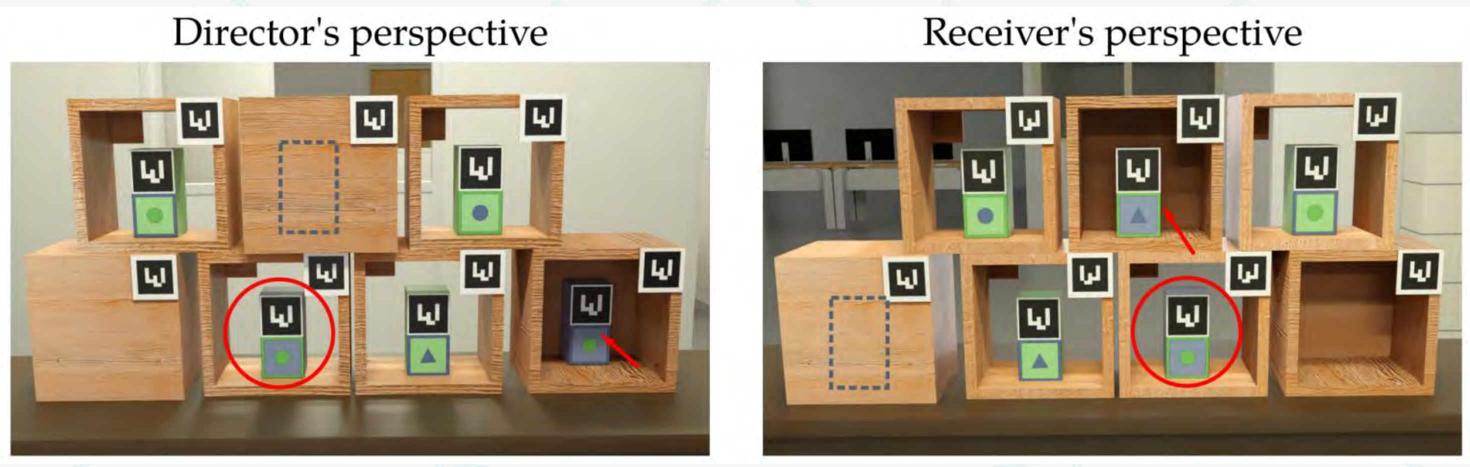
Decoding Error-Related Potentials in HRI: An Other Perspective

Mathias Rihet[1,2], Nadim Saleem[1,2], Aurélie Clodic[2], Raphaëlle N. Roy[1]
[1] *Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, Toulouse, France*
[2] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France*

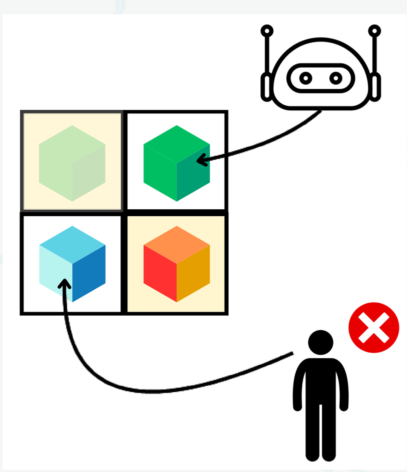
Introduction

Human-Robot Interaction (HRI) involves dynamic exchanges where both agents, human and machine, can commit errors. In the Director task paradigm[1], such errors may be occurred from incorrect instructions, executions failures or even misaligned perspective between both agents. These mismatches between expectation and outcome, during interaction, elicit specific neural responses called Error-related potentials (ErrPs), EEG patterns that emerge when brain detects deviations from expected outcomes[2].

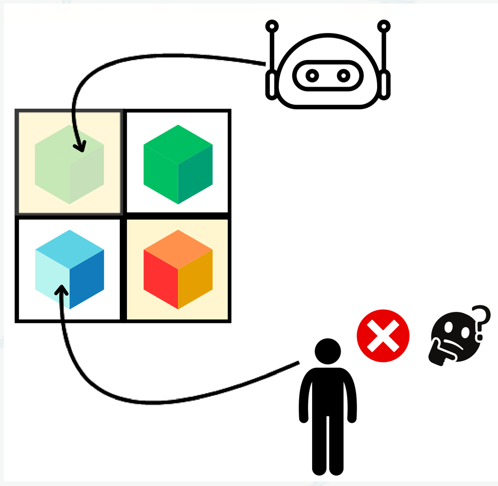


This study presents a protocol designed to evoke then classify neural responses to two principal distinct error types within the Director Task context: **(1)** Misinterpretation errors, **(2)** Perspective-taking failures.

(1) : Robot picks green cube instead of blue



(2) Robot picks a cube only it can see

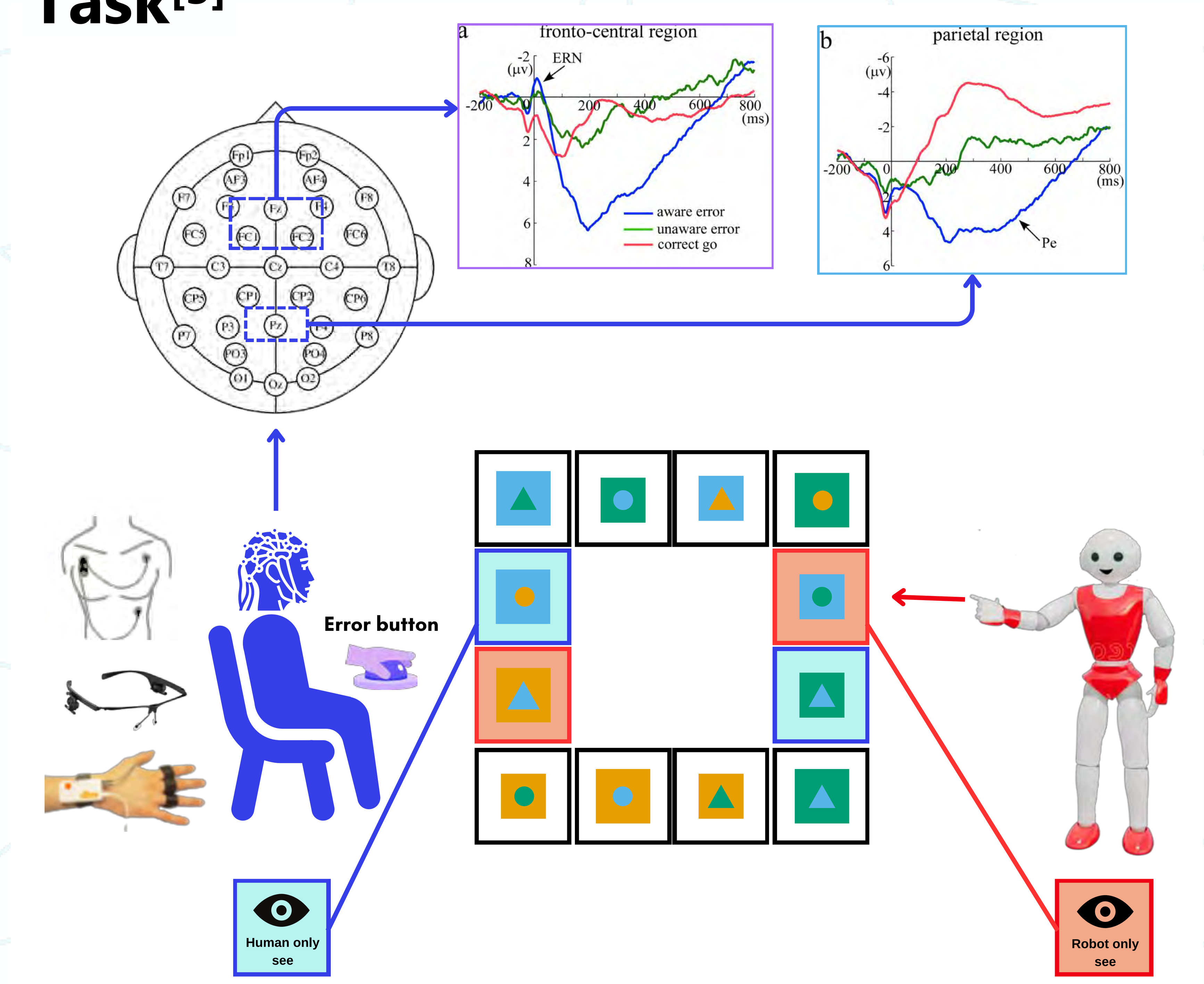


Finally, we adopt a simplified 4*4 grid, in which 8 objects locations are visible to both agents and 4 are private; 2 visible only to human and the other 2 only visible to the robot.

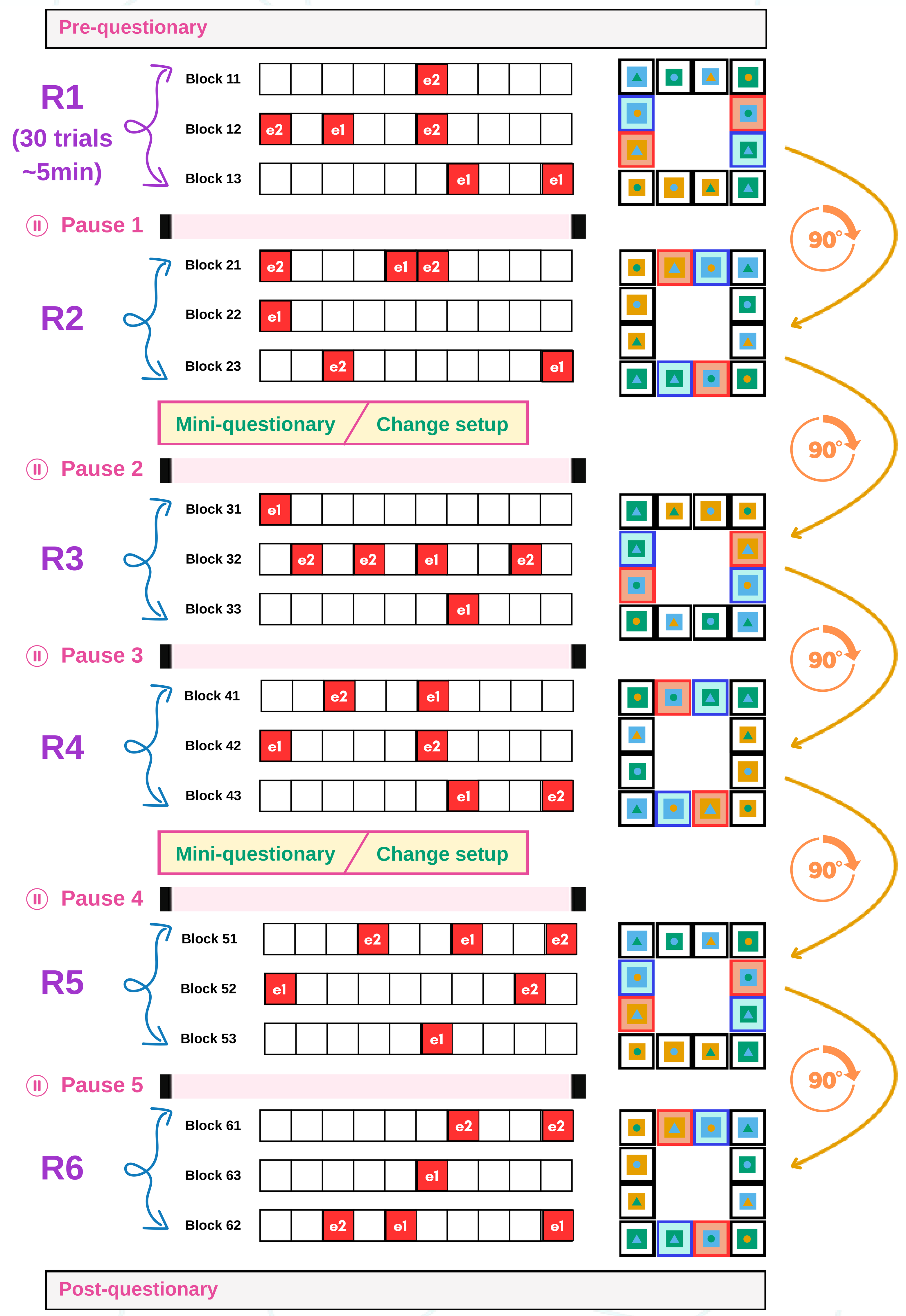
Objectives

1. **Detect** then **classify** ErrPs due to different interactions
2. **Compare** and **differentiate** between misinterpretation errors/perspective-taking failures
3. **Correlate** subjective with objective measurements

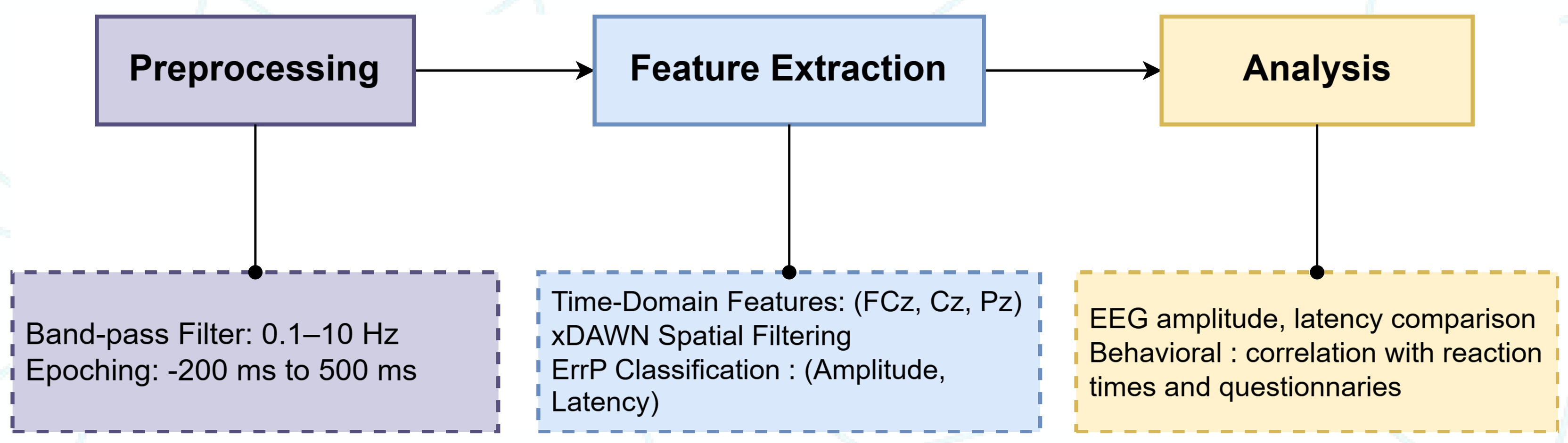
Task[3]



Experimental design[4]



Planned Analysis[5][6]



Why This Matters

- **Beyond binary tasks** : Most ErrPs studies rely on binary decisions. We use a realistic, perspective-based HRI scenario.
- **Richer error typology** : The proposed paradigm captures how humans respond to less obvious, cognitively complex errors.
- **Toward inhanced Quality of Interaction**: By online decoding human brain signals, this work supports future robots that learn and self-correct through implicit neural feedback.

References

[1]. Sarthou, G., Mayima, A., Buisan, G., Belhassein, K., & Clodic, A. (2021). **The Director Task: A psychology-inspired task to assess cognitive and interactive robot architectures**. In 2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 770–775). IEEE.
[2]. Pires, G., Castelo-Branco, M., Guger, C., & Cisotto, G. (2022). **Editorial: Error-related potentials: Challenges and applications**. Frontiers in Human Neuroscience, 16, 984254.
[3]. Rihet, M., Sarthou, G., Clodic, A., & Roy, R. N. (2024). **Electrophysiological measures for human–robot collaboration quality assessment. In Discovering the Frontiers of Human-Robot Interaction** (pp. 363–380). Springer Nature Switzerland.
[4]. Rihet, M., Clodic, A., Sarthou, G., Tula, S., & Roy, R. N. (2024). **A controlled human-robot interaction experimental setup for physiological data acquisition**. EPIIC for Intellect4HRI Workshop, Toulouse, France
[5]. Vidal, F., Burle, B., & Hasbroucq, T. (2022). **On the comparison between the Nc/CRN and the Ne/ERN**. Frontiers in Human Neuroscience, 15, 788167.
[6]. Roy, R. N., Bonnet, S., Charbonnier, S., & Campagne, A. (2015, April 22–24). **Enhancing single-trial mental workload estimation through xDAWN spatial filtering**. In International IEEE EMBS Conference on Neural Engineering (NER). Montpellier, France

isae SUPAERO · LAAS CNRS · cnrs · anr · Défis Clés Occitanie · La Région Occitanie Pyrénées - Méditerranée · Robotique centrée sur l'Humain · ANITI Artificial & Natural Intelligence Toulouse Institute · Université Fédérale Toulouse Midi-Pyrénées