An ethical perspective on Passive BCI: Can we detect information from the human mind that is intended to be hidden?

Diana-Eliza Gherman*, Thorsten O. Zander

Chair for neuroadaptive human-computer interaction, Brandenburg University of Technology, Cottbus – Senftenberg, Germany

*Corresponding author: diana.gherman@b-tu.de

Background. The last decade has seen a significant expansion of possibilities in the Brain-Computer Interface (BCI) research field. BCIs became a potentially useful communication tool between man and machine not only for the population with severe disabilities, but also for the general population. While traditional methods involved voluntary user control (*active* BCI, McFarland and Wolpaw, 2008) or user control through the response to external stimulation (*reactive* BCI, Donchin & Spencer, 2000), mental state monitoring through passive BCI (*passive* BCI) offered novel solutions to brain activity decoding (Zander and Kothe, 2011) and promises exciting applicability in real-life settings (Zander and Krol, 2017).

Passive BCIs (pBCI) give us the ability to uncover inner cognitive and affective states of users, which wouldn't normally be visible to an observer without disruptive and indirect methods such as subjective questionnaires. Besides attention, fatigue and workload (George and Lécuyer, 2010), more complex aspects of user states can also be translated automatically, without additional cognitive effort. For instance, the perception of individual error (Blankertz et al., 2002a) and perceived loss of control over a system (Zander and Jatzev, 2009), or even correlates of human-predictive coding (Zander et al., 2016) have been detected through the use of pBCI**.** Although this technology significantly enhances ergonomics of human-computer interaction, its decoding mechanism doesn't imply the need for users' awareness (Krol, Haselager and Zander, 2020), which raises ethical concerns. More specifically, the lack of consent for cognitive state retrieval poses a privacy problem for human factors and calls for a rigorous assessment of possible solutions.

Our study included a dice game that encouraged bluffing, while EEG data was recorded. Results demonstrate that pBCIs can distinguish truth from bluffing even if users try to actively conceal their true intentions through facial or general expression inhibition. In comparison with other deception detection technologies such as polygraphs which have a sole purpose, BCIs can have multiple purposes and uses at the same time. Hence, users might be misled to believe that their brain data is read for a single desired objective, while multiple other private mental state outputs are obtained by the technology providers. These worries are especially concerning if BCI commercialization becomes widespread. Analysing these ethical concerns requires deep knowledge of the system's capabilities and limits in accessing hidden mental information (Krol and Zander, in press). The high-density data was collected by Team PhyPA (Zander and Krol, 2017) in collaboration with The Technical University of Berlin and the Swartz Center for Computational Neuroscience at UCSD. Neurophysiological details and game dynamics will be discussed in a subsequent paper.

Methods.

Experimental design. The study had 12 participants (6 pairs). In order to obtain mental states corresponding to bluffing or no bluffing decisions, a task inspired by a German drinking game called "Mäxchen" was created. Each pair played 8 rounds.

Players sat in front of computer screens, facing each other, so as to have visibility to the opponent's facial expressions and upper body movements. They had an initial score of 20 points and were supposed to roll two dices with the press of a button, which would indicate a two-digit number only seen by the player rolling the dice. The purpose was to obtain a higher number than the one

announced by the previous player. Conversely, the participant had to bluff by announcing verbally a higher number or to declare quitting. All subjects' utterances were recorded. Each decision could be questioned by the opponent, who would then ask for proof. Given that the detection of bluffs would cost the player 2 points, while quitting would cost 1 point, bluffing was prone to higher loss. A wrong call would cost the opponent 2 points, while a correct one would cost no points. The game was considered won when one of the adversaries lost all of the initially given 20 points and was recompensed with additional monetary bonus.

Data analysis and Classification. To contrast bluffs versus truths, we only looked at the samples of data corresponding to trials in which the players announced the true number or trials in which they bluffed, excluding quitting trials. The pre-processing and feature extraction method followed the steps proposed by Blankertz et al. (2011). The data was sampled down to 250 Hz and a bandpass filter from 0.1 to 8 Hz was used. The features were extracted through a windowed means approach by averaging the EEG data in eight concatenated 50 ms time windows around the button press. For the training and testing classification, we used a regularized linear discriminant analysis (rLDA) and a 5x5 cross-validation method due to the small number of bluffing trials. These steps were repeated three times for each subject: for 32, 64 and 96 channels, respectively.

Results. We obtained an average performance of 75.2% (±10.2%) accuracy, with 91.7% as maximum and 53.4% as minimum for the pBCI system. As suggested by Fang et al. (2004) in their EEG study on deceptive responses in lie detection, the contingent negative variation (CNV) showed a stronger negative peak in bluffing trials than in trials where the truth was told (see *Figure 1*). This CNV usually appears in motor areas prior to a motor response and in the dorsal anterior cingulate cortex. Therefore, motor and frontal electrodes were closely inspected and compared between the two types of trials.

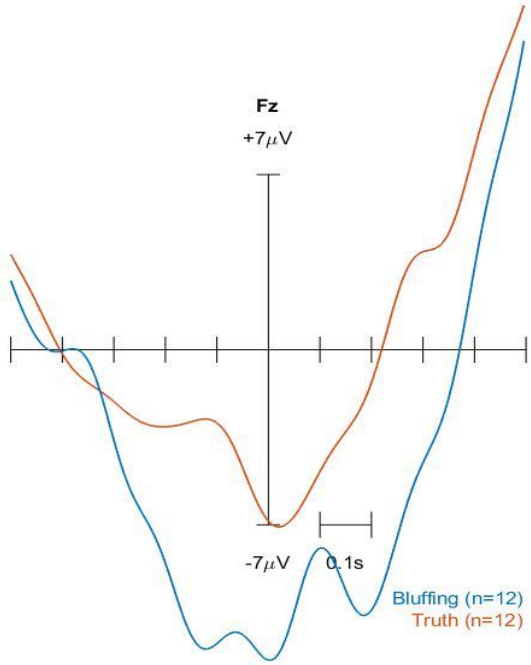

Figure 1: The grand average of the event related potentials (ERPs) of the bluffing (blue) and the truth (orange) condition.

Conclusion. Our results suggest that pBCI could distinguish bluffs from truth with reasonable accuracy, even though players tried to conceal it with conventional behavioural suppression. This potential implicit and one-sided exchange between man and machine yields ethical concerns, as it doesn't entail the user's awareness or consent and it could be used for malicious purposes. However,

before working towards ethical frameworks, more research is needed to see if users can make use of specific strategies to hide their deceiving intentions from the BCI system, similarly to other deception detection technologies.

References.

Blankertz, B., Schäfer, C., Dornhege, G., & Curio, G. (2002, August). Single trial detection of EEG error potentials: A tool for increasing BCI transmission rates. In *International Conference on Artificial Neural Networks* (pp. 1137-1143). Springer, Berlin, Heidelberg.

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), 814-825.

Donchin, E., Spencer, K. M., & Wijesinghe, R. (2000). The mental prosthesis: assessing the speed of a P300-based brain-computer interface. *IEEE transactions on rehabilitation engineering*, 8(2), 174-179.

Fang, F., Liu, Y., & Shen, Z. (2003). Lie detection with contingent negative variation. International Journal of Psychophysiology, 50(3), 247-255.

George, L., & Lécuyer, A. (2010, October). An overview of research on" passive" brain-computer interfaces for implicit human-computer interaction. *In International Conference on Applied Bionics and Biomechanics ICABB 2010-Workshop W1" Brain-Computer Interfacing and Virtual Reality*".

Krol, L. R., Haselager, P., & Zander, T. O. (2020). Cognitive and affective probing: a tutorial and review of active learning for neuroadaptive technology. Journal of neural engineering, 17(1), 012001.

Krol, L. R. & Zander, T.O. (in press). Create no evil? A call to cross ethical boundaries.*Neuroergonomics Conference*.

McFarland, D. J., & Wolpaw, J. R. (2008). Brain-computer interface operation of robotic and prosthetic devices. *Computer*, 41(10), 52-56

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), 025005.

Zander, T. O., Krol, L. R., Birbaumer, N. P., & Gramann, K. (2016). Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity. Proceedings of the National Academy of Sciences, 113(52), 14898-14903.

Zander, T. O., & Jatzev, S. (2009, September). Detecting affective covert user states with passive brain-computer interfaces. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1-9). IEEE.

Zander, T. O., & Krol, L. R. (2017). Team PhyPA: brain-computer interfacing for everyday human-computer interaction. Periodica Polytechnica Electrical Engineering and Computer Science, 61(2), 209-216.