Mind the Traps! Design Guidelines for Rigorous BCI Experiments

Camille Jeunet, Stefan Debener, Fabien Lotte, Jeremie Mattout, Reinhold Scherer, Catharina Zich

Chapter 32
Mind the Traps! Design Guidelines for Rigorous BCI Experiments.

Authors:
* C. Jeunet (Inria, France / EPFL, Switzerland, camille.jeunet@inria.fr)
S. Debener (University of Oldenburg, Germany, stefan.debener@uni-oldenburg.de)
F. Lotte (Inria, France, fabien.lotte@inria.fr)
J. Mattout (Inserm, France, jeremie.mattout@inserm.fr)
R. Scherer (Graz University of Technology, Austria, reinhold.scherer@tugraz.at)
C. Zich (University of Oldenburg, Germany / University of Oxford, UK, catharina.zich@uni-oldenburg.de)

Introduction

Brain-Computer Interfaces (BCIs) enable users to control an application using their brain activity alone. Such control can be achieved following different paradigms. Mainly, BCIs can be active, reactive or passive. Active BCIs rely on modifications of the amplitude of brain rhythms in different frequency bands while users perform an explicit task, such as mental imagery. The most common active BCI paradigm is based on motor imagery: users are asked to perform a motor imagery task (i.e., to imagine movements of their limbs), which lowers the amplitude of sensorimotor rhythms (SMR) in the mu (8-12 Hz) and beta (12-30Hz) frequency bands in the sensorimotor cortex. By detecting this spatio-spectral signal change, the system may be able to infer the mental task performed by the user and in return associate to it a command for the application. For instance, by performing left- and right-hand motor imagery tasks, one can make a wheelchair turn left or right, respectively (Millán 2010). On the other hand, reactive BCIs rely on the detection of brain potentials generated in response to a stimulus. The most popular BCI application is probably the P300 speller (Farwell 1988). The P300 speller was first designed to enable communication for the paralyzed. A matrix of letters is displayed on a screen. The rows and columns of this matrix flash sequentially in a random order. The user has to focus on the letter he/she wants to spell. When this letter flashes, a positive cortical potential is generated around 300ms later. This way, after several repetitions, by intersecting the rows and columns the flashing of which triggered a P300, the system will be able to infer the target letter. Finally, passive BCIs are systems that enable the user's mental state to be measured in order to adapt an application/interface accordingly (Zander 2011). Here, users do not follow specific instructions or voluntary interact with the BCI, in other words, they do not send conscious commands. Instead, their cognitive (e.g., workload), emotional (e.g., frustration) or motivational states are inferred from their ongoing EEG signals, which can also be related to other physiological and behavioral data to identify a particular mental state. A passive BCI application automatically adapts to fluctuations in user states.

BCIs are very promising for a wide range of applications, from assistive technologies to communication devices. Beyond their obvious potential for patients with motor impairments, BCIs also offer new possibilities in different fields, ranging from sports to video games (Lécuyer 2008) through education (Frey 2014) and rehabilitation (Kübler 2013) to name only a few examples. Thus, not only engineers and psychologists are interested in using such technologies, but also medical doctors, neuroscientists, teachers, or specialists in sports science, business, physiotherapy or linguistics, between many others. Yet, designing BCI experiments requires knowledge in many different disciplines: neurosciences to understand the properties of the brain signals used to control the BCI, signal processing to extract the relevant information from these brain signals, machine learning to make the system able to learn which brain patterns correspond to which mental command, psychology to understand the factors influencing users' ability to control a BCI and human-computer interaction to design usable and efficient BCI protocols. However, very few people have skills in all these disciplines. Yet, a lack of knowledge in a single aspect of BCIs is likely to result in flaws in the experimental design, statistical analyses or in the interpretation of the results. Moreover, because the BCI field is relatively young, no widely accepted guidelines are available yet, while at the same time an exponentially increasing number of research teams contributes to this field and thus would benefit from such guidelines.

The objective of this chapter is to propose, in a pedagogical way, step-by-step guidance to design a rigorous BCI experiment. We do not propose a perfect experimental design, but rather name potential pitfalls, and explain how to avoid them. This chapter could be seen as a checklist of points that should be addressed when the aim is to design rigorous and scientifically valid BCI studies and experiments. Therefore, this chapter targets more specifically the numerous researchers, professionals, passionate people or patients and their families who start using BCIs and want to avoid the current traps surrounding BCIs. This chapter does not claim to be exhaustive but rather aims (1) at naming the important elements to be considered while designing a BCI experiment and (2) at guiding the reader towards relevant pieces of literature if they want to investigate further some specific elements.

The next part of this chapter is structured into three categories. The first category focuses on the acquisition of brain signals. Here, we will focus on electroencephalography (EEG) based BCIs, simply because this is the most popular input modality for BCIs. The second category deals with data processing issues, while the third one will introduce important points concerning the experimental design and the target individuals, the BCI users.

To illustrate these different points, we propose to design a BCI experiment all along the chapter, following the recommendations point by point. Let us say that our research question is the following:

We want to evaluate the relevance and reliability of new tasks to control a Mental-Imagery based BCI (MI-BCI), namely "remembering a positive vs. negative emotional souvenir".

The object of providing this example is not to undertake a real research study, nor to answer a specific research question. Thus, no data nor statistical analyses will be described. Rather, the aim is to provide the reader with a clear application so that they know in which context the recommendations could be applied. For specific applications of BCIs for therapeutic purposes, please refer to other chapters of this handbook of Guger et al. (2018) and McFarland (2018).

At the end of each sub-section, we will discuss which tools/methods are the most relevant in the context of this research question.

I - Acquisition of the Signals

1.1 - How to choose the appropriate sensors' type, location and number depending on what we want to measure?

Typical EEG signals refer to voltage fluctuations in the direct current (DC) to approximately 40 Hz frequency range, with amplitudes ranging from a few tens of microvolts to below 1 microvolt. To capture these miniature signals, sensors are placed on the scalp and a conducting gel is applied to lower the skin-electrode impedance to approximately < 20 kohms. However, wet sensors have the disadvantage that individuals have to wash their hair after the EEG recording, to remove the conductive gel. Therefore, new sensor types have been developed, such as active wet electrodes, active dry electrodes or miniature sensors requiring no, or very small amounts of gel only (Debener 2016). Typical EEG recordings require the use of several electrodes (the minimum number is 3, common are 32, 64, 128 or even 256 channel recordings), which are placed with a cap or net on the scalp. Different sensor types have their advantages and disadvantages. The signal quality and wearing comfort of dry electrodes, for instance, is typically inferior to conventional wet electrodes, but the benefits are the faster setup time and that hair washing is not required after the end of the recording. While alternative materials and electrodes may suffice for some EEG signals, we recommend the use of sintered Ag/AgCl electrodes, which, when used as wet electrodes and applied correctly, provide good signal characteristics. They do not generate voltage fluctuations on their own and do not cause frequency distortion, which are two problems that should be avoided when measuring biosignals with microvolt amplitude. The international 10-20 system is the standard for electrode placement. Depending on the type of brain signals that should be recorded, appropriate electrode locations should be used.

Which sensor positions are important? The answer depends on the class of BCI that one wants to implement. However, as outlined in the following section, a good spatial sampling is generally helpful even if only a few channels are used. For BCIs that detect sensory evoked responses, electrodes should be placed such that known topographical representations of sensory evoked potential are captured. This includes placing electrodes over posterior and occipital sites for capturing visual evoked responses, and placing electrodes over fronto-central sites for capturing auditory evoked responses. However, the most discriminative information does not necessarily overlap spatially with locations giving the best signal-to-noise ratio. Moreover, to optimally classify signals it is also helpful to place electrodes away from the signal of interest, in order to cancel out irrelevant activity, as explained by (Blankertz 2011). Accordingly, for BCIs that detect the neural correlates of motor imagery, electrodes should be located over sensorimotor areas, and, in order to disentangle sensorimotor mu from occipital alpha[1], some sensors should be placed over posterior scalp sites as well. In summary, multichannel EEG (32+ channels) acquisition from sensors covering wider parts of the head is beneficial, although for practical applications and for economic reasons, a limited setup is often used, under ideal circumstance, without much loss of performance.

[1] The 8-12 Hz frequency range is in the literature referred to as mu rhythm when related to sensorimotor activity, otherwise it is referred to as alpha rhythm.

And for our experiment? In our case, since we do not have a precise idea of the relevant brain-areas for the tasks (remembering positive/negative emotional souvenirs) we will use at least 32 or 64 electrodes, placed all over the scalp. Also, because we want to maximize signal quality, we will use wet Ag/AgCl electrodes (i.e., with conductive gel).

1.2 - What can we infer from the activation of one electrode? Concept of spatial resolution.

It is important to understand the concept of differential amplification when voltage is measured. EEG signals reflect voltage fluctuations over time, and voltage fluctuations can be best captured as the difference between two locations, say one electrode placed on the top of the head (vertex) and another one behind the ear (mastoid). Even though one electrode is often defined as the reference electrode (mastoid), the recorded signal cannot be regarded as reflecting electrical activity from the patch of brain underneath the other electrode (vertex). Likewise, since the whole body conducts current fairly well (with different tissues having different conductivity properties), it is not required to place electrodes close to the heart to measure the electrocardiogram. The reason for that is that EEG measures the synchronized electrical activity of adjacent pyramidal cells aligned in parallel which rise to electrical fields that are strong enough to be captured with electrode placed on the skin. Accordingly, all brain signals captured by EEG result from relatively large and highly synchronized patches of cortex. Each patch of cortex contributing to the EEG may be best regarded as an equivalent current dipole. If the aim is to capture the electrical activity of such a dipole it is important to understand that dipole orientation is at least as important as dipole location. In fact, two electrodes placed very close to a dipolar generator may not record any activity if placed in the wrong orientation. On the other hand, two electrodes placed further away from the generator may capture its signature nicely.

Given that an unknown number of brain (and non-brain, e.g., electrocardiogram) generators contribute to the scalp-EEG, one can safely assume that the number of generators contributing to the mixed recorded signal is much higher than the number of channels used to record the signals, even if high-density EEG acquisition is performed. The inverse problem means that one cannot determine the brain source of any particular EEG signal for sure. Inferring the number and locations of active brain regions that caused the recorded EEG signals on scalp is sometimes compared to the problem of guessing what 3D shape created an observed shadow on a wall. It is a so-called ill-posed problem which has no unique solution. However, good guesses are possible, and EEG source localization and spatial filtering procedures - which combine the signals from multiple electrodes, see, Blankertz et al. (2008) and chapter “Gentle Introduction to Signal Processing and Classification for Single-Trial EEG Analysis” in this book - can be used to confirm ideas about brain sources contributing to the EEG, with reasonable spatial resolution and precision. To this end, making use of multiple EEG channels can help. The rich spatial detail of multi-channel recordings has two key advantages. First, it is much easier to disentangle brain from non-brain contributions to the measured EEG signal, and second, it is much easier to identify, and disentangle, different brain signals from each other.

And for our experiment? To be able to disentangle the brain area(s) involved in each of the tasks, we will apply, offline, source localization procedures (i.e., source reconstruction algorithms).

1.3 - How muscular activity can interfere with EEG activity and why it should be controlled for to avoid confounds?

As stated in the previous section, EEG sensors are measuring all electrical currents at the sensor location, not only those coming from the brain. In particular, muscle tensions and eye movements generate electrical currents, known respectively as electromyography (EMG) and electrooculography (EOG), that are of much larger magnitude than currents of cortical origin, i.e., signals originating from the brain (Fatourechi 2007) (see Figure 1). EOG signals may dominate the EEG at frontal scalp sites (near the eyes), while EMG signals are common for electrodes placed near muscles (Goncharova 2003), contributing broad band and in particular high frequency activity (Whitham 2007).

EOG/EMG artefacts can corrupt EEG signals and result in poor BCI performance. However, EMG and EOG can also contribute to BCI control and lead to high classification accuracy if their presence or amplitude happens to be correlated with that of the mental states decoded by the BCI. In other words, EOG and EMG are typical confounding factors in BCI experiments. Thus, one may conclude that a given mental state can be decoded from EEG when it is actually decoded from EMG/EOG.

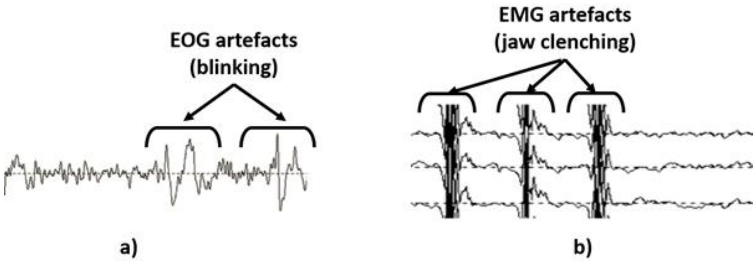

Figure 1: Examples of EOG (a) and EMG (b) measured by EEG sensors (Lotte 2015a).

Such a risk of EMG/EOG confound can for instance be found when decoding emotions from EEG signals (Mühl 2014b). Indeed, a given emotion often appears with a given facial expression, e.g., by frowning when experiencing anger. This would lead to specific facial muscle contractions and thus to specific EMG signatures that would be picked up by EEG sensors. Thus, a BCI classifier based on these EEG signals may actually recognize facial expressions from EMG signals rather than actual emotions from cortical brain signals. As Mühl and colleagues pointed out, if one want to ensure the classification is truly based on cortical signals, it is necessary to control for EMG or EOG activity when classifying emotions from EEG recordings (Mühl 2014b).

This is only one example, and ideally all BCIs, relying solely on brain activity, should controlled for EOG/EMG confounds. While there is no perfect solution, minimizing such confounds can be done by:

- Designing protocols limiting eye and motor/facial movements.
- Using manual or automatic EOG/EMG cleaning techniques (Islam 2016) (note that existing techniques are still improvable).
- Studying and reporting the spatial and spectral EEG topographies for each condition, to ensure they differ from that of EMG/EOG (Goncharova 2003).
- Classifying directly EOG/EMG to see if they contain class-related information. If so, there is a risk of confound
- Performing classification on high frequency EEG signals, likely to contain EMG, to estimate EMG impact (see, e.g., Mühl (2014a)).

- Re-Analyzing BCI data, offline, after removal of artefactual signals

And for our experiment? Here, we will definitely measure EOG and EMG of the face during the experiment as we use emotional tasks that are most likely related to facial expressions. Then, we will apply EOG/EMG cleaning techniques online. Afterwards, we will classify EOG/EMG (offline) to ensure they do not contain class-related information, and also classify artifact-corrected EEG data (offline), to verify BCI performance is above chance (and as high as possible).

1.4 - What kinds of mental states can be used?

The question of what kinds of mental states can be used to operate a BCI goes back to the question of what kinds of mental states can be detected with the neuroimaging technique at hand. Which in turn is simply a matter of the signal-to-noise-ratio. In other words, if the neural signature of a mental state is larger than the background noise, the mental state can be detected and used to operate a BCI. Although some mental states can be detected on a single-trial level, generally the signal-to-noise ratio, and thus the reliability of the BCI output, can be increased with more repetitions. And although more repetitions typically increase the single-trial classification accuracy, repetitions may come at the expense of a lower information transfer rate. Consequently, while theoretically every mental state that can be detected with the neuroimaging technique used can be employed to operate a BCI, the actual range is narrowed down by practical considerations such as the reliability of the neural signature, the speed and accuracy of the BCI system as well as the BCI application.

The variety of types of mental states that can be used within BCIs has led to classify BCIs from active over reactive up to passive (Mühl 2009; Zander 2011). As raised in the introduction, active BCIs require direct and conscious modulation of brain activity, whereby external stimulations serve at most as cues. Motor imagery, the mental imagination of movements, is a prominent active BCI paradigm (Pfurtscheller 1997). Contrariwise, reactive BCIs relay on the indirect modulation of brain activity as a reaction to an external stimulation. Well known examples for reactive BCIs are the P300 speller (De Vos 2014; Farwell 1988) and BCIs that are based on steady state visual / somatosensory evoked potentials (Lalor 2005; Müller-Putz 2005; Müller-Putz 2006). Finally, passive BCIs use brain activity arising without the users' conscious modulation or without external stimulation, such as in the detection of error potentials (Zander 2011). Additionally, different kinds of BCIs can be combined together, to make what is called an hybrid BCI, see (Pfurstcheller et al, 2010) and chapter “Hybrid Brain-Computer Interfaces and Their Applications” in this book. Given a BCI application, it is advisable to use the mental state which optimally balances accuracy and speed for the target application.

And for our experiment? Our experiment aims at investigating whether or not remembering positive/negative emotional souvenirs is reliable enough to control a BCI: thus, we are focusing on active BCIs. In other words, we want to investigate the feasibility of discriminating both these tasks on a single trial basis. Importantly, the length of the trial will be limited: the system has to be able to discriminate the tasks in a few seconds. In other words, the information transfer rate must be at least as good as with standard motor-imagery tasks for these new tasks to be useful for BCI control.

II - Data Processing

2.1 - Which classifier can we use depending on the distribution of the data?

The choice of a machine learning algorithm significantly impacts on the BCI decoding accuracy. To obtain optimal performance, the algorithm capabilities and the data properties have to match. Statistical classifiers are mainly used in BCI (Lotte 2007). In order to discriminate EEG signals into different classes (commands), such classifiers rely on the EEG feature data distribution, i.e., on their probability density function. There are different ways to estimate such density and to infer classifiers from them. Most importantly, the classifier type should be selected based on this data distribution. In general, linear methods such as linear discriminant analysis (LDA) or support vector machines (SVM) are used for classification of EEG signals. Such methods use linear hyperplanes to subdivide the feature space into regions belonging to the different classes/commands. The position and orientation of the hyperplanes are typically computed using the mean and covariance of the EEG features. During subsequent online BCI use, new and unseen data, i.e., independent data, is then assigned to the label of their area, as defined by the hyperplanes. Linear methods are successful when the features density follows a normal distribution. If it does not, the features can be transformed to be so. For example, applying a logarithm transformation converts band power estimates extracted from EEG (the power is the square of the amplitude, and hence not a linear measure) into a normal-like distribution. A linear hyperplane between two normal distributions can therefore potentially lead to reasonable performance. An advantage of linear methods is that they do not need a lot of training data compared to nonlinear methods, and thus a shorter calibration time when used online. Since the lack of sufficient training data is a common issue in BCI (Lotte 2015b), linear methods are leading. As far as LDA is concerned, when little data is available, regularized LDA (notably shrinkage LDA) should be used preferably to LDA, as they were shown to lead to higher accuracy (Blankertz et al, 2011; Lotte 2015b). Nonlinear methods can separate data using curves instead of "just" (hyper)planes, and thus can better capture the shape of the feature density, which may lead to good generalization. Generalization is the ability of a classifier to achieve good results also with independent data. Generalizing is different from memorizing the data (a.k.a overfitting): more details on this point are provided in the next section. Recent nonlinear methods that are somewhat robust against overfitting are Random Forest (Steyrl 2016) and neural networks such as Restricted Boltzmann Machines (Kober 2016). These methods do not make any assumption about the data distribution, and as such, are likely to be increasingly used for BCI in the future, and to give potentially better performances. Proving superior performances nonetheless require a proper evaluation of the algorithms, and notably the use of separate training and testing datasets.

As for classifiers, the other algorithms used in the data processing pipeline, e.g., preprocessing or feature extraction, should also match the data properties. For instance, the Common Spatial Pattern (CSP) spatial filter (Blankertz et al, 2008) can be used to classify oscillatory activity data (EEG rhythm band power) but is suboptimal to classify Event Related Potentials (ERP) such as the P300 (Lotte 2014). For the latter, dedicated spatial filters such as xDAWN should be used (Rivet et al, 2009). For a complete review of the signal processing and machine learning methods used in BCI, please refer to other chapter of this book, e.g., Blankertz (2018) and Chevallier (2018).

And for our experiment? We do not want the calibration of the classifier to be too long. Thus, we will choose a linear classifier, as it requires fewer training data. Moreover, to obtain features respecting the properties of a standard LDA or SVM classifier, we will operate the following transformation of the features in order to obtain a normal-like distribution: (1) compute the power of the signal and (2) take the log of this power.

2.2 - Why is it important to separate the training dataset from the testing dataset?

When evaluating BCIs offline, the classification system should be calibrated using only the data from a training dataset, and the resulting calibrated system should then be evaluated on completely different and independent data, which represents the testing dataset (Lemm 2011). Such evaluation ensures the classifier can indeed generalize to unseen data, and has not just memorized the training dataset class labels. Failure to use distinct and independent training and testing datasets may lead to much higher classification performance than the “real” performance that would be obtained on unseen data (Olivetti 2010), or even to better than chance performance on random data with no class information (Dominguez 2009).

Therefore, not a single parameter of the machine learning algorithm should have been calibrated with the knowledge of the testing dataset. This means the choice of channels, features, hyperparameters and normalization should be done using only the training data. In particular, when using cross-validation (CV)[2], all these should **not** be selected on all the data before applying CV to assess the classifier. Rather, they should be selected separately for each fold of the CV. Misuse of this procedure could result in erroneously concluding that some mental states can be discriminated from brain signals, when they cannot. Such a bias was revealed in (Luu 2008), which claimed that it was possible to predict from functional Near-InfraRed Spectroscopy (fNRIS) which object among two a user was going to choose, before she could see them. The paper claimed they could discriminate the preferred object from the non-preferred one, based on fNIRS signals preceding their presentation, with 80% of average CV classification accuracy (Luu 2008). Unfortunately, feature selection was performed not for each fold of the CV, as it should have, but only once on all the data. Dominguez revealed in (Dominguez 2009) that

[2] CV is a procedure which divides a dataset randomly into K folds of equal size, then spatial filter optimization, feature selection and classifier training are performed on all but one part and the obtained predictions compared to the real labels of the remaining part.

by applying the exact same incorrect CV evaluation procedure on completely randomly generated noise – with no class information – he could obtain the same classification accuracy. Subsequently, when correcting their approach, and performing feature selection only on each training CV fold, the authors of the original paper obtained a classification accuracy as low as 56% (vs. 80% before), i.e., essentially chance level, thus disproving their initial claim (Chau 2009). This stresses once more the need for an independent testing dataset, i.e., a dataset that no part of the machine learning algorithm has ever used.

And for our experiment? In order to assess whether remembering a positive emotional souvenir can be distinguished in EEG signals from remembering a negative one, we will use machine learning algorithms such as LDA/SVM, feature selection and spatial filter optimization. To assess the achievable classification accuracy, we will use K-fold CV (a typical value of K would be 10). Therefore, we will run the LDA/SVM training algorithm, the feature selection algorithm and the spatial filter optimization algorithm on K-1 folds and evaluate the obtained parameters on the remaining fold (i.e., the testing dataset). This training/testing procedure should thus be done K times, for the classifier training, feature selection and filter optimization.

2.3 - How to determine the chance level for the classified data?

In the example introduced just above, we state that 56% of classification accuracy was basically chance level. But how do we know from which accuracy onwards we can consider a result to be above chance? It is indeed an important question to determine whether a classification result, the decoding accuracy, deviates from chance-level or not. Chance-level refers to the rate achieved by random classification. For a 2-class problem, the theoretical chance level is 50%, for a 4-class problem it is 25%, etc. However, achieving a classification result of 70% in a 2–class scenario may or may not indicate a valid above chance classification accuracy. It is important to recognize that the theoretical chance-level is valid only for a large number of samples (or trials). Imagine flipping a coin once, the result could be either head or tails. Now assume you flip the coin four times in a row. By chance alone, it may be that the outcome is four times head. At least, and even though the coin would not be biased, it is not at all certain that the result will be two times head and two times tail. Due to the small number of trials, a strong deviation of the observed rate from the theoretical chance-level can occur. Only with a sufficient number of trials the outcome of heads only becomes more and more unlikely – and for a sufficient number of observations the frequency of heads and tails will approach the theoretical chance-level of 50%. So, if the theoretical chance-level can be exceeded by chance alone, how could it be determined whether a particular classification result is significantly above chance? Analytical (binomial statistic) and empirical approaches (permutation tests) are available to answer this question, as summarized in (Müller-Putz 2008, Combrisson 2015). By assuming that classification errors follow a cumulative binomial distribution, one can apply the inverse binomial cumulative distribution to figure out whether a particular classification result is above a particular significance threshold. The critical number of correctly classified trials that could arise by chance alone is determined by:

Crit_trials = binoinv(1 – p, n, 1/c) * 100/n

Here, p refers to the significance threshold (e.g., p = 0.05), n indicates the number of trials, and c the number of classes (assuming equal class occurrence). The binoinv function is available for instance in Matlab (Mathworks Inc., MA, USA). According to this equation, for a 2-class problem, a p = 0.05 and n = 20 trials, the critical number of correctly classified trials is 70%. Hence, only a classification accuracy exceeding 70% can be interpreted as a result significant above chance level. A corresponding look-up table is provided by Müller-Putz et al. (Müller-Putz 2008) and Combrisson & Jerbi (Combrisson 2015) for different p-values, 2-, 4- and 8-class scenarios and different trial counts. As discussed by Combrisson & Jerbi (Combrisson 2015), the analytical approach has some theoretical limitations, whereas the empirical approach comes at the expense of high computational costs. However, for random noise data, both suggest similar thresholds.

The procedure just described refers to random data, typically resulting in balanced class labels for a sufficient number of trials. However, in unbalanced 2-class situations where one class occurs in the majority of all trials (e.g. 90% non target class trials, 10% target class trials in paradigms of reactive BCIs such as the P300 speller, when measuring single trial P300 detection performances), the calculation of accuracies across all trials can be highly misleading. If one always goes with the majority vote the resulting accuracy may falsely indicate very high recognition rate: Imagine a classifier always voting 1, then the accuracy would be incorrectly calculated as 90%! In those cases, the confusion or error matrix should be reported, which makes it easy to evaluate which classes are confused. One simple solution is the calculation of a corrected accuracy given as the mean across all recognition rates for all classes. In the above example, for class 1 the recognition rate would be 100%, for class 2 the recognition rate would be 0%, and the average of both is then 50% - the corrected classification accuracy. Another frequently used performance metric in that case is the area under the receiver operating characteristic curve (Bradley 1997). The interested reader can refer to the tutorial in (Thompson 2014) for more information on classification performance metrics. Alternatively or additionally, the end-application performances can also be reported, e.g., the percentage of correctly selected characters in the P300-speller, which is a balanced problem.

And for our experiment? Once our cross-validation process performed, we will have to answer the question: Can remembering a positive emotional souvenir be distinguished from remembering a negative one, in EEG? The answer depends on whether the obtained classification accuracy is above or below chance level. If it is above it means that these two tasks can actually be distinguished, otherwise it means the classifier did not manage to separate them. To know the chance level, it is enough to have a look at the tables introduced here-above and look at the chance level for 2 classes (because here we have positive vs. negative emotional souvenirs). The chance level will also depend on the number of trials per class. This should thus encourage to collect as many trials as possible (the higher the number of trials, the lower the chance level).

2.4 - To which extent can commercial algorithms be trusted?

Consumer-grade EEG and BCI systems, such as the popular Neurosky (www.neurosky.com), Emotiv (www.emotiv.com) or many other devices, are increasingly used. Many commercial BCI systems come with ready-to-use algorithms to detect mental states such as attention, emotions or meditation. Such algorithms are often used as they are, as a ground truth value, notably in Human-Computer Interaction (HCI) research. This could sometimes be an issue to design rigorous BCI experiment for a number of reasons.

First, many (but fortunately not all) of these algorithms claim to be able to measure such mental states but have never been scientifically validated. It does not mean that they do not work, but rather that one does not know if they work. As such, if they were not independently and rigorously validated in a scientific journal publication (whose rigor standards are usually higher than that of conferences), such algorithms cannot be used as reliable measures. For instance, an independent evaluation of the Emotiv emotion recognition algorithms revealed that “the data is unreliable and incoherent” (Jorgensen 2012).

Second, such algorithms are most often black boxes, meaning one does not know how they work and which features they use. This makes a study using them potentially unreproducible with other EEG devices. Moreover, it also prevents us from assessing whether these algorithms are really BCI, based only on signals from cortical origin, or whether they are based on confounding factors such as EOG/EMG (see also section 1.3). For instance, Neurosky and Emotiv algorithms are believed to or even admit they use EMG/EOG, and thus are not real BCI (Singer 2008).

Even for algorithms that are scientifically validated and purely based on signals from cortical origin, it should be noted that EEG signals are changing heavily due to their context of use (Mühl 2014a, Brandl 2016). As such, they should be validated again in their target context, if this context is different from the one in which the validation was performed. Finally, even for such algorithms, like any other BCI system, they are not perfect, and make frequent mistakes when estimating the mental state being performed. Therefore, they should be treated as such, and not as perfect mental state decoders.

Overall, using commercial algorithms is thus not a problem per se, and can even be very useful and convenient, but it should be done with care. In particular, they should be used in rigorous scientific BCI experiments only if they were scientifically validated, including in the target context of use, and if the algorithm is known (i.e., not a black box).

And for our experiment? Since some commercial algorithms claim to be able to recognize various emotion-related mental states, or even any custom-made mental state in EEG, we might be tempted to compare the performances they obtain for positive vs negative emotional souvenir, to the one we obtain with our own data and algorithms. However, we should refrain from doing so, at least with algorithms being black box, because we do not know whether they rely on pure EEG signals, or whether they use

EMG/EOG as well. In the latter case, any comparison would be unfair and meaningless since we use only EEG.

2.5 - How to determine the relevance of neurophysiological markers?

One crucial issue when designing BCIs is the choice of features used to encode messages. Features that do not contain relevant information add noise to the system. If the machine learning algorithm is not robust, then adding noise or redundant information may decrease the BCI decoding accuracy. One way to prevent this from happening is feature reduction or selection. This means that only features that lead to high decoding performance are given to the machine learning algorithm. Note that to ensure that such features do not lead to overfitting, i.e., that they can generalize to new unseen data, it is necessary to assess them on a different dataset than the one on which they were selected, as indicated in Section 2.2. The most crucial point, however, is that the selected features are neurophysiological meaningful. Machine learning methods generally cannot judge whether the used features are neurophysiological meaningful. For instance, artifacts not originating from the brain (e.g., EMG) may be strongly correlated with the BCI task, and would be easier to detect given their higher amplitudes. An algorithm exploiting them would however not be considered as a pure BCI, as discussed in section 1.3. A relevant concept in machine learning is "Garbage in, Garbage out", which means that if the used features are meaningless, even the best machine learning algorithm will not be able to find patterns that can be discriminated.

One way to check whether there are significant differences between conditions in the spontaneous EEG is to compute time/frequency Event-Related Desynchronization (ERD - relative amplitude decrease in a specific frequency band over defined brain areas) and Event-Related Synchronization (ERS - relative amplitude increase) maps (Pfurtscheller 1999). These maps show statistically significant changes as function of time. In other words, they show which oscillatory components undergo significant amplitude changes. If identified components are in agreement with patterns reported in the literature, then the process was successful. For instance, for a motor imagery experiment, in agreement with the literature would mean ERD in alpha and beta range over sensorimotor areas (Pfurtscheller 1997), while for a Steady-State Visual Evoked Potential (SSVEP) experiment, it would mean ERS at the stimulation frequencies over occipital areas (Vialatte 2010). As mentioned before, if the experiment consists in exploring a new mental task for BCI, then such analyses are even more necessary to ensure the machine learning algorithms is not in fact using artefacts (see also section 1.3).

Finally, it should be stressed that the weights obtained by training the machine learning algorithms cannot necessarily be used directly to identify the involved brain areas. Indeed, most classifier weights are actually "filters", and can thus give high weights to sources of noise in order to cancel them (Haufe 2014). Rather, the weights should be transformed into "patterns" before interpretation, see notably (Haufe 2014) for the linear case.

And for our experiment? In addition to the offline classification of EOG/EMG introduced in section 2.1, we will perform a time/frequency ERD/ERS analysis in order to determine the neurophysiological features involved. We hope that some neurophysiological features related to emotions (positive vs. negative valence) based on the literature, such as the frontal asymmetry (Dolcos 2004; Schmidt 2001), will be involved whereas frontal and occipital high frequencies (gamma) will not be involved, as they are most likely related to EMG (Goncharova 2003).

2.6 - Why is it important to correct for multiple comparisons?

The aim of most experiments is to test and compare alternative hypotheses. In BCI, typical hypotheses may arise from questions like: When is the sensorimotor desynchronization significantly different from baseline? Where on scalp can the beta rebound be best captured after motor imagery? In which frequency band can we observe a significant difference between two mental states? Or what signal features best discriminate between the desired alternative commands?

Although quite different, those questions share a common goal (feature identification) and a common procedure (statistical hypothesis testing). Importantly, when the data space is large, such as with EEG data that unfold in space, time and frequency, feature identification involves multiple hypothesis testing. As the size of the space of possible features increases, the number of tests or comparisons to be performed increases. And as the number of tests increases, so does the risk of concluding to a significant effect in at least one of those comparisons, by mistake. The latter means when this effect does not truly exist, or equivalently when the null hypothesis is actually true.

Controlling for that risk (called the risk alpha or type-I error) is very important since wrongly identifying an effect would typically yield a choice of BCI features that would not work in practice. The risk alpha at the level of multiple tests relates to the risk alpha at the level of a single comparison. In classical statistics, there is a wide agreement to keep that risk below 5%. This number has been chosen somewhat arbitrarily and could be chosen otherwise. What is important is to define that limit prior to the testing, so as not to bias this choice and the ensuing conclusion. 5% means that the probability p of mistakenly rejecting the null hypothesis is equal to or less than 0.05.

A classic example of multiple comparisons is the two-sided t-test, when comparing the means of two populations, say and . The null hypothesis states that . The test to reject this null is two-sided when one tests for both alternatives: and . In that case, the risk alpha for each comparison is typically set to 2.5% so that the risk alpha of mistakenly rejecting the null is 5%. Hence the family risk is equal to the number of comparisons ($n = 2$) times the risk alpha at the single test level (2.5%).

The same rationale applies to n multiple comparisons with $n > 2$ so that the risk for a single test is set to $0.05/n$ and guarantees that the family risk remains 5%. This is known as the Bonferroni correction. However, this correction is often too conservative, since it relies on the a priori assumption that the multiple comparisons are mutually independent which is rarely true in practice, at least when dealing with neurophysiological or neuroimaging data. For instance, scalp EEG data are known to be spatially blurred so that nearby sensors will likely display highly similar activities. These spatial correlations should be accounted for in order to increase the sensitivity of the statistical analysis. Similarly, EEG signals exhibit strong temporal correlations and most reliable significant effects typically expand over several tens of milliseconds. Methods have been developed so as to optimize the corrections for multiple comparisons in the particular context of brain functional data. The most popular ones are available in main academic software packages (see for instance (Litvak 2011), (Oostenveld 2011) and other articles in that same special issue). These are the corrections based on random field theory (RFT, see (Worsley 2006)) to control for the Familywise Error Rate (FWER) or on approaches to control for the False

Discovery Rate (FDR, see (Nichols 2006)). Note that the former is typically made for statistical inference on parametric images. However, non-parametric or permutation approaches are also very much used for the analysis of electrophysiological data (Maris & Oostenveld 2007).

To sum up, as the number of comparisons increases, the risk of mistakenly rejecting the null hypothesis in one of these comparisons increases. A correction is needed to reliably identify a useful feature for BCI. Be it at the individual or group level, feature identification will often require multiple testing. This reminds us very importantly, that whenever prior knowledge is available to reduce the search space to the most likely relevant features, the correction for multiple comparisons will be less drastic and the ensuing identification will be more sensitive.

And for our experiment? Here, to complete the time/frequency analysis, we would like to know if, in accordance to the literature, the frontal asymmetry varies depending on the valence of the emotional memory and if this asymmetry enables to classify reliably enough our data. Thus, we will divide the frequency range and sub-bands, for instance: delta (1-4 Hz), theta (4-8 Hz), low-alpha (8-10 Hz), high-alpha (10-12 Hz), low-beta (12-24 Hz) and high beta (24-30 Hz). Then, we will compare the frontal asymmetry for each of these frequency bands. Because we perform several comparisons, we will apply a correction, for instance the False Discovery Rate, in order to adjust the significance threshold.

III - Experimental Design & the User Component

3.1 - What to have in mind when designing a new BCI experiment?

Along the long road to develop and validate a new application, BCI experiments may have different purposes. At an early stage, offline experiments may be required to explore the neural correlates of some targeted mental processes and their potential usefulness. For instance, significant ERDs or ERSs in specific frequency bands will be investigated in a population of subjects, between two conditions that we wish to distinguish (e.g. low vs. high mental workload).

Then, online experiments are typically designed to evaluate or compare the performance of a BCI or of one of its component (a neurophysiological marker, an algorithm, a feedback…). In that case typically, the same subjects will be tested under two or more conditions in order to answer questions such as: is it useful to include the early N170 visual component in the classification to improve P300-based spelling? Which classifier provides the best performance? Does it make a different if we move from a 2D to a 3D visual feedback? Note that some of those questions can partially be answered with offline experiments in which many tests can be performed a posteriori, based on the same data. However, a full evaluation will require an online study where the different conditions to be compared will have to be evaluated in separate trials.

BCI experiments may also aim at assessing a learning curve over several sessions or at comparing groups of users. Although not mutually exclusive, these questions are very different from each other and point towards different design parameters. In particular, the former will require defining the number of sessions for each subject, while the latter will require defining the number of subjects in each group.

Finally, at a later stage of development, validation of a BCI may take the form of a randomized controlled trial (e.g. for Neurofeedback training applications) which will have to be carefully designed to efficiently demonstrate and help quantifying the desired effect. An important question, in particular, will be to control for putative confounders and ensure that the observed effect is indeed produced by the intended manipulation (e.g. in BCI, that the control is based on brain and not muscular activity, or in Neurofeedback, that the effect is specific to the modulation of the targeted neural activity).

Hence BCI experiments cover pretty much the whole spectrum of possible designs that one may encounter in empirical science. The crucial question of how to optimally design a BCI experiment can thus rely on principles derived from applied statistical works in the fields of experimental psychology, cognitive neurosciences and neuroimaging (e.g. (Henson 2006, Daunizeau 2011)).

Put simply, designing an experiment first requires to clearly state the alternative hypothesis to be tested. If properly done, this greatly constraints the experimental conditions one should consider and naturally points towards confounds that should be carefully controlled. In other words, this early and seemingly simple first step is essential to enable finding natural and proper answers to most important design questions that come next about the control group (Section 3.2), the control condition (Section 3.3) and the appropriate statistical tests (Section 3.4).

An often overlooked aspect in BCI though, like in many other fields, relates to the sampling issue. How many subjects should I test? And how many trials per subject should I record? Answering those questions is crucial for guaranteeing the reproducibility of BCI results.

Interestingly, the theoretical field of design optimization is still very active and BCI already motivated methodological innovations in that field, be it for maximizing design efficiency with respect to hypothesis testing or parameter estimation, by optimizing the stimuli or the number of samples (trials and subjects) (see for instance (Sanchez 2016) and (Melinscak & Montesano 2016)).

And for our experiment? Here, our objective is to "explore the neural correlates of some targeted mental processes and their potential usefulness", the mental process being remembering positive/negative emotional souvenirs. Given the high between and intra-subject variability in terms of BCI performance, in order to obtain a statistically significant response to our hypothesis, we need an important number of participants (at least 20 would be good) and many trials (a typical BCI session would be 4 runs of 20 trials per class, i.e., 80 trials per class in total).

3.2 - Why and how to have a good control group?

Broadly speaking one can distinguish between feasibility studies and controlled studies. As the name indicates, feasibility studies, also referred to as proof-of-concept studies, analyze the viability of an idea. Often feasibility studies are designed to pave the way for future controlled studies. To take but out one example, Gharabaghi et al. (Gharabaghi 2014) demonstrated that closing the loop between mental states, cortical stimulation and haptic feedback is feasible. Building on this, larger (clinical) controlled studies are necessary to evaluate the utility of this approach for the rehabilitation of lost motor function after stroke. It is not unusual that at the beginning of a scientific achievement a higher rate of feasibility studies is performed, which, at success, are followed by controlled studies. Contrary to most feasibility studies, controlled studies comprise an experimental group and a control group, whereby the nature of the control group largely depends on the research question. For instance, if one wants to assess the effect of age on the accuracy of a motor imagery based BCI, the experimental group could comprise older individuals and the control group younger individuals. If one is, however, interested in the consequences of stroke on the ability to steer a motor imagery based BCI, the experimental group could consist of stroke patients and the control group of healthy individuals. If, to name another example, one wants to examine the rehabilitative effects of motor imagery based BCI training after stroke, both, experimental and control group, should consists of stroke patients, whereby the experimental group receives real feedback and the control group sham feedback during the BCI training (for details on the aspects of sham feedback, see next section). In all cases, the experimental group and the control group differ ideally only with regard to the independent variable. This can be achieved by matching individuals on variables of putative importance but of non-interest, such as gender, age, education, or ideally, if possible, employ a within-subject design. Taken together, well-controlled studies have the potential to isolate the effect of the independent variable.

And for our experiment? We are proposing a feasibility study: the goal being to investigate whether or not it is possible to control a BCI using the remembrance of positive/negative emotional souvenirs. This step being preliminary, a control group is not mandatory. Nonetheless, we also aim to assess the relevance of these tasks in comparison to more standard motor-imagery tasks such as imagining left- and right-hand movements. This is why we will propose a within-subject design with 2 conditions: in some runs, participants will perform the new tasks (remembering positive vs. negative emotional souvenirs), while in other intermixed runs, they will perform standard left- and right-hand motor-imagery tasks. To keep a high number of trials per class (as required, see section 3.2), participants will take part in two sessions of 5 runs.

3.3 - How to avoid biases? Concepts of counterbalancing, sham control, double blindness and randomisation.

In order to avoid biases, a couple of concepts can be useful. One of these is to compare the effects of real feedback with the effects of sham feedback. The so-called sham-controlled designs enable to better evaluate the effectiveness and specificity of the feedback. In other words, the inclusion of a sham-control group is crucial to control for non-specific factors such as motivation, expectancy and practice effects. Based on these principles, there seems to be general agreement that the inclusion of a sham-control group is of advantage. However, at present there are no common criterions for the optimal sham-control condition. The existing sham-control conditions can be mainly assigned to five groups: (1) no feedback (Cohen Kadosh 2016; Zich 2015); (2) feedback based on activity stemming from a different brain region (Harmelchen 2015; Paret 2016; Yoo 2012; Yao 2016; Zotev 2016); (3) feedback based on the activity from a different point in time, e.g. different trial or session (Braun 2016; Okazaki 2014); (4) feedback based on activity from a different user (Chiew 2012; Engelbregt 2016; Escolano 2014; Kober 2014; Ros 2013; Witte 2013) and (5) feedback based on artificially created irrelevant randomized signals (Arnold 2012; Mihara 2012; Mihara 2013). Furthermore, Gevensleben et al. (Gevensleben 2014) designed for their learning study a particular exceptional sham-control condition. In brief, feedback was based on data from a previous study, providing a variety of different feedback curves, which were additionally weighted by coefficients to control the development over time (Gevensleben 2014). To the best of our knowledge, there is no ideal sham-control condition[3] at the present time, which is why it is even more important to indicate in detail what kind of sham-control condition was used. Furthermore, the ethical concerns of sham-control conditions should be considered, this particular applies for clinical research were standard treatment exists (La Vaque 2001; Vernon 2004).

For sham-control designs, but also other group comparisons (e.g. two different mental strategies) the question arises whether to employ a within-subject design or a between-subject design. Each has its advantages and disadvantages, for instance while a between-subject design avoids order and carryover effects inter-individual variation introduce non-specific difference between the experimental groups. In each case, it is advisable to (pseudo-)randomize and counterbalance the conditions, ideally in a double blind manner. While no randomization and counterbalancing can introduce order effects no blinding can compromise the objectivity of the evaluation.

[3] Indeed, the above-mentioned control conditions do not control exactly for the same apects/effects; therefore, it would be hard if not impossible to imagine a sham condition that would control for all of them.

And for our experiment? Here, we are performing offline analyses. In other words, we do not classify online the data and therefore do not propose a closed-loop BCI. Thus, participants will not be provided with a feedback. Furthermore, as we propose a within-subject design with the comparison of 2 pairs of mental-imagery tasks, we have to choose to either randomize or counterbalance the conditions in order to avoid order effects. Given the low number of participants, it will be more relevant to counterbalance the conditions. The experiment will be conducted in a single-bind manner: the experimenter is blind to the order of the 2 mental-imagery tasks.

3.4 - How to select the appropriate statistical tests?

The type of statistical test to use depends on the research question, i.e., more precisely on the hypothesis. Typically, a dependency analysis is performed: either we want to find differences (univariate analyses) or correlations (bivariate analyses). There are two cases if one looks for differences. Either the samples are independent or related (paired). For instance, if we have a variable with two modalities, A and B, in (1) a between-subject design group 1 will use the modality A and group 2 the modality B: the samples will be independent; (2) a within-subject design, all the participants will use both modalities: the samples will be related/paired. Then, it is also important to pick the method based on the distribution of the data: if the data have a normal distribution, it will be possible to use parametric tests, otherwise, non-parametric tests should be used (although it is worth noting that parametric tests are fairly robust to deviations from the normal distribution). Concerning bivariate analyses, if the data's distribution is normal, a Pearson correlation analysis should be performed, otherwise, a Spearman rank correlation should be used. Concerning univariate analyses, for a better readability, we propose a table to find the appropriate statistical test depending on (1) the number of variables, (2) the fact the samples are independent of not and (3) the distribution of the data.

Parametric tests:

	INDEPENDENT/UNPAIRED SAMPLES	PAIRED SAMPLES

2 SAMPLES	*Variances assumed equal:* Student t-test *Otherwise:* Welch t-test	Paired t-test
3+ SAMPLES	*Variances assumed equal:* n-way ANOVA *Otherwise:* Krustall-Wallis test	ANOVA for repeated measures

Non parametric tests:

	INDEPENDENT/UNPAIRED SAMPLES	PAIRED SAMPLES
2 SAMPLES	Mann & Whitney U-test	Wilcoxon signed rank test
3+ SAMPLES	Krustall Wallis	Friedman

Then, when we have 3 or more groups, if the analysis shows significant effects (for instance $p<0.05$), post-tests can be performed for which correction for multiple comparisons should be applied (see Section 2.6).

And for our experiment? In our experiment, we want to investigate the difference of BCI performance between the tasks “remembering positive/negative emotional souvenirs” and “imagining left/right-hand movements”. Thus, we will do a univariate analysis. Also, we have 1 variable “MI tasks” with 2 modalities: “remembering emotional souvenirs” and “performing motor-imagery”; we use a within-subject design and counterbalance the conditions to avoid order effects. Thus, we will average the performance obtained at the five runs of each condition in order to obtain one measure of performance for each pair of tasks, for each participant. We will analyze the distribution of the data, but given the small sample (20 participants) we will most likely have to use a non-parametric test. Thus, we will do a univariate non-parametric test for 2 paired samples: a Wilcoxon signed rank test.

Summary & Conclusion

To conclude, we propose, in the following table, a summary of the key points to be considered to design a rigorous BCI experiment. Once again, we do not claim to be exhaustive, but hope that we tackled the key points that will enable the reader to understand how to avoid common pitfalls when designing a BCI experiment. Nonetheless, it has to be noted that designing a rigorous experiment is often not enough to guarantee the significance and relevance of the latest. It is also of the utmost importance to question the scientific relevance of the study, to carefully acknowledge the impact of the user training and feedback (Kleih & Kübler, 2018; Mladenovic et al., 2018), as well as to consider the impact of social and relational aspects, i.e., of the way the study is performed, the relationship between the experimenter and the participant/patient.

	TO BE REMEMBERED
I – Acquisition of the signals	
1.1 – How to choose the appropriate sensors' type, location and number depending on what we want to measure?	Currently, a trade-off has to be done between the comfort/ease of use and the quality of the signal. Ag/AgCl wet electrodes offer the highest quality of signal. Using the 10-20 system with at least 32 electrodes enables a large coverage of the scalp. Nonetheless, for economical/practical reasons, lighter set-ups can be used.
1.2 – What can we deduct from the activation of one electrode? Concept of spatial resolution.	Given that an unknown number of brain (and non-brain) generators contribute to the scalp-EEG, one cannot determine the brain source of any particular EEG signal for sure only based on the activity measure at one electrode location. EEG source localization procedures can be used to confirm ideas about brain sources contributing to the EEG.
1.3 – How muscular activity can interfere with EEG activity and why it should be controlled for to avoid confounds?	Ocular (EOG) and muscular (EMG) activity are also measured by EEG sensors. They are typical confounding factors in BCI experiments, when correlated with the EEG patterns used by the BCI. They should thus be controlled for in any study.
1.4 – What kinds of mental states can be used?	Theoretically, any mental state can be used while its neural signature is larger than the background noise. Then, the mental state should be selected depending on the information-transfer rate and on the classification accuracy required for the target application.
II – Data processing	
2.1 – Which classifier can we use depending on the distribution of the data?	When a few data are available for the calibration, linear classifiers such as LDA or SVM should be used (to avoid overfitting). It should be reminded that to be allowed to

	use these classifiers, a transformation of the features must be performed to obtain a normal-like distribution.
2.2 – Why is it important to separate the training dataset from the testing dataset?	During offline analyses, cross validation procedures should use testing datasets that are independent from the training dataset to ensure the algorithm can indeed generalize to unseen data, and has not just memorized the training dataset class labels.
2.3 – How to determine the chance level for the classified data?	The chance level depends on the number of classes and of the number of trials per class. To know the chance level, it is enough to have a look at the tables presented in the papers cited in this section.
2.4 – To which extent can commercial algorithms be trusted?	Using commercial algorithms can be very useful and convenient, but it should be done with care. In particular, these algorithms should be used in rigorous scientific BCI experiments only if they were scientifically validated, including in the target context of use.
2.5 – How to determine the relevance of neurophysiological markers?	One way to determine the relevance of neurophysiological markers is to do time/frequency analyses of ERD and ERS and to compare the highlighted features to the literature.
2.6 – Why is it important to perform corrections for multiple comparisons?	As the number of tests increases, so does the risk of concluding to a significant effect in at least one comparison by mistake. Controlling for that risk (called the risk alpha or type-I error) is very important since wrongly identifying an effect would typically yield a choice of BCI features that would not work in practice.
III – Experimental design & the user component	
3.1 – What to have in mind when designing a new BCI experiment?	Designing an experiment first requires to clearly state the alternative hypotheses to be tested. If properly done, this greatly constraints the experimental conditions one should consider and naturally points towards confounds that should be carefully controlled.
3.2 – Why and how to have a good control group?	Control groups are required to prove the efficiency of a new paradigm/approach. The nature of this control group should be defined depending on the hypotheses and on the goal of the experiment.
3.3 – How to avoid biases? Concepts of counterbalancing, sham control, double blindness and randomization.	The so-called sham-controlled designs enable to better evaluate the effectiveness and specificity of the feedback. Also, it is advisable to randomize or counterbalance the conditions, ideally in a double blind manner. While no randomization and counterbalancing can introduce order effects, no blinding can compromise the objectivity of the evaluation.
3.4 – How to select the appropriate statistical tests?	The type of statistical test to use depends on the research question, and more precisely on the hypothesis. Typically, a dependency analysis is performed: either we want to find differences or correlations. Then, to determine which test to perform, 3 questions should be answered: (1) is the distribution of the data normal-like?,

	(2) are the samples paired?, (3) how many groups are involved?

Acknowledgements

CJ and FL were supported by the French National Research Agency with the REBEL project (grant ANR-15-CE23-0013-01). JM is supported by the LabEx Cortex ("Construction, Function and Cognitive Function and Rehabilitation of the Cortex", grant ANR-11-LABX-0042) of Université de Lyon.

References

Arnold, L.E., Lofthouse, N., Hersch, S., Pan, X., hurt, E., Bates, B., Kassouf, K., Moone, S., Grantier, C. 2012. “EEG Neurofeedback for ADHD: Double-Blind Sham-Controlled Randomized Pilot Feasibility Trial”, *Journal of Attention Disorders*, 17(5) 410-419.

Baillet, S., Mosher, J.C., and Leahy, R.M. 2001. Electromagnetic brain mapping. *IEEE Signal Process. Mag.*, 18 (6), 14–30.

Blankertz, B. 2018. Gentle Introduction to Signal Processing and Classification for Single-Trial EEG Analysis. Book: Brain-Computer Interfaces Handbook – Technological and Theoretical Advances.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K. R. 2011. Single-trial analysis and classification of ERP components—a tutorial. *NeuroImage*, *56*(2).

Brandl, S., Frølich, L., Höhne, J., Müller, K. R., & Samek, W. 2016. Brain–computer interfacing under distraction: an evaluation study. *Journal of neural engineering*, *13*(5), 056012.

Braun, N., Emkes, R., Thorne J.D., Debener, S. 2016. Embodied neurofeedback with an anthropomorphic robotic hand, *Sci Rep*, 6, 37696.

Chau, T., Damouras, S. 2009. Reply to: On the risk of extracting relevant information from random data. *Journal of neural engineering*, 6(5).

Chevallier, S. 2018. Riemannian Classification for SSVEP Based BCI: Offline Versus Online Implementations. Book: Brain-Computer Interfaces Handbook – Technological and Theoretical Advances.

Chiew, M., LaConte, S.M., Graham, S.J., 2012. Investigation of fMRI neurofeedback of differential primary motor cortex activity using kinesthetic motor imagery, *NeuroImage*, 61(1), 21-31.

Combrisson, E., & Jerbi, K. 2015. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods*, *250*, 126-136.

Daunizeau, J., Preuschoff, K., Friston, K., & Stephan, K. 2011. Optimizing Experimental Design for Comparing Models of Brain Function. *PLoS Comput. Biol.,* 7 (11), e1002280.

Debener, S., Emkes, R., De Vos, M., & Bleichner, M. (2015). Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear. *Scientific reports*, *5*, 16743.

De Vos, M., Kroesen, M., Emkes, R., & Debener, S. 2014. P300 speller BCI with a mobile EEG system: comparison to a traditional amplifier. *Journal of neural engineering*, *11*(3), 036008.

Dolcos, F., LaBar, K. S., Cabeza, R. 2004. Dissociable effects of arousal and valence on prefrontal activity indexing emotional evaluation and subsequent memory: an event-related fMRI study. *Neuroimage*, *23*(1), 64-74.

Dominguez, L. G. 2009. On the risk of extracting relevant information from random data. *Journal of neural engineering*, 6(5), 058001.

Engelbregt, H.J., Keeser, D., van Eijk, L., Suiker, E.M., Eichhorn, D., Karch, S., Deijen J.B., Pogarell, O. 2016. Short and long-term effects of sham-controlled prefrontal EEG-neurofeedback training in healthy subjects, *Clinical Neurophysiology*, 127(4), 1931-1937.

Escolano, C., Navarro-Gil, M., Garcia-Campayo, J., Minguez, J. 2014 The effects of a Single Session of Upper Alpha Neurofeedback for Cognitive Enhancement: A Sham-controlled study, *Applied Psychophysiology and Biofeedback*, 39(3), 227-236.

Farwell, L., Donchin, E. 1988. Talking off the top of your head: towards a mental prosthesis utilizing event-related brain potentials, *Electroencephalography and Clinical Neurophysiology*, 70, 510-523.

Fatourechi, M., Bashashati, A., Ward, R., Birch, G., 2007. EMG and EOG artifacts in brain computer interface systems: A survey, 118, 480-494.

Frey, J., Gervais, R., Fleck, S., Lotte F., and Hachet, M., 2014c. Teegi: tangible EEG interface. In: *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, pp. 301–308.

Gevensleben, H., Albrecht, B., Lütcke, H., Auer, T., Dewiputri, W.I., Schweizer, R., Moll, G., Heinrich, H., Rothenberger, A. 2014. Neurofeedback of slow cortical potentials: neural mechanisms and feasibility of a placebo-controlled design in healthy adults, *Front Hum Neurosci.,* 8, 990.

Gharabaghi, A., Kraus, D., Leao, M.T., Spüler, M., Walter, A., Bogdan, M., Rosenstiel, W., Naros, G., Ziemann, U., 2014. Coupling brain-machine interfaces with cortical stimulation for brain-state dependent stimulation: enhancing motor cortex excitability, *Frontiers in Human Neuroscience*, 8 (122).

Goncharova, I., McFarland, D., Vaughan, T., Wolpaw, J. 2003. EMG contamination of EEG: spectral and topographical characteristics, *Clinical Neurophysiology*, 114, 1580 – 1593.

Guger, C., Spataro, R., Annen, J., Ortner, R., Irimia, D., Allison, B., La Bella, V., Cho, W., Edlinger, G., & Laureys, S. 2018. *Brain-Computer Interfaces for Motor Rehabilitation, DOC Assessment and Communication.* Book: Brain-Computer Interfaces Handbook – Technological and Theoretical Advances.

Harmelech, T., Friedman, D., Malach, R. 2015. Differential Magnetic Resonance Neurofeedback Modulations across Extrinsic (Visual) and Intrinsic (Default-Mode) Nodes of the Human Cortex, *The Journal of Neuroscience,* 35(6), 2588-2595.

Heingartner, D. 2009. *Mental block IEEE Spectrum*, 46, 42-43.

Henson, R. 2006. *Efficient experimental design for fMRI*. In Karl Friston, John Ashburner, Stefan Kiebel, Thomas Nichols, and William Penny, editors, Statistical Parametric Mapping: The Analysis of Functional Brain Images, pages 193 – 210. Elsevier.

Islam, M. K., Rastegarnia, A., & Yang, Z. 2016. Methods for artifact detection and removal from scalp EEG: A review. *Neurophysiologie Clinique/Clinical Neurophysiology*, *46*(4), 287-305.

Jorgensen, M., Bakland, T., Thorsen, E. 2017. *Satisfaction Measured Emotiv*, University of Oslo Technical Report, INF2260, 2012.

Kadosh, K.C., Luo, Q., de Burca, C., Sokunbi, M.O, Feng, J., Linden, D.E., Lau, J.Y. 2016. Using real-time fMRI to influence effective connectivity in the developing emotion regulation network *NeuroImage,* 15(125), 616-626.

Kleih, S. C. & Kübler, A. 2018. Why User-Centered Design is Relevant for Brain-Computer Interfacing and How It can be Implemented in Study Protocols. Book: Brain-Computer Interfaces Handbook – Technological and Theoretical Advances.

Kobler, R., Scherer, R. 2016. Restricted Boltzmann Machines in Sensory Motor Rhythm Brain-Computer Interfacing: A Study on Inter-Subject Transfer and Co-Adaptation. In *IEEE International Conference on Systems, Man and Cybernetics SMC 2016*.

Kober, S.E., Wood, G., Kurzmann, J., Friedrich, E.V.C., Stangl, M., Wippel, T., Väljamäe, A., Neuper, C. 2014. Near-infrared spectroscopy based neurofeedback training increases specific motor imagery related cortical activation compared to sham. *Biological Psychology*, 95, 21-30.

Kübler, A., D. Mattia, R. Rupp, and M. Tangermann 2013. Facing the challenge: Bringing brain-computer interfaces to end-users. In: *Artificial intelligence in medicine.*

La Vaque, T.J., Rossiter, T. 2001. The Ethical Use of Placebo Controls in Clinical Research: The Declaration of Helsinki. *Applied Psychophysiology and Biofeedback*, 26(1), 23-37.

Lalor, E. C., Kelly, S. P., Finucane, C., Burke, R., Smith, R., Reilly, R. B., & Mcdarby, G. 2005. Steady-state VEP-based brain-computer interface control in an immersive 3D gaming environment. *EURASIP journal on applied signal processing*, *2005*, 3156-3164.

Lécuyer, A., Lotte, F., Reilly, R., Leeb, R., Hirose, M. and Slater, M. 2008. Brain-Computer Interfaces, Virtual Reality and Videogames. In: *IEEE Computer* 41.10, pp. 66–72.

Lee, J. H., Kim, J., & Yoo, S. S. 2012. Real-time fMRI-based neurofeedback reinforces causality of attention networks. *Neuroscience research*, *72*(4), 347-354.

Lemm, S., Blankertz, B., Dickhaus, T., Müller, K. R. 2011. Introduction to machine learning for brain imaging. *Neuroimage*, 56(2), 387-399.

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

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. 2007. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of neural engineering*, *4*(2), R1.

Lotte, F., Bougrain, L. & Clerc, M. 2015. *Electroencephalography (EEG)-based Brain-Computer Interfaces* Wiley Encyclopedia on Electrical and Electronics Engineering, Wiley.

Lotte, F. 2015. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces. *Proceedings of the IEEE*, 103(6), 871-890.

Luu S., Chau, T. 2008. Decoding subjective preference from single-trial near-infrared spectroscopy signals, *Journal of Neural Engineering*, vol 6, no 1.

Maris, E., Oostenveld, R. 2007. Nonparametric statistical testing of EEG- and MEG-data, *Journal of Neuroscience Methods*, vol. 164, no 1, 177-190, 2007.

McFarland, D. J. 2018. *Therapeutic Applications of BCI Technologies*. Book: Brain-Computer Interfaces Handbook – Technological and Theoretical Advances.

Melinscak, F., and Montesano, L. 2016. Beyond p-values in the evaluation of brain–computer interfaces: A Bayesian estimation approach. *J. Neurosci. Methods*, 270, 30–45.

Mihara, M., Miyai, I., Harrori, N., Hatakenaka, M., Yagura, H., Kawano, T., Okibayashi, M., Danjo, N., Ishikawa, A., Inoue, Y., Kubota, K. 2012. Neurofeedback using real-time near-infrared spectroscopy enhances motor imagery related cortical activation, *PLoS One*, 7(3), e32234.

Mihara, M., Hattori, N., Hatakenaka, M., Yagura, H., Kawano, T., Hino, T., Miyai, I. 2013. Near-infrared Spectroscopy-mediated Neurofeedback Enhances Efficacy of Motor IMagery-based Training in Poststroke Victims. *Stroke,* 44, 1091-1098.

Millán, J. D. R., Rupp, R., Mueller-Putz, G., Murray-Smith, R., Giugliemma, C., Tangermann, M., ... & Neuper, C. 2010. Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges. *Frontiers in neuroscience*, *4*, 161.

Mladenovic, J., Mattout, J., Lotte, F. 2018. A Generic Framework for Adaptive EEG-Based BCI Training and Operation. Book: Brain-Computer Interfaces Handbook – Technological and Theoretical Advances.

Mühl, C., Bos, P.D., Thurlings, M.E., Scherffig, L., Duvinage, M., Elbakyan, A.A., … Heylen D. 2009. Bacteria Hunt: A multimodal, multiparadigm BCI game, In *Workshop Report for the Interface Workshop* (pp. 1-22). Genova, Italy.

Mühl, C., Jeunet, C., Lotte, F. 2014. EEG-based Workload Estimation Across Affective Contexts *Frontiers in Neuroscience section Neuroprosthetics*, 8, 114.

Mühl, C., Allison, B., Nijholt, A., Chanel, G. 2014. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, 1-19.

Müller-Putz, G. R., Scherer, R., Brauneis, C., Pfurtscheller, G. 2005. Steady-state visual evoked potential (SSVEP)-based communication: impact of harmonic frequency components. *Journal of Neural Engineering*,*2*(4), 123–30.

Müller-Putz, G. R., Scherer, R., Neuper, C., Pfurtscheller, G. 2006. Steady-state somatosensory evoked potentials: suitable brain signals for brain-computer interfaces? *Neural Systems and Rehabilitation Engineering IEEE Transactions*, *14*, 30–37.

Müller-Putz, G., Scherer, R., Brunner, C., Leeb, R., Pfurtscheller, G. 2008. Better than random: a closer look on BCI results. *International Journal of Bioelectromagnetism*, *10,* 52-55.

Nichols, T. 2006. *False Discovery Rate procedures*. In Karl Friston, John Ashburner, Stefan Kiebel, Thomas Nichols, and William Penny, editors, Statistical Parametric Mapping: The Analysis of Functional Brain Images, pages 246 – 252. Elsevier.

Okazaki, Y.O., Horschig, J.M., Luther, L., Oostenveld, R., Murakami, I., Jensen, O. 2015. Real-time MEG neurofeedback training of posterior alpha activity modulates subsequent visual detection performance, *NeuroImage*, 107, 323-332.

Olivetti, E., Mognon, A., Greiner, S., Avesani, P. 2010. Brain decoding: biases in error estimation. In Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD), *IEEE First Workshop on* 40-43.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J-M. 2011. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Comput. Intell. Neurosci.,* Articles ID 156869.

Paret, C., Ruf, M., Gerchen, M.F., Kluetsch, R., Demirakca, T., Jungkunz, M., Bertsch, K., Schmahl, C., Ende, G. 2016. fMRI neurofeedback of amygdala response to aversive stimuli enhances prefrontal-limbic brain connectivity, *NeuroImage* 15(125), 182-188.

Pfurtscheller, G., Neuper, C., Flotzinger, D., Pregenzer, M. 1997. EEG-based discrimination between imagination of right and left hand movement, *Electroencephalography and Clinical Neurophysiology*, 103(6), 642-651.

Ros, T., Theberge, J., Frewen P.A., Kluetsch R., Densmore, M., Calhoun, V.D., Lanius, R.A. 2013. Mind over chatter: Plastic up-regulation of the fMRI salience network directly after EEG neurofeedback, *NeuroImage*, 65, 324-335.

Sanchez, G., Lecaignard, F., Otman, A., Maby, E., Mattout, J. 2016. Active SAmpling Protocol (ASAP) to Optimize Individual Neurocognitive Hypothesis Testing: A BCI-Inspired Dynamic Experimental Design. *Front. Hum. Neurosci*., 10.

Schmidt, L. A., Trainor, L. J. 2001. Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions. *Cognition & Emotion*, *15*(4), 487-500.

Singer, E. 2008. Brain games. *Technology Review*, 111(4), 82-84.

Steyrl, D., Scherer, R., Faller, J., Müller-Putz, G. R. 2016. Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: a practical and convenient non-linear classifier. *Biomedizinische Technik. Biomedical Engineering*, 61(1), 77–86.

Vernon, D., Frick, A., Gruzelier, J. 2004. Neurofeedback as a Treatment for ADHD: A Methodological Review with Implications for Future Research, *Journal of Neurotherapy*, 8(2), 53-82.

Whitham, E. M., Pope, K. J., Fitzgibbon, S. P., Lewis, T., Clark, C. R., Loveless, S., Broberg, M., Wallace, A., DeLosAngeles, D., Lillie, P. & others. 2007. Scalp electrical recording during paralysis: quantitative evidence that EEG frequencies above 20Hz are contaminated by EMG, *Clinical Neurophysiology*, Elsevier, 118, 1877-1888.

Witte, M., Kober, S.E., Ninaus, M., Neuper, C., Wood., G. 2013. Control beliefs can predict the ability to up-regulate sensorimotor rhythm during neurofeedback training, *Frontiers in Human Neuroscience*, 7, 478.

Worsley, K. 2006. *Random Field Theory*. In Karl Friston, John Ashburner, Stefan Kiebel, Thomas Nichols, and William Penny, editors, Statistical Parametric Mapping: The Analysis of Functional Brain Images, pages 232 – 236. Elsevier.

Yao, S., Becker, B., Geng, Y., Zhao, Z., Xu, X., Zhao, W., Ren, P., Kendrick, K.M. 2016. Voluntary control of anterior insula and its functional connections is feedback-independent and increases pain empathy, *NeuroImage*, 130, 230-240.

Zander, T.O., Kothe, C. 2011. Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general, *Journal of Neural Engineering*, 8(2), 25005.

Zich, C., Debener, S., De Vos., M, Frerichs, S., Maurer, S., Kranczioch, C. 2015. Lateralization patterns of covert but not overt movements change with age: An EEG neurofeedback study, *NeuroImage*, 1(116), 80-91.

Zotev, V., Yuan, H., Misaki, M., Phillips, R., Young, K.D., Feldner, M.T., Bodurka, J. 2016. Correlation between amygdala BOLD activity and frontal EEG asymmetry during real-time fMRI neurofeedback training in patients with depression, *NeuroImage*, 11, 224-38.