Robust EEG-based cross-site and cross-protocol classification of states of consciousness

Journal: Brain

Manuscript ID: BRAIN-2018-00220.R2

Manuscript Type: Original Article

Date Submitted by the Author: 03-Aug-2018

Complete List of Authors:
Engemann, Denis; Inria Centre de Recherche Saclay Ile-de-France; INSERM CEA Cognitive Neuroimaging Unit; Institut du cerveau et de la moelle epiniere
Raimondo, Federico; Institut du cerveau et de la moelle epiniere; Universidad de Buenos Aires
King, Jean-Remi; CEA/Saclay; New York University; Frankfurt Institute for Advanced Studies
Rohaut, Benjamin; ICM Research Center, UMRS 975,
Louppe, Gilles; New York University
Faugeras, Frederic; INSERM, ICM Research Center,
Annen, Jitka; Universite de Liege, Giga Research, Coma Science Group;
Centre hospitalier universitaire de Liege, Neurology
Cassol, Helena; Universite de Liege, Giga Research, Coma Science Group
Liege, BE
Gossieres, Olivia; University of Liège, Cyclotron Research Centre, Coma Science Group
Fernández Slezak, Diego; Universidad de Buenos Aires
Laureys, Steven; Universite de Liege, Coma Science Group
Naccache, Lionel; INSERM, ICM Research Center,
Dehaene, Stanislav; INSERM CEA Cognitive Neuroimaging Unit
Sitt, Jacobo; Institut du cerveau et de la moelle epiniere

Subject category: CNS injury and stroke

To search keyword list, use whole or part words followed by an *:
Consciousness < SYSTEMS/DEVELOPMENT/PHYSIOLOGY, Stroke: imaging < CNS INJURY AND STROKE, Traumatic brain injury < CNS INJURY AND STROKE, Anoxia < CNS INJURY AND STROKE, EEG < EPILEPSY AND SLEEP, Coma < CNS INJURY AND STROKE
Title:

Robust EEG-based cross-site and cross-protocol classification of states of consciousness

Denis A. Engemann*1,2,3, Federico Raimondo*3,4,5,6, Jean-Rémi King2,7,8, Benjamin Rohaut3,9, Gilles Louppe7, Frédéric Faugeras3, Jitka Annen10, Helena Cassol10, Olivia Gossières10, Diego Fernandez-Slezak4,5, Steven Laureys10, Lionel Naccache4, Stanislas Dehaene2,11, Jacobo D. Sitt2,3

1 Parietal project-team, INRIA Saclay – Île de France, France
2 Cognitive Neuroimaging Unit, CEA DSV/I2BM, INSERM, Université Paris-Sud, Université Paris-Saclay, NeuroSpin center, 91191 Gif sur Yvette, France
3 Institut du Cerveau et de la Moelle épinière, ICM, PICNIC Lab, F-75013, Paris, France
4 Laboratorio de Inteligencia Artificial Aplicada, Departamento de Computación FCEyN, UBA, Argentina
5 CONICET – Universidad de Buenos Aires, Instituto de Investigación en Ciencias de la Computación, Godoy Cruz 2290, C1425FQB, Ciudad Autónoma de Buenos Aires, Argentina
6 Sorbonne Universités, UPMC Université Paris 06, Faculté de Médecine Pitié-Salpêtrière, Paris, France
7 New York University, 6 Washington Place, New York, USA
8 Frankfurt Institute for Advanced Studies, Frankfurt, Germany
9 Department of Neurology, Columbia University, New York, NY
10 Coma Science Group, GIGA Consciousness, University and University Hospital of Liège, Liège, Belgium
11 Collège de France, Paris, France

*equal contributions

Correspondence to: Dr. Denis A. Engemann,
1 Rue Honoré d'Estienne d'Orves,
91120 Saclay, France
E-mail: denis-alexander.engemann@inria.fr

Correspondence may also be addressed to: Federico Raimondo, E-mail: fraimondo@dc.uba.ar and Jacobo Sitt, E-mail: jacobo.sitt@inserm.fr
Abstract

Determining the state-of-consciousness in patients with disorders-of-consciousness is a challenging practical and theoretical problem. Recent findings suggest that multiple markers of brain activity extracted from the electro-encephalogram (EEG) may index the state of consciousness in the human brain. Furthermore, machine learning has been found to optimize their capacity to discriminate different states of consciousness in clinical practice. However, it is unknown how dependable these EEG signatures markers are in the face of signal variability due to different EEG-configurations, EEG-protocols and subpopulations from different centers encountered in practice. In this study we analyzed a total of 327 recordings of patients with disorders of consciousness (148 UWS and 179 MCS) and 66 healthy controls obtained in two independent research centers (Paris Pitié-Salpêtrière and Liège). We first show that a non-parametric classifier based on ensembles of decision trees provides high and robust out-of-sample performance on unseen data with a predictive AUC of about 0.77 that was only marginally affected when using across various alternative EEG-configurations (different numbers and positions of sensors, numbers of epochs, average AUC = 0.750 +/- 0.014) by enhancing the impact of robust EEG signatures. In a second step, we observed that classifiers based on multiple as well as single EEG signatures features generalize to recordings obtained from different patient cohorts, EEG protocols and different centers. However, the multivariate model always performed best with a predictive AUC of 0.73 for generalization from Paris1 to Paris2 and an AUC of 0.78 from Paris to Liège. Using simulations, we subsequently demonstrate that multivariate pattern classification has a decisive performance advantage over univariate classification as the stability of EEG signatures features decreases as different EEG.
configurations are used for feature-extraction or as noise is added. Moreover, we show that the generalization performance from Paris to Liège remains stable even if up to 20 percent of the diagnostic labels are randomly flipped. Finally, consistent with recent literature, analysis of the learned decision rules of our classifier suggested that markers related to the most robust EEG markers of consciousness relate to dynamic fluctuations in theta- and alpha frequency bands carried independent information and were most influential, which turn out to carry complementary information. Our findings demonstrate that EEG signatures markers of consciousness can be reliably, economically and automatically identified with machine learning in various clinical and acquisition contexts.

Keywords
electroencephalography, disorders of consciousness, biomarker, machine learning, diagnosis

Abbreviations

ECG Electrocardiography
EOG Electrooculography
DOC Disorders of consciousness
UWS Unresponsive wakefulness syndrome
MCS Minimally conscious state
CMD Cognitive motor dissociation
PET Positron emission tomography
fMRI Functional magnetic resonance imaging
AUC  Area under the curve
FFT  Fast Fourier transform
SNR  Signal-to-noise ratio
SVM  Support vector machine
Patients suffering from disorders of consciousness (DOC) demonstrate that it is possible to be awake in the absence of behavioral evidence of consciousness (Laureys et al., 2010). Despite best efforts for consistency, current diagnostic procedures rely on human interaction and are, therefore, error-prone (Rohaut & Claassen, 2018). The degree of misdiagnosis in DOC patients may exceed 40 percent when relying on the clinician’s judgment without standardized behavioral assessments (Schnakers et al., 2009). Even when using diagnostic instruments such as the Coma Recovery Scale-Revised (Giacino, Kalmar, & Whyte, 2004), misdiagnosis can remain high if patients are not assessed repeatedly within a short time window (Wannez et al., 2017). Furthermore, in some cases evidence of conscious processing in these patients can only be obtained using functional neuroimaging where patients sometimes demonstrate willful modulations of their brain activity (Monti et al., 2010; Owen et al., 2006). These patients have been labelled as ‘covert awareness’ or ‘cognitive motor dissociation (CMD)’ patients (Curley, Forgacs, Voss, Conte, & Schiff, 2018; Gosseries, Zasler, & Laureys, 2014; Nicholas D. Schiff, 2015).

Among the DOC one distinguishes the comatose state, the unresponsive wakefulness syndrome (UWS, historically vegetative state), and the minimally conscious state (MCS) (Giacino et al., 2002; Laureys et al., 2010). The presence of eye-opening helps to distinguish UWS patients from comatose ones (Jennett & Plum, 1972). Additionally, MCS but not UWS patients show signs of awareness (i.e. visual pursuit in MCS- and command following in MCS+) (Bruno, Vanhaudenhuyse, Thibaut, Moonen, & Laureys, 2011) while neither achieving functional communication nor object-use. It is nevertheless believed that these patients can have a partial and fluctuating awareness of themselves and their surroundings and are more likely to recover (Faugeras et al., 2018; Luauté et al., 2010), which emphasizes the importance of reliable diagnostic tools.

In the last two decades, non-invasive brain imaging has supplemented behavioral assessments for detection of consciousness. Sleep studies and neurological assessments have early on revealed preferentially altered electroencephalogram (EEG) amplitudes in the delta (2-4 Hz), theta (4-8Hz) and alpha (8-12 Hz)
frequency ranges (Emmons & Simon, 1956; Rosenberg, Johnson, & Brenner, 1977). Positron Emission Tomography (PET) revealed globally decreased glucose uptake in DOC patients as compared to healthy controls (Stender et al., 2014). Several functional Magnetic Resonance Imaging (fMRI) studies have documented disruption of functional connectivity along diverse subcortical and neocortical pathways in DOC patients (Demertzi et al., 2014). Ever since, advances in cognitive science have allowed to infer consciousness from increasingly fine-grained patterns of brain activity. Accordingly, recurrent interactions between higher-order neocortical networks, as well as the morphology and complexity of brain dynamics in response to stimulation have been related to the states-of-consciousness (Casali et al., 2013; Dehaene & Naccache, 2001; Iotzov et al., 2017; Tononi & Edelman, 1998), which has led to various types of putative signatures of consciousness.

Following recent trends in neuroimaging, the increasing number of neural markers of consciousness is likely to be best approached with multivariate pattern analysis (MVPA) (Claassen et al., 2016; J. R. King et al., 2013; Naci et al., 2012). Indeed, machine learning algorithms can be trained to best predict the medical status of individual patients from unknown combinations of physiological markers (Chang et al., 2005). Typically, a classifier is trained to optimally discriminate clinical labels based on brain data. Generalization performance is then assessed by comparing the predictions of the classifier to the actual diagnosis when presented with unseen data. In the absence of independent datasets, cross-validation is performed to estimate the out-of-sample performance by subdividing the data into training and testing sets and averaging over testing set scores. It is, however, noteworthy that cross-validation tends to be too optimistic when sample sizes are small (Varoquaux, 2017; Varoquaux et al., 2016; Woo, Chang, Lindquist, & Wager, 2017), rendering face-value interpretation of scores futile for a significant proportion of neuroimaging studies. Examples of MVPA for the study of DOC patients include the analysis of patterns of resting state fMRI functional connectivity (Demertzi et al., 2015), spectral responses to command following (Cruse et al., 2012; Goldfine, Victor, Conte, Bardin, & Schiff, 2011) and cerebral metabolism to distinguish locked-in patients from UWS (Phillips et al., 2011).
In this context, EEG is particularly interesting as this neurophysiological technique conveys rich temporal information on cognitive operations and can be economically operated in a wide range of situations, potentially enabling bedside or home assessment. The challenge of processing large amounts of EEG-data at scale can nowadays be addressed using automated EEG processing methods (e.g., Engemann et al., 2015; Jas, Engemann, Bekhti, Raimondo, & Gramfort, 2017). However, preferences for cognitive theories and EEG-methodologies are heterogeneous across laboratories, which significantly obstructs the development of large-scale data resources well suited for high-fidelity machine learning. Thus, emerging EEG-markers keep emerging in various local flavors, so far, falling into four conceptual families. Evoked markers are based on time-locked event-related analysis of cognitive experiments. The other families contain markers defined independently from protocols, including, connectivity markers exploiting brain-network interactions, information theory markers capitalizing on information properties of time series and spectral markers quantifying neuronal oscillations or stochastic band-limited dynamics. Yet, the situation is further complicated by the fact that DOC reflect several cognitive and neurological components rather than a single dimension, motivating the consideration of marker-profiles (Bayne, Hohwy, & Owen, 2016; Sergent et al., 2017). In a recent study, using a Support Vector Machine (SVM) classifier, Sitt, King and colleagues (2014) analyzed dozens of EEG-markers from more than 150 high-density EEG recordings during an auditory novelty task. Interestingly, combinations of markers synergistically outperformed single markers. Similarly, using graph-theoretical summaries of alpha-band connectivity, Chennu et al. (2017) presented an alternative SVM-approach cross-validated on 104 patients with severe brain injury (among those 89 with DOC).

Nevertheless, a generalized large-scale attempt for cross-laboratory predictions of state of consciousness in brain-injured patients is missing, and several practical questions remain unanswered: What is the optimal duration for individual EEG-recordings? Which task should the patient undergo, if any? How many sensors should be used, and where should they be located? Can a single machine learning algorithm perform on data from different clinical centers? Do models based on current EEG-markers achieve prospective
generalization on independent data (Woo et al., 2017)? Are single markers sufficiently powerful and when does multivariate classification provide the clearest advantage?

To address these questions, we rigorously probed the robustness and validity of EEG-signatures markers of consciousness. Using the robust Extra-Trees algorithm (Geurts, Ernst, & Wehenkel, 2006) we developed a classifier trained (named “DOC-forest”) to differentiate UWS from MCS patients. This classifier was trained and tested using 28 EEG potential markers of consciousness from 249 patients recorded at the Paris Pitié-Salpêtrière and 78 patients from the University Hospital of Liège. We first show that different EEG-configurations (sensor number, sensor position and numbers of epochs) and experimental protocols (auditory stimulation or resting state) induce significant changes in the distribution and performance of the EEG markers. Yet, we found that the DOC-Forest is relatively immune to such variations by exploiting the information conveyed by reliable EEG-markers. We subsequently demonstrate out-of-sample generalization to two independent datasets: a new cohort of 107 previously not analyzed task-EEG recordings from Paris and 78 resting state EEG recordings from the University Hospital of Liège. Moreover, we show that our DOC-Forest’s generalization performance is decisively superior to univariate markers. Finally, by investigating the influence of individual markers on the decisions of DOC-Forest, we found that alpha band power, theta-band connectivity and time-series complexity carry complementary information about states of consciousness.

Materials and Methods

Ethics Statement

This research project was approved by the ethical committee of the Pitié-Salpêtrière hospital under the code ‘Recherche en soins courants’ (routine care research). All investigations were carried out in accordance with the Declaration of Helsinki on ethical principles for medical research involving human subjects. For the
dataset from the Coma Science Group, the family of the patient gave their informed consent for participation in the study, and the Ethics Committee of the University hospital of Liège approved the study.

Participants

In total, 327 EEG recordings from 268 distinct patients from our expert centers were included in the current study (Table 1). Patients were assessed at variable delays (sub-acute or chronic stage following the brain injury) in order to try clarifying the actual state of consciousness. Clinical assessments were performed at least three times in the Paris dataset and five times in the Liège dataset, in all cases on different days by trained clinicians (see acknowledgments) and included systematically the Coma Recovery Scale–Revised (CRS-R). CRS-R scoring ranges from 0 to 23 and is based on the presence or absence of response on a set of hierarchically ordered items testing auditory, visual, motor, oromotor, communication, and arousal functions (Giacino et al., 2004). According to the best assessment, each patient was diagnosed with unresponsive wakefulness syndrome (UWS) or minimally conscious state (MCS). The data acquisition protocol included, in all centers, multiple clinical assessments and at least one EEG recording. For some patients, several EEG recordings were available, which we later on accounted for by statistical modelling. The number of recordings varied considerably across datasets, however, the ratio of MCS to UWS patients was roughly balanced. Across all datasets more male than female patients were observed. Age distributions were similar, however, the delay from accident was visibly higher for the resting state dataset. Likewise, the distribution of etiologies was different for the resting state dataset while proportions were consistent with the literature.

Table 1 Patient characteristics in the three datasets.

<table>
<thead>
<tr>
<th></th>
<th>Auditory Local Global Task</th>
<th>Resting State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Paris 1</td>
<td>Paris 2</td>
</tr>
<tr>
<td>( n_{(EEG)} )</td>
<td>142</td>
<td>107</td>
</tr>
<tr>
<td>( n_{(patients)} )</td>
<td>98</td>
<td>92</td>
</tr>
<tr>
<td>( n_{(UWS)} )</td>
<td>75</td>
<td>52</td>
</tr>
<tr>
<td>(n_{(MCS)})</td>
<td>676</td>
<td>556</td>
</tr>
<tr>
<td>----------------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Gender ratio (male/female)</td>
<td>2.06</td>
<td>1.93</td>
</tr>
<tr>
<td>Age (M[SD]), in years</td>
<td>46.5[17.8]</td>
<td>45.4[17.7]</td>
</tr>
<tr>
<td>Delay (M[SD]), in day</td>
<td>126.0[372.9]</td>
<td>299.6[823.6]</td>
</tr>
<tr>
<td>Delay (min to max), in days</td>
<td>6 to 2611</td>
<td>8 to 6570</td>
</tr>
<tr>
<td>Anoxia (%)</td>
<td>29.6</td>
<td>30.4</td>
</tr>
<tr>
<td>Stroke (%)</td>
<td>29.6</td>
<td>15.2</td>
</tr>
<tr>
<td>Traumatic brain injury (%)</td>
<td>23.5</td>
<td>28.2</td>
</tr>
<tr>
<td>Other (%)</td>
<td>18.4</td>
<td>29.4</td>
</tr>
</tbody>
</table>

**Experimental paradigm:** In Paris 1 & 2 datasets, task-related EEG signals were obtained using the ‘Local-Global’ protocol (Bekinschtein et al., 2009) designed to study unconscious and conscious auditory processing. In the Liège dataset EEG recordings were task-free (see supplementary materials for details).

**Selection and computation of putative EEG markers of consciousness:** We extracted 28 putative EEG-biomarkers detailed in (Sitt et al., 2014). The markers can be grouped into four conceptual families, i.e., information theory, connectivity, spectral, and evoked response markers (See Table 2). Among several connectivity metrics described in Sitt, King et al. (2014), we only considered the Weighted Symbolic Mutual Information (wSMI) metric in theta frequency band as previous research had suggested that the long-range connectivity patterns theoretically related to consciousness are most robustly and accurately assessed by this metric (King et al., 2013). Note that for the analysis of resting state EEG we did not make use of the evoked response markers as those are only defined for the task used in the Paris datasets. For a detailed description and discussion of the markers see (Sitt et al., 2014).

The markers commonly used in clinical neuroscience are often defined at a general level and can be observed over multiple electrodes, time points or frequency bands. To delineate low-level features, each marker, four summary statistics were computed four summary statistics from each marker (Figure 1). To summarize epochs, we either computed the 80% trimmed mean, or the standard deviation. The sensor dimension was then summarized using a mean or the standard deviation, yielding four combinations in total.
Throughout the manuscript we refer to these marker subtypes as “mean,mean”, “std,mean”, “mean,std” and “std,std” and in figures, for brevity, “m,m”, “s,m”, “m,s”, “s,s”. For a full list and abbreviations see Table 2.

Computation was carried out using a designated Python software library implementing the biomarker extraction functionality from in Sitt, King et al 2014. The extracted markers closely matched the original values and group results for the reference datasets were qualitatively reproduced (Engemann et al., 2015).

Table 2 Potential EEG Biomarkers of consciousness

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Marker</th>
<th>Conceptual Family</th>
<th>Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE Θ</td>
<td>Permutation Entropy</td>
<td>Information theory</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>Kolmogorov Complexity</td>
<td>Information theory</td>
<td></td>
</tr>
<tr>
<td>wSMI Θ</td>
<td>Weighted Symmetrical Symbolic Mutual Information</td>
<td>Connectivity</td>
<td></td>
</tr>
<tr>
<td>α</td>
<td>Alpha PSD</td>
<td>Spectral</td>
<td></td>
</tr>
<tr>
<td></td>
<td>α</td>
<td></td>
<td>Normalized Alpha PSD</td>
</tr>
<tr>
<td>β</td>
<td>Beta PSD</td>
<td>Spectral</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td></td>
<td>Normalized Beta PSD</td>
</tr>
<tr>
<td>δ</td>
<td>Delta PSD</td>
<td>Spectral</td>
<td></td>
</tr>
<tr>
<td></td>
<td>δ</td>
<td></td>
<td>Normalized Delta PSD</td>
</tr>
<tr>
<td>γ</td>
<td>Gamma PSD</td>
<td>Spectral</td>
<td></td>
</tr>
<tr>
<td></td>
<td>γ</td>
<td></td>
<td>Normalized Gamma PSD</td>
</tr>
<tr>
<td>θ</td>
<td>Theta PSD</td>
<td>Spectral</td>
<td></td>
</tr>
<tr>
<td></td>
<td>θ</td>
<td></td>
<td>Normalized Theta PSD</td>
</tr>
<tr>
<td>MSF</td>
<td>Median Power Frequency</td>
<td>Spectral</td>
<td></td>
</tr>
<tr>
<td>SE90</td>
<td>Spectral Entropy 90</td>
<td>Spectral</td>
<td></td>
</tr>
<tr>
<td>SE95</td>
<td>Spectral Entropy 95</td>
<td>Spectral</td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>Spectral Entropy</td>
<td>Spectral</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Contingent Negative Variation</td>
<td>Evoked Task</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>short-latency auditory Potential to the first sound</td>
<td>Evoked</td>
<td></td>
</tr>
<tr>
<td>P3a</td>
<td>mid-latency auditory Potential to the first sound</td>
<td>Evoked</td>
<td></td>
</tr>
<tr>
<td>P3b</td>
<td>mid-latency auditory Potential to the first sound</td>
<td>Evoked</td>
<td></td>
</tr>
<tr>
<td>GD-GS</td>
<td>full contrast</td>
<td>Evoked</td>
<td></td>
</tr>
<tr>
<td>LD-LS</td>
<td>full contrast</td>
<td>Evoked</td>
<td></td>
</tr>
<tr>
<td>LGSD-LDGS</td>
<td>full contrast</td>
<td>Evoked</td>
<td></td>
</tr>
<tr>
<td>LSGS-LDGD</td>
<td>full contrast</td>
<td>Evoked</td>
<td></td>
</tr>
<tr>
<td>ΔMMN</td>
<td>Contrasted MNN (Local Deviant vs Local Standard)</td>
<td>Evoked</td>
<td></td>
</tr>
<tr>
<td>ΔP3a</td>
<td>Contrasted P3a (Local Deviant vs Local Standard)</td>
<td>Evoked</td>
<td></td>
</tr>
<tr>
<td>ΔP3b</td>
<td>Contrasted P3b (Global Deviant vs Global Standard)</td>
<td>Evoked</td>
<td></td>
</tr>
</tbody>
</table>

Note. PSD = power spectral density; GS = global standard; GD = global deviant; LS = local standard, LD = local deviant;

Statistical Analysis

**Classification of DOC from EEG-markers:** Diagnosis was classified based on EEG-markers using a univariate and a multivariate machine learning strategy. To enable comparisons across studies, we also computed model-free performance on single markers as in Sitt, King et al (2014). Performance was assessed using the Area Under the Curve (AUC). For details consider the section *Area under the Curve metric* in the supporting information. For multivariate and univariate pattern analysis, we chose the *Extra-Trees* algorithm (Geurts, Ernst, & Wehenkel, 2006) whose non-parametric design facilitates robust classification. To complement insights from univariate classification, we extracted the so-called variable importance metric from the *Extra-Trees* following best practice recommendations for enhanced interpretability (Louppe, 2014; Louppe, Wehenkel, Sutera, & Geurts, 2013). Accordingly our variable importance scores reflect mutual information between a variable and the diagnosis while conditioning out the other variables. For background information on parameters and model tuning, see the section on *Multivariate Pattern Classification* in the supporting information. To use a common currency when comparing univariate with multivariate marker
performance, we turned single markers into fully functional classification models by using the identical recipe as for the DOC-Forest, effectively only changing the variables features passed to the classifier. This allowed us to predict the probability of DOC diagnosis from single markers using the identical framework as for multivariate analysis.

**Statistical Inference:** We extended our visualizations into hypothesis tests by employing the percentile bootstrap (Efron & Tibshirani, 1993) (see supplementary materials for details). To assess out-of-sample generalization we used two complementary approaches: A conservative validation on independent data (new cohorts, different protocols and laboratories) and cross-validation (see supplementary materials for details).

**Software**

All data were processed using the Python programming language. To simplify preprocessing and feature extraction for machine learning, we developed a designated software library\(^1\) built on top the open source software libraries MNE (Gramfort et al., 2014) and scikit-learn (Pedregosa et al., 2011). The DOC-Forest recipe is publicly available\(^2\) to encourage community efforts into building predictive models of DOC patients’ state of consciousness.

**Results**

Robust detection of state-of-consciousness from EEG signatures

**Multivariate classification of UWS versus MCS diagnosis is robust using across different EEG configurations:** We assessed how each marker discriminated between UWS and MCS patients as a function of EEG configurations using the task-EEG dataset recorded in Pitié-Salpêtrière (Paris 1 dataset). Comparing the cross-validated model-based AUC for each marker with the model-free AUC (Figure S3A in supplementary materials) revealed a tight positive correlation ($\rho_{\text{Spearman}} = 0.94$, 95% CI [0.905, 0.962], $p < 0.001$). This finding

---

1 available at https://github.com/nice-tools/nice

2 URL will be made available upon request and upon acceptance of the article.
suggests that our univariate forest models performed equivalently to the previous model-free method. Considering discrimination performance between UWS and MCS across EEG configurations revealed fluctuations according to marker subtype, conceptual family (Figure S3, Figure S4 in supplement) and variance induced by the EEG configuration itself (cf. Figure 2A, 3B). Indeed, we found a positive correlation between fluctuations of marker estimates and discrimination performance across EEG configurations ($\rho_{\text{spearman}} = 0.46$, 95% CI[0.289, 0.6]), $p < 0.001$, suggesting that performance was more stable when the marker itself was stable too (Figure S3B and S4 in supplementary materials). Interestingly, evoked response markers appeared more severely affected than other markers.

The DOC-Forest classifier exhibited an average performance of AUC = 0.75 (SD = 0.014) and performed better and more robustly than most other markers did individually (Figure 2A-B, Figure S1-S24 in supplementary materials). Subsequent analyses suggested that, on average, over all EEG, the DOC-Forest performance assumed at least the 98th percentile compared to other markers (PercentileDOC-Forest = 0.985, 95% CI[0.981, 0.989]) and its fluctuation over configurations was situated around the 83rd percentile (PercentileDOC-Forest = 0.832, 95% CI[0.796, 0.894]). Moreover, its discrimination performance increased with the number of sensors ($\rho_{\text{spearman}} = 0.803$, 95% CI[0.646, 0.891] $p < 0.001$) and epochs ($\rho_{\text{spearman}} = 0.46$, 95% CI[0.07, 0.668] $p < 0.05$) (Figure 2B) but was already strong with 16 sensors and 5% of epochs. Importantly, using the full EEG configuration, the performance closely resembled previous results reported by Sitt, King et al 2014 and beat any other marker (Figure S24 in supplementary materials). These results suggest that the DOC-Forest preferentially tracks information conveyed by a few robust markers over a variety of EEG configurations.

We subsequently assessed, using the full configuration, the consistency of classification success for different etiological groups and different levels of chronicity (See section Consistency of classification results in diagnostic sub-groups in SM). Comparable results were obtained for the chronic (delay > 30 days) and acute (delay ≤ 30 days) groups. The classification performance was significant for all the
etiology groups (i.e., anoxia, stroke and TBI). Yet, in the case of TBI patients the performance was slightly lower suggesting that the heterogeneity of this group makes it more difficult to classify. For additional fine-grained comparisons between single markers and the DOC-Forest, see section Detailed comparison between individual markers and DOC-Forest in the supplementary materials.

Classification is preferentially driven by distinct theta and alpha band dynamics: While it is not convenient to reason separately about each of the 2000 decision trees grown inside our DOC-Forest, we can still analyze the relative contributions of EEG markers to classification performance by considering the variable importance (See Materials and Methods). This multivariate metric approximates the mutual information between a marker and the diagnosis while controlling for the contribution of other markers. The variable importance can deviate systematically from the univariate AUC whenever information is shared between markers or the model has identified non-linear interaction effects. Inspecting all 36 DOC-Forest classifiers, we observed that markers contributing most strongly on average belonged to different conceptual families (Figure 2C). Specifically, permutation entropy and long-range connectivity also in the theta band as well as alpha frequency band power were top ranked, both, in terms of univariate discrimination and variable importance. In contrast, evoked markers, on average, often assumed values below 0.89%, which is less than would be expected if all markers were equally influential. We observed a positive but non-linear relationship between average AUC and average variable importance ($\rho_{\text{Spearman}} = 0.817$, 95% CI[0.727, 0.880], $p < 0.001$). It can be seen that highly performing markers were disproportionally more important than expected for a linear association (Figure 2C).

These results suggest that the robust performance of our pattern classifier across EEG configurations was enabled by preferentially enhancing the impact of the better performing markers, such as theta wSMI, theta permutation entropy or alpha frequency power. Interestingly, these markers also turned out to be
among the markers that were more robust to changing EEG configurations (Figure S3B in supplementary materials).

Exploiting invariant EEG signatures features of consciousness for generalization

**Generalization to independent data, protocols and configurations:** Here we considered two independent cohorts, 107 task-EEG recordings from the Paris Pitié-Salpêtrière Hospital (Paris 2) and 78 resting state EEG recordings by an independent research group (Coma Science Group, Liège, Belgium, see Table 2 Liège dataset for an overview). When training the DOC-Forest on the Paris 1 dataset, and testing this algorithm on the Paris 2 dataset, each time using the full EEG configuration, we observed significant classification performance with an AUC around 0.73 (SEM = 0.05, 95% CI[0.63,0.82]) (Figure 3A). Likewise, when trained on all available data from Paris (Paris 1 and Paris 2) but ignoring the evoked markers (Table 1, Figure 1A), the DOC-Forest scored an AUC of 0.78 (SD = 0.06, 95% CI[0.66,0.89]) on the Liège resting state data (Figure 3B).

We subsequently assessed generalization of our classifier trained on the Paris dataset to distinguish UWS vs MCS to a dataset of 66 conscious controls. The DOC-Forest classified 94% of the controls (Paris local-global paradigm: 34 out of 36, Liège resting state: 28 out of 30) as MCS. This result suggests that the patterns used by the classifier to distinguish UWS vs MCS patients extrapolate to normal controls.

Furthermore, we detected two cognitive-motor dissociation patients in the Liège dataset. These patients were originally labeled as UWS from their behavior but showed evidence of conscious processing using an active fMRI paradigm (see supplementary materials for a brief description of the two cases). Both cases were classified as MCS by DOC-Forest. This further suggests the utility of the proposed tool to characterize the state-of-consciousness of DOC patients.

**Generalization using univariate markers:** Less consistent results were obtained when using univariate forests based on the markers from the connectivity, information theory and spectral families that showed the highest cross-validation performance on the training set. For Paris 1 (Figure 3A) these were wSMI
(mean, mean), theta permutation entropy (mean, mean) and normalized alpha power (std, mean) (see Table 2 for abbreviations) with scores of 0.75, 0.74 and 0.77, respectively. For the combined Paris 1&2 dataset these were: theta wSMI (std, mean), theta permutation entropy (mean, mean) and alpha band power (mean, mean) with cross-validated scores of 0.69, 0.69 and 0.73, respectively. All univariate models showed lower generalization performance [0.04 to .14 AUC points] compared to the DOC-Forest and only the alpha band classifiers performed meaningfully better than a dummy classifier (Figure 3, middle panels). Comparing the variable importance to each marker’s out-of-sample performance, again, revealed positive non-linear associations (Figure 3A&B, right-most panel, \( \rho_{\text{Spearman Paris 1->2}} = 0.477, 95\% \text{ CI} [0.312, 0.620], p < 0.001; \)
\( \rho_{\text{Spearman Paris -> Liège}} = 0.521, 95\% \text{ CI} [0.309, 0.684], p < 0.001 \). The display reveals that several univariate models showed reasonable generalization performance with AUC values beyond .70. Highly performing markers were disproportionally more important for the DOC-Forest than would have been expected assuming a linear relationship. Again, these findings suggest that the DOC-Forest achieves generalization by preferentially enhancing the influence of reliable markers.

Strikingly, generalization was even successful when different EEG-configurations were combined, e.g., training with 100% of the epochs and 32 sensors and testing with 50% of the epochs and 8 sensors, although this induced decodable differences between training and testing sets (Figure S3 in supplementary materials). On average, the DOC-Forest performed significantly higher than any of the three corresponding univariate forests (See Table 3). Inspection of the cross-configuration generalization patterns revealed that the performance changes were far from random, favoring specific but distinct combinations of sensors and epochs for both generalization tasks (Figure S46 in supplementary materials).

**Table 3** Average generalization performance over different EEG-configurations

<table>
<thead>
<tr>
<th>Generalization</th>
<th>Contrast</th>
<th>Difference</th>
<th>95% CI</th>
</tr>
</thead>
</table>

ScholarOne, 375 Greenbrier Drive, Charlottesville, VA, 22901  Support (434) 964 4100
Paris 1 -> 2
DOC-Forest - wSMI Θ (m,m)  
\[ D = 0.124^{***} \]  
\[ [0.122, 0.125] \]

Paris 1 -> 2
DOC-Forest - PE Θ (m,m)  
\[ D = 0.097^{***} \]  
\[ [0.096, 0.098] \]

Paris 1 -> 2
DOC-Forest - |α| (s,m)  
\[ D = 0.035^{***} \]  
\[ [0.033, 0.037] \]

Paris -> Liège
DOC-Forest - wSMI Θ (s,m)  
\[ D = 0.140^{***} \]  
\[ [0.139, 0.142] \]

Paris -> liège
DOC-Forest - PE Θ (m,m)  
\[ D = 0.118^{***} \]  
\[ [0.115, 0.120] \]

Paris -> liège
DOC-Forest - α (m,m)  
\[ D = 0.035^{***} \]  
\[ [0.034, 0.037] \]

*Note.* *** p < 0.001

Robustness to noise:

As the DOC-Forest seemed resilient to mismatching EEG-configurations, we conducted a computational stress-tests by adding noise to the markers in the testing set until classification broke down (Figure 5AS7 in supplementary materials). Unsurprisingly, across generalization tasks, the univariate classifiers collapsed earlier at signal to noise ratios (SNR) between 1/10 and 1/100, whereas the DOC-Forest endured longer, eventually failing at SNR values of 1/1000.

Another concern potentially limiting generalization performance is the quality of the diagnostic information. We empirically assessed in a second computational stress-test the stability of generalization from Paris to Liège in the face of increasingly inaccurate diagnostic training labels (Figure 5S8B in supplementary materials). By design, this simulation forced the DOC-Forest to collapse and eventually yield systematically wrong predictions. However, the classifier still delivered reasonable predictions even if up to 30 percent of the diagnostic labels were flipped. Moreover, the literature would predict between 6% to 17% of misdiagnoses...
(Wannez et al., 2017) for the three to five CRS-R repetitions used in this study and, here, fall into the range of resilient generalization. These results demonstrate that the DOC-Forest is not only relatively robust to noise in the data but also to noise in the diagnostic labels.

The results suggest that the DOC-Forest is robust to a certain amount of noise in the labels and that its advantage over univariate classifiers was especially pronounced when systematic differences occurred between training and testing sets.

### Table 3: Average generalization performance over different EEG configurations

<table>
<thead>
<tr>
<th>Generalization</th>
<th>Contrast</th>
<th>Difference</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paris 1 -&gt; 2</td>
<td>DOC-Forest - wSMI (m,m)</td>
<td>$\Delta = 0.124^{***}$</td>
<td>[0.122, 0.125]</td>
</tr>
<tr>
<td>Paris 1 -&gt; 2</td>
<td>DOC-Forest - PE (m,m)</td>
<td>$\Delta = 0.097^{***}$</td>
<td>[0.096, 0.098]</td>
</tr>
<tr>
<td>Paris 1 -&gt; 2</td>
<td>DOC-Forest - $</td>
<td>\alpha</td>
<td>$(s,m)</td>
</tr>
<tr>
<td>Paris -&gt; Liège</td>
<td>DOC-Forest - wSMI (s,m)</td>
<td>$\Delta = 0.140^{***}$</td>
<td>[0.139, 0.142]</td>
</tr>
<tr>
<td>Paris -&gt; Liège</td>
<td>DOC-Forest - PE (m,m)</td>
<td>$\Delta = 0.118^{***}$</td>
<td>[0.115, 0.120]</td>
</tr>
<tr>
<td>Paris -&gt; Liège</td>
<td>DOC-Forest - $\alpha$(m,m)</td>
<td>$\Delta = 0.035^{***}$</td>
<td>[0.034, 0.037]</td>
</tr>
</tbody>
</table>

Note. $^{***} p < 0.001$

**Discussion**
We evaluated the robustness to different EEG configurations and recording conditions of univariate and multivariate pattern based on 28 putative EEG biomarkers of consciousness using the Extra-Trees algorithm. To the best of our knowledge, our study represents the most extensive validation of a machine learning approach to diagnose UWS versus MCS patients for two reasons. Our findings are based on the currently biggest EEG dataset of patients suffering from DOC, encompassing, in total, 327 recordings. Second, in the context of DOC, the present study is the first to demonstrate prospective generalization of multivariate pattern classification between different centers, EEG configurations, and protocols. We demonstrated that robust generalization can be achieved despite nontrivial changes in the spatiotemporal configuration of the EEG and that this generalization can be resistant to certain degree of uncertainty in the training labels (up to 20%). We showed that by relying on a robust classification algorithm, meaningful generalization could be achieved even if the performance of individual markers varied systematically between datasets. While certain EEG-signatures markers, i.e., alpha band power and its fluctuations turned out to be useful as stand-alone classifiers we found that the advantage of multivariate over univariate classification was most striking when systematic differences between the training and testing sets were present. Moreover, we found the DOC-Forest to preferentially base its predictions on diverse aspects of alpha and theta frequency band dynamics. Importantly, our results show that EEG-signatures markers of consciousness can be accessed equivalently from task and resting state EEG.

Robust learning of UWS vs MCS diagnosis from EEG-signatures markers of consciousness

Our results demonstrate that UWS vs MCS patients diagnosis can be robustly inferred from multivariate pattern classification using a wide array of EEG configurations (Figure 2A-B). This was also the case with a minimum of sensors (about 16) and epochs (10-50) and even when EEG configurations differed on the training and testing data (Figure 4, S3-S46 in supplementary materials), e.g., when training on 10% of the epochs with 8 sensors and testing on all epochs with 256 sensors. We observed
that many individual markers were highly variable (Figure 2A, S13-24). Nonetheless, our DOC-Forest fluctuated narrowly between AUC scores of 0.72 and 0.77 (Figure 2D). Inspection of our classifier in terms of the variable importance revealed a striking pattern (Figure 2C, S13B in supplementary materials). Markers that were most influential for its classifications not only were the ones with the greatest individual discrimination performance, but also turned out to be less susceptible to changes in the EEG configuration, noise on the EEG features and noise in the diagnostic labels (Figures 4-5). Interestingly, the overall relationship between univariate performance and variable importance was not linear. As univariate marker performance increased, marker importance increased disproportionally, i.e., at the top of the distribution, a change in univariate AUC lead to a much bigger change in importance than at the bottom of the distribution. Our findings, therefore, suggest that our DOC-Forest provides robust learning of UWS vs MCS diagnosis by enhancing robust and efficient EEG markers.

In this context, it may be interesting to reconsider the recently issued notion that predictive variables are not necessarily the ones that differ significantly (Bzdok, Engemann, Grisel, Varoquaux, & Thirion, 2018; Lo, Chernoff, Zheng, & Lo, 2015). As the AUC can be regarded as a rescaled Mann-Whitney-U test (See supporting information), significant univariate classification as in Sitt, King et al. (2014) implies significant differences in a marker between the diagnoses. The presence of univariate classification success and its positive correlation with multivariate variable importance suggests that, herein the present study, more significant variables were more predictive while less predictive variables were less significant, thus, providing little support for the claim by Lo and colleagues.

Robust classification was driven by distinct alpha and theta frequency band dimensions.

Our findings suggested that protocol-general markers were, overall, more reliable. Strikingly, these markers, belonging to different conceptual families, were all related to neuronal dynamics in the theta and alpha range (Figures 3-4). The robustness of these markers may be explained by the fact that no excessive
averaging is needed for their extraction and their characteristic EEG-topographies are simple and easy to capture with few sensors. However, the tight relationship between variable importance and conditional mutual information (Louppe, 2014) suggests that these top-performing markers carry independent information. Indeed, recent research has suggested a rather complex picture of functional and pathophysiological landscapes. The complexity of theta-band signals and their long-range interactions could reflect distinct memory processes underlying consciousness, such as access and maintenance (Axmacher et al., 2010). Similarly, alpha band power may reflect global arousal and demands for dynamic inhibition required for functional encapsulation of cortical networks (For an overview see Sadaghiani & Kleinschmidt, 2016).

Moreover, intact consciousness has been related to the peak frequency of alpha and theta band oscillations originating from distinct cerebral generators (ND Schiff, 2010; Williams et al., 2013). In fact, the mesocircuit model predicts that the down-regulation of the thalamo-cortical circuits following a brain injury should be directly associated to changes in the interactions within these frequency bands observed in this study (Schiff, Nauvel, & Victor, 2014; Victor, Drover, Conte, & Schiff, 2011). Yet, this is further complicated by the fact that these generators can be selectively disrupted for different etiologies and can show a variety of regional effects during anesthesia (Purdon et al., 2013). While future experimental research is desirable to disentangle these facets, our findings suggest that the presence of independent physiological sources of information may enhance generalization as it is unlikely that all of them will be jointly corrupted at the same time on new data samples.

But do our results imply that less important variables were useless? Not necessarily. Many evoked markers enjoy a high degree of neuroscientific validation and intuitively support clinical reasoning. The P3 markers, for example, belong to the most studied indices of consciousness in the EEG literature and are commonly used in brain computer interfaces settings (Lulé et al., 2013). They have been related to processing novelty in bottom-up information, the global neuronal workspace, access consciousness, and context-updating (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006; Donchin & Coles, 1988; Pins, 2003; Polich, 2007; Sergent,
Baillet, & Dehaene, 2005). Considering such markers for MVPA may, thus, improve interpretability.

Additionally, evoked markers indexing auditory novelty have been shown to be rather specific than sensitive (e.g., King et al., 2013). Like-wise, it could be the case that candidate signatures markers of conscious access, e.g., P3b, may be more relevant to distinguish MCS+ from MCS- patients (Naccache, 2017). Although being deemphasized by the DOC-Forest, evoked markers may still have contributed positively. Indeed, excluding all evoked markers from the Paris 1 to Paris 2 generalization actually reduced DOC-Forest-performance marginally (AUC = 0.71, 95% CI[0.618,0.807], SD= 0.049). One could, therefore, argue that, evoked markers should be considered for MVPA of DOC whenever available, alongside a few robust markers.

EEG signatures markers of consciousness are shared between protocols and contexts

In the field of clinical neuroscience, cross-validation is commonly used to assess MVPA performance. However, it has been shown that cross-validation can give positively biased performance estimates (Saeb, Lonini, Jayaraman, Mohr, & Kording, 2016; Varoquaux, 2017; Varoquaux et al., 2016; Woo et al., 2017). Beyond cross-validation, here, we demonstrated significant, positive generalization to independent EEG data from a different EEG protocol recorded by an independent research group (Figure 4A-B) and we did not observe considerable deviations from cross-validation scores. Generalization from the Paris to the Liège dataset even showed marginal improvements over cross-validation. As noted previously, this could not be explained by the absence of evoked markers. Precluding the possibility of random selection bias, this may suggest that either the signal quality or the diagnostic information may have been more favorable on the Liège data. Interestingly, compared to the best markers, i.e., alpha band power and its fluctuations, the advantage of the DOC-Forest was only marginal by a few AUC points. In contrast, the other remaining univariate models (based on theta band Permutation Entropy and theta wSMI) did not generalize significantly. Thus, our findings demonstrate that single markers can yield reasonable stand-alone classifiers but also expose the difficulty of anticipating which marker will actually succeed. Fortunately, MVPA potentially solves this selection problem with greater
success by learning predictive profiles of markers. Indeed, we observed that DOC-Forest was more robust than individual markers when using different combinations of EEG-configurations for training and testing. Likewise, we observed that univariate classifiers collapsed earlier and faster than the DOC-Forest as we experimentally corrupted the training data (Figure 5S7 in supplementary materials).

The significant generalization from task to resting state EEG deserves separate consideration. It is well conceivable that EEG markers related to the so-called functional axis of consciousness (Sergent et al., 2017), are accessible during task and resting state EEG. Accordingly, changing states of consciousness should impact markers of global house-keeping functions such as alpha band power, global long-range connectivity or signal complexity, irrespective of the context. For instance, for a patient with locked-in syndrome we observed EEG-patterns similar to healthy persons during rest (Rohaut et al., 2017) and here we also demonstrate the discrimination of two cognitive motor dissociation patients from UWS patients from their resting state EEG. This can be explained by that fact that we observed significant generalization from task to resting state EEG by several EEG makers, principally for alpha band power (Figure 3B right panel).

Practical Implications and Suggestions

**How long should EEG recordings be to yield a useful feature space for machine learning?** Our results suggest that reasonable results can be achieved with a short duration EEG recording (30 seconds to 3 minutes). This potentially broadens the scope of protocols usable in practice and encourages development of fast, time-resolved, economic screening tasks.

**How many EEG sensors should be used?** When high-density nets are available, using the full configurations turns out to be beneficial for model fitting. However, results based on 16 sensors from a 10-20 montage scheme are already encouraging. As a consequence, this supports the idea that data can be successfully pooled over various EEG systems even when the number of electrodes differs.
**Which EEG protocol should be used?** Both, univariate and multivariate analysis suggested that EEG-signatures-markers of consciousness are accessible using task and resting state data. This suggests that protocols can be liberally combined in clinical practice and encourages the development of simpler and faster screening routines as compared to a full-blown cognitive experiment encompassing hundreds of trials.

**Can classification models generalize to data from other sites?** Our findings demonstrate prospective generalization to new data from younger cohorts and data from other research laboratories. The use of robust methods is particularly recommended to alleviate problem of changing marker distributions between datasets.

**When should multivariate analysis be preferred to predict diagnosis?** Multivariate classification is more resilient to changes of marker distributions across datasets, be it due to noise in the signals or in the training labels, differences of populations or differences in EEG-configurations and protocols. Beyond optimizing accuracy, multi-variate classification models therefore yield more dependable classification performance.

**How to extract biological insight from machine learning models?** Here we demonstrate how the careful inspection of multivariate variable importance scores supplements the univariate analysis in qualifying interdependencies between EEG-markers. While such insight may also be obtained from model coefficients of linear models, the variable importance metric as used in this study, is not limited to linear relationships and does not necessitate explicit definition of non-linear effects or interaction effects.

Besides these specific points, we want to emphasize that we did not find one single globally best biomarker and that using machine learning tools to robustly combine theoretically heterogeneous markers is the recommended strategy.

**Conclusions**
In the current study, we demonstrated that electrophysiological signatures markers of consciousness can be robustly exploited across contexts and protocols by relying on robust machine learning techniques. In this context, the proposed feature-extraction method based on multiple summary statistics turned out particularly useful as it permits to abstract away specific sensor layouts, recording protocols and local EEG methodologies. Future work will have to demonstrate if the here proposed "robust tool for detecting state-of-consciousness in brain-injured patients" can be extended to a “robust neurophysiological marker of conscious state”. It will have to be demonstrated that the proposed model can generalize to other loss of consciousness scenarios, such as sleep or anesthesia. We wish that our findings and our publicly released recipe for classification will contribute to building large datasets that could eventually enable intensely data-driven, cross-center approaches to treatment of severely brain-injured patients and understanding the neural-underpinnings of conscious processing.

Acknowledgments

This work was supported by an ERC proof of concept grant issued to Stanislas Dehaene, Institut National de la Santé et de la Recherche Médicale (France), the James S. McDonnell Foundation, the Institut du Cerveau et de la Moelle Épinière (France) to Lionel Naccache, Consejo Nacional de Investigaciones Científicas y Técnicas (Argentina), the FRM Equipe 2015 grant to Lionel Naccache, STIC-AmSud grants Complexity as a neural marker: applications to EEG and natural language processing and RTBRAIN - Towards Real-time processing of brain signals, the Belgian Funds for Scientific Research (FRS-FNRS), the European Commission, the Human Brain Project (EU-H2020-fetflagship-hbp-sga1-ga720270), the Luminous project (EU-H2020-fetopen-ga686764), the Center-TBI, the European Space Agency, Belspo, "Fondazione Europea di Ricerca Biomedica", the BIAL Foundation, Wallonia-Brussels Federation Concerted Research Action and the Mind Science Foundation. Stanislas Dehaene gratefully acknowledges additional support from CIFAR. Denis Engemann acknowledges support by the Amazon Web Services research grant. Olivia Gosseries is post-doctoral fellow
and Steven Laureys is research director at FRS-FNRS. We thank Charlene Aubinet, Olivier Bodart, Manon Carrier, Athena Demertzi, Charlotte Martial and Sarah Wannez for their contributions with to clinical evaluation of the patients. We would like to thank the UNICOG and Parietal team at Neurospin for repeated fruitful and stimulating discussion on this research project. We would also like to express our gratitude to the MNE community. The current study would not have been possible without our collaborative software development efforts. We specifically acknowledge helpful discussions and comments on this study by Alexandre Gramfort, Benjamin de Haas, Danilo Bzdok, Johan Stender, Stefania de Vito and Virginie van Wassenhove (alphabetical order). This study is dedicated to the patients and to their close relatives.

Figure Legends

**Figure 1: Extraction of EEG-signatures features of consciousness.** (A) The EEG markers fell into four conceptual families, i.e., spectral, information theory, connectivity and evoked responses. When computing the markers from the preprocessed EEG, we obtained several observations for channels, epochs, time points and frequency bins, depending on the family. Following Sitt, King et al. (2014), we compressed extracted each four features from each marker into four single numbers (indicated by the red dots), by summarizing the observations systematically: we computed either the mean or the standard deviation first across epochs (1) and then across sensors (2). If a third dimension was present (3), we summarized it using the mean. We, hence, referred to the ensuring four marker types features as “mean,mean”, “mean,std”, “std,mean” and “std,std”. (B) We repeated this process using six alternative sensor configurations (256,128, 64, 32, 16, 8) and six alternative percentages of consecutive epochs (1, 5, 25, 50, 75, 100) with about 7 epochs at 1 percent and about 700 epochs at 100 percent. Sensors were selected such that they approximated realistic EEG caps.
respecting the international 10-20 system. Selection of epochs respected the relative proportions of conditions used in the task. This allowed us to compute markers based on experimental contrasts at any point. In total, this yielded 36 alternative EEG configurations. Abbreviations: L = local, G = global, S = standard, D = deviant, freq. = frequency, sens. = sensor, std = standard deviation.

Figure 2: Performance of EEG-signatures markers of consciousness across different EEG-configurations. (A) Performance distribution over markers (grey: model-free in-sample performance, blue: cross-validation with univariate forests) and the multivariate DOC-Forest pattern classifier (red) across 36 EEG configurations on the Paris 1 dataset. (B) DOC-Forest tended to improve as more epochs and sensors were used. Although optimal performance was achieved with 128 electrodes, reasonable performance could still be obtained with only 16 electrodes and a minimum of epochs. (C) Cross-validated univariate performance as a function of multivariate variable importance in the DOC-Forest, both averaged across EEG-configurations. Marker subtype and conceptual family are indicated by shape and color, respectively. A positive but non-linear relationship emerged. The best univariate markers were disproportionally more important to the DOC-Forest as a linear relationship would predict. It is noteworthy that markers from the spectral, connectivity and information theory families had the highest univariate performance and were assigned the highest importance by the classifier while the evoked markers systematically less important.
Figure 3: Generalization between datasets and protocols. (A) Generalization from the Paris 1 cohort to 107 new EEG recordings from Paris (task-EEG in both cases). The left panel shows the ROC curves for the multivariate DOC-Forest and three univariate forests based on the markers.feature that performed best (cross-validation) on the training set corresponding to the connectivity, information and spectral families. The middle panel depicts bootstrap distributions of improvements over a dummy classifier based on paired differences, ordered by performance. Positive values indicate performance better than the dummy model. Boxplot whiskers show the 95% confidence intervals. The right panel plots the generalization performance of each marker against training-set importance. The 10 most important markers features are labeled for convenience. (B) Generalization from 249 task-EEG (Paris 1 + Paris 2) to 78 resting state EEG recordings (Liège) depicting an equivalent analysis as in (A) but not including the evoked markers.response features. The results suggest meaningful prospective generalization for the DOC-Forest while the univariate models were overall less successful.

Figure 4: Generalization between datasets and protocols when EEG configurations differ. (A) Generalization from Paris 1 to Paris 2 when 1296 different combinations of EEG-configurations were used for training and testing (6 sensors x 6 epochs configurations for each set). The same univariate forest models as in Figure 3 were considered next to the multivariate DOC-Forest. The distribution of AUC scores is indicated by the histograms, single observations are indicated by the rug plot. The orange solid lines indicate the mean of the distribution, the orange dotted line the performance when the reference configuration of 100% epochs and 256 sensors is used on both training and testing. (B) presents the same analysis for the generalization from the joint Paris 1 & 2 dataset to the Liège dataset. It can be seen that, on average, the DOC-Forest outperforms any of the univariate models.
Figure 5: Computational stress tests. (A) The generalization performance of the DOC Forest and three univariate models as signal-to-noise ratio (SNR) is gradually reduced on the testing set. The noise was generated independently from Gaussian distributions with mean and variance parameters from each feature with 50 realizations, scaled by the SNR parameter and added to the testing set, such that at 1/10 the noise was 10 times stronger than the signal. The standard deviation of performance over realizations is indicated by the shaded areas. It can be readily seen that the DOC-Forest survives longest while at the same time decreasing its performance more slowly than each of the three univariate models. In general, univariate models did not survive a signal to noise ratio of 1 / 100 or smaller while the DOC-Forest still showed meaningful generalization performance beyond such low SNR values. (B) We estimated the impact of misdiagnosis on generalization empirically by flipping the diagnosis labels for an increasing percentage of patients (0 to 100 in steps of 5). To avoid bias and estimate variability, we randomly draw patients at each percentage level and repeated the process 50 times. The median generalization performance is depicted by the boxplots (whiskers show the 2.5 and 97.5 percentiles) and the mean performance by the super imposed red circles. The performance at 0% and 100% flipping is shown by the red circles. For convenience, the percentage of mis-diagnoses predicted from the number of CRS-R assessments reported by Wannez et al (2017) is superimposed by the colored dotted lines. It can be seen that the mean generalization performance drops more slowly between 10 and 30 percent than between 30 and 50 percent and remains reasonable even if up to 30% of the diagnoses are flipped.

References


Giacino, J. T., Kalmar, K., & Whyte, J. (2004). The (JFK) Coma Recovery Scale-Revised: Measurement characteristics and diagnostic utility11No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the authors or upon . *Archives of Physical Medicine and Rehabilitation, 85*(12), 2020–2029.


https://doi.org/http://dx.doi.org/10.1016/j.neuroimage.2013.10.027


https://doi.org/10.1001/jamaneurol.2015.2899


Figure 1: Extraction of EEG-features. (A) The EEG markers fell into four conceptual families, i.e., spectral, information theory, connectivity and evoked responses. When computing the markers from the preprocessed EEG, we obtained several observations for channels, epochs, time points and frequency bins, depending on the family. Following Sitt, King et al. (2014), we extracted four features from each marker (indicated by the red dots) by summarizing the observations systematically: we computed either the mean or the standard deviation first across epochs (1) and then across sensors (2). If a third dimension was present (3), we summarized it using the mean. We, hence, referred to the ensuring four features as "mean,mean", "mean,std", "std,mean" and "std,std". (B) We repeated this process using six alternative sensor configurations (256,128, 64, 32, 16, 8) and six alternative percentages of consecutive epochs (1, 5, 25, 50, 75, 100) with about 7 epochs at 1 percent and about 700 epochs at 100 percent. Sensors were selected such that they approximated realistic EEG caps respecting the international 10-20 system. Selection of epochs respected the relative proportions of conditions used in the task. This allowed us to compute markers based on experimental contrasts at any point. In total, this yielded 36 alternative EEG configurations. Abbreviations: L = local, G = global, S = standard, D = deviant, freq. = frequency, sens. = sensor, std= standard deviation.
Figure 2: Performance of EEG-markers of consciousness across different EEG-configurations. (A) Performance distribution over markers (grey: model-free in-sample performance, blue: cross-validation with univariate forests) and the multivariate DOC-Forest pattern classifier (red) across 36 EEG configurations on the Paris 1 dataset. (B) DOC-Forest tended to improve as more epochs and sensors were used. Although optimal performance was achieved with 128 electrodes, reasonable performance could still be obtained with only 16 electrodes and a minimum of epochs. (C) Cross-validated univariate performance as a function of multivariate variable importance in the DOC-Forest, both averaged across EEG-configurations. Marker subtype and conceptual family are indicated by shape and color, respectively. A positive but non-linear relationship emerged. The best univariate markers were disproportionally more important to the DOC-Forest as a linear relationship would predict. It is noteworthy that markers from the spectral, connectivity and information theory families had the highest univariate performance and were assigned the highest importance by the classifier while the evoked markers systematically less important.
Figure 3: Generalization between datasets and protocols. (A) Generalization from the Paris 1 cohort to 107 new EEG recordings from Paris (task-EEG in both cases). The left panel shows the ROC curves for the multivariate DOC-Forest and three univariate forests based on the feature that performed best (cross-validation) on the training set corresponding to the connectivity, information and spectral families. The middle panel depicts bootstrap distributions of improvements over a dummy classifier based on paired differences, ordered by performance. Positive values indicate performance better than the dummy model. Boxplot whiskers show the 95% confidence intervals. The right panel plots the generalization performance of each marker against training-set importance. The 10 most important features are labeled for convenience.

(B) Generalization from 249 task-EEG (Paris 1 + Paris 2) to 78 resting state EEG recordings (Liège) depicting an equivalent analysis as in (A) but not including the evoked response features. The results suggest meaningful prospective generalization for the DOC-Forest while the univariate models were overall less successful.
Figure 4: Generalization between datasets and protocols when EEG configurations differ. (A) Generalization from Paris 1 to Paris 2 when 1296 different combinations of EEG-configurations were used for training and testing (6 sensors x 6 epochs configurations for each set). The same univariate forest models as in Figure 3 were considered next to the multivariate DOC-Forest. The distribution of AUC scores is indicated by the histograms, single observations are indicated by the rug plot. The orange solid lines indicate the mean of the distribution, the orange dotted line the performance when the reference configuration of 100% epochs and 256 sensors is used on both training and testing. (B) presents the same analysis for the generalization from the joint Paris 1 & 2 dataset to the Liège dataset. It can be seen that, on average, the DOC-Forest outperforms any of the univariate models.
Figure 5: Computational stress tests. (A) The generalization performance of the DOC Forest and three univariate models as signal-to-noise ratio (SNR) is gradually reduced on the testing set. The noise was generated independently from Gaussian distributions with mean and variance parameters from each feature with 50 realizations, scaled by the SNR parameter and added to the testing set, such that at 1/10 the noise was 10 times stronger than the signal. The standard deviation of performance over realizations is indicated by the shaded areas. It can be readily seen that the DOC-Forest survives longest while at the same time decreasing its performance more slowly than each of the three univariate models. In general, univariate models did not survive a signal to noise ratio of 1 / 100 or smaller while the DOC-Forest still showed meaningful generalization performance beyond such low SNR values. (B) We estimated the impact of misdiagnosis on generalization empirically by flipping the diagnosis labels for an increasing percentage of patients (0 to 100 in steps of 5). To avoid bias and estimate variability, we randomly draw patients at each percentage level and repeated the process 50 times. The median generalization performance is depicted by the boxplots (whiskers show the 2.5 and 97.5 percentiles) and the mean performance by the super imposed red circles. The performance at 0% and 100% flipping is shown by the red circles. For convenience, the percentage of mis-diagnoses predicted from the number of CRS-R assessments reported by Wannez et al (2017) is superimposed by the colored dotted lines. It can be seen that the mean generalization performance drops more slowly between 10 and 30 percent than between 30 and 50 percent and remains reasonable even if up to 30% of the diagnoses are flipped.
Supporting Information

Supplementary Methods

Experimental procedure

Hierarchical auditory oddball task (Paris 1 & 2 dataset): Task-related EEG signals were obtained from the ‘Local-Global’ protocol (Bekinschtein et al., 2009) designed to study unconscious and conscious auditory processing. Accordingly, brain responses to two types of auditory events were recorded: automatically processed short-time-range violations and long-time-range violations whose recognition depends on explicit working memory effort. For optimal cognitive performance, patients underwent recordings at least 24 hours after sedation discontinuation. Medication that could potentially modify the EEG, e.g., myorelaxants and anti-epileptic drugs were not controlled. At the beginning of each recording session, EEG signal quality was assessed. If seizures or any other clearly identifiable abnormal activity were observed, the recording was stopped. EEG recordings were sampled at 250 Hz with a 256-electrode geodesic sponge sensor net (EGI) referenced to the vertex. Recordings were band-pass filtered (from 0.5 to 45Hz using a 6 and 8 order FFT-based Butterworth filter). Data were then epoched from -200ms to 1336ms relative to the onset of the first sound. Epochs were excluded based on adaptive outlier detection as in Engemann et al. (2015). Subsequently, data were re-referenced using an average reference and baseline correction was applied.

Task-free recordings (Lige dataset): Data were chunked into 1536ms pseudo-epochs, matching the length of the task-based epochs, with random intervals between epochs matching the inter-trial intervals of the auditory task. Otherwise, the same acquisition preprocessing scheme was applied.

Statistical Analysis
**Statistical Inference:** We extended our visualizations into hypothesis tests by employing the percentile bootstrap (Efron & Tibshirani, 1993). Accordingly, we generated 2000 bootstrap samples by drawing with uniform probability and replacement $n$ samples from the dataset. The test-statistic of interest was then evaluated on each bootstrap sample. Two-sided 95% confidence intervals were obtained by querying the 2.5 and 97.5 percentiles and the significance-level was then obtained by inversion of the confidence interval that excluded the value under H0. We denoted difference-statistics by $D$. For correlation analysis, we relied on the non-parametric Spearman’s Rank correlation coefficient. In the latter case, we obtained the significance level from analytical p-values.

To assess out-of-sample generalization we used two complementary approaches: A conservative validation on independent data (new cohorts, different protocols and laboratories) and cross-validation. For cross-validation we used a group Monte Carlo sampling scheme with a training set size of 80 percent, a testing set size of 20 percent and 50 iterations. The Monte Carlo procedure is known to minimize estimation variance and has been shown to yield low positive cross-validation bias (Varoquaux et al., 2016). The grouping consisted in exclusively assigning subjects to either the test or the train sets. For assessing the generalization capacity of the DOC-Forest on new datasets we contrasted the performance against empirically estimated chance levels. These were obtained from comparisons against a dummy classifier that generated random predictions based on the observed class-probabilities.

**Area under the Curve metric:** Univariate and multivariate discrimination performance was summarized with Area Under the Curve (AUC) calculated from the receiver operator characteristic (ROC). For a binary classification system, the ROC pits the detection probability, commonly referred to as *sensitivity* against the probability of false alarm ($1 - sensitivity$). These probabilities are empirically estimated by moving the decision cut-off along the sorted values of a continuous variable, e.g. a score, and evaluating its relation to the true label. In the case of traditional model-free univariate analysis, the score is the EEG-marker itself, in the
case of univariate or multivariate machine learning it is the predicted probability of a given sample to belong to the target class. The AUC can then be conveniently used to summarize the performance, where a score of 0.5 is uninformative and equals to random guessing whereas a score of 1 amounts to perfect classification and 0 to total confusion, indicating negative correlation between the score and the label. We used the AUC in two different contexts, once to summarize the class-probabilities issued by our classification models, once in a direct fashion on single marker values without a classification model. Note that in the latter context, markers often show univariate AUC scores smaller than chance level (0.5) because they are conceptually related to absence of consciousness. To avoid confusion, we rectified the direct AUC in that case by \( \text{abs}(\text{AUC} - 0.5) + 0.5 \).

Contrastingly, for classification models, AUC values smaller than 0.5 can indicate inconsistencies of patterns between training and testing data or result from low performance of the learned decision rules.

Furthermore, note that the AUC of a clinical score regarding two groups of patients’ amounts is closely related to computing the Mann-Whitney-U statistic and can be directly obtained from dividing U by the product of the two groups sample sizes:

\[
AUC = \frac{U}{n_1 n_2}
\]

**Multivariate Pattern Classification:** We chose the *Extra-Trees* algorithm (Geurts, Ernst, & Wehenkel, 2006) because this algorithm is well-established and belongs to the most popular machine learning techniques for regression and classification problems, next to Random Forests (Breiman, 2001), Support Vector Machines (SVM) and penalized linear models. We chose this algorithm to improve the robustness of classification. Moreover, we found randomized classification trees to achieve, ad-hoc, without tuning hyper-parameters and without feature-selection a performance equivalent to an SVM with tuned regularization parameter and with explicit feature-selection (Engemann et al., 2015). *Extra-Trees* are non-parametric and robust by design and are not sensitive to the measurement scale of the input data. This algorithm can handle so-called wide
datasets in which more variables than samples are available. Moreover, *Extra-Trees* belong to the family of adaptive algorithms capable of scaling the complexity of the learned model to the amount of data available. The *Extra-Trees* algorithm achieves its efficiency by generalizing the non-linear decision tree approach. Single decision trees are non-parametric rule-based models that automatize variable selection and can be thought of as learning a “regression surface” from the data by recursive orthogonal partitioning (Efron & Hastie, 2016). In other words, decision trees map joint value ranges of the input variables to values of the outcome variable. However, decision trees poorly generalize to new data. The *Extra-Trees* retains all benefits of decision trees while mitigating their excessive variance and poor generalization capability. This is achieved by averaging over many randomly constructed, hence uncorrelated, decision trees, each of which is “grown” on randomly drawn subsets of \( m \) input variables when looking for candidate splits at internal nodes (typically, \( m = \sqrt{p} \)). Unlike Random Forests, their historical predecessor, the Extremely Randomized Trees algorithm does not make use of randomization across samples via bootstrapping. It instead achieves additional randomization at the level of the thresholds used when constructing the trees, which can improve approximation of a fully randomizes tree and can facilitate its interpretation (Louppe, Wehenkel, Sutera, & Geurts, 2013). This principle, effectively, permits a random search through a combinatorial space of variables. Therefore, the *Extra-Trees* algorithm lends itself to exploit interactions between variables if sufficient data is available. Consequently, the ensuing model often consists of thousands of trees which practically renders interpretation difficult.

To improve the interpretability of randomized classification trees, the so-called variable importance metric has been proposed and can be readily computed for a fitted model. Intuitively, the importance score of a variable can be understood as the weighted reduction of entropy of the outcome, over all the internal nodes where that variable has been used for making a split (Louppe et al., 2013). Variable importance can be shown to correspond to a weighted average of the mutual information between that variable and the outcome, conditionally over any possible configuration of any subset of the other variables if entropy is used as impurity criterion. Variable importance is of considerable interest, as its conditional (multivariate) nature complements
marginal (univariate) statistics. In other words, variable importance is multivariate and accordingly a variable can be important either because of a univariate (marginal) correlation with the outcome, or because its interdependencies with other variables are informative about the outcome. This potentially facilitates discovery of interesting variables which would be overlooked otherwise and is a welcomed remedy to the problem that predictive variables are not necessarily significant (Bzdok, Engemann, Grisel, Varoquaux, & Thirion, 2018; Lo, Chernoff, Zheng, & Lo, 2015). Some cautious interpretation is advised as masking effects can occur and not all necessarily influential variables are captured by the variable importance under default settings, especially when using Random Forests (Louppe et al., 2013). However, this effect can be mitigated by using only one variable for splitting (K or max_features = 1), by constraining the maximum tree depth to 3-5 and by approximating full randomization by using Extremely Randomized Trees (Louppe, 2014). This approximation of full randomization has the advantage that the variable importance is only driven by relevant variables, not irrelevant ones and, when combined with entropy as impurity criterion, its interpretation as mutual information holds (Louppe et al., 2013). As performance usually stabilizes at a certain point as trees are added to the model a trade-off between performance and speed has to be made (Geurts, Ernst, & Wehenkel, 2006). Profiling suggested stable and reasonable performance after 1000-2000 trees.

In the present study, we hence used the Extremely Randomized Trees with 2000 trees, a tree depth of four, entropy as impurity criterion, and a single feature for splitting. Otherwise we used default parameter values which have been shown to be optimal in most situations and are typically not recommended to be tuned to not jeopardize the computational benefits of Extra Trees over other randomized tree techniques (Geurts et al., 2006). For a detailed description of default values please consider the documentation of the Scikit-Learn software for machine learning (Pedregosa et al., 2011).
Case report of the cognitive motor dissociation patients

**Patient P1** This male 40 years old patient was admitted to the hospital with the diagnosis of UWS 11 months after traumatic brain injury resulting from a car accident. The patient behaviorally assessed with the CRS-R five times within a week, during which he showed auditory (1/5) and oral (4/5) reflexive behavior. The patient showed eye opening 3 out of 5 assessments and was diagnosed as comatose the other two assessments. The results of the structural neuroimaging highlighted micro-hemorrhages and diffuse axonal injury. The ventricles were moderately enlarged and atrophy was pronounced in the midbrain and around the cortical sulci. Functional MRI results suggested the presence of brain activity consistent with covert response to command as tested with the tennis paradigm, yet, no activation of resting state networks could be detected.

**Patient P2** This female, 36 years old patient was admitted to the hospital with the diagnosis of UWS six years and two months after a post-ischemic vertebrobasilar stroke. The patient was assessed with the CRS-R 5 times within the week of hospitalization, during which she showed motor (4/5) and oral (5/5) reflexes. The patient showed eye opening during all assessments. Structural MRI revealed the presence of severe ischemic lesions. Atrophy was most pronounced at the upper part of the cerebellum, while the frontal, parietal and striatum seemed relatively spared. Ventricles were enlarged. Functional MRI results suggested the presence of brain activity consistent with covert response to command as tested with the tennis paradigm, in addition, activation of the default mode resting state network could be detected.

**Supplementary Results**

Consistency of classification in diagnostic sub-groups

In order to assess classification accuracy for sub-groups we performed cross-validation on the Paris 1 dataset as described in the main text based on the full EEG-configuration. In each fold we subdivided the test set
according to diagnostic groups and computed separately the average AUC metric across all 50 folds. We first considered the three largest etiological groups (anoxia, stroke and TBI). We obtained an AUC of 0.77 (SEM=0.014) for anoxia patients (n = 34), an AUC of 0.83 (SEM = 0.024) for stroke patients (n = 47) and an AUC of 0.68 (SEM = 0.036) for TBI patients (n = 38). We then subdivided the dataset in acute patients (delay <= 30, n = 71) and chronic patients (delay > 30, n = 71). For acute patients we obtained an AUC of 0.78 (SEM = 0.023) and for chronic patients an AUC of 0.80 (SEM = 0.017).

Detailed comparison between individual markers and DOC-Forest

We assessed how each marker discriminated between UWS and MCS patients as a function of EEG-configurations using the task-EEG dataset recorded in Pitié-Salpêtrière (Paris 1 dataset). Comparing the cross-validated model-based AUC for each marker with the model-free AUC (Figure S1A) revealed a tight positive correlation (\(\rho_{\text{Spearman}} = 0.94\), 95% CI[0.905, 0.962], \(p < 0.001\)). This finding suggests that our univariate forest models performed equivalently to the previous model-free method. Subsequent analyses suggested that, on average, over all EEG, the DOC-Forest performance assumed at least the 98th percentile compared to other markers (Percentile\(_{\text{DOC-Forest}} = 0.985\), 95% CI[0.981, 0.989]) and its fluctuation over configurations was situated around the 83rd percentile (Percentile\(_{\text{DOC-Forest}} = 0.832\), 95% CI[0.796, 0.894]).

Tracking the discrimination performance between UWS and MCS across the EEG configurations revealed fluctuations according to marker subtype, conceptual family (Figure S1, Figure S2) and variance induced by the EEG configuration itself (cf. Figure 2A and 3B in the main text). Indeed, we found a positive correlation between fluctuations of marker estimates and discrimination performance across EEG configurations (\(\rho_{\text{Spearman}} = 0.46\), 95% CI[0.289, 0.6], \(p < 0.001\), suggesting that performance was more stable when the marker itself was stable too (Figure S1B and S2). Interestingly, evoked response markers appeared more severely affected than other markers.

Supplementary Figures
**Figure S1: Classification with univariate models.** (A) Comparison between traditional model-free classification on the training data using the univariate AUC metric as in Sitt, King et al. 2014 and estimation of out-of-sample performance using univariate classifiers and cross-validation. A positive relationship between model-free in-sample performance and univariate out-of-sample performance was observed. Markers that performed better as measured by traditional model-free AUC also showed higher cross-validated performance. This relationship was less tight at lower performance. Note that markers performing close to chance level tended to score below a cross-validated AUC of 0.5. The 10 best markers are indicated with labels for convenience. (B) Relationship between relative marker variance and performance over 36 EEG configurations displayed on double logarithmic scale. Markers at each configuration were standardized to the reference configuration of 100% epochs and 256 sensors. A non-linear positive relationship emerged, suggesting that markers whose values were more strongly changed by the EEG configuration also showed stronger fluctuations in performance. The 10 least variable markers are labeled for convenience.
Figure S2: Discrimination performance of EEG-markers across experiments. The cross-validated AUC for all 112 EEG-markers of consciousness and the DOC-Forest across all 36 EEG configurations on the Paris 1 dataset ordered by average performance. The marker family is indicated by the colors. The marker subtype by the letter suffix, e.g., mean over epochs and standard deviation over channels reads “m,s”. The boxplot whiskers indicate value ranges. The black notches indicate performance at the reference configuration of 100% epochs and 256 sensors. It can be seen that some markers rather improved their performance as sensors and epochs were changed (e.g., θ m,m) while others rather decreased their performance (e.g., α m,s). Note that performance fluctuated less when markers performed higher on average.

Figure S3: Distribution shifts from mismatching EEG-configuration. To ensure that using different combinations of sensors and epochs for training and testing induces nontrivial differences between the datasets, we applied the DOC-Forest to the origin of an EEG-recording instead of the diagnosis of the patient using 5-fold cross-validation over all combinations of EEG-configurations (A Paris 1 versus 2, B Paris versus Liège). It can be seen that the origin is almost perfectly decodable when EEG-configurations maximally diverge, suggesting that changing the number of sensors and epochs impacts the distribution of the datasets.
**Figure S4: Cross-configuration generalization.** (A) DOC-Forest generalization performance when using different EEG configurations for training on Paris 1 (rows) and testing on Paris 2 (columns). The 36 combinations of sensors and epochs are ordered such that, from the left corner, epochs grow slowly (1% to 100%) while sensors repeat (1 to 256). We observed reasonable generalization performance for majority of combinations of EEG-configuration. Note the horizontal and vertical stripes in the upper right part of the matrix, which point out peak performance with fewer sensors (8-32) and at least 25% of the epochs. (B) The same analysis for generalization from the combined Paris 1 & 2 dataset to the Liège dataset. Again, reasonable generalization performance was observed for the majority of combinations. However, the overall generalization pattern was markedly different, suggesting peak performance when less epochs were used on for training but more epochs and at least 16 sensors for testing.
References


