ETIENNE Elodie (Applicant)
CLAVEL Chloé (Advisor)
March 2025

Mitigating bias and enhancing explainability during collaboration between humans and socially interactive agents groups

1. Project Description

Context

Socially Interactive Agents (SIAs) are increasingly used to facilitate collaboration within hybrid teams composed of both humans and agents. Research highlights their role in teamwork, where they support decision-making, task coordination, and social cohesion in dynamic environments (van Wissen *et al.*, 2011). Studies have shown that perceiving SIAs as part of the team enhances group performance and trust, with experiments on social robots indicating that trust in both the group and the agent itself plays a key role in its integration and acceptance as an in-group member (Oberhofer, 2025). Additionally, multimodal communication strategies, such as combining verbal cues with non-verbal behaviours like gestures and gaze, have been found to improve SIAs' persuasiveness and perceived politeness, fostering better humans-agents interactions (Zojaji *et al*. 2020 and 2023).

However, these technologies are built on real-world data, which is inherently biased (Choi Kristy *et al*., 2020) in aspects such as gender, cultural norms, economic disparities, and professional hierarchies. As a consequence, SIAs tend to replicate these biases (Van Niekerk *et al.*, 2024; Frankel and Edward, 2020), and humans interacting with them may inadvertently inherit and propagate these stereotypes (Vincente and Matute, 2023). This reinforcing cycle, where biased data produces biased SIAs that perpetuate real-world biases, is particularly problematic as it continuously amplifies societal stereotypes if left unaddressed.

Although significant progress has been made in mitigating bias in Artificial Intelligence (AI) in general (e.g., González-Sendino *et al.*, 2024; Rajabi and Garibay, 2021), large language models (LLMs) (e.g., Lin *et al.*, 2024), and non-verbal behaviour (e.g., Delbosc *et al*., 2024), biases in group collaboration remain largely unaddressed. For example, biases in turn-taking, gaze direction, and posture reinforce gendered power dynamics when SIAs replicate these patterns. Mitigating these biases is essential for fair and effective collaboration.

Our project addresses this gap by leveraging machine learning (ML) to identify biases (specifically gender biases) within group interactions using multimodal interaction corpora. We examine both verbal and non-verbal behaviours while accounting for the dynamic flow of interactions. Building on this analysis, we aim to develop a computational model capable of modifying, filtering, or mitigating biased behaviours exhibited by SIAs during group interactions, ensuring more equitable and inclusive engagement. We will also measure the effect of newly adapted behaviours on reducing or reshaping user stereotypes.

Challenges

One of the key challenges in our project is the automatic identification of gender biases in both human-human and human(s)-agent(s) group interactions. While existing research has established that biases shape group dynamics (such as women being interrupted more often or receiving less recognition for their contributions (Bryant *et al*., 2020) or individuals from minority backgrounds being less frequently acknowledged as experts (Ely *et al*., 2001)), we still lack robust methods to systematically detect and quantify these biases in multimodal interactions. The challenge is to develop ML approaches that can identify biases in verbal and non-verbal behaviours, as well as conversation dynamics while maintaining model transparency and explainability. Given that SIAs often reproduce human biases (e.g. female agents being assigned subordinate roles and displaying more deferential behaviours, while male agents are framed as experts (Deutsch *et al.*, 1987)), it is crucial to build models that not only detect but also interpret these patterns in a way that is understandable and actionable.

A second major challenge is mitigating these biases in SIAs while preserving both the naturalness of their behaviour and the performance of interaction models. Addressing bias should not result in artificial or

robotic communication patterns, which could negatively impact user engagement. Instead, we aim to develop algorithms that adapt SIAs' behaviours, filtering or modifying biased responses without compromising the fluidity and effectiveness of group interactions. For this last challenge, we focus on mitigating biases in LLMs, which are central to generating verbal behaviours.

Objectives

By placing the human user at the centre of our approach, we aim to promote more ethical, transparent, and inclusive humans-SIAs collaborations in a group.

Our project seeks to develop comprehensive evaluation frameworks that leverage automatic tools and methodologies to identify gender stereotypes embedded in SIAs, with a focus on multimodal behavioural datasets capturing interactions used during pre-training and model development.

At the same time, we will mitigate detected biases by generating balanced and modified datasets that exclude stereotypes while preserving behavioural naturalness. To ensure the effectiveness of these bias mitigation strategies, we will design and implement robust experimental protocols that diverge from conventional performance metrics. Drawing inspiration from established methodologies for the evaluation of generated contents (Guo *et al.* 2024), we will develop an automated protocol that rigorously assesses bias reduction at the system level by employing novel evaluation metrics tailored to capture stereotypical behaviours. Furthermore, subjective evaluations will examine the impact of these strategies on user perception.

Finally, we aim to enhance explainability by equipping SIAs with advanced tools that not only clarify decision-making processes but also reveal how stereotypes are embedded in the behaviours of agents, empowering users to understand and challenge the underlying factors that lead to stereotypical outputs.

Approach

For our project, we plan to leverage several established corpora that capture multimodal interactions between SIAs and humans. For instance, the Niki and Julie Corpus provides collaborative multimodal dialogues between humans, robots, and virtual agents (Artstein *et al.,* 2018), offering rich insights into human-robot and human-agent interactions. Additionally, RDG-Map (Paetzel *et al.*, 2020) contributes valuable data on such interactions. To complement these, we also consider human-human dialogue datasets like Multissimo (Koutsombogera and Vogel, 2018) and Trueness (Ochs *et al*., 2023), a corpus specifically designed to raise discrimination awareness. Moreover, the AMI corpus (McCowan *et al.*, 2016) provides extensive multimodal data from meeting scenarios. These diverse datasets will underpin our multimodal models, enabling us to automatically extract verbal, non-verbal behavioural and conversational dynamics features. Specifically, we will extract turn-taking patterns, gaze and posture features through OpenSmile, OpenFace and OpenPose tools. These features are processed using interpretable ML models, such as decision trees, recurrent or convolutional neural networks. LLM models will also be investigated as they play a central role in generating verbal content in SIAs. The models will then be analysed with methods for interpretability (like SHAPLEY values (Lundberg and Lee, 2017) or Hemamou *et al.* (2021b)) to identify and explain the presence of stereotypes.

For bias mitigation, our approach integrates indirect adversarial techniques, inspired by Hemamou *et al.* (2021a), to remove sensitive information from neural representations while preserving model coherence.

Originality

This project contributes novel insights into applying multimodal analysis to address stereotypes in SIAs. Unlike previous efforts, it focuses on reproducing and analysing multimodal behaviours as a foundation. The feasibility of the approach is ensured by leveraging proven methodologies and well-established corpora, which provide a robust basis for the proposed work. By integrating explainability tools and advancing bias mitigation techniques, the project contributes uniquely to the field, filling a critical gap in multimodal, ethical AI design.

Anticipated Results

Our work is expected to yield both concrete deliverables and a lasting positive impact on society. We aim to develop novel bias mitigation techniques for SIA, accompanied by new corpora and benchmarks that enable rigorous evaluation of stereotype reduction. In addition, our project will provide the community with de-stereotyped multimodal models for SIAs that can be readily reused and further improved. We will also establish comprehensive frameworks and guidelines for ethical, fair, and adaptive AI-human collaborations, ultimately reducing the influence of stereotypes in SIAs over the long term.

2. Project Organisation

Our project builds on established methods, datasets, and state-of-the-art techniques in explainability and bias mitigation, supported by robust partnerships with leading AI experts and multimodal behaviour researchers. The research is structured over a 24-month period divided into four phases. In the initial phase (months 1-6), we will conduct data analysis and stereotype identification. This will be followed by the development of bias mitigation methods and explainability tools during months 7-12. In the subsequent phase (months 13-18), experimental validation and user studies will be performed, culminating in a final phase (months 19-24) dedicated to disseminating our findings and establishing ethical guidelines. Key milestones include the completion of stereotype identification by month 6, the development of bias mitigation techniques by month 12, validation of our tools by month 18, and the final report by month 24. Ultimately, our deliverables will comprise innovative models and tools for efficient systems, principles for adapting systems to user diversity, and ethical best practices for SIA-human interactions, all reinforced by our strategic collaborations.

3. Adequacy with the Objectives of PEPR eNSEMBLE and the Selected PC

The MATCHING PC framework emphasises creating adaptive human-computer collaborations that enhance user capabilities while safeguarding autonomy. Our project aligns closely with Axis 2, i.e., maintaining understanding and control with complex intelligent systems, addressing its key objectives as outlined in WP2.1 (Modeling & understanding of collaborative or competitive interactions) and WP2.2 (Integrating and mitigating users' diversity and agency).

Indeed, our project contributes to WP2.1 by developing models and frameworks to measure, understand, and adapt to user states using multimodal signals. It leverages data-driven insights from human-human interaction and human-machine interaction to improve human-machine collaboration, ensuring the SIA adapts dynamically to the user whatever its characteristics. This includes analysing verbal and non-verbal patterns and turn-taking behaviours to better align AI responses with human expectations. Furthermore, the project aligns with WP2.2 by ensuring that the system design addresses explainability and fairness. By generating bias-free datasets and employing explainability tools, it mitigates unintended negative consequences, ensuring equitable outcomes for all stakeholders.

Our project also addresses a critical concern of Theme 3 (Impact of intelligent systems on expertise and loss of competence): the impact of stereotypes on human critical thinking and collaboration. While we do not focus on prolonged interactions per se, we recognise that unchecked stereotypes embedded in SIAs can distort users' perceptions and impair decision-making processes. Such biases may not only affect individual judgments but can also deteriorate the quality of collaboration among humans interacting with these systems. By mitigating these stereotypes and enhancing explainability, our approach aims to counteract the negative influence of biased AI on human analytical skills and collaborative efficacy. In doing so, our work contributes to preventing the erosion of human expertise and supports the establishment of equitable, shared authority in hybrid interactions, thus aligning our proposal with the objectives of Theme 3.

By addressing these priorities, the project ensures that intelligent systems remain comprehensible, adaptable, and inclusive, fostering user empowerment and ethical AI-human collaboration. This alignment underscores the project's contribution to PEPR eNSEMBLE's overarching goals of creating human-centred intelligent systems.