A controlled human-robot interaction experimental setup for physiological data acquisition

Mathias Rihet
ISAE-SUPAERO
Université de Toulouse,
France
mathias.rihet@isae-supaero.fr

Aurélie Clodic
LAAS-CNRS
Université de Toulouse, CNRS
Toulouse, France
aurelie.clodic@laas.fr

Guillaume Sarthou
LAAS-CNRS
Université de Toulouse, CNRS
Toulouse, France
guillaume.sarthou@laas.fr

Sridath Tula
LAAS-CNRS
Université de Toulouse, CNRS
Toulouse, France
sridath.tula@laas.fr

Raphaëlle N. Roy
ISAE-SUPAERO
Université de Toulouse
France
raphaelle.roy@isae-supaero.fr

***Abstract*—Physiological measurements are promising tools to perform an online evaluation of human-robot interaction. In this study, a controlled human-robot interaction experimental setup for physiological data acquisition is presented, focusing on the elicitation of two cognitive states: cognitive effort and automation surprise. Using various physiological sensors along with subjective and behavioral measures, this setup allowed to collect data from 16 subjects over 2 sessions. Subjective and behavioural data confirm the induction of cognitive effort but not automation surprise yet. Physiological data processing is currently underway. Several challenges are discussed concerning this implementation, including the elicitation of the target cognitive states, the synchronization of all the devices and the need for repeated measures.**

***Index Terms*—HRI, evaluation, physiology, data acquisition, cognitive state, session effect**

I. Introduction

As for any tool, robots need to add value to the task they are used for in order to be useful. Yet, both complexity and the increasing number of expectations related to this field can lead to a wide range of metrics. In this context, it seems important to develop a framework of shareable metrics, both subjective and objective ones, to make findings' comparison easier and allow for the development of state-of-the-art evaluation toolkits [1], [2].

This question is even more important in collaborative human-robot interaction (HRI) where social robots are part of dynamic and non-deterministic interaction. Usefulness can, and should, be evaluated not only regarding the session but also regarding each task and/or each action that occurred.

Among the methods commonly used by the Human-Robot Interaction (HRI) community [3], self-assessments and interviews add crucial subjective information but they cannot be used during the interaction without interrupting it, while behavioral measures and performance metrics can be computed during the interaction but do not directly reflect the variations in humans' mental state. Thus, from our perspective, physiological measures are a very promising complementary method to perform an online estimation of human's mental state during HRI.

To our knowledge, the HRI literature has mostly focused on users' affective states, while users' cognitive state has seldom been investigated. Two cognitive states that are reflected by variations in electrophysiological activity seem, in particular, relevant to monitor in HRI settings [4]. On the one hand, cognitive effort -a.k.a. mental workload or engagement-, is a well documented state with robust metrics [5], [6]. On the other hand, automation surprise, which is the phenomenon of the user being surprised by the behavior of the automated system [7], can be of great interest to catch flaws in the fluency of the interaction. In addition, both of these cognitive states can be directly measured and assessed to a certain extent using portable, cheap, and non-invasive recording methods.

Nevertheless, a single measure is not sufficient to evaluate any interaction. Various kinds of metrics collected at various stages of the interaction are required to perform a sound assessment, particularly when it comes to psychophysiological inference. An ideal setup would then use a comprehensive approach with subjective, behavioural and physiological metrics.

Hence, this experiment's main objective is to design and implement a controlled HRI experimental setup for physiological data acquisition to elicit and measure two cognitive states: cognitive effort and automation surprise. In addition, this setup would need to be tested on a reasonable number of participants and over two sessions in order to deal with the reproducibility and replicability problems that affect both the HRI and neuroscience communities.

II. Methods

A. Materials

Robot

The PR2 (Personal Robot 2) is a research and development platform built by Willow Garage. Its software system, written using ROS, allows researchers to use all the tools already developed in this popular middleware to ensure that the robot suits their specifics needs. Because this experimental campaign focuses on the PR2's ability to grasp and manipulate objects in human environments, the robot was only used in a static position. In addition, its speakers were used to explicitly sequence the interaction with dialogues using text-to-speech google API.

Physiological Sensors

Electroencephalography (EEG) measures brain electrical activity with a very high temporal resolution (i.e. ms). EEG data were collected using a 20-channel Enobio (Neuroelectrics) with dry electrodes located on a cap according to the 10/20 international system and sampled at 500 Hz. Electrocardiography (ECG), photoplethysmography (PPG), galvanic skin response (GRS) and eye-tracking, respectively measure cardiac activity, blood volume changes, electrodermal activity and ocular behavior. ECG data sampled at 125 Hz were collected using a Faros 360°. GSR and PPG data were both collected using a Shimmer3 GSR+ Unit and sampled at 50 Hz. Eye-tracking was performed with a Pupil core at 200 Hz.

Software

Recording time series through multiple devices raises many issues, and particularly synchronization issues, the open-source ecosystem lab streaming layer (LSL) [8] was used to unify the recordings. Although this ecosystem allows for online signal processing, only offline processing was performed. Hence, LabRecorder allows to save the synchronized time series in a file following the open source xdf standards [9].

Tasks

In order to induce various cognitive effort levels, a digit span task was used. This typical memory-load task has already been applied with success to robotic tasks monitored via EEG [10]. Here, the digits were displayed sequentially on a screen located on the wall near the robot. Sequences of 1, 3 and 7 digits were expected to result respectively in a low, medium and high cognitive effort.

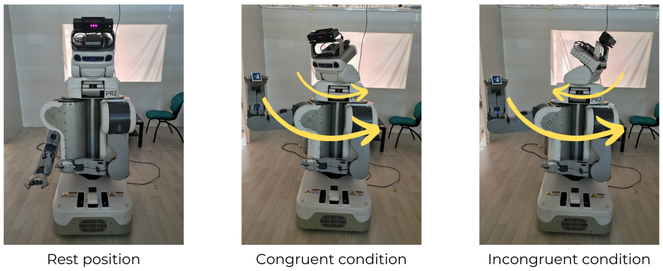

Fig. 1. PR2 main behaviours

The Joint manipulation task itself consists in piling cubes with a robot and begin with the participants seated at a low table and the robot (PR2) standing on the other side, facing them in its "rest position" (see Figure 1, Rest position). A dialog indicates that PR2's turn is beginning (”My Turn”). During this turn, PR2 picks with its right hand the cube at its right and place it at the center of the table. This sequence of actions is entirely scripted, in order to avoid any unexpected behaviour, but also to ensure that it always last approximatively the same time (25 seconds). Then, another dialog indicates that participant's turn is beginning (”Your Turn”). During this turn, participants are instructed to place, with their right hand, the cube at their right on the top of the cube situated at the center.

To induce automation surprise, 2 versions of this cube piling task were designed. In most trials, PR2's head followed the cube while moving it (Fig. 1, congruent condition). Though, in some trials, PR2's head faced the opposite side of the table, doing the symmetrically opposite head movement while moving the cube (Fig. 1, incongruent condition). These incongruences were scripted and occurred a fixed amount of times.

In addition, a resting state was performed before and after the main task to be used as a baseline for data analysis.

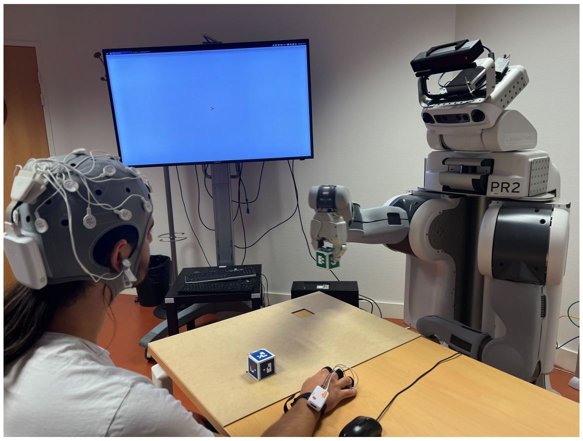

Fig. 2. PR2 Moving the cube in front of the participant during the cube piling task

B. Participants

Seventeen volunteers took part in the experiment. One participant's data were excluded due to completing only 1 out of 2 sessions. The remaining 16 participants (10 females, 6 males) had an average age of: 25.5 ± 3.1 years. Prior knowledge with robots scored 2.5 ± 0.9 on a scale of 1 "not at all" to 5 "very much" and, on the same scale, prior knowledge with physiological sensors scored 2.5 ± 1.7. The participants were instructed about the experimental protocol and provided an informed consent. The study was approved by the institutional ethics review board of the Federal University of Toulouse (project n°2022_525).

C. Experimental protocol

Sessions were scheduled one week apart and occurred at the same time of day. As detailed in Figure 3, upon arrival at the laboratory for their first session, participants were informed about the purpose and the procedure of the study. The researchers asked if there were any unclear questions or if the participants wanted to withdraw at this point. Then,

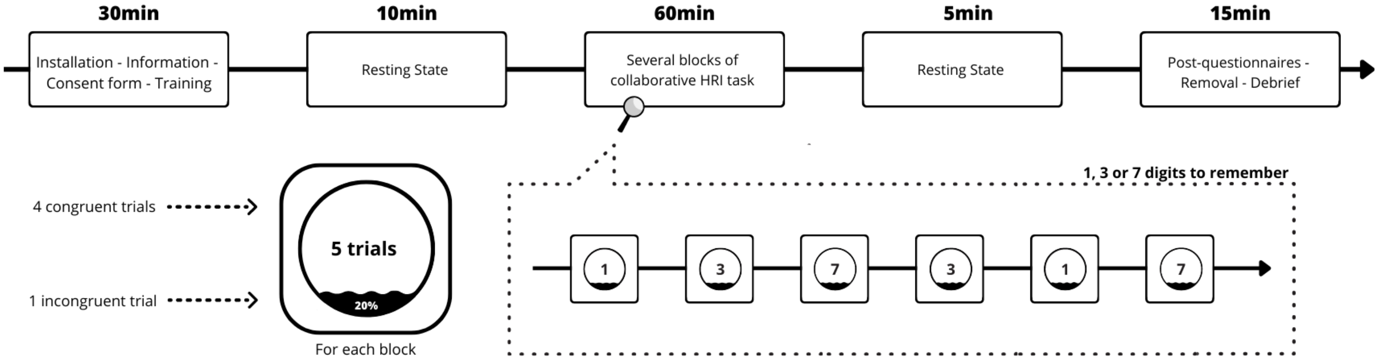
30min
Installation - Information - Consent form - Training
10min
Resting State
60min
Several blocks of collaborative HRI task
5min
Resting State
15min
Post-questionnaires - Removal - Debrief
4 congruent trials
1 incongruent trial
5 trials
20%
For each block
1, 3 or 7 digits to remember
1
3
7
3
1
7

Fig. 3. Overview of the experimental protocol, including the global timeline (top), the structure of a block (bottom-left corner) and an example of the order that blocks can follow (bottom-right corner).

they were asked to sign the informed consent form, the form for the inclusion and exclusion criteria, the demographic's questionnaire and the Edinburgh handedness inventory [11].

Next, participants were introduced to the task and the robot (see Figure 2). A short training sequence for the task composed of six trials (3 in the low cognitive effort condition and 3 in the high cognitive effort condition) and which lasted approximately 5 minutes was provided. Among these trials, only one (in the low cognitive effort condition) belonged to the incongruent condition in order to check participant's reaction while maintaining this condition infrequent. Participants were free to ask questions at any time and the researcher ensured that the task was understood by the participants.

After completion of the training, most of the sensors (ECG, EEG, GSR, EOG, PPG) were set up. Then, participants completed a resting state period during which they were instructed to do nothing. Because eyes were closed half the time, eye-trackers were never used during resting state.

Once the resting state and the Amsterdam Resting State Questionnaire (or ARSQ, that allows to study the link between ongoing brain activity and cognition [12]) were completed, the eye-tracker was set up and calibrated, allowing to begin the experimental task. This task lasted approximately one hour and was composed of 60 trials gathered in 12 blocks of 5 (see Figure 3). Each trial began with a dialogue ("New trial") followed by a random sequence of digits displayed on a screen. Next, they had to remember this sequence while performing the joint manipulation task. Finally, they had to recall the digits in the correct order by typing them on a virtual keyboard.

All trials of a same block belonged to the same cognitive effort condition, thus the cognitive effort condition only varied between blocks. In addition, one out of the five trials of each block was incongruent. These blocks were completed by the participants in a pseudo-randomized order.

Between each block the participants completed the ISA questionnaire [13], [14] and gave subjective feedback on how natural the behaviour of the robot appeared to them (on a 5-point likert scale), they were also able to take a short break if needed. The Human–Robot Interaction Evaluation Scale (or HRIES, an approach of anthropomorphism in HRI through four componentes : Sociability, Disturbance, Agency and Animacy [15]) was only completed once, at the end of the 12 blocks, to capture their subjective perception of the task and the robot.

At least two researchers remained present at all times, monitoring the task in order to ensure that each trial was completed accurately. One was in charge of the well-being of the participants, paying a particular attention to any discomfort that could occur because of the sensors. The other was carefully monitoring the robot, both on the computer and in the room, ready to stop it in case of an unplanned behaviour or if requested by the participant.

Next, participants took off the eye-tracker and completed another five-minute resting state period. Finally, researchers removed all the sensors and asked if participants had any questions regarding the research, informing them that, should any questions arise, they can contact the researchers via the contact information given on the information sheet. Participants were also financially compensated for their participation in the study.

The second session followed the same procedure but omitted the consent form as well as the demographics questionnaires and Edinburgh Handedness Inventory. Including arrival, questionnaires, installation/removal of the sensors and task completion, a usual session lasted two hours.

III. Protocol validation

A. Subjective data

Perceived effort significantly varied according to digit sequence's length (Friedman's test, $\chi^2_F(2) = 29.75$, $p < .001$) but not across sessions. In particular, it increased when the number of digits increased (Wilcoxon signed-rank test with Bonferroni correction, Mdn(1) = 1.5, Mdn(3) = 2, Mdn(7) = 4, at least $p < .01$ for each pair).

Perceived naturalness did not significantly vary according to digit sequence's length but was significantly higher in session 1 than in session 2 (Wilcoxon signed-rank test, Mdn(S1) = 3, Mdn(S2) = 3, $\eta^2 = .001$, $p < .001$), although slightly.

HRIES dimensions did not vary significantly across sessions.

B. Behavioral data

Accuracy significantly varied according to digit sequence's length (Friedman's test, $\chi^2_F(2) = 25.28$, $p < .001$) but not across sessions. In particular, it was lower for 7-digit sequences (Mdn(7) = 80) than for both 1-digit (Wilcoxon signed-rank test with Bonferroni correction, Mdn(1) = 100, $p < .001$) and 3-digit sequences (Wilcoxon signed-rank test with Bonferroni correction, Mdn(3) = 100, $p < .01$)

Response Time did not significantly vary according to digit sequence's length but was significantly higher in session 1 than in session 2 (Wilcoxon signed-rank test, Mdn(S1) = 1.69, Mdn(S2) = 1.35, $\eta^2 = .05$, $p < .001$).

IV. Discussion & conclusion

This experiment's main objective was to design and implement a controlled human-robot interaction experimental setup for physiological data acquisition. It requires, at first, to successfully induce the targeted mental states, namely cognitive effort and automation surprise.

Concerning cognitive effort, participants perceived an increasing effort when sequence size increased and performed less accurately with 7-digit to recall than with 1 or 3 digits. Thus, this state seemed to vary according to the sequence's size. In addition, none of these two metrics varied significantly across sessions, making this induction process robust to session effect.

Concerning automation surprise, the design of the study did not allow for direct behavioural and/or subjective validation of a successful induction. To limit the number of conditions to balance (and recruit only a reasonable number of participants), automation surprise occurrence did not vary between blocks. Hence, only a variation in response time could have reflected an automation surprise effect, but such variation was not observed. In addition, an interaction between automation surprise and cognitive effort could affect the perceived naturalness of the robot (if participants only noticed incongruent trials under some cognitive effort condition). Yet, such an interaction was not observed either. The current results are thus not sufficient to expect that automation surprise was robustly induced by this protocol.

Further, concerning the session effect, response time was longer in session 1 than in session 2 – which may reflect a learning effect or a more relaxed state for session 2, while naturalness decreased in a small but significant way. No significant session effect was observed on HRIES' dimensions.

Hence, the subjective and behavioral results validate only in part the experimental setup, yet this remains to ascertain by the ongoing analyses which focus on physiological metrics related to cognitive effort. At this stage, a strong impact of the session effect has been observed. This intersession variability stresses the importance of doing multiple sessions in HRI research, particularly with physiological measures [16], but also to work on adequate data processing pipelines.

Indeed, another constraint of physiological measures comes from their variability across time, even during the same session. Many measures need to be collected in identical contexts for each condition investigated. Applied to HRI, it leads to a robot repeating many times the same sequence of movements for hours, rising endurance issues such as overheating. Thus, this kind of setup needs to find the balance between an HRI long enough to be meaningful and short enough to allow many iterations in an acceptable amount of time, from both technical and human perspectives.

Along with these processing challenges, this implementation also highlighted several design challenges. At first, physiological acquisition, and particularly EEG, needs a high temporal synchronization between signals. While Lab Streaming Layer (LSL) [8] is a great framework that addresses this kind of issue, it does not help for the delay between the robot's computation and actions. Thus, particular focus needs to be put on the minimization of this delay or an external device used to catch the exact timing of robot actions.

Finally, participants are not necessarily focused on the robot during the whole HRI, especially when this interaction occurs many times. If possible, stimuli designed to induce specific mental states need to be presented directly by the robot in order not to shift participants' focus. In addition, this focus should be required by the HRI when stimuli are presented to minimize the chance that they miss them.

This article was written in the hope to stimulate discussion around the topic of users' cognitive state characterization during HRI. The next milestone towards an online HRI adaptation based on users' state assessment is to perform cognitive state estimation thanks to machine learning methods, a field known as physiological computing [17] and passive brain-computer interfaces when using brain activity as the main input [18].

Acknowledgment

This work was supported by the French National Research Agency (ANR) with project EPIIC (grant 21-CE33-0005), project AI4HRI (ANR-20-IADJ-0006) and by the Artificial and Natural Intelligence Toulouse Institute - Institut 3iA ANITI ANR-19-PI3A-0004.

References

[1] J. A. Marvel, S. Bagchi, M. Zimmerman, and B. Antonishek, "Towards effective interface designs for collaborative hri in manufacturing: metrics and measures," *ACM Trans. Hum.-Robot Interact. (THRI)*, vol. 9, no. 4, pp. 1–55, 2020.

[2] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich, "Common metrics for human-robot interaction," in *Proc. ACM SIGCHI/SIGART Conf. human-robot interaction*, 2006, pp. 33–40.

[3] C. L. Bethel and R. R. Murphy, "Review of human studies methods in hri and recommendations," *Int. J. Soc. Robot.*, vol. 2, no. 4, pp. 347–359, 2010.

[4] R. N. Roy, N. Drougard, T. Gateau, F. Dehais, and C. P. Chanel, "How can physiological computing benefit human-robot interaction?" *Robotics*, vol. 9, no. 4, p. 100, 2020.

[5] B.-W. Hsu, M.-J. J. Wang, C.-Y. Chen, and F. Chen, "Effective indices for monitoring mental workload while performing multiple tasks," *Perceptual and motor skills*, vol. 121, no. 1, pp. 94–117, 2015.

[6] P. Antonenko, F. Paas, R. Grabner, and T. Van Gog, "Using electroencephalography to measure cognitive load," *Educational psychology review*, vol. 22, no. 4, pp. 425–438, 2010.

[7] N. B. Sarter, D. D. Woods, C. E. Billings *et al.*, "Automation surprises," *Handbook of human factors and ergonomics*, vol. 2, pp. 1926–1943, 1997.

[8] C. Kothe, D. Medine, C. Boulay, M. Grivich, and T. Stenner, “Lab streaming layer,” 2014. [Online]. Available: https://github.com/sccn/labstreaminglayer

[9] A. Ojeda and C. Kothe, “Extensible data format,” 2015. [Online]. Available: https://github.com/sccn/xdf

[10] Y. Liu, M. Habibnezhad, and H. Jebelli, “Brainwave-driven human-robot collaboration in construction,” *Automation in Construction*, vol. 124, p. 103556, 2021.

[11] J. F. Veale, “Edinburgh handedness inventory–short form: a revised version based on confirmatory factor analysis,” *Laterality: Asymmetries of Body, Brain and Cognition*, vol. 19, no. 2, pp. 164–177, 2014.

[12] B. A. Diaz, S. Van Der Sluis, S. Moens, J. S. Benjamins, F. Migliorati, D. Stoffers, A. Den Braber, S.-S. Poil, R. Hardstone, D. Van’t Ent *et al.*, “The amsterdam resting-state questionnaire reveals multiple phenotypes of resting-state cognition,” *Front. Hum. Neurosci.*, vol. 7, p. 446, 2013.

[13] A. J. Tattersall and P. S. Foord, “An experimental evaluation of instantaneous self-assessment as a measure of workload,” *Ergonomics*, vol. 39, no. 5, pp. 740–748, 1996.

[14] G. D. Flumeri, G. Borghini, P. Aricò, A. Colosimo, S. Pozzi, S. Bonelli, A. Golfetti, W. Kong, and F. Babiloni, “On the use of cognitive neurometric indexes in aeronautic and air traffic management environments,” in *Int. Workshop on Symbiotic Interaction*. Springer, 2015, pp. 45–56.

[15] N. Spatola, B. Kühnlenz, and G. Cheng, “Perception and evaluation in human–robot interaction: The human–robot interaction evaluation scale (hries)—a multicomponent approach of anthropomorphism,” *Int. J. Soc. Robot.*, vol. 13, no. 7, pp. 1517–1539, 2021.

[16] S. Saha and M. Baumert, “Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: a review,” *Front. Comput. Neurosci.*, vol. 13, p. 87, 2020.

[17] S. H. Fairclough, “Fundamentals of physiological computing,” *Interacting with computers*, vol. 21, no. 1-2, pp. 133–145, 2009.

[18] F. Lotte and R. N. Roy, “Brain–computer interface contributions to neuroergonomics,” in *Neuroergonomics*. Elsevier, 2019, pp. 43–48.