PhD Proposal in AI and Neuroscience

Towards Robust identification and Unraveling of human-System Trust correlates (TRUST)

Nicolas DROUGARD – ISAE-SUPAERO
Bertille SOMON – ONERA

- Start date of the thesis: October 2023 – Duration: 36 months.
- Key words: *Trust, Human-System Interaction, EEG, Passive Brain-Computer Interface, Machine Learning, Signal Processing.*

1 Introduction

The increasing technology development during the last decades brings us to interact daily with automated systems. These systems are becoming more and more reliable, relegating operators to the role of passive supervisors and excluding them from the control loop [2]. Moreover, their increasing complexity has constrained the designers of these systems to make them more and more obscure, imposing on the operators a blind trust in the decisions made [4]. However, trust in automated systems poses major problems because it is regularly subject to calibration errors. Numerous research studies, both fundamental and in ecological environments, have demonstrated that this over-trust in highly reliable systems can be the cause of operational difficulties such as the inability to detect (infrequent) system errors when they appear [14]. Similarly, under-trust, or even mistrust, can also be deleterious and reduce the operational performance of the operator-automated system couple [10]. However, the mechanisms supporting the emergence of trust are still unclear and little studied. In particular, the neurophysiological markers associated with trust are poorly identified and validated [1, 5]. The literature shows that research on these correlates has not focused on characterizing trust, but on identifying its impact on specific cognitive processes, such as error detection [3] or the feeling of control [7].

One of the difficulties related to this characterization comes from the field of Brain-Computer Interfaces (BCI), where promising advances have been made in the last ten years [9], especially concerning passive BCIs which remain a major challenge in Machine Learning (ML) based on brain activity measurements. Indeed, ML techniques are hard to implement on neurophysiological data, which are generally offered in limited amounts, often very noisy, and which are subject to a very high temporal and inter-individual variability. An important field of BCI research consists in using features that have a physiological meaning, in

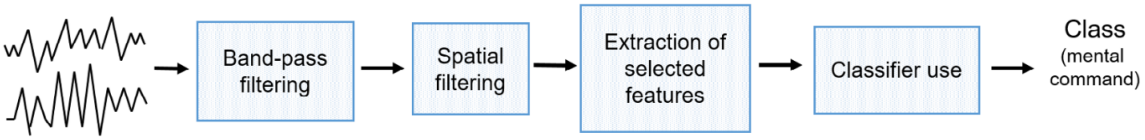

Figure 1: Typical classification process in EEG-based BCI systems, from [9].

addition to their classification performance. Obtaining this type of features enables a better understanding of the processes underlying complex phenomena.

2 Project

In this thesis, we seek to identify the brain correlates of trust in highly automated reliable systems, as well as their variations over time, based in particular on measurements of brain activity (electroencephalography, or EEG). A major emphasis is put on the characterization, understanding and evaluation of trust in a transversal way independently of the type of task or cognitive process involved. We therefore wish to determine brain correlates of trust, in the general sense of the term, and which could be measured in real-time on operators.

The contributions of this thesis concern three challenges at the interface between neuroscience and artificial intelligence:

- The brain correlates of trust that we wish to identify must have good generalization properties, *i.e.* they must account for the level of trust of the operator, with robustness, independently of the task performed (invariance challenge);
- The selected markers must allow a certain level of explainability of the mechanisms of emergence and variation of trust, and thus have a physiological meaning and plausible explanations from the neuroscience literature, in addition to their classification performances (transparency challenge);
- The algorithms for identifying and classifying markers must be usable on online and real-time measurements in order to estimate the temporal evolution of trust (immediacy challenge).

For short, the goal of this thesis is to identify new objective markers and techniques that are, at the same time, robust to non-trust related variability, defined in a rather fundamental way, and usable in real time in an operational framework.

2.1 Why?

As it has been demonstrated for other cognitive processes, or "operator states", we hypothesize that selection and classification of neurophysiological markers could allow us to better

define the trust mechanism and explain its drifts, thanks to the monitoring of its evolution in real time, and in different operating contexts. The estimation of this mental state, or cognitive process, is essential to obtain a dynamic mental map of the operator, *i.e.* the estimation of mental states of the operators in real time, in order to allow a refined analysis of the human-system missions, or even to ensure a context favorable to the success of such missions. Indeed, the use of passive brain-computer interfaces [13] able to provide this kind of dynamic mental maps, allows to develop and integrate countermeasures in the system (such as alarms, interaction mode changes, or more generally reactions), in order to extract the operator from mission-deleterious mental states, such as over- or under-trust.

2.2 How?

Measuring and identifying brain correlates and computational markers of trust in automated systems involves leveraging current available resources in terms of neuroscience knowledge, ML and BCI techniques, and EEG datasets. Indeed, an in-depth analysis of the neuroscience literature will be conducted in order to determine the currently recognized markers of trust, and their explainability in terms of the cognitive processes involved, as well as the experimental tasks performed in these studies. Also, the selection of open access EEG datasets, associated with scientific publications dealing with the identification of trust correlates, will allow to test, under various conditions and tasks, the reliability of the identified markers through statistical analyses. After defining the associated experimental protocols, new datasets will be created from the execution of experiments with electroencephalographic (EEG) data acquisition which will be conducted at ONERA and/or ISAE-SUPAERO. This data collection will allow the reproduction of the results of the literature as well as the identification of new markers. Machine learning algorithms from the literature will be trained on these two types of datasets, those retrieved from the internet and those recorded during the lab experiments, using appropriate libraries (e.g. scikit-learn [11] and MNE [6]) and evaluation (e.g. MOABB [8]). Figure 1 illustrates classical BCI pipelines from raw EEG signals to the mental state estimation, *i.e.* the trust level in our case. Newest techniques, including robust features extraction (e.g. TDA features [15]) and data augmentation [12], will be implemented.

The statistical analysis of the data as well as the Machine Learning tools implemented will be part of the results of this thesis, as well as the documented datasets resulting from the experiments, that could be shared using platforms like MOABB [8].

3 PhD candidate profile

Candidates should have a Master's degree in Cognitive Science, or a Master's degree in Machine Learning (or equivalent) with experience in Signal Processing and Statistics. Programming skills (Python and MATLAB) are required. Experience in electroencephalographic (EEG) data acquisition would be appreciated but is not mandatory.

4 Application procedure

Formal applications should include a detailed resume, a motivation letter and transcripts of master's degree. Samples of published research by the candidate and reference letters are appreciated but not necessary. Applications should be sent by email to

- Nicolas DROUGARD (`firstname.lastname@isae-supaero.fr`);
- and Bertille SOMON (`firstname.lastname@onera.fr`).

References

[1] Ighoyota Ben Ajenaghughrure, Sonia Da Costa Sousa, and David Lamas. Measuring trust with psychophysiological signals: a systematic mapping study of approaches used. *Multimodal Technologies and Interaction*, 4(3):63, 2020.

[2] Bruno Berberian, Bertille Somon, Aïsha Sahaï, and Jonas Gouraud. The out-of-the-loop brain: a neuroergonomic approach of the human automation interaction. *Annual Reviews in Control*, 44:303–315, 2017.

[3] Ewart J De Visser, Paul J Beatty, Justin R Estepp, Spencer Kohn, Abdulaziz Abubshait, John R Fedota, and Craig G McDonald. Learning from the slips of others: Neural correlates of trust in automated agents. *Frontiers in human neuroscience*, 12:309, 2018.

[4] Sidney WA Dekker and David D Woods. Maba-maba or abracadabra? progress on human–automation co-ordination. *Cognition, Technology & Work*, 4:240–244, 2002.

[5] James R Elkins. Comparison of machine learning techniques on trust detection using eeg. 2021.

[6] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013.

[7] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.

[8] Vinay Jayaram and Alexandre Barachant. Moabb: trustworthy algorithm benchmarking for bcis. *Journal of neural engineering*, 15(6):066011, 2018.

[9] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018.

[10] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] Cédric Rommel, Joseph Paillard, Thomas Moreau, and Alexandre Gramfort. Data augmentation for learning predictive models on eeg: a systematic comparison. *Journal of Neural Engineering*, 19(6):066020, 2022.

[13] Raphaëlle N Roy, Marcel F Hinss, Ludovic Darmet, Simon Ladouce, Emilie S Jahanpour, Bertille Somon, Xiaoqi Xu, Nicolas Drougard, Frédéric Dehais, and Fabien Lotte. Retrospective on the first passive brain-computer interface competition on cross-session workload estimation. *Frontiers in Neuroergonomics*, 3, 2022.

[14] Christopher D Wickens, William S Helton, Justin G Hollands, and Simon Banbury. *Engineering psychology and human performance*. Routledge, 2021.

[15] Xiaoqi Xu, Nicolas Drougard, and Raphaëlle N Roy. Topological data analysis as a new tool for eeg processing. *Frontiers in Neuroscience*, 15:761703, 2021.