Mutually Adaptive Trust Calibration in Human-AI Teams

Ewart J. de Visser\textsuperscript{1,2}, Ali Momen\textsuperscript{1}, James C. Walliser\textsuperscript{1}, Spencer C. Kohn\textsuperscript{2}, Tyler H. Shaw\textsuperscript{2} and Chad C. Tossell\textsuperscript{1}

\textsuperscript{1} U.S. Air Force Academy, 2354 Fairchild Drive, Colorado Springs, CO 80840, USA
\textsuperscript{2} George Mason University, 4400 University Drive, Fairfax, VA 22030, USA

Abstract
This paper presents the idea of mutually adaptive trust calibration in Human-AI Teams (HATs). Mutually adaptive trust calibration in HATs is established when both the human agent and the machine agents continuously adapt to one another, in terms of beliefs, attitudes and behaviors, to optimize trust and team performance. This goal requires new concepts, definitions, models, and measures. We highlight our past and recent studies that advance this important objective.

Keywords
Trust, Human-Autonomy Teams, Trust Calibration

1. Introduction

Much current work with regards to establishing “common ground” in relation to trust has been to either 1) enhance the human’s understanding of the machine or 2) enhance the machine’s understanding of the human \cite{1,2}. The first approach, to create human-readable machines, encompasses efforts to enhance the transparency of the machine and the explainability of its actions to the human by enhancing interface design and training materials, amongst other methods. The second approach, to create machine-readable humans, attempts to measure the intent of a human to then provide adaptive feedback and adjustments to enhance overall understanding and collaboration (see Figure 1). Yet, these methods overlook an important factor that is naturally evident in human-human communication: flexible real-time and fluent coordination. For example, real-time communication is naturally messy and requires quick adjustments, adaptation to errors, and recognition of mutual understanding. A more flexible and dynamic approach is needed that can account for errors, can adjust on the fly and can calibrate in real-time. The ultimate version of success with these approaches is to establish bi-directional and adaptive mutual trust calibration.

![Figure 1: Approach for human-autonomy collaboration](image-url)
2. Mutually Adaptive Trust Calibration

2.1. Trust Calibration Definitions

Mutual trust is a fundamental property and predictor of good performing teams. We define trust as “the continuous process of setting and updating a discrete interaction policy towards another agent in risky situations.” Some people trust AI a lot which can lead to over-trusting, leading to misuse and potentially disastrous outcomes. Others distrust AI which can cause under-trust, leading to disuse and unnecessary additional workload [3]. For instance, people tend to apply broad heuristics to trusting other agents such as the system-wide strategy, which occurs when one faulty system “bleeds over” negatively into the perception of similar, but well performing systems [4]. Ideally, people have calibrated trust, an optimized state when the perceived trust matches actual machine trustworthiness (see Figure 2). Early on, we identified and studied teams composed of people and unmanned vehicles and found that trust increased with experience even though the robot made many mistakes [5]. Over-trust can be adjusted downwards by dampening expectations and under-trust can be adjusted upwards through trust repair strategies [6,7].

![Figure 2: Trust calibration for over- and under trust with Artificial Intelligence](image)

2.2. A Model for Mutually Adaptive Trust Calibration

Recently, we have established a model for mutually adaptive trust calibration in HATs [8]. This work presents a new model explaining the role and process of establishing mutually adaptive longitudinal social trust calibration throughout the life cycle of a HAT (see Figure 3). The HAT Trust Model describes the development and role of trust calibration in HAT collaboration. HAT consists of four parts including 1) Relationship Equity, 2) Social Collaborative Processes, 3) Perceptions of Team Partner, and 4) Perceptions of Self. Central to our model is the idea of relationship equity which describes the cumulative result of the cost and benefit relationship acts that are exchanged during repeated collaborative experiences (including social and/or emotional interactions) between two actors. The middle part of the model describes the collaborative task performance between the teammates. Together, they perform a joint activity with the purpose of achieving a common goal. Collaboration is risky: actions may fail and circumstances may change. Therefore, the individual actors monitor the behavior and collaboration of themselves and their teammates. Based on their observations, actors aim to establish appropriate trust stances towards one another (A in B and B in A) to mitigate the potential risks involved in accomplishing the joint task. One part of the model includes passive trust calibration process: Based on team...
members’ perceptions of one another, actors predict one another’s trustworthiness. Taking into account their current formal work agreements and informal way of collaboration, they then (sub)consciously assess the risk involved in the collaboration as it currently is, and decide upon a trust stance towards one another. Another process of the model is the active trust calibration process. This process is based on an actor’s awareness concerning their involvement in team trust calibration. This awareness enables both actors to engage in deliberate attempts to influence, aid, or hamper the trust calibration process.

![Diagram](image-url)

**Figure 3:** The human-AI team (HAT) trust model

### 3. Measurement of Trust in Human-Autonomy Teams

To validate models of mutually adaptive trust calibration, it is essential to develop and empirically assess trust building, development and repair in human-autonomy teams. Essential in this endeavor is good measures of trust and trustworthiness [9-12]. Recently, we have catalogued the state of the art of trust measurement [13] and we have mapped this to Mayer’s original trust model [14]. We divide measures of trust into self-report, behavioral and physiological measures and show some examples of how these measures are used in the next sections.

#### 3.1. Trusting the Moral Judgments of a GPT-Enabled Robot

We explored how a machine agent trained to respond to moral queries is perceived and trusted by human questioners [15]. Participants were tasked with querying the human-like agent with the goal of figuring out whether the agent, presented as a humanlike robot or a web client, was morally competent and could be trusted. Participants rated the moral competence and perceived morality of both agents as high yet found it lacking because it could not provide justifications for its moral judgments. While both agents were also rated highly on trustworthiness, participants had little intention to rely on such an agent in the future. This work presents an important evaluation of a morally competent algorithm integrated with a human-like platform that could advance how moral robot advisors are trusted in the future [16-18].
3.2. System-wide Trust Effects in Human-Autonomy Teams

To demonstrate the effectiveness of interventions on trust calibration we provide an example of a recent UAV supervisory control study that countered a bias known as the system-wide trust effect [4]. This bias is the tendency for operators to apply trust broadly rather than exhibiting specific trust in each component of the system when one of these systems is faulty. Our study assessed the effectiveness of two trust calibration interventions to counter this bias including transparency feedback and scenario-based training. Results showed that by providing both system transparency feedback and training resulted in the most optimal verification rates and response times (see Figure 4). This finding affects how HATs should be designed and trained [19-20].

![Figure 4](image)

**Figure 4**: Verification rates (left) and subjective trust (right) monitoring automation recommendations in a UAV supervisory setting. The Faulty UAV1 is indicated by a red asterisk. By providing transparency feedback (informed condition), participants were more calibrated but showed the system-wide trust effect.

3.3. Neural Correlates of Automated Agents

We also investigated the neural underpinnings and mechanisms of trust in automated agents [21]. We used two event-related potentials measured by electroencephalography as an indicator of trustworthiness (see Figure 5). We demonstrated that this marker could distinguish between high and low reliability.

![Figure 5](image)

**Figure 5**: Observed error related negativity (oERN) and observed positivity (oPE) as indicators of trustworthiness of an automated agent.

4. Future Directions
New efforts are under way to realize the vision of mutually adaptive trust calibrating HATs. New developments in generative AI will enable novel types of communication that can allow us to improve human-AI teaming, which is not as good as human-human teaming [22]. We are also investigating how trust calibration efforts trade-off with situation awareness and workload to improve teaming performance. Other efforts are focusing on quantifying the relationship equity construct that some have found useful for the prediction of cross-task and long-term AI teaming [23]. Furthermore, recent reviews have found trust repair, dampening and explanation efforts are not always effective and determined there is a need to develop better, predictive models to enhance such interventions [24, 25]. Lastly, we are examining how AI teammates can align with human decision-makers in terms of values and decision styles, which focuses on the integrity (process) and benevolence (purpose/intent) aspects of trustworthiness (as opposed to the ability / performance dimension of trustworthiness). Combined, these efforts may help to further advance and improve trust calibration in human-AI teams.

References


