Neuroscience xxx (2017) xxx–xxx

MECHANISMS OF NEUROFEEDBACK: A COMPUTATION-THEORETIC APPROACH

EDDY J. DAVELAAR

Department of Psychological Sciences, Birkbeck, University of London, Malet Street, WC1E 7HX London, United Kingdom

Abstract—Neurofeedback training is a form of brain training in which information about a neural measure is fed back to the trainee who is instructed to increase or decrease the value of that particular measure. This paper focuses on electroencephalography (EEG) neurofeedback in which the neural measures of interest are the brain oscillations. To date, the neural mechanisms that underlie successful neurofeedback training are still unexplained. Such an understanding would benefit researchers, funding agencies, clinicians, regulatory bodies, and insurance firms. Based on recent empirical work, an emerging theory couched firmly within computational neuroscience is proposed that advocates a critical role of the striatum in modulating EEG frequencies. The theory is implemented as a computer simulation of peak alpha upregulation, but in principle any frequency band at one or more electrode sites could be addressed. The simulation successfully learns to increase its peak alpha frequency and demonstrates the influence of threshold setting – the threshold that determines whether positive or negative feedback is provided. Analyses of the model suggest that neurofeedback can be likened to a search process that uses importance sampling to estimate the posterior probability distribution over striatal representational space, with each representation being associated with a distribution of values of the target EEG band. The model provides an important proof of concept to address pertinent methodological questions about how to understand and improve EEG neurofeedback success.

This article is part of a Special Issue entitled: Neurofeedback. **© 2017 IBRO. Published by Elsevier Ltd. All rights reserved.**

Key words: neurofeedback, electroencephalography, computational neuroscience, computer model.

INTRODUCTION

There are an increasing number of media reports about controlling electronic devices through brainwaves, whether it is for therapeutic reasons or for pure entertainment. These brain-computer interfaces utilize the ability of people to learn to voluntarily change their brain rhythms when provided with corrective feedback. This process is called neurofeedback and understanding how it works is the topic of this paper. Neurofeedback is a goal-directed process of modulating one's own neural dynamics by means of feedback-induced learning. The feedback that is obtained can either be internal phenomenological experiences or external provided stimulation of which visual and auditory stimulations are the most common modalities. The neural dynamics that are being influenced can be measured through electroencephalography (EEG), magnetoencephalography (MEG), blood-oxygen-level-dependent (BOLD) responses or any other direct or indirect methods.

This paper focuses on EEG neurofeedback. The aim here is to provide a proof of concept that using computational methods from neuroscience can pave the way for understanding how neurofeedback works. In several papers and books authored by leading practitioners and researchers, a common call is expressed to develop a theoretical understanding of neurofeedback (e.g., Evans and Abarbanel, 1999; Budzynski et al., 2009). Although the neural circuitry involved in some brain rhythms are understood at a descriptive level, how these rhythms are influenced through feedback-based learning at a mechanistic level is still unclear.

Why would we need to seek to understand neurofeedback? Knowledge of the mechanisms underlying neurofeedback at the neural level provides a critical foundation for (1) interpreting findings (from the lab and clinic), (2) guiding research efforts, (3) developing new protocols, (4) improving existing protocols, (5) quality assurance, (6) risk assessment and management, and (7) approval of protocols. Currently, the research on neurofeedback is making great steps in validating the efficacy of neurofeedback training (see for recent special issues Gruzelier et al., 2014a,b; van Boxtel and Gruzelier, 2014). However, there is no validated standardized methodology or a standard way of reporting the methods that have been used, leading to high study-exclusion rates in systematic reviews or meta-analyses (e.g., Emmert et al., 2016). To circumvent this, computational methods can be used to test whether certain choices, such as threshold settings, integrating neural activity across electrodes, the time window over which the neural signal is calculated, and the maximum

Abbreviations: ALIC, anterior limb of the internal capsule; BOLD, blood-oxygen-level-dependent; EEG, electroencephalography; MEG, magnetoencephalography; MSN, medium spiny neuron; PAF, peak alpha frequency; rt, real-time; SCP, slow cortical potential; UAF, upper alpha frequency.

E-mail address: e.davelaar@bbk.ac.uk

http://dx.doi.org/10.1016/j.neuroscience.2017.05.052

feedback rate, have an effect of neurofeedback success and if so what the optimal parameters are. Given that properly conducted experimental studies are costly in terms of investment of money, time, and labor, dry-testing a protocol using a computational model can make the research efforts more efficient. At present, such an opportunity does not exist.

To facilitate the process, this paper puts forward an initial step toward a computational theory that explicates the neural mechanisms underlying neurofeedback training. By way of illustration, a computational model using spiking neurons is implemented that shows successful neurofeedback training and allows an analysis of the learning process. To demonstrate its benefit, the influence of threshold setting on learning is addressed in a mathematical abstraction of the model. Future work will address how each of the aforementioned points can be supported by using a computational modelling approach.

There are many detailed descriptions of the neuroanatomical circuitry involved in the generation of brain frequencies (e.g., SCP: Birbaumer et al., 1990; SMR: Sterman, 1996; Hughes and Crunelli, 2005; Ritter et al., 2009; theta: Gruzelier, 2009; Hsieh and Ranganath, 2014), but only some reports provide suggestions of what neurofeedback might be doing to this circuitry at the neurophysiological level (e.g., Sterman, 1996; Koralek et al., 2012). Here, a general theory of neurofeedback learning is proposed that is articulated at the neurophysiological level and addresses the contributions of the striatum and the thalamus. The theory is assumed to be applicable to any neurofeedback modality (EEG, MEG, BOLD) and breaks neurofeedback learning down into three stages, of which the first stage is then implemented in a computer simulation model. In order to address how the first stage unfolds a second simulation model is analyzed in which the threshold setting for positive feedback is systematically varied. The results support the view that during stage 1, the striatum performs a search through representational space using importance sampling, i.e., maintaining only those sampled representations that lead to reward.

THE IMPORTANCE OF THE STRIATUM

Several studies employing a variety of methodologies confirm the critical contribution of the striatum in neurofeedback learning. Neuroimaging studies have shown the involvement of the entire striatum in neurofeedback (ventral striatum: Johnston et al., 2010; putamen: Hinterberger et al., 2005; caudate: Levesque et al., 2006). Johnston et al. (2010) trained participants using a real-time (rt) fMRI-neurofeedback to increase the activation in the emotion network, as defined by the collection of brain regions that was maximally responsive to negative versus neutral stimuli. One of the non-target areas that was activated during the learning process was the ventral striatum. Hinterberger et al. (2005) demonstrated the involvement of the putamen and thalamus in regulating the slow cortical potential (SCP) over Cz. Levesque et al. (2006) observed increased functional activation of the caudate in ADHD children after a theta/SMR/beta1 protocol (20 sessions SMR increase with theta decrease followed by 20 sessions beta1 increase with theta decrease) over Cz. Furthermore, in a structural MRI study, Ghaziri et al. (2013) found increased white matter density in the anterior limb of the internal capsule (ALIC) after increasing beta1 at F4 and P4 with EEG neurofeedback. Increases in fractional anisotropy in the left ALIC, which includes cortico-striatal as well as frontal cortico-thalamic fibers, was correlated with enhanced visual attention. Finally, in a critical experiment involving rats, Koralek et al. (2012) measured the activity of motor neurons and transformed this activity into an auditory signal. They showed that rats lacking cortico-striatal plasticity could not learn to control the auditory pitch. In a series of experiments they also demonstrated that cortico-striatal plasticity is necessary for neuroprosthetic control to occur. These studies provide strong support for the view that the entire striatum is involved in neurofeedback learning with lasting functional and structural consequences. Whether there is specificity in the recruitment of striatal subregions in relation to the EEG frequency, learning direction, and electrode site is yet unclear.

This is not to say that it is impossible for EEG spectrum modification to occur through synaptic changes at cortical sites only in the absence of a striatal contribution, but the current literature provides compelling evidence for a striatal theory of neurofeedback learning. In a recent meta-analysis of 12 rt-fMRI studies, Emmert et al. (2016) observed that the striatum and the anterior insula were non-target regions that were consistently activated during the neurofeedback learning. They suggested the existence of a "regulating network" of which the striatum and the anterior insula contribute through their involvement in reward-based learning and self-awareness processes, respectively. These findings are critical building blocks of the proposed theory to which we turn next.

A MULTI-STAGE THEORY OF NEUROFEEDBACK LEARNING

What happens in the brain of a person from the first training session to demonstrable voluntary control over EEG brainwaves? In the theory advanced here three stages are discerned (see Fig. 1). In the first stage (indicated by the red parts in Fig. 1), trainees may try different things, such as the strategies provided by the trainer, strategies read from the internet, or idiosyncratic strategies that come to mind during the training session. Examples of strategies are trying to relax, focus on breathing, counting numbers, thinking back of positive events, trying to become angry or positive, staring at a point on the computer screen, and many more. This stage is the problem solving or exploration stage and involves performing various mental acts and evaluating their consequences on the feedback signal. It is expected that frontal brain areas are critically involved in this stage, as it requires retrieval, creation, and maintenance of goal representations (i.e., possible strategies), execution of these strategies as response

Please cite this article in press as: Davelaar EJ. Mechanisms of neurofeedback: A computation-theoretic approach. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.05.052