Socratic: Enhancing Human Teamwork via AI-enabled Coaching

Sangwon Seo
Rice University
Houston, TX, USA
sangwon.seo@rice.edu

Bing Han
Rice University
Houston, TX, USA
bing.han@rice.edu

Rayan E. Harari
Harvard Medical School
Boston, MA, USA
rharari@bwh.harvard.edu

Roger D. Dias
Harvard Medical School
Boston, MA, USA
rdias@bwh.harvard.edu

Marco A. Zenati
Harvard Medical School
Boston, MA, USA
marco_zenati@hms.harvard.edu

Eduardo Salas
Rice University
Houston, TX, USA
eduardo.salas@rice.edu

Vaibhav Unhelkar
Rice University
Houston, TX, USA
vaibhav.unhelkar@rice.edu

ABSTRACT

Coaches are vital for effective collaboration, but cost and resource constraints often limit their availability during real-world tasks. This limitation poses serious challenges in life-critical domains that rely on effective teamwork, such as healthcare and disaster response. To address this gap, we propose and realize an innovative application of AI: task-time team coaching. Specifically, we introduce Socratic, a novel AI system that complements human coaches by providing real-time guidance during task execution. Socratic monitors team behavior, detects misalignments in team members' shared understanding, and delivers automated interventions to improve team performance. We validated Socratic through two human subject experiments involving dyadic collaboration. The results demonstrate that the system significantly enhances team performance with minimal interventions. Participants also perceived Socratic as helpful and trustworthy, supporting its potential for adoption. Our findings also suggest promising directions both for AI research and its practical applications to enhance human teamwork.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Intelligent agents**; *Planning under uncertainty*; *Multi-agent planning*; *Machine learning*.

KEYWORDS

Teamwork, Mental Models, Decision Support, Imitation Learning

ACM Reference Format:
Sangwon Seo, Bing Han, Rayan E. Harari, Roger D. Dias, Marco A. Zenati, Eduardo Salas, and Vaibhav Unhelkar. 2025. Socratic: Enhancing Human Teamwork via AI-enabled Coaching. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Consider your favorite sports team — whether it is soccer, cricket, basketball, or another team sport — working together to achieve a common goal. Even though all the team members are trained professionals, some teams consistently outperform others. Indeed, *a team of individual experts does not necessarily make for an expert team* [3]; building a successful team requires the confluence of multiple factors [40]. Human factors research has identified key drivers of team effectiveness, including capability, coordination, communication, and coaching [82]. Through targeted training and interventions, human teams can significantly improve coordination and enhance their performance in collaborative tasks.

Coaches play a crucial role in both team training and interventions. Rather than performing tasks themselves, they enhance collaboration by offering expert insights. These insights are provided both during task execution, such as in games, and during training sessions, such as in practice. While coaches are common in professional sports, integrating them into life-critical fields presents significant challenges [52, 73]. Resource constraints and a shortage of experts make it difficult to employ coaches during task execution. For example, in surgical teamwork, a coach could be invaluable in reducing preventable medical errors [39, 73, 91]. Reducing these errors would significantly improve patient health outcomes. However, due to the shortage of medical professionals, it is not feasible for a specialist to continuously serve in this coaching role. Similarly, in aviation, coaches assist with simulation-based training, but they cannot accompany a flight crew on every flight [34].

Recognizing the need for coaching assistance in life- and safety-critical applications, we propose an innovative use of artificial intelligence (AI): task-time team coaching. Specifically, we envision an AI agent that complements a human coach by monitoring a team during task execution and providing real-time guidance to improve teamwork, particularly in situations where the human coach may be busy or unavailable. While coaches offer a variety of feedback before, during, and after tasks, in this work, we limit our scope to delivering task-time feedback in time-critical tasks. For this setting,

An extended version of this paper, which includes supplementary material mentioned in the text, is available at http://tiny.cc/socratic-appendix

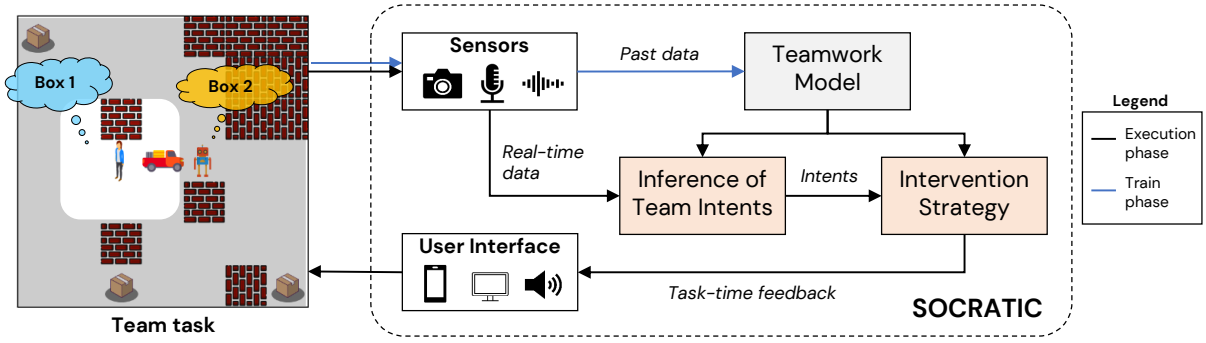


Figure 1: Schematic of Socratic: an AI coach for enhancing teamwork during task execution. Blue arrows represent the workflow during the training phase, whereas black arrows indicate the workflow during the execution phase.

we present a novel proof-of-concept AI agent, Socratic, designed to complement a human coach and thereby enhance teamwork.

Illustrated in Fig. 1, the overall design of Socratic is grounded in the extensive literature on human team training. Specifically, Socratic operates by observing task execution and identifying points where the team's mental models regarding shared plans may become misaligned. When such a misalignment is detected, Socratic prompts the team to pause, reflect on their plans, and offers suggestions for improvement. By encouraging the team to reconsider future actions that could lead to inefficiencies or errors, Socratic aims to enhance collaborative decision-making.

From an AI perspective, Socratic leverages recent advances in imitation learning and multi-agent systems. First, it employs multi-agent imitation learning to model team behavior based on demonstrations from previously executed tasks. Using this model and data from an ongoing task, it builds on TIC– a recent algorithm for agent-based teamwork – to algorithmically detect points of misalignment and generate recommendations. Finally, through an interactive user interface, Socratic delivers the automatically generated interventions aimed at aligning the team's understanding and improving overall performance.

To evaluate Socratic, we conducted two human subject experiments: one focused on training and the other on validation. Both experiments involved two collaborative tasks and dyadic teams. In the training experiment, we curated a novel dataset of human demonstrations annotated with intents and used it to train Socratic. In the validation experiment, we conducted a randomized controlled trial to evaluate both Socratic's objective performance and the users' subjective perceptions of the system. The experimental results show that Socratic significantly improves team performance with minimal interventions. Equally important for its adoption, participants perceive Socratic as helpful to improving teamwork. The evaluations also suggest promising directions for both AI research and the proposed applications, highlighting the potential of AI agents to support human teamwork.

2 BACKGROUND

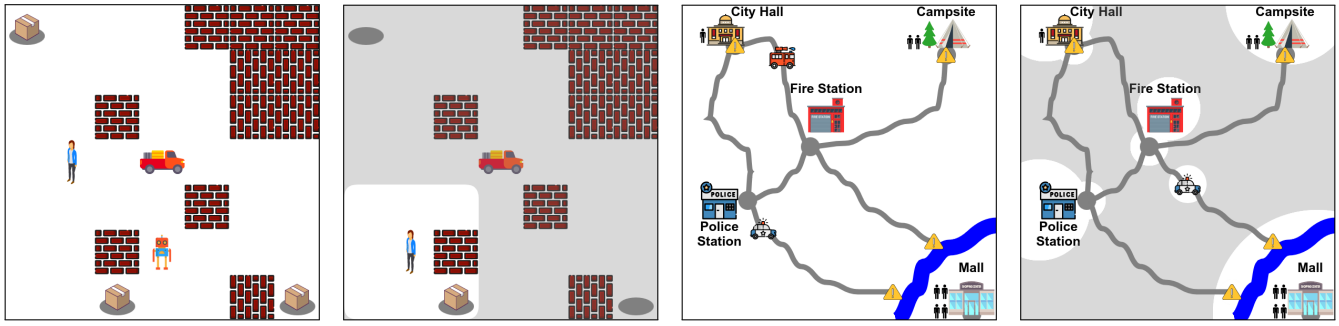
Before describing Socratic, we present the concepts and related research that inform our approach.

2.1 Collaborative Tasks

Teamwork is fundamental to many human endeavors, spanning scenarios such as sports, healthcare, aviation, and more. Our focus is on *time-critical scenarios*, such as healthcare and disaster response, where effective teamwork is crucial for mission success. Teamwork occurs at various levels, ranging from large organizations to small ad-hoc teams. We focus on *mission-oriented, sequential tasks*, where an established team works toward a clearly defined mission (e.g., Fig. 2). Although the mission is well-defined, there are often multiple ways to achieve the task. Real-world challenges, such as uncertainty, information asymmetry, and partial observability, can create barriers to efficient teamwork and task completion. Finally, we consider teams composed of human members, either in human-only teams or hybrid human-AI teams.

To develop an AI agent capable of supporting such teamwork, the first step is to mathematically model the task and team dynamics. Fortunately, research in multi-agent systems offers several established formalisms for modeling collaborative tasks, including belief-desire-intention frameworks, Markov models, and game theory [2, 7, 15, 19, 20, 24, 43, 46, 47, 70, 79, 81, 84]. In our work, we leverage decentralized multi-agent partially observable Markov decision processes (Dec-POMDPs) [51]. This choice is motivated by their ability to model tasks with well-defined missions, structured teams, time constraints, action uncertainties, and partial observability, as well as their prior use in modeling time-critical collaboration scenarios like disaster response [8, 12, 36, 37, 86].

We define a task as the tuple $\mathcal{M} = (n, S, A, \Omega, T, O, R, \gamma, h)$, where n is the number of agents, S, $A \doteq \times_i A_i$ and $\Omega \doteq \times_i \Omega_i$ denote a state space, an action space and an observation space, respectively, $T(s'|s, a)$ denotes a probability of a state s transitioning to another state s' given a joint action $a \doteq (a_1, \cdots, a_n)$, $O(o|s', a)$ is a probability of a joint observation $o \doteq (o_1, \cdots, o_n)$ given a state s' and a joint action a, R is the task reward (objective), γ is the discount factor, and h is the task horizon. In theory, a team could act optimally by computing a decentralized policy using Dec-POMDP solvers based on the task model. However, it is unrealistic to expect human team members to compute and execute such a policy flawlessly and without errors. Therefore, we draw on human factors research to model team behaviors and identify strategies for improving teamwork.



(a) Map of `Movers` (b) `Movers`: Alice's perspective (c) Map of `Flood` (d) `Flood`: Police's perspective

Figure 2: `Movers` and `Flood` domains, detailed in Sec. 4.1. Team members can observe only the unshaded region of the environment.

2.2 The Science of Human Teamwork

The science of human teamwork focuses on the question: *What makes teams work?* [68]. Over the past four decades, psychologists and human factors researchers have systematically identified the factors that make teamwork challenging and developed methods to improve it [9, 10, 65, 67]. We briefly review key insights that inform our work, while directing readers to recent survey by Tannenbaum and Salas [82] for more details. There is broad consensus that teamwork is especially challenging in time-critical scenarios, where success depends on the convergence of multiple factors. A major challenge is that humans often make suboptimal decisions due to bounded rationality [29, 30, 77] and limited situational awareness [13, 14, 54, 78], especially under time constraints. Hence, Socratic does not assume perfect rationality or situational awareness from team members. Even teams composed of experts may not function optimally due to a lack of shared mental models, leading to poor coordination and even fatal errors [6, 22, 28, 40, 41, 89]. Thus, Socratic explicitly considers team members' intent and allows for potential misalignment, which can lead to suboptimal teamwork.

To enhance teamwork, the science of teamwork recommends several methods and best practices, including effective communication, simulation-based training, and coaching – the latter being the focus of this paper. Coaches play a crucial role by assessing teamwork and providing feedback to improve it. While human coaches rely on their expertise and experience for these activities, the science of teamwork has developed principled methods and formalized best practices for coaching. Researchers have established robust methods for assessing teams [11, 16, 17, 32, 68] and generating targeted insights to enhance teamwork [4, 21, 55, 66, 93]. However, these assessments are typically post-hoc, lack automation, and are limited to contexts where a human coach is available. Thus, we explore the design of an AI coach capable of operationalizing these insights, detecting misalignments in team members' shared intents, and providing real-time feedback during task execution.

2.3 AI-Assisted Teamwork

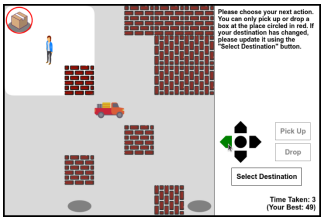
AI-assisted human teamwork is an emerging area of research with applications being explored across various domains [23, 31, 56, 57, 62, 71, 73, 90]. For instance, DeepMind and Liverpool FC are investigating data-driven approaches to analyze and enhance team strategies in football [85]. For applications in healthcare and disaster response, researchers have applied AI to analyze team conversations and improve extended-duration teamwork [1, 33]. Closer to our focus on time-critical scenarios, domain-specific methods for automated teamwork assessment have been developed [16, 35]. However, these methods, to our knowledge, provide only post-hoc support, and AI has not yet been used for task-time coaching.

Approaches for assessing and improving teamwork in human-robot or robot-only teams are also relevant to our work [63, 64, 83, 86, 94]. Research in human-robot collaboration introduces metrics for evaluating teamwork [26, 38, 50] and algorithms for improving it [5, 45, 49, 69, 87]. However, these methods focus on training robots to work with humans. In contrast, our work centers on an AI agent that provides coaching and decision support, without directly performing the task. Closest to our work are the recent frameworks TIC [72] and TARS [97], which generate task-time interventions to enhance multi-agent teamwork. TARS uses Dynamic Epistemic Logic-POMDP to generate interventions through planning algorithms [97]. TIC employs Dec-POMDPs and multi-agent imitation learning to generate interventions through a learned model [72, 76]. However, these methods have not been applied or evaluated in settings with human team members.

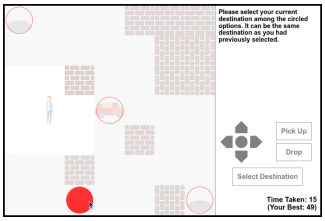
Our work builds on these methods but differs in key ways. First, we adopt a systems perspective to develop Socratic that includes both an intervention algorithm and a user interface, enabling interaction with and coaching for human users. Second, our methodology incorporates mechanisms to collect training data on human teamwork, including their cognitive states. Finally, we validate the effectiveness of the solution through human subject experiments.

3 SOCRATIC

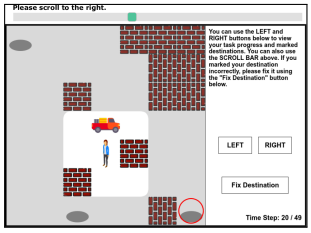
We now describe Socratic: the *System for Objective Coaching through Automated Task-time Interventions for Collaboration*. Drawing on multiple disciplines (Sec. 2), we begin by outlining the system's design requirements and architecture. We then detail its key components: a module for monitoring team performance, an algorithm for learning team behavior models, another for generating task-time interventions, and a user interface to communicate these interventions to the team. We illustrate Socratic using two human-AI collaboration tasks, detailed in Sec. 4.1 and inspired by real-world scenarios, implemented on a web-based simulation platform.



(a) Task scene and UI for Alice.



(b) UI for intent annotation.



(c) UI for after-action review.

Figure 3: Snapshots of the `Movers` task from the first study (larger images are available in the appendix).

3.1 System Overview

3.1.1 Scope. We limit our scope to collaborative tasks modeled as Dec-POMDPs (Sec. 2.1) and teams that include at least one human member. Importantly, we do not make assumptions about team members' rationality or expertise levels. As reviewed in Sec. 2.2, the science of teamwork identifies several key drivers of effective teamwork. In this proof-of-concept work, we focus on team alignment[1] — ensuring that the team is "on the same page." Misalignment is particularly common in time-critical scenarios, where teams may lack sufficient time to communicate and coordinate shared plans. Additionally, real-world factors such as partial observability, fatigue, and uncertainty can further degrade team member's understanding of each other's beliefs, desires, and intentions.

3.1.2 Design Requirements. With this scope defined, we design SoCRATIC: an AI-enabled coaching agent to improve teamwork during task execution. The design process began by identifying system requirements through brainstorming sessions with an interdisciplinary team of researchers in human factors, team training, AI, and usability. We determined that an AI agent capable of *detecting misalignments in team members' intents* and *alerting the team to pause, reflect, and adjust their plans* is both feasible to develop and can significantly enhance collaboration. For the successful realization and adoption of such an agent, we distilled key requirements (**Rx**); namely, SoCRATIC must be:

- **R1.** able to sense and monitor teamwork;
- **R2.** able to accurately infer intents of the team members;
- **R3.** able to accurately anticipate future actions of the team;
- **R4.** able to generate effective task-time interventions;
- **R5.** able to effectively deliver the interventions; and
- **R6.** perceived as useful by the team members.

[1]Investigating other drivers of effective teamwork, intervention mechanisms, and teamwork settings is an important avenue for future research.

3.1.3 System Architecture. To meet the design requirements, SoCRATIC leverages recent advancements in imitation learning and multi-agent systems, incorporating an interactive user interface to monitor the team and deliver interventions. For **R1** (sensing and monitoring teamwork), we assume SoCRATIC is equipped with sensors to observe both the team and task environment. Similar to sport scenarios, where team members may have partial observability, the coach has full visibility of the environment. To meet **R2** (inferring intents) and **R3** (anticipating future actions), SoCRATIC employs a recent multi-agent imitation learning algorithm BTIL that explicitly models team members' intents and learns a generative model of team behavior [76]. Building on this model, SoCRATIC utilizes a specialized instance of the TIC framework to generate task-time interventions to meet **R4** [72]. Lastly, SoCRATIC includes a user interface to deliver these interventions to the team, addressing **R5**.

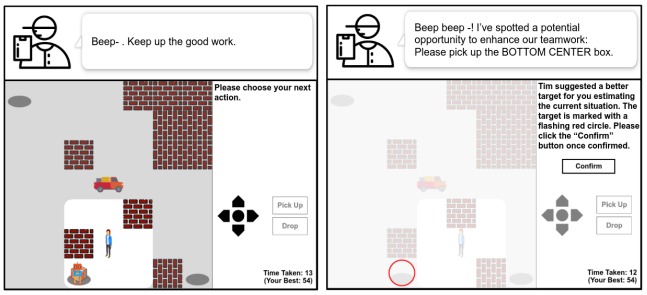
3.1.4 System Operation. SoCRATIC operates in two phases: training and execution. During the training phase, SoCRATIC observes the team performing tasks in practice sessions, collecting teamwork data and learning generative models of team behavior. During task execution, SoCRATIC uses the learnt generative model to infer team intents, detect misalignments, and compute and deliver effective interventions. We now detail each system component.

3.2 Training Phase

3.2.1 Team Model. To effectively monitor the team, SoCRATIC builds upon a mathematical model of the task and team behavior. Having described the task model in Sec. 2.1, we now formalize the model of team behavior. Human decision-making often depends on factors beyond the task state, such as cognitive states corresponding to beliefs and intents [25, 48]. Hence, SoCRATIC explicitly models the influence of team members' intent – a latent variable – on their behavior. More specifically, following the Agent Markov Model (AMM) [88], j-th team member's behavior is defined by the tuple $\mathcal{H}_j = (X, \pi_j, \zeta_j; \mathcal{M})$, where X represents the set of possible task-specific intents, $\pi_j(a|x,s)$ denotes the team member's policy, and $\zeta_j(x'|s',a,x)$ represents the intent transition model.[2] While this model is well-defined, it is not trivial for domain experts to specify. Therefore, SoCRATIC leverages imitation learning to learn the model parameters from demonstrations collected during training sessions.

3.2.2 Model Learning. In particular, SoCRATIC uses BTIL to learn the unknown parameters of the team behavioral model: $\pi(a|s,x)$ and $\zeta(x'|s',x,a)$. BTIL is a multi-agent imitation learning algorithm that explicitly models latent decision factors, such as intents [76]. By leveraging a Bayesian approach, BTIL has been shown to attain sample- and label- efficient model learning from team demonstrations. Additionally, BTIL can learn from both optimal and sub-optimal demonstrations. This is especially important for SoCRATIC, as it learns the team model from demonstrations collected during practice sessions, where team behavior may not always be optimal. In practice, SoCRATIC's model learning begins with the collection

[2]Although team members' behavior may also depend on other latent factors, such as cognitive states and beliefs about unobserved parts of the environment, the decision to model only intent simplifies the system design. Our experiments confirm that this modeling choice is valid for the domains considered. However, we believe that performance of future AI-enabled coaching systems could be further enhanced by incorporating additional decision factors and more sophisticated behavioral models.



(a) Example: no intervention **(b) Example: Socratic-generated intervention**

Figure 4: Snapshots of Socratic's interactive user interface (larger images are available in the appendix).

of data on observable features of team demonstration, specifically (s, a)-trajectories. With the assistance of a human annotator, a subset of these trajectories is annotated with the values of team intent (x). Using this combination of trajectory data and intent annotations, Socratic utilizes the semi-supervised variant of BTIL to learn the team behavioral model $\mathcal{H}_j \forall j = 1 : n$.

3.2.3 Team Monitoring. Equally critical to team modeling are the mechanisms for monitoring the team and collecting teamwork data: specifically, (s, a)-trajectories and annotations of (x) for a subset of the training data. In this proof-of-concept, we focus on collaborative tasks conducted through a web-based interface and develop methods for data collection and annotation specific to this setting, illustrated in Fig. 3 and detailed in Sec. 4.2. For real-world applications, we recommend using multimodal sensors to monitor and gather teamwork data. We leave the exploration of related perception challenges for future work, with relevant research directions discussed in Sec. 5. Socratic uses the same monitoring mechanisms during the task execution phase, which we describe next.

3.3 Execution Phase

3.3.1 Intent Detection. Socratic monitors the team during task execution, identifying potential misalignments in team members' intents and computing timely interventions. This capability is enabled by TIC, a framework that has been experimentally shown to generate task-time interventions that enhance teamwork among AI agents [72]. We extend this framework to develop an AI-enabled coaching system for teams that include human members. During task execution, Socratic can observe team members' states and actions, but their intents (a latent variable) remain unobservable. While Socratic leverages a human annotator to obtain partial intent annotations during the training phase, involving a human in the loop during task execution is impractical. Therefore, to infer team members' intents, Socratic frames the problem as one of Bayesian filtering. Specifically, given the learned model of team behavior $(\mathcal{H}_j \forall j = 1 : n)$ and the partial (s, a)-trajectory of the team's task execution, Socratic employs the forward-backward algorithm to infer each team member's current intent $\hat{x}$.

3.3.2 Intervention Generation. Socratic next uses the inferred intents to assess whether the team is aligned. If the intended plans of the team members are likely to lead to suboptimal task performance, Socratic intervenes by weighing the costs and benefits of the intervention. Under the TIC framework, determining this balance requires an intervention strategy, which can be hand-crafted or learned. For Socratic, we opt for a learned, value-based strategy to minimize human effort in intervention generation. Specifically,[3]

- Socratic first computes the expected return (g) conditioned on the inferred intent: $g(\hat{x}|s) = E_{\mathcal{H}}[\sum_t \gamma^t r_t | s, \hat{x}]$.
- Next, Socratic computes the intent values and return for a hypothetical fully aligned team as $x^* = \arg\max_x g(x|s)$ and $g(x^*|s) = E_{\pi,\zeta}[\sum_t \gamma^t r_t | s, x^*]$, respectively. We define the benefit of an intervention as the difference between the optimal and estimated return: $g(x^*|s) - g(\hat{x}|s)$.
- Finally, if the benefit of an intervention exceeds its cost c by a pre-defined threshold (i.e., $g(x^*|s) - g(\hat{x}|s) > c + \delta$), then Socratic prompts the team to pause, reflect on their plans, and recommends the optimal plan corresponding to x^*.

Choosing an appropriate cost (c) and threshold (δ) for interventions is crucial, as unnecessary or incorrect interventions could impair team performance and reduce human trust in, and adoption of, Socratic. Sec. 4.3 outlines the approach for selecting these hyperparameters for our implementation and evaluation of Socratic.

3.3.3 Intervention Delivery. To assist human team members, in addition to generating interventions, Socratic requires effective mechanisms for delivering these instructions. In this work, we utilize an interactive user interface for delivering interventions, as illustrated in Fig. 4 and detailed in Sec. 4.2. Since human team members can choose whether to accept the AI-generated recommendations, Socratic incorporates a hyperparameter p_a, which models the probability of a human accepting its recommendation.

4 FEASIBILITY STUDIES

We conducted human subject evaluations to assess the feasibility of AI-enabled coaching in enhancing collaborative task execution. The IRB-approved experimental protocols were designed to evaluate:

Q1. Is Socratic capable of learning a useful team model?
Q2. Is Socratic capable of improving team performance?
Q3. Is Socratic perceived as useful by human users?

The evaluations consisted of two studies: training and validation. Both studies involved dyadic teams completing two collaborative tasks. In the training study, we curated a novel dataset of human demonstrations, annotated with intents, to train Socratic. In the validation study, we conducted a randomized controlled trial[4] to evaluate the objective performance of Socratic and gather subjective feedback from participants regarding AI-enabled coaching.

4.1 Domains

We first describe the collaborative tasks used in our evaluations: `Movers` and `Flood`. Introduced in [72], these dyadic tasks require teams to maintain a shared plan for effective execution. However, due to partial observability and lack of communication, achieving coordination and high task performance is challenging.

[3] Since this computation relies on observations, task model, and the learned model of team behavior, it requires no additional human input or domain-specific knowledge.
[4] Validation Study: The experimental group received coaching from Socratic, while the control group completed tasks without any AI-enabled coaching.

4.1.1 `Movers`. As shown in Fig. 2a, Alice and Rob are tasked with moving three boxes to the truck as quickly as possible. The boxes are heavy and require both teammates to lift them together. Teamwork is effective as long as the teammates agree on which box to move and act accordingly, regardless of the order. However, as depicted in Fig. 2b, each team member has a limited view of the environment and cannot communicate with the other during task execution, making coordination challenging. The task ends after 150 time steps or when all boxes are moved to the truck, whichever comes first. The cumulative team reward is defined as 150 minus the time step at which the task terminates.

4.1.2 `Flood`. The second task is inspired by time-critical disaster response scenarios. As shown in Fig. 2c, the environment includes victims at three sites: one at City Hall, two at the Campsite, and four at the Mall. A rescue team, consisting of a police car and a fire truck, must save all victims within a time limit of 30 time steps. While victims at City Hall and the Campsite can be rescued by a single vehicle, rescuing those at the Mall requires both vehicles to collaborate in repairing one of two bridges. Teamwork in this task is more complex: sometimes the team must work together (e.g., at the Mall), while in other cases, dividing sub-tasks is more efficient (e.g., at City Hall and the Campsite). As depicted in Fig. 2d, team members can only observe each other when at the same location or a landmark, complicating coordination. The total team reward is defined as the number of victims rescued within the time limit.

4.2 Study 1: Training

The first study focused on the training phase of Socratic to collect training data and evaluate **Q1**. Forty participants (20 females, 20 males, mean age: 28.5±4.9 years) completed the `Movers` and `Flood` tasks with a robot teammate, while also providing annotations of their task-relevant intent ($x \in X$). For `Movers`, intent is defined as the box a team member plans to pick up or drop next. For `Flood`, intent refers to the site a team member plans to approach next.

4.2.1 Materials and Setup. We developed a website using the Flask framework [18] that included the two tasks, complete with a user interface for task execution and intent labeling (Fig. 3). This platform enabled participants to perform the experiment remotely. Each participant was paired with a robot teammate, forming a dyadic human-robot team. Following Sec. 3.2.1, behavior of each teammate was modeled as $\mathcal{H}_j = (X, \pi_j, \zeta_j; \mathcal{M})$. The robot (denoted as R) had its policy π_R pre-trained using value iteration, and its intent dynamics ζ_R were manually specified. The experiment aimed to collect data on the human teammate's (denoted as H) behavior in order to learn their policy π_H and intent dynamics ζ_H. Both teammates had to make decisions under partial observability and infer the intent of their teammate to complete the task successfully.

4.2.2 Procedure. Upon providing informed consent, participants were introduced to the experiment and completed a demographic survey. They were then instructed to complete the dyadic tasks with the robot, following the same process for both `Movers` and `Flood`. This process included an interactive tutorial and four task trials. The tutorial introduced participants to the task and trained them on how to navigate the user interface (UI). The tutorial featured a guided scenario that mirrored the actual task. For each domain, participants proceeded through four task trials after completing the tutorial. Each trial was followed by a simplified *after-action review* [44, 60, 80]. During each trial, the website displayed a task scene and a task control UI, allowing participants to control their character to complete the task (Fig. 3a). The experiment collected data on task states (s) and team actions (a) while generating human intent annotations (x). Intent annotations were generated during the task and refined via the after-action reviews, as described in Sec. 4.2.3. After completing four trials for both the `Movers` and `Flood` tasks, the experiment concluded with a post-experiment survey, where participants provided open-ended feedback about their experience.

Table 1: Survey Statements

#	Statement (rated on a 5-point Scale)
1	The team worked fluently together.
2	The robot contributed to the fluency of the interaction.
3	The team improved over time.
4	During the task, I followed the AI Coach suggestions in general.
5	The AI Coach was intelligent.
6	The AI Coach was trustworthy.
7	The AI Coach's suggestions were effective.
8	The AI Coach's suggestions were timely.
9	The AI Coach contributed to the fluency of the interaction.

4.2.3 Annotation. Training Socratic requires both observable (s, a)-trajectories and time series data of team members' intents (x), which are latent and must be manually annotated. In this study, we collected intent data through participant reports, supported by user-centered annotation mechanisms to ensure reliable data collection. Recall that in both domains, intent is tied to a physical location in the task scene, such as a box or a rescue site. To streamline reporting, we developed a "Destination Selection" UI, allowing participants to report their intended destination during task execution (Fig. 3b). Potential destinations are highlighted, and participants select their intended location with a mouse click. Participants are encouraged to update their intent when it changes and are prompted if five time steps pass without a report. Selected intents are visually indicated with a flashing red circle. Additionally, key actions like "Pick Up," "Drop," or "Rescue" are restricted to the selected destination, ensuring alignment between reported intents and actions. After each task trial, participants use the "after-action review" UI to verify and, if needed, correct their annotations (Fig. 3c). This interface replays the task execution, displaying both team actions and selected intents, allowing participants to confirm their reports. If discrepancies are found, participants can adjust incorrect intents using the "Fix Destination" button, improving the accuracy of the dataset used to train and validate Socratic.

4.2.4 Data Analysis. We collected 160 demonstrations per domain and trained Socratic using a semi-supervised approach. Recognizing that intent annotation is resource-intensive, we used only 30% of the intent labels for training and reserving the rest for validation. This approach enables evaluating Socratic in a more realistic setting, where only partial intent annotations are available.

Table 2: Success Rate of the Learned Model

Domain	Intent	Success (%)	Wrong (%)	Nowhere (%)
Movers	Box 1	75.4	5.9	18.7
	Box 2	72.5	12.5	15.0
	Box 3	69.3	13.8	16.9
	Truck	99.6	0.0	0.4
	Mean	79.2	8.05	12.75
Flood	City Hall	88.3	0.3	11.4
	Campsite	38.2	9.2	52.6
	Bridge 1	59.4	2.2	38.4
	Bridge 2	45.4	3.1	51.5
	Mean	57.8	3.7	38.5

4.3 Study 2: Validation

After collecting the training data, we conducted a second study to evaluate Socratic's performance (**Q2**) and perceived usefulness (**Q3**). The study was a randomized control trial, where only the experimental group received coaching from Socratic.

4.3.1 Participants. We recruited participants via Prolific [53]. Of the 73 users who accessed the experiment, 61 completed it. To ensure balanced group sizes, we used the first 30 participants from each group. The control group consisted of 13 females and 17 males (age: 28.7±8.4 years), while the experimental group included 11 females, 17 males, and 2 non-binary participants (27.8±9.1 years).

4.3.2 Materials and Setup. Similar to the first study, we developed a website featuring the two tasks with an interactive user interface. However, instead of intent annotation mechanisms, this version incorporated Socratic on the backend and its user interface on the frontend for interacting with the team during task execution. As shown in Fig. 4, the interface features an AI coach icon with a speech balloon above the task screen. During task execution, the speech balloon nominally displays: *"Keep up the good work."* However, if Socratic detects misaligned intents and decides to intervene, it pauses the task, prompts the team to reflect on their plans, and recommends an optimal course of action corresponding to x^*. As illustrated in Fig. 4b, the speech balloon displays:

"I've spotted a potential opportunity to enhance our teamwork: Please ⟨recommendation⟩".

The suggestion is highlighted with a red circle, and participants must click the "Confirm" button to resume the task. While Socratic offers recommendations, participants ultimately decide whether to accept them. To account for the fact that not all recommendations will be followed, we set $p_a = 0.9$ for this study. Socratic utilizes two additional hyperparameters: the cost of intervention c and the threshold δ. For `Movers`, the intervention cost is set to 1, representing the loss of one time step to pause and reflect on the recommendation. In contrast, for the life-critical `Flood` task, the cost is considered negligible ($c = 0$) as any small delays caused by interventions are justified if they assist in rescue efforts. The threshold δ was determined through a grid search over the hyperparameter space, with values set to 5 for `Movers` and 0.1 for `Flood` based on simulated experiments with the learned teamwork model.

4.3.3 Procedure. The overall structure of this experiment closely mirrors that of the first study. It is web-based and includes a study overview, a demographic survey, `Movers` and `Flood` domains, and a post-experiment survey. For each domain, participants completed an interactive tutorial followed by four task trials. While the tutorials and trials were similar to the first study, intent annotation features were removed. Only for the experimental group, Socratic's features were integrated into the tutorial and task trials. Each domain involved one practice trial to help participants familiarize themselves with the task and the robot teammate, followed by three test trials. Neither group received assistance from Socratic during the practice trial. In the test trials, the control group performed the task without coaching, while the experimental group received task-time interventions from Socratic. After the trials, participants completed the survey described next.

4.3.4 Measures. We assess **Q1** by quantifying the intent-conditioned success rate of the learned model. For **Q2**, team performance is evaluated using task scores. Beyond improving teamwork, the perceived usefulness of Socratic is essential for its adoption by human users. Hence, to address **Q3**, we use subjective statements adapted from a widely used scale [26]. The first three questions solicited participants perception regarding the robot teammate, while the rest regarding Socratic the AI coach. Control group rated the first three statements listed in Table 1, while the experimental group rated all statements. Responses were recorded on a 5-point scale, ranging from *strongly disagree* (1) to *strongly agree* (5).

4.4 Experimental Results

4.4.1 Socratic learns intent-driven models of team behavior. To address **Q1**, Socratic first learns models of team behavior using the training data. We then evaluate if the learned model captures intent-driven behaviors by simulating the policy 1000 times for each intent x and measuring its success rate in completing the intended sub-task within 20 time steps. For instance, if the specified intent is to rescue victims at City Hall, we check how often the model succeeds. Table 2 presents the success rates of the learned model for each intent. Failures are categorized as either *Wrong* (where the model accomplishes a sub-task associated with a different intent, such as rescuing victims at the Camp Site when the specified intent was City Hall) or *Nowhere* (where the model fails to complete any sub-task within the time limit). On the challenging task of modeling team behaviors from human data, the model achieved an average success rate of 79% for `Movers` and 58% for `Flood`. Most failures belong to the *Nowhere* category, suggesting that model learns intent-driven models of team behavior. Through the second study, we find that this model learning performance is sufficient for Socratic to deliver effective task-time interventions to improve teamwork.

4.4.2 Socratic improves teamwork via targeted interventions. To answer **Q2**, we compared the performance of the two groups, using a cost-adjusted score that accounts for the time spent on processing and responding to interventions. Specifically, *Score* is defined as $R - C$, where R is the cumulative team reward and C is the total cost of interventions. For the control group, *Score* is equal to the task score, as no interventions took place. As shown in Fig. 5, the experimental group outperformed the control group. In `Movers`,

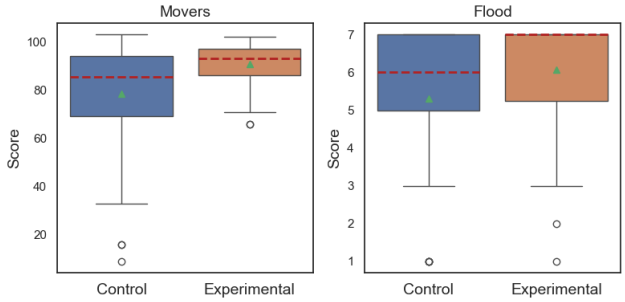


Figure 5: Team Scores: with and without Socratic.

teams coached by Socratic scored on average 90.8(±8.4) compared to 78.6(±21.7) for the control group, with a statistical significance of $p < 0.001$. Similarly, for `Flood`, the average score of the experimental group was 6.1(±1.5) versus 5.3(±1.7) for the control group, with $p < 0.01$. The average number of interventions was 3.9 (±1.7) for `Movers` and 2.3 (±2.1) for `Flood`. These results highlight that Socratic effectively enhanced teamwork with minimal interventions. This targeted approach to delivering interventions is crucial for improving team performance and building human users' trust in AI-enabled coaching.

4.4.3 Socratic is perceived as useful by human users. To answer **Q3**, we analyzed participants' survey responses. Fig. 6 displays the percentage of positive, neutral, and negative assessments for each statement. Responses to statements #1-3, which evaluated the robot teammate, were largely similar across both groups, indicating that participants had comparable perceptions of the robot teammate's capabilities. This consistency ensures a fair comparison between the groups, allowing us to accurately evaluate the AI coach's utility.

Statements #4-9, which evaluated Socratic and were rated only by the experimental group, indicate that participants perceived Socratic as useful, effective, intelligent, and trustworthy. Based on these statements, the average rating of Socratic was 3.81(±1.03) for `Movers` and 3.28(±1.32) for `Flood` on a 1 – 5 scale. Except for statement #8 for `Flood` task, the positive responses outweighed the negative ones for all statements.

Open-ended feedback suggested that Socratic was seen as more helpful in the `Movers` task, while participants found `Flood` more challenging. Regarding statement #8, which asked about the timeliness of Socratic 's recommendations, one participant commented:

> *"There was one occasion when the AI's suggestion came a bit late, causing me to waste a few moves."*

While Socratic is already designed to provide proactive guidance using a predictive model of teamwork, participants' responses suggest that they value this proactivity and may expect even more planning support from an AI Coach. Informed by these findings, we conclude by summarizing our contributions and discussing their implications for both team training and AI research.

5 CONCLUSION

We introduce Socratic: a system that provides AI-enabled coaching to teams with human members during task execution. Through

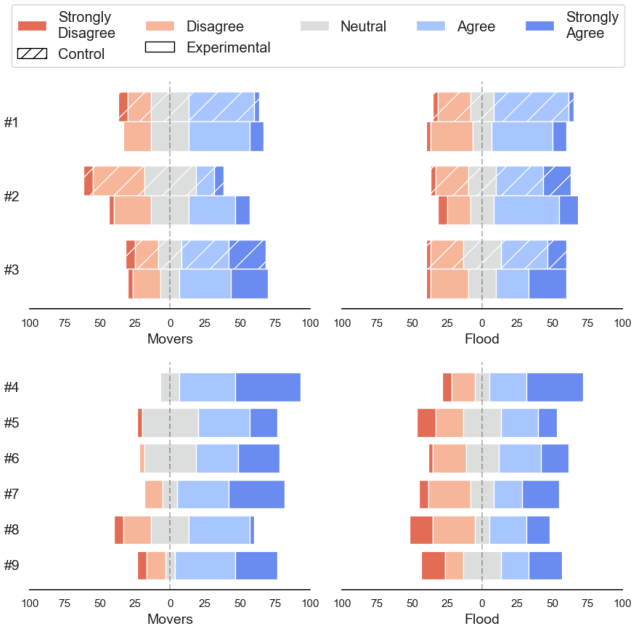


Figure 6: Participant Responses to Survey Statements

human subject experiments on challenging dyadic tasks, we demonstrated that Socratic not only enhances team performance but is also perceived as useful by participants. Since Socratic does not perform the tasks itself, it has the potential to assist in various domains, including those where AI agents may lack the capability to act but can still analyze and enhance human task execution.

Along with its strengths, we also highlight the limitations of this proof-of-concept work, which suggest exciting future research directions. First, while the experimental tasks captured challenging elements of real-world collaboration, they were conducted in a web-based environment. Future work should investigate AI-enabled coaching in more complex scenarios that include dynamic environments, larger teams, multiple objectives, ad-hoc collaboration, or members with diverse expertise [74, 75, 79]. Additionally, expanding AI coaching to physical teamwork settings using multimodal perception is an important next step [92, 95].

Second, from a human-centered perspective, our study opens up new opportunities to examine how AI can support team training and complement human coaches. Informed by the science of teamwork, expanding Socratic's interventions to include more varied recommendations would further enhance its utility [4, 55, 93]. To enhance usability, interventions could be delivered through user-friendly interfaces such as screens, audio systems, or augmented/virtual reality. Lastly, given AI coaches can make errors, a critical area for further investigation is ensuring the safe and responsible deployment of AI coaches [27, 42, 58, 59, 61, 96]. This includes examining how trust in AI coaches can be effectively built, calibrated, and maintained to foster successful human-AI collaboration.

ACKNOWLEDGMENTS

This research was supported by NSF award #2205454.

REFERENCES

[1] Ofra Amir, Barbara J Grosz, and Krzysztof Z Gajos. 2016. Mutual influence potential networks: Enabling information sharing in loosely-coupled extended-duration teamwork. In *Proceedings of IJCAI*.

[2] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research* 27, 4 (2002), 819–840.

[3] Tiffany Bisbey, Allison Traylor, and Eduardo Salas. 2021. Transforming teams of experts into expert teams: eight principles of expert team performance. *Journal of expertise* 4, 2 (2021).

[4] Jennifer Jane Britton. 2015. Expanding the coaching conversation: Group and team coaching. *Industrial and Commercial Training* 47, 3 (2015), 116–120.

[5] Guilhem Buisan and Rachid Alami. 2021. A human-aware task planner explicitly reasoning about human and robot decision, action and reaction. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*. 544–548.

[6] Janis A Cannon-Bowers, Eduardo Salas, and Sharolyn Converse. 1993. Shared mental models in expert team decision making. *Individual and group decision making: Current issues* 221 (1993), 221–46.

[7] Crystal Chao and Andrea Thomaz. 2016. Timed Petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration. *The International Journal of Robotics Research* 35, 11 (2016), 1330–1353.

[8] Shaofei Chen, Feng Wu, Lincheng Shen, Jing Chen, and Sarvapali D Ramchurn. 2015. Decentralized patrolling under constraints in dynamic environments. *IEEE transactions on cybernetics* 46, 12 (2015), 3364–3376.

[9] Nancy J Cooke, Jamie C Gorman, Christopher W Myers, and Jasmine L Duran. 2013. Interactive team cognition. *Cognitive science* 37, 2 (2013), 255–285.

[10] Nancy J Cooke and William F Lawless. 2021. Effective human–artificial intelligence teaming. *Systems Engineering and Artificial Intelligence* (2021), 61–75.

[11] Dana Milanovich Costar and Kendall K Hall. 2020. Improving team performance and patient safety on the job through team training and performance support tools: a systematic review. *Journal of patient safety* 16, 3 (2020), S48–S56.

[12] Chongwu Dong, Yanbin Sun, Muhammad Shafiq, Ning Hu, Yuan Liu, and Zhihong Tian. 2023. Optimizing Mobility-Aware Task Offloading in Smart Healthcare for Internet of Medical Things Through Multi-Agent Reinforcement Learning. *IEEE Internet of Things Journal* (2023).

[13] MR Endsley. 2000. *Situation awareness analysis and measurement*. Lawrence Erlbaum Associates.

[14] Mica R Endsley. 2021. Situation awareness. *Handbook of human factors and ergonomics* (2021), 434–455.

[15] David Fridovich-Keil, Ellis Ratner, Lasse Peters, Anca D Dragan, and Claire J Tomlin. 2020. Efficient iterative linear-quadratic approximations for nonlinear multi-player general-sum differential games. In *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 1475–1481.

[16] Dennis Granåsen. 2019. Towards automated assessment of team performance by mimicking expert observers' ratings. *Cognition, Technology & Work* 21, 2 (2019), 253–274.

[17] David AP Grimm, Jamie C Gorman, Nancy J Cooke, Mustafa Demir, and Nathan J McNeese. 2023. Dynamical measurement of team resilience. *Journal of Cognitive Engineering and Decision Making* 17, 4 (2023), 351–382.

[18] Miguel Grinberg. 2018. *Flask web development*. O'Reilly Media, Inc.

[19] Barbara J Grosz, Luke Hunsberger, and Sarit Kraus. 1999. Planning and acting together. *AI magazine* 20, 4 (1999), 23–23.

[20] Barbara J Grosz and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence* 86, 2 (1996), 269–357.

[21] J Richard Hackman and Ruth Wageman. 2005. A theory of team coaching. *Academy of management review* 30, 2 (2005), 269–287.

[22] Ryan Harari, Roger D Dias, Eduardo Salas, Vaibhav Unhelkar, Theodora Chaspari, and Marco Zenati. 2024. Misalignment of Cognitive Processes within Cardiac Surgery Teams. In *The Hamlyn Symposium on Medical Robotics: proceedings*, Vol. 16. NIH Public Access, 33.

[23] Rayan Ebnali Harari, Roger D Dias, Lauren R Kennedy-Metz, Giovanna Varni, Matthew Gombolay, Steven Yule, Eduardo Salas, and Marco A Zenati. 2024. Deep Learning Analysis of Surgical Video Recordings to Assess Nontechnical Skills. *JAMA Network Open* 7, 7 (2024), e2422520–e2422520.

[24] Maaike Harbers, MB van Riemsdijk, and CM Jonker. 2012. Measuring sharedness of mental models and its relation to team performance. In *Proceedings 14th International Workshop on Coordination, Organisations, Institutions and Norms*. 106–120.

[25] Laura M Hiatt, Cody Narber, Esube Bekele, Sangeet S Khemlani, and J Gregory Trafton. 2017. Human modeling for human–robot collaboration. *The International Journal of Robotics Research* 36, 5-7 (2017), 580–596.

[26] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.

[27] Robert R Hoffman, Matthew Johnson, Jeffrey M Bradshaw, and Al Underbrink. 2013. Trust in automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88.

[28] Catholijn M Jonker, M Birna Van Riemsdijk, and Bas Vermeulen. 2010. Shared mental models: A conceptual analysis. In *International workshop on coordination, organizations, institutions, and norms in agent systems*. Springer, 132–151.

[29] Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American economic review* 93, 5 (2003), 1449–1475.

[30] Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 99–127.

[31] Ece Kamar, Ya'akov Gal, and Barbara J Grosz. 2009. Incorporating helpful behavior into collaborative planning. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Springer Verlag.

[32] Lauren R Kennedy-Metz, Heather M Conboy, Anna Liu, Roger D Dias, Rayan E Harari, Ajami Gikandi, Alexander Shapeton, Lori A Clarke, Leon J Osterweil, George S Avrunin, et al. 2024. A novel multimodal, intraoperative cognitive workload assessment of cardiac surgery team members. *The Journal of Thoracic and Cardiovascular Surgery* (2024).

[33] Joseph Kim and Julie A Shah. 2016. Improving team's consistency of understanding in meetings. *IEEE Transactions on Human-Machine Systems* 46, 5 (2016), 625–637.

[34] Candace K Kolander. 2019. Flight and cabin crew teamwork: improving safety in aviation. In *Crew resource management*. Elsevier, 407–420.

[35] Igor Kotlyar, Tina Sharifi, and Lisa Fiksenbaum. 2023. Assessing Teamwork Skills: Can a Computer Algorithm Match Human Experts? *International Journal of Artificial Intelligence in Education* 33, 4 (2023), 955–991.

[36] Hyun-Rok Lee and Taesik Lee. 2021. Multi-agent reinforcement learning algorithm to solve a partially-observable multi-agent problem in disaster response. *European Journal of Operational Research* 291, 1 (2021), 296–308.

[37] Miao Liu, Kavinayan Sivakumar, Shayegan Omidshafiei, Christopher Amato, and Jonathan P How. 2017. Learning for multi-robot cooperation in partially observable stochastic environments with macro-actions. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1853–1860.

[38] Lanssie Mingyue Ma, Martijn Ijtsma, Karen M Feigh, and Amy R Pritchett. 2022. Metrics for human-robot team design: A teamwork perspective on evaluation of human-robot teams. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 3 (2022), 1–36.

[39] Martin A Makary and Michael Daniel. 2016. Medical error—the third leading cause of death in the US. *Bmj* 353 (2016).

[40] John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. 2000. The influence of shared mental models on team process and performance. *Journal of applied psychology* 85, 2 (2000), 273.

[41] Sara McComb and Vicki Simpson. 2014. The concept of shared mental models in healthcare collaboration. *Journal of advanced nursing* 70, 7 (2014), 1479–1488.

[42] Francisco Meneses. 2024. *Perceptions and Preferences Towards AI and Human Coaching in Fitness*. Master's thesis. Universidade Catolica Portuguesa (Portugal).

[43] Pragnesh Jay Modi, Wei-Min Shen, Milind Tambe, and Makoto Yokoo. 2005. ADOPT: Asynchronous distributed constraint optimization with quality guarantees. *Artificial Intelligence* 161, 1-2 (2005), 149–180.

[44] John E Morrison. 1999. *Foundations of the after action review process*. Vol. 42. United States Army Research Institute for the Behavioral and Social Sciences.

[45] Debasmita Mukherjee, Kashish Gupta, Li Hsin Chang, and Homayoun Najjaran. 2022. A survey of robot learning strategies for human-robot collaboration in industrial settings. *Robotics and Computer-Integrated Manufacturing* 73 (2022), 102231.

[46] Ranjit Nair and Milind Tambe. 2005. Hybrid BDI-POMDP framework for multiagent teaming. *Journal of Artificial Intelligence Research* 23 (2005), 367–420.

[47] Ranjit Nair, Milind Tambe, Makoto Yokoo, David Pynadath, and Stacy Marsella. 2003. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *IJCAI*, Vol. 3. 705–711.

[48] Catherine Neubauer, Kristin E Schaefer, Ashley H Oiknine, Steven Thurman, Benjamin Files, Stephen Gordon, J Cortney Bradford, Derek Spangler, and Gregory Gremillion. 2020. Multimodal Physiological and Behavioral Measures to Estimate Human States and Decisions for Improved Human Autonomy Teaming. *CCDC Army Research Laboratory Aberdeen Proving Ground United States* (2020).

[49] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. 2017. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research* 36, 5-7 (2017), 618–634.

[50] Adam Norton, Henny Admoni, Jacob Crandall, Tesca Fitzgerald, Alvika Gautam, Michael Goodrich, Amy Saretsky, Matthias Scheutz, Reid Simmons, Aaron Steinfeld, et al. 2022. Metrics for robot proficiency self-assessment and communication of proficiency in human-robot teams. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 3 (2022), 1–38.

[51] Frans A Oliehoek and Christopher Amato. 2016. *A concise introduction to decentralized POMDPs*. Vol. 1. Springer.

[52] Liubove Orlov Savko, Zhiqin Qian, Gregory Gremillion, Catherine Neubauer, Jonroy Canady, Vaibhav Unhelkar, and Catherine Neubauer. 2024. RW4T Dataset: Data of Human-Robot Behavior and Cognitive States in Simulated Disaster Response Tasks. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 924–928.

[53] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of behavioral and experimental finance* 17 (2018), 22–27.

[54] Avi Parush, Chelsea Kramer, Tara Foster-Hunt, Kathryn Momtahan, Aren Hunter, and Benjamin Sohmer. 2011. Communication and team situation awareness in the OR: Implications for augmentative information display. *Journal of biomedical informatics* 44, 3 (2011), 477–485.

[55] Jacqueline Peters and Catherine Carr. 2013. Team effectiveness and team coaching literature review. *Coaching: An International Journal of Theory, Research and Practice* 6, 2 (2013), 116–136.

[56] David PYNADATH, Nikolas GURNEY, Sarah KENNY, Rajay KUMAR, Stacy MARSELLA, Haley MATUSZAK, Hala MOSTAFA, Pedro SEQUEIRA, Vokan USTUN, and WU Peggy. 2023. Improving Teamwork through a Decision-Theoretic Coach in a Minecraft Search-and-Rescue Game. In *International Conference on Computers in Education*.

[57] David V Pynadath, Nikolos Gurney, Sarah Kenny, Rajay Kumar, Stacy C Marsella, Haley Matuszak, Hala Mostafa, Pedro Sequeira, Volkan Ustun, and Peggy Wu. 2023. Effectiveness of Teamwork-Level Interventions through Decision-Theoretic Reasoning in a Minecraft Search-and-Rescue Task. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 2334–2336.

[58] Peizhu Qian, Harrison Huang, and Vaibhav Unhelkar. 2024. PPS: Personalized Policy Summarization for Explaining Sequential Behavior of Autonomous Agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 1167–1179.

[59] Peizhu Qian and Vaibhav Unhelkar. 2022. Evaluating the role of interactivity on improving transparency in autonomous agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 1083–1091.

[60] Zhiqin Qian, Liubove Orlov Savko, Catherine Neubauer, Gregory Gremillion, and Vaibhav Unhelkar. 2024. Measuring Variations in Workload during Human-Robot Collaboration through Automated After-Action Reviews. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 852–856.

[61] Carlos Quintero-Pena, Peizhu Qian, Nicole M Fontenot, Hsin-Mei Chen, Shannan K Hamlin, Lydia E Kavraki, and Vaibhav Unhelkar. 2023. Robotic Tutors for Nurse Training: Opportunities for HRI Researchers. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 220–225.

[62] Alen Rajšp and Iztok Fister Jr. 2020. A systematic literature review of intelligent data analysis methods for smart sport training. *Applied Sciences* 10, 9 (2020), 3013.

[63] Patrick Riley and Manuela Veloso. 2004. Advice generation from observed execution: Abstract Markov decision process learning. In *Proceedings of the 19th national conference on Artifical intelligence*. 631–636.

[64] Patrick Riley, Manuela Veloso, and Gal Kaminka. 2002. An empirical study of coaching. In *Distributed Autonomous Robotic Systems 5*. Springer, 215–224.

[65] Eduardo Salas, Nancy J Cooke, and Michael A Rosen. 2008. On teams, teamwork, and team performance: Discoveries and developments. *Human factors* 50, 3 (2008), 540–547.

[66] Eduardo Salas, Deborah DiazGranados, Cameron Klein, C Shawn Burke, Kevin C Stagl, Gerald F Goodwin, and Stanley M Halpin. 2008. Does team training improve team performance? A meta-analysis. *Human factors* 50, 6 (2008), 903–933.

[67] Eduardo Salas, Rylee Linhardt, and Gabriela Fernández Castillo. 2024. The Science (and Practice) of Teamwork: A Commentary on Forty Years of Progress... *Small Group Research* (2024), 10464964241274119.

[68] Eduardo Salas, Denise L Reyes, and Susan H McDaniel. 2018. The science of teamwork: Progress, reflections, and the road ahead. *American Psychologist* 73, 4 (2018), 593.

[69] Francesco Semeraro, Alexander Griffiths, and Angelo Cangelosi. 2023. Human–robot collaboration and machine learning: A systematic review of recent research. *Robotics and Computer-Integrated Manufacturing* 79 (2023), 102432.

[70] Elham Semsar-Kazerooni and Khashayar Khorasani. 2009. Multi-agent team cooperation: A game theory approach. *Automatica* 45, 10 (2009), 2205–2213.

[71] Sangwon Seo. 2024. AI-Assisted Human Teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23415–23416.

[72] Sangwon Seo, Bing Han, and Vaibhav Unhelkar. 2023. Automated Task-Time Interventions to Improve Teamwork using Imitation Learning. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*. 335–344.

[73] Sangwon Seo, Lauren R Kennedy-Metz, Marco A Zenati, Julie A Shah, Roger D Dias, and Vaibhav V Unhelkar. 2021. Towards an AI coach to infer team mental model alignment in healthcare. In *2021 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 39–44.

[74] Sangwon Seo and Vaibhav Unhelkar. 2024. IDIL: Imitation Learning of Intent-Driven Expert Behavior. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 1673–1682.

[75] Sangwon Seo and Vaibhav Unhelkar. 2025. Hierarchical Imitation Learning of Team Behavior from Heterogeneous Demonstrations. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*.

[76] Sangwon Seo and Vaibhav V. Unhelkar. 2022. Semi-Supervised Imitation Learning of Team Policies from Suboptimal Demonstrations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. International Joint Conferences on Artificial Intelligence Organization, 2492–2500. https://doi.org/10.24963/ijcai.2022/346

[77] Herbert Alexander Simon. 1997. *Models of bounded rationality: Empirically grounded economic reason*. Vol. 3. MIT press.

[78] Neville A Stanton, Paul M Salmon, Guy H Walker, Eduardo Salas, and Peter A Hancock. 2017. State-of-science: situation awareness in individuals, teams and systems. *Ergonomics* 60, 4 (2017), 449–466.

[79] Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 24. 1504–1509.

[80] Michael Taberski, Kristi Davis, Kristin E Schaefer, and Ralph Brewer. 2021. Visualizing Human-Autonomy Team Dynamics through the Development of a Global After-Action Review Technology. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 46–53.

[81] Milind Tambe, Emma Bowring, Hyuckchul Jung, Gal Kaminka, Rajiv Maheswaran, Janusz Marecki, Pragnesh Jay Modi, Ranjit Nair, Stephen Okamoto, Jonathan P Pearce, et al. 2005. Conflicts in teamwork: Hybrids to the rescue. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. 3–10.

[82] Scott Tannenbaum and Eduardo Salas. 2020. *Teams that work: the seven drivers of team effectiveness*. Oxford University Press.

[83] Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. 2016. Computational human-robot interaction. *Foundations and Trends® in Robotics* 4, 2-3 (2016), 105–223.

[84] Ran Tian, Liting Sun, Andrea Bajcsy, Masayoshi Tomizuka, and Anca D Dragan. 2022. Safety assurances for human-robot interaction via confidence-aware game-theoretic human models. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 11229–11235.

[85] Karl Tuyls, Shayegan Omidshafiei, Paul Muller, Zhe Wang, Jerome Connor, Daniel Hennes, Ian Graham, William Spearman, Tim Waskett, Dafydd Steel, et al. 2021. Game Plan: What AI can do for Football, and What Football can do for AI. *Journal of Artificial Intelligence Research* 71 (2021), 41–88.

[86] Vaibhav Unhelkar and Julie Shah. 2016. Contact: Deciding to communicate during time-critical collaborative tasks in unknown, deterministic domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[87] Vaibhav V Unhelkar, Shen Li, and Julie A Shah. 2020. Decision-making for bidirectional communication in sequential human-robot collaborative tasks. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 329–341.

[88] Vaibhav V Unhelkar and Julie A Shah. 2019. Learning models of sequential decision-making with partial specification of agent behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2522–2530.

[89] Piet Van den Bossche, Wim Gijselaers, Mien Segers, Geert Woltjer, and Paul Kirschner. 2011. Team learning: building shared mental models. *Instructional science* 39 (2011), 283–301.

[90] Maaike Van Roy, Pieter Robberechts, Wen-Chi Yang, Luc De Raedt, and Jesse Davis. 2023. A Markov framework for learning and reasoning about strategies in professional soccer. *Journal of Artificial Intelligence Research* 77 (2023), 517–562.

[91] Joyce A Wahr, Richard L Prager, JH Abernathy Iii, Elizabeth A Martinez, Eduardo Salas, Patricia C Seifert, Robert C Groom, Bruce D Spiess, Bruce E Searles, Thoralf M Sundt III, et al. 2013. Patient safety in the cardiac operating room: human factors and teamwork: a scientific statement from the American Heart Association. *Circulation* 128, 10 (2013), 1139–1169.

[92] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. 2019. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In *Proceedings of the 27th ACM international conference on multimedia*. 374–382.

[93] Sallie J Weaver, Sydney M Dy, and Michael A Rosen. 2014. Team-training in healthcare: a narrative synthesis of the literature. *BMJ quality & safety* 23, 5 (2014), 359–372.

[94] Haochen Wu, Amin Ghadami, Alparslan Emrah Bayrak, Jonathon M Smereka, and Bogdan I Epureanu. 2022. Evaluating emergent coordination in multi-agent task allocation through causal inference and sub-team identification. *IEEE Robotics and Automation Letters* 8, 2 (2022), 728–735.

[95] Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, et al. 2025. SportU: A Comprehensive Sports Understanding Benchmark for Multimodal Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

[96] X Jessie Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 408–416.

[97] Yuening Zhang, Paul Robertson, Tianmin Shu, Sungkweon Hong, and Brian C Williams. 2024. Risk-Bounded Online Team Interventions via Theory of Mind. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 12964–12970.