 temporal density of the significant coefficients and instead, the values of the different
402 temporal responses are assessed in our study through the four different defined
403 sampling regimes.

404 Our results indicate that in both the EEG and MEG data, the Shannon surprise is
405 always significantly better decoded than the Bayesian surprise for all temporal
406 sampling regimes, while it is significantly better decoded than the Confidence-
407 corrected surprise only for the Middle, Late, or Entire temporal segments of the EEG
408 data. As for comparing the temporal parts of the recorded data in decoding surprise,
409 in EEG both the middle and late segments yield high decoding powers not
410 significantly different from each other, while in the MEG data only the Middle
411 components have relatively high decoding powers. Generally, for both data types,
412 the most powerful decoding time instance always resides in the Middle segment,
413 suggesting a good placeholder for the MMN readout. In the EEG data, the best time
414 point in the Late segment, while not as powerful as the best point in the Middle
415 segment, serves as a time point for extracting the P300 amplitude. In the MEG data,
416 the Late components are not powerful enough for decoding surprise. In addition, the
417 results show that in surprise decoding, augmenting the Middle segment with

418 components of the Late segment does not improve the decoding power noticeably.

419 These observations are true for all three definitions of surprise used in this study.

420 **Evidence for Bayesian brain and ideal observer**

421 In this work, we have constructed our decoding model based on the Bayesian model
 422 of the brain by presuming a generative model for the world [39-43]. The brain is
 423 assumed to update its perception of the world after receiving inputs according to the
 424 Bayesian updating model. Our surprise decoding approach assumes an ideal observer
 425 model for the brain, which postulates that the brain attempts to find the distribution
 426 from which the input sequence is generated (parameterized by a transition
 427 probability matrix), and updates its estimate of this distribution using the Bayes rule
 428 [12-15, 27, 29, 38]. Therefore, the assumptions of the Bayesian brain and the ideal
 429 observer are embedded in the way we have calculated the theoretical surprise of each
 430 stimulus. Although there are three different approaches for defining this surprise, all
 431 are based on the parameters learned following the Bayesian brain and ideal observer
 432 assumptions.

433 Our results demonstrate the feasibility of decoding these three quantifications of the
 434 theoretical surprise on a trial-by-trial basis with noticeable decoding power, and
 435 hence provide new evidence for the Bayesian brain and ideal observer assumptions.

436 **Finding the best model for surprise**

437 Studying the phenomena associated with the brain's surprise response is an
438 important aspect of modeling the process of learning in the brain. In these models,
439 surprise has been regarded as a parameter which the brain tries to minimize during
440 learning [14, 16, 17]. One of this paper's objectives has been to propose an approach
441 to best quantify surprise as a parameter varying with every incoming stimulus that
442 fits the recorded data. Finding the best description of surprise can offer evidence on
443 whether the temporal components of EEG (or MEG) data more strongly reflect the
444 unlikelihood of the input's value or corroborate the underlying belief updating
445 processes.

446 The Shannon surprise is a quantification that only considers how unlikely the
447 appearance of a certain input value is. It does not consider the estimated probability
448 of receiving other values for the stimulus. The Bayesian quantification of surprise
449 considers the whole distribution that generates the stimulus and updates estimates of
450 its parameters before and after the arrival of each stimulus.

451 The Shannon surprise was shown earlier to better correlate with the P300
452 fluctuations of the response than the Bayesian surprise [12]. Our work demonstrated
453 that the Shannon surprise fits (statistically) significantly better with the data than the
454 Bayesian surprise when the entire epoch or certain segments of the recorded data are
455 used to decode surprise, and that this advantage exists in both auditory and visual
456 sensory modalities in EEG records as well as the auditory modality in MEG data.

457 The superiority of the Shannon surprise can be explained by noting that the
 458 underlying processes associated with the different manifestation of surprise in EEG
 459 recording modulate the activity of different brain regions [21-23]. The Shannon
 460 surprise has been associated with the activity of the posterior parietal cortex (human
 461 lateral intraparietal area), while the Bayesian surprise has been linked to updating
 462 processes which occur in the anterior cingulate cortex [22]. As the former activity
 463 occurs closer to the brain's surface beneath the scalp, its presence may be more
 464 strongly reflected in the EEG data and hence, be better decoded.

465 Furthermore, we are using a “Bayesian sampler” approach to describe the behavior
 466 of the brain, employing a limited sample size to estimate the parameters of the input
 467 distribution [44]. Hence, a possible explanation for the significant difference
 468 between the decoding powers of the Shannon and Bayesian surprises might be the
 469 effect of limited sample size when estimating the parameters involved in these two
 470 models. Further analysis may reveal the particular differences stemming from
 471 estimating the mean of a distribution (Shannon model) from a sequence of samples
 472 versus comparing an entire distribution to its update (Bayesian update) when a
 473 limited number of previous observations is used. This analysis can be the subject of
 474 future work related to our results.

475 **A general consideration of the definition of surprise**

476 It is important to distinguish between the surprise in the recorded response of the
 477 brain and the underlying surprise of the sequence of the input stimuli. The former is

478 studied extensively as a parameter to be extracted from the recorded data; for
 479 example, in MEG [38, 45], and in EEG, as the amplitude of the P300 component [13,
 480 20, 24, 25, 46], or as the MMN component [27, 28]. The latter terminology for
 481 surprise is extracted from the sequence of input stimuli based on an assumption
 482 about how the brain estimates the parameters of the input stimuli distribution (for
 483 example, the ideal observer assumption used in this work). This surprise is hence not
 484 dependent on the actual recorded data from the brain. There are two major
 485 approaches for quantifying the surprise of the stimuli which have been reported to be
 486 related to the surprise of the brain: the Shannon surprise and the Bayesian surprise.

487 Aside from these two well-known models for the stimuli surprise, we have also
 488 considered a new approach of quantifying the surprise of the stimuli called the
 489 Confidence-corrected surprise [14], the correlation of which with the EEG recorded
 490 data has not yet been examined (to the best of our knowledge). The dataset we
 491 selected for this work contains only two kinds of stimuli; therefore, if an observer
 492 considers the probability p for the deviant stimuli, the probability of the other
 493 stimulus type will be $1 - p$. Hence, the commitment parameter defined by Faraji,
 494 2018 will not show its importance in our study. Another study with more than two
 495 stimuli types might be able to compare the power of this new definition of
 496 puzzlement surprise with that of the Shannon surprise.

497 A remarkable distinction of our work is that we have not considered any single EEG
 498 (or MEG) temporal component as a representative for the surprise of the brain.

499 Extracting a reliable value from each epoch (even after epoch averaging, which is a
500 common practice in ERP analysis) is a complex and rather ad-hoc procedure [25, 47-
501 50]. In our approach, we use the entire recorded response to derive the surprise of the
502 brain. Similar approaches to decoding the brain's surprise have been reported
503 recently in the works of Modirshanechi et al. (2019) and Maheu et al. (2019). In our
504 analysis, for every subject, every sensory modality, and each of the three definitions
505 of theoretical surprise, we have calculated the surprise of the brain as a linear
506 combination of selected components of the recorded data.

507 The outcome of our data-driven approach provided new evidence for the value of the
508 MMN and P300 components of the EEG epoch as placeholders for the brain's
509 surprise when decoding any of the three quantifications of surprise. The importance
510 of samples in the Middle and Late components of the ERP response in reflecting the
511 brain's surprise was established through rigorous statistical analysis of their
512 decoding power for all three definitions of theoretical surprise.

513 **A general consideration of the location of surprise**

514 Ostwald et al. (2013) postulate that the secondary somatosensory cortex represents
515 the processes associated with the Bayesian surprise. Also, based on fMRI data
516 analysis, O'Reilly et al. (2013) and Schwartenbeck et al. (2016) associate the
517 Shannon and Bayesian surprises with the activities of the posterior parietal cortex
518 and the anterior cingulate cortex, respectively. In addition, the well-known surprise-
519 related components of the ERP signal such as the MMN and P300 have been shown

520 to emanate from the right frontal cortex and the fronto-central regions of the brain,
 521 respectively [24, 51]. In the current study, we have not performed any spatial
 522 selection among the EEG (or MEG) electrodes or among the ICs with spatial
 523 distributions close to the known sources of the different surprise processes in the
 524 brain. While this choice offers generality to our analysis through employing all
 525 available data and letting the decoders capture all the relevant information during the
 526 training procedure, a more targeted approach can also be pursued based on selecting
 527 subsets of either the electrode channels or the ICs which are spatially associated with
 528 the brain regions known to modulate the Shannon or Bayesian surprise mechanisms.
 529 Such an extension to our work might reveal whether the Bayesian theoretical
 530 surprise can achieve a higher decoding power than the Shannon surprise in decoding
 531 the recorded EEG (or MEG) data in spatially-selected configurations of channels or
 532 ICs.

533 **MMN versus P300 for describing surprise**

534 Earlier studies have used one of the definitions of surprise to describe the trial-by-
 535 trial fluctuations of one of the MMN or P300 components [12, 27, 29]. We have
 536 considered the role of both of these components with each of the different
 537 quantifications of theoretical surprise. We observed that both the Middle and Late
 538 components, containing the MMN and P300, respectively, have noticeable surprise-
 539 decoding power even when considered separately. However, while we did not
 540 observe a significant difference between the decoding powers of the Middle and Late

541 components, our analysis revealed that the Middle components of the ERP can
542 decode the surprise better than the Late components for all three definitions of
543 surprise. As a result, when using only one component of the ERP signal, the MMN
544 value fits better than P300 to describe the stimuli's theoretical surprise. An extension
545 to this study could be to evaluate whether a decoder trained using the MMN
546 component can efficiently decode the occurrence of a P300 component [52], or vice
547 versa.

548 **Optimizing the timescale of integration**

549 The timescale of integration, or the integration coefficient denoted by w in our
550 model, is a parameter representing the number of recent stimuli values kept
551 effectively (with considerable coefficients) in the memory for estimating the
552 distribution that generates the stimuli [13, 15, 24, 25, 53]. The timescale of
553 integration works between the trials. For each subject and for each of the definitions
554 of surprise, we optimized for the timescale which provided the best decoding power
555 for each of the temporal sampling regimes.

556 According to the results of our optimization, a large timescale of integration is not
557 suitable for describing the Bayesian surprise. In other words, the best description for
558 the Bayesian surprise derived from the brain's response occurs when a rather short
559 window of integration is used. This behavior stems from the very definition of the
560 Bayesian surprise, which measures the difference between the distribution estimated
561 for the parameters of interest before and after the arrival of each stimulus. This

562 divergence amount constantly decreases as we increase the timescale of integration
563 since the two distributions become closer to each other.

564 Given the rather short window of integration involved in keeping track of the
565 Bayesian surprise, this quantification of surprise tends to be more sensitive to the
566 fluctuations of the recorded data compared to the Shannon and Confidence-corrected
567 surprises, which use longer windows of integration, making them more robust to
568 such fluctuations.

569 Considering the oddball tests used in our study as examples to make this case, if the
570 sequence of stimuli is generated by a random process with fixed statistics (e.g. the
571 probability of the occurrence of a deviant), a longer window of integration allows for
572 estimating the probability of the deviant stimuli with higher accuracy, and hence is
573 optimal for offering a high decoding power.

574 However, in the case of the Bayesian surprise, on the one hand a long integration
575 window results in diminished divergence in the consecutive updates and hence
576 cannot produce surprise values at par with the brain's data. On the other hand, a short
577 integration window tends to follow local fluctuations in the statistics and not provide
578 accurate estimates of the underlying statistics of the input sequence generation
579 process. This lack of performance for both long and short timescales of integration
580 may offer an explanation for the inferior decoding power of the Bayesian surprise
581 across all of our experiments.

582 **Motor response effect**

583 In this work, we have studied the EEG recordings with a motor response by the
 584 subjects to the deviant stimuli. In order to examine whether the motor response of the
 585 participants had a confound effect on the decoding result, here we analyze another
 586 auditory EEG dataset which includes no motor responses. This analysis allows us to
 587 evaluate the power of our decoder when motor responses are not part of the oddball
 588 task. Earlier, the work of Modirshanechi et al., 2019 reported no significant
 589 difference between the decoding powers of pairs of oddball EEG recordings as
 590 follows: 1) Between the decoding power of the dataset with asymmetric motor
 591 response (the EEG dataset used in our study in which subjects pressed a button only
 592 in response to the deviant stimuli) and the decoding power of another EEG dataset
 593 with symmetric motor responses (in which the subjects pressed a button for both the
 594 standard and deviant stimuli); 2) Between the decoding powers of the dataset with
 595 asymmetric motor response and another EEG dataset with no motor response.
 596 Despite this observation, it is still probable that the strong decoding power of the late
 597 components of the EEG data may be due to the motor responses even though the
 598 surprise can still be decoded based on the late components that do not have any
 599 motor response confound.

600 To address this, we applied our decoding methodology to the EEG dataset with no
 601 motor response in the *Samples* regime to evaluate the decoding power of the Late
 602 components in comparison to that of the Middle components.

603 This dataset includes four different auditory oddball tasks and is openly accessible at
 604 www.bnci-horizon-2020.eu/database/data-sets with the name “auditory oddball
 605 paradigm during hypnosis”. There are data from two healthy subjects, one female
 606 and one male, in this dataset. Two of the tasks are performed in normal awake
 607 condition, and during the other two tasks the subjects were hypnotized. We used the
 608 data in which the subjects passively listen to the auditory stimuli.

609 The subjects were presented a random sequence of 420 short complex high tones
 610 (440 + 880 + 1760 Hz) and 60 short complex low tones (247 + 494 + 988 Hz). The
 611 duration of each stimulus was 50 ms. The EEG data were recorded with a sampling
 612 rate of 512 Hz using active electrodes in 27 channels and two pairs of electrodes
 613 recorded EOG (Electrooculography).

614 The steps of EEG preprocessing described in Materials and Methods were similarly
 615 applied of this dataset. The dimension of each feature vector was $p = 80 \times$
 616 $27 = 2160$ (80 samples in every epoch multiplied by the number of channels), and
 617 after removing noisy epochs approximately $N = 440$ features remained.

618 Fig 9 depicts the decoding power of each temporal sample attained using our surprise
 619 decoding method regressed the Shannon surprise. It can be observed that the Late
 620 component responses still possess a strong decoding power comparable to the
 621 decoding power of the Middle component responses. It can be hence deducted that

the relatively strong decoding power of the Late components of the main EEG dataset in our analysis plotted in Fig 1 is not due to the motor response confound.

Fig 9. The power of every single temporal Sample for decoding Shannon Surprise in both active and passive tasks of benchmarking EEG dataset.

MEG and EEG comparison

Better signal-to-noise ratio and readability of the MEG recording compared to EEG [35, 54] offer opportunities for further examination of the mechanisms that generate surprise in the brain. A larger number of recording sensors distributed more densely across the head provides better coverage of local activity beneath the scalp. In our study, we observed that the Late components of the MEG epoch had noticeably less decoding powers than the Middle components. This was a distinctly different behavior from the EEG analysis results in which both the Middle and Late components of an epoch offered significant decoding powers. A likely explanation

for the lower performance of the Late components in the MEG analysis can be that while each EEG sensor collects and integrates data from a rather distributed set of sources in the brain, each MEG sensor can only capture the activities of sources in its close proximity beneath the scalp. The surprise generation mechanisms of the brain transmit signals to a number of different regions of the brain, which in turn produce the Late components of surprise which are distributed and diffused. Since these Late components are generated by distributed sources, MEG sensors may not be able to adequately capture them [34, 55].

Limitations and extensions

Maheu et al. (2019) argue that the early brain responses to an oddball experiment are best explained by the learning of the deviant-to-standard frequency, while the late responses are best explained by the learning of transition probabilities between the two input types. They further propose the presence of both long-term and short-term sequence learning mechanisms in the brain, with the long-term learning process resulting in the early to middle responses observed in an oddball experiment while the short-term learning process being active during the late stage response period [38].

In the datasets analyzed in our work, the probability of the occurrence of the deviants is fixed and small ($1/5$ in the EEG datasets and $1/3$ in the MEG dataset). This probability, called the item frequency in some literature [13, 38], along with a second statistical parameter, the alternation probability, fully determine the characteristics of

the input sequence generation. To assess the performance of the two-parameter decoding models discussed earlier, experiments in which these two probabilities take different values would be needed. In addition, the decoding model based on the set of two transition probabilities between the two stimuli types can be analyzed using such experiments. The datasets analyzed in our work did not include variations of the statistical parameters and hence allowed for only a partial study on the role of the sequence generation parameters. As a result, we have not compared the decoding power of calculating surprise based on the item frequency with the one based on transition probabilities as in [13, 38]. In order to analyze models based on both the simple parameters like the item frequency as well as models utilizing higher-level parameters like transition probability matrix, a more diversified dataset with higher probabilities for the deviants, and including various alternation and repetitions probabilities would be required. One might analyze the other three blocks of MEG data that was used in this paper to be able to compare the two ideal observers estimating transition probability matrix and item frequency [38]. Furthermore, analyzing surprise on a task including more than two types of stimuli (like the tasks reported in Mars et al. (2008) and Seer et al. (2016)), with parameters more complex than a mere item frequency, can offer new insight into understanding the mechanisms of surprise in the brain.

The decoder utilized in this paper was proposed by Modirshanechi et al. (2019) based on using single-trial EEG (or MEG) data to regress the surprise of the stimuli.

678 To train the decoder, each epoch of the EEG (or MEG) data is used without
 679 considering its temporal order in the experiment. As such, the effect of experience by
 680 the participant during the presentation of each block of trials was ignored in our
 681 analysis. In an extended analysis, one may consider the order of the current stimulus
 682 in selecting the best definition of surprise and extracting the temporal components
 683 with highest decoding powers. Such analysis may reveal advantages for the Bayesian
 684 quantification of surprise by taking into account the learning process that occurs
 685 during the experiment, as the brain learns the generative distribution of the incoming
 686 sequence of stimuli.

687 **Materials and methods**

688 Fig 10 provides an overview of the overall flow of data and the decoding approach
 689 used in our analysis.

690

691 **Fig 10. The overall diagram of the decoding model.** a) The scheme of the decoding model and
692 machine learning tools. b) The processes performed on the recorded raw EEG to acquire features for
693 regression. c) Surprise calculation module using the oddball sequence of stimuli as input and
694 generating labels for training the regression model.

696 Dataset and task

697 **EEG dataset.** Our analysis is applied to two EEG oddball test datasets with visual
698 and auditory tasks. Each dataset contains simultaneously recorded EEG and fMRI
699 signals and has been published [56-58]. We applied our analysis to the EEG signals
700 of both the auditory and visual datasets. For further clarity and in order to better
701 motivate our results, the unique features of these datasets relevant to our study are
702 briefly described below.

703 In both sensory modalities, the rare stimuli occurred with a probability of 0.2 and the
704 repeated stimuli occurred with a probability of 0.8 in three different runs for each

705 participant, each run consisting of a sequence of approximately 125 stimuli. Each
 706 stimulus lasted about 200 milliseconds. The interval between two consecutive stimuli
 707 had a uniform distribution between 2 and 3 seconds. The visual stimuli types
 708 consisted of a small green circle as the standard and a large red circle as the deviant.
 709 In the auditory mode, the standards were single tones of 390 Hz and the deviants
 710 were generated as a broadband “laser gun” sound. The first two stimuli of each run
 711 were always fixed to be standards.

712 The EEG signal was sampled at a frequency of 1000 Hz by a conventional MRI-
 713 compatible EEG system using bipolar electrode pairs in 49 different channels, re-
 714 referenced to a 34-electrode space. The precision of recording was $1\mu\text{V}$.

715 ○ There were 17 participating adults in the experiments with labeled age and
 716 gender (six females, eleven males; age mean: 27.7 years, age range: 20-40
 717 years). All participants took part in both auditory and visual tasks. They
 718 were asked to press a certain button only at the time of observing deviants
 719 and their response latency was also recorded.

720 **MEG dataset.** Our study is also applied to the recorded MEG data of an auditory
 721 oddball task recently published by [38] et al., 2019. In this dataset MEG was
 722 recorded for oddball tasks conducted with four different probabilistic models of
 723 stimuli generation. We selected the frequency-biased model for stimuli sequence
 724 generation, for which the statistical model of the resulting oddball sequence is
 725 similar to that of the stimuli sequence of the EEG dataset. In the MEG task, the

standard and deviant stimuli are two different French syllables randomly drawn from a binomial distribution with the probability of the frequent syllable being 2/3 and that of the deviant syllable being 1/3. Each syllable lasts about 200 milliseconds and the interval between two successive stimuli is 1400 milliseconds. Each block consists of 400 to 409 stimuli.

Participants include 11 females and 9 males, aged between 18 and 25. The data of two subjects were removed because of their excessive head movements. To ensure that the participants paid attention to the task, they were asked every 12 to 18 sounds to predict the next stimuli (being a standard or a deviant) using one of two buttons.

The brain activity was recorded by 306 channels (102 magnetometers and 204 gradiometers) with whole-head Elekta Neuromag MEG system using a sampling rate of 1 kHz and a hardware-based bandpass filter of 0.1 to 330 Hz. Raw MEG data was corrected for head movement and bad channels, and was also cleaned from powerline and muscle movement artifacts. Then, a low pass filter below 30 Hz and a 250 Hz down-sampler was applied to the data. Eye blinks and cardiac artifacts were removed using ICA (Independent Component Analysis). Finally, the data was baseline corrected using a window of 250ms before the stimulus onset [38].

Preprocessing

EEG preparation. The raw EEG signal was processed in our analysis through the following steps (Fig 10.b):

746 In the first step, the InfoMax ICA algorithm [59] was applied to the data of each
 747 subject using the EEGLAB toolbox [60]. Our aim was neither the rejection of
 748 artifacts nor selecting among the resulting ICA components. Instead, we assumed a
 749 general model and utilized ICA just in order to separate different independent
 750 sources of the EEG signal, and make the features of the subsequent surprise
 751 regression model as independent from each other as possible. We anticipated that the
 752 regression coefficients of surprise-irrelevant components would be zero, and hence a
 753 sparse model would be a suitable fit as described further below. Data of each
 754 participant for each sensory modality was transformed from the electrode space to
 755 the independent component (IC) space by the ICA method. The EEG signals were
 756 re-referenced to 34 electrodes, so the resulting IC space was also of dimension 34.

757 In the second step, a bandpass filter [0.5-38 Hz] was applied to the transformed EEG
 758 signals to eliminate undesired frequency components. Then, considering an 800ms
 759 period for the acquired EEG data in each epoch, consisting of 200ms baseline before
 760 the stimulus onset and 600ms after the onset, subsampling by a factor of 10 produced
 761 a sample vector of 80 points for each electrode in each epoch (i.e. a trial in the
 762 oddball terminology). Then, the sample vectors of all the 34 electrodes were
 763 concatenated to construct a feature vector with the length of $p = 80 \times 34 = 2720$ for
 764 each epoch. After removing noisy epochs, around $N = 360$ epochs (from a total of
 765 nearly $3 \times 125 = 375$ epochs) were left to be used for training a decoder in the

subsequent steps of the analysis. Therefore, the input of our regression model would be a $N \times p$ feature matrix denoted by $S_{N \times p}$ (Fig 11).

Fig 11. The feature matrix and the surprise label vector used to train the decoding machine.

As will be described later, we defined four different regimes of temporal component selection within each epoch in order to assess the relative values of various time points or temporal segments in the EEG epoch in offering surprise decoding power in the decoding model we used. Hence, not all the 80 time points defined above as feature vector were used in the different regimes, and p , the dimension of the feature vector, varied between the four regimes.

MEG preparation. The MEG data was initially preprocessed by its developer for head movement correction, bandpass filtering, down sampling and ICA-based artifact rejection [38]. However, in order to employ independent sources as features of the regression model, we performed ICA analysis using the entire 306 channels of the MEG data using the FastICA algorithm [61]. We have chosen FastICA for this data because of the high number of channels (306) which could render the InfoMax algorithm excessively slow. We ended up with an average of 69 independent

783 components for the whole sensors using FastICA. Although in some earlier studies
 784 [34, 38] the analysis is done separately on the data of the magnetometers and
 785 gradiometers of the MEG recording, we have run the ICA method on the entire 306
 786 sensor set composed of 102 magnetometers and 204 gradiometers in order to use all
 787 the available information and to take advantage of the two types of sensors that
 788 record different directions of the magnetic field and the spatial gradients [36]. We
 789 also considered the interval of [-200ms, 600ms] as the response period and reduced
 790 the number of samples by downsampling to 80. Therefore, for the MEG data, the
 791 number of features used for training was $N \in [400, 409]$ (equal to the number of
 792 stimuli, which varied between the participants), and the dimension of every feature
 793 was around $p = 80 \times 69 = 5520$ (equal to the number of time samples multiplied by
 794 the number of independent components).

795 **Ideal observer model**

796 A fundamental question in the Bayesian brain literature is how the brain learns the
 797 distribution of the sensory stimuli. The brain is assumed a near-optimal estimator for
 798 the probability of the input sequence based on a generative model with Bayesian
 799 inference [12, 13, 15, 19, 62, 63]. To be more precise, the brain uses prior belief
 800 about the environment, and updates it after each stimulus arrives. In addition, in
 801 order to initialize the inference process, it is presumed that the brain begins with the
 802 assumption of equally probable input types despite exposure to any possible previous
 803 blocks of stimuli [11, 13, 64, 65].

Here, two crucial questions are what exactly constitutes the statistics that the brain attempts to learn from the recent history of observations, and what mechanism is employed by it to arrive at an optimal estimate of this probability. Rubin et al. (2016) proposed that in order to have maximum predictive power for the future stimuli while maintaining a certain complexity level, the best single-parameter model would be to describe the past sequence of stimuli with the number of occurrences for every stimulus type. For example, in a binary oddball paradigm task, the number of standard (or deviant) stimuli up to a certain time is a sufficient statistics deduced as a single-parameter representation of the past for making an estimate for the probability of the next stimulus being a standard or a deviant [19].

Transition probabilities. In an oddball experiment, every stimulus can be denoted by a binary random variable x^i for $i = 1, \dots, T$, where T is the length of the stimuli sequence. We consider $x^i = 0$ if the i^{th} stimulus is standard and $x^i = 1$ otherwise. This variable follows a Binomial distribution with parameters p_0 and $p_1 = 1 - p_0$ as the probabilities of the standard and deviant stimuli, respectively. Based on the hypothesis that the sequence of items has been generated by a “Markovian” generative process, the sequence can be modeled by the probabilities of transition between the stimuli types. For a binary oddball sequence, the transition probabilities can be stated as a 2×2 matrix, which can be estimated by counting the number of successive transitions [13]. It has been demonstrated that utilizing the transition probability matrix for describing the stimuli sequence will statistically outperform

the single-parameter approach to describe the brain's response [13]. For a binary oddball sequence, the definition of the model parameter θ can be stated in the form of a 2×2 matrix:

$$\theta \triangleq \begin{bmatrix} p_{0|0} & p_{0|1} \\ p_{1|0} & p_{1|1} \end{bmatrix};$$

where $p_{a|b}$ is the probability of transition from stimulus type a to stimulus type b . Since the sum of each column of this matrix is equal to 1, we can reduce the model parameter's definition to a vector θ :

$$\theta = \begin{bmatrix} \theta_{0|1} \\ \theta_{1|0} \end{bmatrix} = \begin{bmatrix} p_{0|1} \\ p_{1|0} \end{bmatrix}.$$

Based on this definition, the likelihood of a sequence of observations $\mathbf{X}^j$ with length j will be:

$$p(\mathbf{X}^j | \theta^j) = 0.5 \left(\theta_{0|1}^{n_{0|1}^j} (1 - \theta_{0|1})^{n_{1|1}^j} \right) \left(\theta_{1|0}^{n_{1|0}^j} (1 - \theta_{1|0})^{n_{0|0}^j} \right); \quad (1)$$

where θ^j with elements $\theta_{0|1}^j$ and $\theta_{1|0}^j$ is the estimated parameter vector after receiving j inputs denoted by the vector $\mathbf{X}^j$, the probability of the first stimulus is assumed to be $1/2$, and $n_{a|b}^j$ is the number of transitions from stimulus type a to stimulus type b in the j observations up to the present sample.

The parameter $n_{a|b}^j$ can be computed in different ways [13]. In this paper we have selected the leaky integration method to account for earlier observations. In this method, the most recent stimulus is given a maximum weight and the weights of the

843 preceding observations decrease exponentially with a parameter w (the integration
844 coefficient) moving backwards toward earlier observations.

845 Eq. (1) is the product of two Binomial distributions, each representing one of the two
846 elements of the vector θ . Using the Beta distribution notation to represent the prior
847 probability of these elements as the conjugate prior of Binary distribution, the
848 posterior distribution of θ^j after j inputs will be the multiple of two new Beta
849 distributions:

$$850 \quad p(\theta^j | \mathbf{X}^j) = \text{Beta}(1 + n_{0|1}^j, 1 + n_{1|1}^j) \text{Beta}(1 + n_{1|0}^j, 1 + n_{0|0}^j); (2)$$

851 To sum up, the posterior probability of the stimulus-generating Binomial distribution
852 parameter is obtained with a two-dimensional descriptor parameter in Eq. (2). The
853 next step is to use this equation to calculate the theoretical surprise inherent in the
854 stimuli sequence.

855 **Surprise definition and calculation**

856 Our analysis compares three definitions of surprise for their suitability in modeling
857 the brain's response in binary oddball experiments. In our decoding model, surprise
858 labels $Y_{N \times 1}$ are employed as labels in the process of training the decoder (Fig 10.c,
859 and 12).

860

Fig 12. Calculating the three different quantifications of surprise.

861

862 Estimation of the stimulus-generating distribution leads to a prediction about the next
863 stimuli, which if violated may produce a “surprise” response by the brain [12, 13, 15,
864 27] reflecting the prediction error [19, 27].

865 **Shannon surprise.** The Shannon surprise provides a measure of surprise based on
866 the predicted probability for an observed event [10, 11, 13, 15]. The more unlikely
867 an observation is, the more surprising it will be. The Shannon surprise can be
868 computed for the case of the j^{th} stimulus having a value c :

869
$$\text{Shannon surprise}(x^j = c) = -\log_2 \hat{p}_c;$$

870 where c is 0 or 1 in an oddball test and $\hat{p}_c$ is the predicted probability of stimulus j to
871 be c and is obtained using this formula:

872
$$\hat{p}_c = \int p(x^j = c | \theta^{j-1}) p(\theta^{j-1} | X^{j-1}) d\theta^{j-1}.$$

873 For the single-parameter definition of θ (i.e. when θ is chosen as the item
874 frequency), we have:

$$875 \quad p(x^j = 0|\theta^{j-1}) = \theta^{j-1}; p(x^j = 1|\theta^{j-1}) = 1 - \theta^{j-1};$$

876 where θ^{j-1} is the estimated value of θ at the preceding sample. In the two-parameter
877 model, θ is a vector of the two probabilities of transition. It can be shown that the
878 probability of the j^{th} item having a value of 0 or 1 can be estimated using the
879 estimated value of θ at the preceding time (i.e. $\theta^{j-1} = [\theta_{0|1}^{j-1} \theta_{1|0}^{j-1}]^T$) via:

$$880 \quad p(x^j = 0|\theta^{j-1}) = \frac{\theta_{0|1}^{j-1}}{\theta_{0|1}^{j-1} + \theta_{1|0}^{j-1}};$$

$$881 \quad p(x^j = 1|\theta^{j-1}) = \frac{\theta_{1|0}^{j-1}}{\theta_{0|1}^{j-1} + \theta_{1|0}^{j-1}}.$$

882 **Bayesian surprise.** It measures how much difference is made in the estimated
883 generative distribution after observing every stimulus. In fact, the surprise can be
884 calculated using Kullback–Leibler (KL) divergence of two distributions assessed
885 before and after observing every observation [66, 67]. For two-dimensional θ
886 parameter, the surprise after observing j^{th} stimulus is:

$$887 \quad \text{Bayesian surprise} \triangleq \text{KL}(p(\theta^{j-1}|\mathbf{X}^{j-1}), p(\theta^j|\mathbf{X}^j));$$

888 which is the distance between the distributions estimated exactly before stimulus j ,
889 having $j-1$ observations, and after it.

890 **Confidence-corrected surprise.** Assume a choice reaction time task, with four types
 891 of stimuli $[A, B, C, D]$, created two models of the world for two observers. The first
 892 observer believes that all stimuli types are equally likely (i.e. the probabilities of the
 893 four types are $[0.25, 0.25, 0.25, 0.25]$), while the second observer estimated the
 894 distribution as $[0.75, 0.25, 0, 0]$. After receiving an input B , both observers will have
 895 the same puzzlement surprise equal to $-\log_2^{0.25} = 2$ in the Shannon's model of
 896 surprise. But it is obvious that the first person is surprised less than the second one
 897 due to the indifference in the predicted probability of occurrence for all types. In
 898 other words, the first person has no commitment to the world.

899 The most recent definition for surprise, calculating it in the puzzlement sense,
 900 considers how observers commit to their belief about the world, and attempts to
 901 address the above issue [14]. It measures the distance between the distributions
 902 estimated for model parameters after observing all the $j - 1$ samples, and the
 903 posterior belief of a naïve observer for parameters after observing the j^{th} sample, x^j .
 904 The latter is the posterior probability of parameters considering a uniform prior, after
 905 receiving x^j as the first observation. Therefore,

$$\begin{aligned} & \text{confidence - corrected surprise} \\ 906 & = \text{KL}(p(\theta^{j-1} | \mathbf{X}^{j-1}), p(\theta | x^j, \text{uniform prior belief})); \end{aligned}$$

907 where θ can be a number or a vector. For the binary oddball paradigm case, this can
 908 be modeled as:

$$p(\theta|x^j, \text{uniform prior belief}) = \text{Beta}(1 + I_{0,1}^j + I_{1,1}^j);$$

where I_k^j is the identity function (I_k^j is one if $x^j = k$, and zero otherwise) in the single-parameter model. In the two-parameter transition matrix model:

$$p(\theta|x^j, \text{uniform prior belief}) = \text{Beta}(1 + I_{0|1}^j + I_{1|1}^j)\text{Beta}(1 + I_{1|0}^j + I_{0|0}^j);$$

with $I_{k|l}^j$ defined as one if $x^j = k$ and $x^{j-1} = l$, and zero otherwise.

Each of these three definitions of surprise yields a vector $Y_{N \times 1}$, with elements to be used as surprise labels for each of the N trials in training a surprise decoder. The results of using these definitions of surprise will be compared in terms of their ability to characterize trial-by-trial fluctuations of the EEG record after a decoder is trained (Fig 10.c). The analysis is performed for a binary oddball paradigm task in terms of decoding surprise responses, and comparisons between the predictive powers of the surprise definitions are made based on the two-parameter model described earlier.

EEG temporal samples and segments

In each epoch of the oddball task, four basic time intervals are often identified for coarse-level segmentation of the response profile: From -200ms pre-stimulus to time 0 (Baseline), from time 0 to 100ms (Early Components), from 100ms to 250ms (Middle Components), and from 250ms to 600ms (Late Components). In our study, we seek to identify the significant single time instances or time intervals which can best regress the surprise level. Hence, we define four different regimes of sample

929 selection from the temporal data record, and apply our decoder operation in each
 930 regime to find out the most significant set of points in time which can be used to
 931 predict the surprise level of the input stimuli (Fig 13). The four regimes of temporal
 932 data selection are described below. Note that the entire set of data covering one
 933 epoch is 80 points (see “Preprocessing of Data”), and there are 34 channels of data
 934 after the ICA function is applied to the electrode signals.

- 935 • **Entire epoch:** The total response time (-200ms to 600ms) is used for
 936 regression to identify all significant coefficients..
- 937 • **Segments (Baseline, Early, Middle, and Late):** To evaluate the regression
 938 power of the target time interval, only a pre-defined range (for example late
 939 components) is used as input feature vector to the decoder.
- 940 • **Intervals:** To evaluate the significance of an interval of accumulated temporal
 941 samples from the beginning of the epoch to the current target time, the interval
 942 of -200 to time t is used as input to the decoder. This operation is repeated for
 943 all values of t .
- 944 • **Samples:** A single sample at time t is regarded as the decoder’s input, and this
 945 operation is repeated for all values of t . Our study applies each of these four
 946 temporal sample selection regimes to three decoders, each using one of the
 947 three definitions of surprise described earlier. This is done independently for
 948 each subject and for each of the two input modalities.

949 **Fig 13: Different EEG/MEG temporal component selection regimes are used to define feature**
950 **vectors as inputs to the decoder. a)** All temporal components of a trial are used (All). **b)** Pre-
951 defined temporal segments are used (Segments). **c)** Temporal components in the range of $[-200, t]$
952 are used (Intervals). **d)** Each single temporal point t is used (Samples).

953

954 At the end of each decoding analysis task, to judge the resulting R-squared values,
955 we tested the hypothesis that $S_{N \times p}$ and $Y_{N \times 1}$ are independent of each other. This was
956 done by making random permutations in the vector $Y_{N \times 1}$ and acquiring the R-
957 squared value of the resulting regression each time as chance level. This analysis was
958 performed for every subject and every task for each type of surprise.

959 **MEG temporal samples and segments**

960 The definitions of the four regimes of temporal data selection are similar to those for
961 the EEG analysis. The four basic time segments in each epoch of the MEG oddball
962 task are defined as follows: from -200ms pre-stimulus to time 0 (Baseline), from

time 0 to 100ms (Early Components), from 100ms to 250ms (Middle Components),
and from 250ms to 600ms (Late Components).

Acknowledgments

We thank Jennifer M. Walz and her colleagues for making the EEG oddball dataset publicly available [56]. We also thank Maxime Maheu and his colleagues for making the MEG oddball dataset and the Bayesian observer code publicly available [38].

References

1. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*. 1999 Jan;2(1):79.
2. Friston K. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*. 2005 Apr 29; 360(1456):815-36.
3. Doya K, Ishii S, Pouget A, Rao RP, editors. *Bayesian brain: Probabilistic approaches to neural coding*. MIT press; 2007.
4. Kok P, de Lange FP. Predictive coding in sensory cortex. In *An introduction to model-based cognitive neuroscience 2015* (pp. 221-244). Springer, New York, NY.
5. Baldi P. A computational theory of surprise. In *Information, Coding and Mathematics 2002* (pp. 1-25). Springer, Boston, MA.

6. Baldi P, Itti L. Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks*. 2010 Jun 1;23(5):649-66.
7. Itti L, Baldi P. Bayesian surprise attracts human attention. *Vision research*. 2009 Jun 2;49(10):1295-306.
8. Friston K. The free-energy principle: a rough guide to the brain?. *Trends in cognitive sciences*. 2009 Jul 1;13(7):293-301.
9. Seer C, Lange F, Boos M, Dengler R, Kopp B. Prior probabilities modulate cortical surprise responses: a study of event-related potentials. *Brain and cognition*. 2016 Jul 1;106:78-89.
10. Shannon CE. A mathematical theory of communication. *Bell system technical journal*. 1948 Jul;27(3):379-423.
11. Strange BA, Duggins A, Penny W, Dolan RJ, Friston KJ. Information theory, novelty and hippocampal responses: unpredicted or unpredictable?. *Neural Networks*. 2005 Apr 1;18(3):225-30.
12. Mars RB, Debener S, Gladwin TE, Harrison LM, Haggard P, Rothwell JC, Bestmann S. Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience*. 2008 Nov 19;28(47):12539-45.
13. Meyniel F, Maheu M, Dehaene S. Human inferences about sequences: A minimal transition probability model. *PLoS computational biology*. 2016 Dec 28;12(12):e1005260.

14. Faraji M, Preuschoff K, Gerstner W. Balancing new against old information: the role of puzzlement surprise in learning. *Neural computation*. 2018 Jan;30(1):34-83.
15. Modirshanechi A, Kiani MM, Aghajan H. Trial-by-trial surprise-decoding model for visual and auditory binary oddball tasks. *NeuroImage*. 2019 Aug 1;196:30217.
16. Schmidhuber J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*. 2010 Jul 12;2(3):230-47.
17. Friston KJ, Frith CD. Active inference, communication and hermeneutics. *cortex*. 2015 Jul 1;68:129-43.
18. Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G. Active inference: a process theory. *Neural computation*. 2017 Jan;29(1):1-49.
19. Rubin J, Ulanovsky N, Nelken I, Tishby N. The representation of prediction error in auditory cortex. *PLoS computational biology*. 2016 Aug 4;12(8):e1005058.
20. Kolossa A, Kopp B, Fingscheidt T. A computational analysis of the neural bases of Bayesian inference. *Neuroimage*. 2015 Feb 1;106:222-37.
21. Visalli A, Capizzi M, Ambrosini E, Mazzonetto I, Vallesi A. Bayesian modeling of temporal expectations in the human brain. *NeuroImage*. 2019 Nov 15;202:116097.

22. O'Reilly JX, Schüffelen U, Cuell SF, Behrens TE, Mars RB, Rushworth MF. Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*. 2013 Sep 17;110(38):E3660-9.
23. Schwartenbeck P, FitzGerald TH, Dolan R. Neural signals encoding shifts in beliefs. *Neuroimage*. 2016 Jan 15;125:578-86..
24. Squires KC, Wickens C, Squires NK, Donchin E. The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science*. 1976 Sep 17;193(4258):1142-6.
25. Kolossa A, Fingscheidt T, Wessel K, Kopp B. A model-based approach to trial-by-trial P300 amplitude fluctuations. *Frontiers in human neuroscience*. 2013 Feb 8;6:359.
26. Näätänen R. Mismatch negativity (MMN): perspectives for application. *International Journal of Psychophysiology*. 2000 Jul 1;37(1):3-10.
27. Lieder F, Daunizeau J, Garrido MI, Friston KJ, Stephan KE. Modelling trial-by-trial changes in the mismatch negativity. *PLoS computational biology*. 2013 Feb 21;9(2):e1002911.
28. Garrido MI, Kilner JM, Stephan KE, Friston KJ. The mismatch negativity: a review of underlying mechanisms. *Clinical neurophysiology*. 2009 Mar 1;120(3):453-63.

29. Ostwald D, Spitzer B, Guggenmos M, Schmidt TT, Kiebel SJ, Blankenburg F. Evidence for neural encoding of Bayesian surprise in human somatosensation. *NeuroImage*. 2012 Aug 1;62(1):177-88.
30. Nieuwenhuis S, Aston-Jones G, Cohen JD. Decision making, the P3, and the locus coeruleus--norepinephrine system. *Psychological bulletin*. 2005 Jul;131(4):510.
31. Patel AD, Iversen JR, Chen Y, Repp BH. The influence of metricality and modality on synchronization with a beat. *Experimental brain research*. 2005 May 1;163(2):226-38.
32. Barcelo F, Escera C, Corral MJ, Periáñez JA. Task switching and novelty processing activate a common neural network for cognitive control. *Journal of cognitive neuroscience*. 2006 Oct;18(10):1734-48.
33. Duncan CC, Barry RJ, Connolly JF, Fischer C, Michie PT, Näätänen R, Polich J, Reinvang I, Van Petten C. Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology*. 2009 Nov 1;120(11):1883-908.
34. Wacongne C, Labyt E, van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S. Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences*. 2011 Dec 20;108(51):20754-9.

35. Strauss M, Sitt JD, King JR, Elbaz M, Azizi L, Buiatti M, Naccache L, Van Wassenhove V, Dehaene S. Disruption of hierarchical predictive coding during sleep. *Proceedings of the National Academy of Sciences*. 2015 Mar 17;112(11):E1353-62.
36. King JR, Gramfort A, Schurger A, Naccache L, Dehaene S. Two distinct dynamic modes subtend the detection of unexpected sounds. *PloS one*. 2014 Jan 27;9(1):e85791.
37. Symonds RM, Lee WW, Kohn A, Schwartz O, Witkowski S, Sussman ES. Distinguishing neural adaptation and predictive coding hypotheses in auditory change detection. *Brain topography*. 2017 Jan 1;30(1):136-48.
38. Maheu M, Dehaene S, Meyniel F. Brain signatures of a multiscale process of sequence learning in humans. *Elife*. 2019 Feb 4;8:e41541.
39. Knill DC, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*. 2004 Dec 1;27(12):712-9.
40. Behrens TE, Woolrich MW, Walton ME, Rushworth MF. Learning the value of information in an uncertain world. *Nature neuroscience*. 2007 Sep;10(9):1214.
41. Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JJ. Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience*. 2012 Jul;15(7):1040.

42. Nassar MR, Wilson RC, Heasley B, Gold JI. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*. 2010 Sep 15;30(37):12366-78.
43. Mathys C, Daunizeau J, Friston KJ, Stephan KE. A Bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*. 2011 May 2;5:39.
44. Sanborn AN, Chater N. Bayesian brains without probabilities. *Trends in cognitive sciences*. 2016 Dec 1;20(12):883-93.
45. Heilbron M, Chait M. Great expectations: is there evidence for predictive coding in auditory cortex?. *Neuroscience*. 2018 Oct 1;389:54-73.
46. Verleger R, Jaskowski P, Wauschkuhn B. Suspense and surprise: On the relationship between expectancies and P3. *Psychophysiology*. 1994 Jul;31(4):359-69.
47. Amini Z, Abootalebi V, Sadeghi MT. Comparison of performance of different feature extraction methods in detection of P300. *Biocybernetics and Biomedical Engineering*. 2013 Jan 1;33(1):3-20.
48. Turnip A, Hong KS, Jeong MY. Real-time feature extraction of P300 component using adaptive nonlinear principal component analysis. *Biomedical engineering online*. 2011 Dec;10(1):83.

49. Cecotti H, Graser A. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE transactions on pattern analysis and machine intelligence*. 2010 Jun 28;33(3):433-45.
50. Debener S, Makeig S, Delorme A, Engel AK. What is novel in the novelty oddball paradigm? Functional significance of the novelty P3 event-related potential as revealed by independent component analysis. *Cognitive Brain Research*. 2005 Mar 1;22(3):309-21.
51. Giard MH, Perrin F, Pernier J, Bouchet P. Brain generators implicated in the processing of auditory stimulus deviance: a topographic event-related potential study. *Psychophysiology*. 1990 Nov;27(6):627-40.
52. Sassenhagen J, Fiebach CJ. Finding the P3 in the P600: Decoding shared neural mechanisms of responses to syntactic violations and oddball targets. *NeuroImage*. 2019 Jun 20.
53. Huettel SA, Mack PB, McCarthy G. Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. *Nature neuroscience*. 2002 May;5(5):485.
54. Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa OV. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*. 1993 Apr 1;65(2):413.

55. Ilmoniemi RJ, Sarvas J. Brain Signals: Physics and Mathematics of MEG and EEG. Mit Press; 2019 May 3.
56. Walz JM, Goldman RI, Carapezza M, Muraskin J, Brown TR, Sajda P. Simultaneous EEG-fMRI reveals temporal evolution of coupling between supramodal cortical attention networks and the brainstem. Journal of Neuroscience. 2013 Dec 4;33(49):19212-22.
57. Walz JM, Goldman RI, Carapezza M, Muraskin J, Brown TR, Sajda P. Prestimulus EEG alpha oscillations modulate task-related fMRI BOLD responses to auditory stimuli. NeuroImage, 2015.
58. Conroy BR, Walz JM, Sajda P. Fast bootstrapping and permutation testing for assessing reproducibility and interpretability of multivariate fMRI decoding models. PloS one. 2013 Nov 14;8(11):e79271.
59. Bell AJ, Sejnowski TJ. An information-maximization approach to blind separation and blind deconvolution. Neural computation. 1995 Nov;7(6):1129-59.
60. Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods. 2004 Mar 15;134(1):9-21.
61. Haykin S. Neural Networks and Learning Machines, 3/E. Pearson Education India; 2010.

62. Daunizeau J, Den Ouden HE, Pessiglione M, Kiebel SJ, Stephan KE, Friston KJ. Observing the observer (I): meta-bayesian models of learning and decision-making. PLoS One. 2010 Dec 14;5(12):e15554.
63. Friston K. The history of the future of the Bayesian brain. NeuroImage. 2012 Aug 15;62(2):1230-3.
64. Harrison LM, Duggins A, Friston KJ. Encoding uncertainty in the hippocampus. Neural Networks. 2006 Jun 1;19(5):535-46.
65. Bestmann S, Harrison LM, Blankenburg F, Mars RB, Haggard P, Friston KJ, Rothwell JC. Influence of uncertainty and surprise on human corticospinal excitability during preparation for action. Current Biology. 2008 May 20;18(10):775-80.
66. Kullback S. Information theory and statistics. Courier Corporation; 1997 Jul 7.
67. Cover TM, Thomas JA. Elements of information theory. John Wiley & Sons; 2012 Nov 28.

Task: Auditory sample

Task: Visual sample

Task: Auditory interval

Task: Visual interval

Task: Auditory interval

Task: Visual interval

Task: Auditory sample

Task: Visual sample

Heatmap on ω
interval KL2 Auditory transition

Heatmap on ω
sample KL2 Auditory transition

Heatmap on ω
interval PUZ Auditory transition

Heatmap on ω
sample PUZ Auditory transition

Heatmap on ω
interval log2 Auditory transition

Heatmap on ω
sample log2 Auditory transition

Heatmap on ω
interval KL2 Visual transition

Heatmap on ω
sample KL2 Visual transition

Heatmap on ω
interval PUZ Visual transition

Heatmap on ω
sample PUZ Visual transition

Heatmap on ω
interval log2 Visual transition

Heatmap on ω
sample log2 Visual transition

Heatmap of R^2 on ω
interval KL2 4

Heatmap of R^2 on ω
sample KL2 4

Heatmap of R^2 on ω
interval PUZ 4

Heatmap of R^2 on ω
sample PUZ 4

Heatmap of R^2 on ω
interval log2 4

Heatmap of R^2 on ω
sample log2 4

New Feature					
Trials					
Samples of IC 1	Samples of IC 2	Samples of IC 3	Samples of IC 4		Samples of IC 14

Surprise Values

Trials