Galland Lucie (Applicant), OCHS Machalie (Supervisor) October 2025

Improving Group Problem-Solving Task Abilities by Calibrating Trust

1. Project Description

Context

In collaborative problem-solving, group performance depends not only on participants' actual skills but also on the degree to which they trust each other's contributions. When trust is miscalibrated and an individual either overrely or underrely on their partners, it leads to a decrease in group efficiency. The problem is amplified when one of the group members is an autonomous system: under-trusting robots can limit their usefulness, while over-trusting them can lead to errors or even accidents (Nesset, 2021). For example, people who over-rely on information generated by a large language model (LLM) risk basing decisions on false or misleading outputs. Prior work demonstrates that the behaviors of autonomous agents can significantly influence team communication and coordination (Demir et al., 2019; Shah & Breazeal, 2010). Achieving calibrated trust is therefore crucial. However communicating appropriately the level of confidence of a system to achieve such calibrated trust is challenging. Research indicates that interaction transparency enhances decision-making (Nesset, 2021), but merely informing users of system capabilities is insufficient; trust must also be established through subtle verbal and non-verbal behaviors (Metzger, 2024). However, LLMs and many AI systems struggle with poorly calibrated confidence expressions, often overstating or understating their certainty (Mielke, 2022). While ongoing work focuses on correcting these behaviours in a dyadic setting (Tomsett, 2020; Okamura, 2020) the study of trust within groups of human and AI agents is less explored. In human groups, trust dynamics are very different from dyadic trust, and new phenomena can emerge, such as trust spirals where diversity in individuals' propensities to trust can lead to downward spirals of trust over time (*Ferguson, 2015).* This is something that does not manifest in dyadic trust because there's only one relationship to consider. Therefore, the design of a system, focused towards group interaction, that can communicate its level of confidence appropriately through adapted verbal and non-verbal behavior, while taking into account differences between group members, would be a valuable step towards calibrating trust and improving overall interaction efficiency in groups.

Challenges

A central challenge of this project is that trust calibration has been studied primarily in dyadic human–AI interactions but remains largely unexplored in group collaboration settings. Yet, miscalibrated trust can lead to misuse, disuse, or even accidents, thereby undermining collective performance (Lee & See, 2004). One difficulty lies in designing verbal and non-verbal behaviors that allow the agent to transparently communicate its own capabilities to different group members despite differences between them. Studies show that both overconfident and underconfident AI systems reduce team performance, by reducing cohesion and collaboration in case of underconfidence or using inappropriate inputs in case of overconfidence, underscoring the need to align expressed confidence with actual

competence (Li et al., 2024). In groups, this trust dynamic becomes even more complex. Group trust encompasses not only the dyadic relationships among members but also the collective perception*s* of the group as a unit (Sapp, 2019). Moreover, divergent levels of trust toward the same system may disrupt coordination, while trust in one agent can also spill over to teammates and shape interpersonal evaluations (Zhang et al., 2024). Therefore in group settings, calibration of trust also might require being able to estimate each teammate's level of trust towards the system to adjust accordingly. Finally, research demonstrates that if a single team member changes its behaviors to ensure trust calibration, the team as a whole improves its performance, cohesion and collaboration (Johnson et al., 2021), which opens the door for agents to act as trust facilitators within groups. Therefore, studying trust dynamics in group settings to design an agent able to calibrate every group member's trust towards them would be a valuable contribution as it would improve overall team cohesion, collaboration and overall performance.

Objectives

The first objective of this project is to understand trust calibration dynamics in human–AI collaboration group and how trust is communicated and calibrated within groups. While prior work has produced corpora on dyadic human–AI trust interactions, there is currently no dataset that captures how trust is calibrated within groups that include multiple humans and an AI agent, To address this gap, we will first collect a multimodal dataset of group collaborative tasks, enabling analysis of verbal and non-verbal behaviors linked to trust dynamics in a group. Building on these data, we will develop methods to detect over-trust and under-trust in group contexts toward AI systems, taking into account dyadic trust toward the system and overall group trust level, through supervised machine learning. We will then design an AI agent capable of transparently communicating its competence and confidence in ways that support the calibration of appropriate trust, being able to adapt to different level of trust from the group members. Finally, we will evaluate whether such adaptive behaviors improve group performance and whether trust in the AI agent also shapes trust among human teammates. This evaluation will be conducted through a subjective experiment in which a group of humans interacts with an AI agent to solve a collaborative task. The agent will be able to express its confidence appropriately to achieve trust calibration, as well as detect miscalibration of trust and adjust its behavior accordingly. We will measure and study the impact of trust calibration on the success of group collaboration, group cohesion, as well as the impact of the system on the level of trust humans have towards each other.

Originality

This project is original in its focus on trust calibration within multi-party communication, an area that has been limitedly studied. While prior work has primarily examined trust calibration in dyadic human–AI interactions, few studies have addressed the issue of miscalibrated trust as a source of inefficiency and inequity in teamwork. Our approach is novel in several ways. First, it combines analysis of trust calibration in group settings, rather than just one-on-one interactions. Second, it incorporates adaptive dialogue and non-verbal behavior generation, tailored to the trust levels of multiple human participants simultaneously. Finally, it explicitly investigates whether trust calibration between humans and an AI agent can influence the trust dynamics among human teammates, potentially improving both team efficiency and fairness.

Anticipated Results

We anticipate that introducing socially interactive agents capable of calibrated trust communication will encourage human participants to adjust their reliance on the system more appropriately, avoiding both overtrust and undertrust. This improved alignment is expected to enhance group decision-making efficiency and foster more equitable participation, as individuals who previously over- or under-valued their own or others' contributions recalibrate their behavior. Additionally, we expect spillover effects, whereby trust calibration with the system positively influences the trust and evaluation between human teammates, improving the team's overall efficiency.

2. Project Organisation

Our project builds on established computational models for trust calibration in group problem-solving tasks, with the ability to calibrate human trust towards the system. The research is structured over a 24-month period divided into four phases. In the initial phase (months 1-6), we will conduct data collection, analysis, and identify how humans communicate to calibrate trust in a group setting. This will be followed by the development of an agent capable of calibrating human trust levels in a group setting through adaptive verbal and non-verbal behaviors. 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 investigating how the introduction of a trust calibrating agent impacts the trust calibration between the Human team members. Key milestones include the completion of data collection and trust levels detection model by month 6, the development of an adaptive agent by month 12, the validation of our tools through a user study by month 18, and the final report by month 24. Ultimately, our deliverables will comprise innovative models and tools for efficient systems, principles for improving trust calibration in group solving problems.

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

The project falls within the framework of the Targe project MATCHING, which aims to model and understand collaborative or competitive interactions between humans and AI-driven entities. This project will enrich this theme with an interdisciplinary methodological approach, including the collection and study of multimodal interaction data for understanding trust calibration within group interactions and generating group behaviours using state-of-the-art machine learning models. The objective of the project to enhance collaborations between humans through training with virtual agents simulating real-life collaborative situations through trust calibration is closely linked to Axis 1 the MATCHING target project. This project is an opportunity to develop models of socially interactive agents, centred on user perception, taking into account intra and inter-individual variability in the production and generation of behaviours for collaborative tasks answering question relevant to WP1.1. It also focuses on improving trust calibration for the agent to take an appropriate space in the group wich aligns with issues raises with WP1.2. By addressing trust calibration in mixed groups of humans and AI-agent this project is a step towards better human agent collaboration in groups and therefore could be a contribution to PEPR eNSEMBLE's goals of creating human centered intelligent systems.