A Framework for Intervention Based Team Support in Time Critical Tasks

Dana Hughes[1], Huao Li[2], Max Chis[2], Ini Oguntola[1], Simon Stepputtis[1],
Keyang Zheng[2], Joseph Campbell[1], Katia Sycara[1], Michael Lewis[2]

***Abstract*— In this paper we describe the intervention framework of ATLAS, an artificial socially intelligent agent that advises teams. The framework treats interventions as atomic components, and manages the lifecycle of each intervention through presentation, as well as followups to interventions. The key benefit of this framework is that it allows for rapid development of scenario-specific Interventions that leverage scenario-agnostic team models. The implementation of this framework is reported for three player teams in a Search and Rescue task simulated in Minecraft. Low competence teams advised by ATLAS improved more between first and second trials than those with a human advisor while the reverse was found for high competence. Four times as many interventions were proposed as were presented. 15% of advice was withheld to avoid repetitive advice, excessive rate of advice, and needlessly advising high performing teams, while a Theory of Mind model and delay for confirmation mechanism filtered out other unnecessary advice.**

I. Introduction

Human work often involves teams of individuals coordinating their efforts to complete a set of tasks. In such scenarios, the performance of the team depends not only on the ability of individual team members to perform specific taskwork, but also on the processes that team members use to facilitate work [1]. The recognition that team processes influence team performance has resulted in efforts to determine the utility of team training and team building interventions on various aspects of team performance. Training and intervention has positive influence on several aspects of team effectiveness, including affective aspects, cognitive measures, task-based skills, and team skills [2]. Importantly, training to improve *both taskwork and teamwork* have the most positive outcomes on improving team performance [3].

Despite these benefits, team training and intervention is expensive, and individuals providing team interventions, such as coaches or mission commanders, may not be available in certain scenarios, such as emergency response or disaster relief. Moreover, time criticality in various scenarios may prohibit or limit human ability to intervene. As such, developing agents to provide training and/or intervention is highly desirable [4]. While autonomous agents and systems have been developed to provide intelligent tutoring [5] or act as synthetic confederates in training simulation [6], developing autonomous agents to provide advice to ad-hoc and potentially inexperienced teams, such as may be needed for emergency or disaster response, remain largely unexplored.

Existing software assistants, such as Siri, Alexa, intelligent tutors, etc., aim to provide support for *individuals*. Focusing on *teams* introduces several unique challenges for development of an assistive agent. In a team setting, agents not only need to monitor each individual in the team, but also the interaction among teammates. This aspect requires an agent to not only determine *what* advice to provide, but to *whom* to provide the advice. Additionally, assistive agents operating in emergency or disaster response scenarios need to operate with impoverished information, and must rely on information as it is discovered. This assumption does not hold with existing software assistants (e.g., intelligent tutors have a library of problems and corresponding solutions to present). Finally, while many existing software agents operate in an informational space, where available data is machine-understandable, assistive agents in a team setting may not easily understand all available observation modalities, such as visual scenes or natural language discourse among team members, which presents additional challenges.

The primary purpose of this paper is to describe the team intervention/advice aspects of ATLAS (Assisting Teamwork via Learning and Advising System) [7], an agent designed to monitor team behavior and provide advice to improve or correct deficiencies in observed team processes. We detail the lifecycle of our intervention framework, and highlight its desirable characteristics. Using data collected as part of an ongoing DARPA program[1], we provide an initial analysis of the intervention capabilities of ATLAS in a simulated Urban Search and Rescue (USAR) task. We demonstrate that interventions generated by ATLAS have a beneficial influence on team performance, especially those with low initial task competency; and compare teams advised by ATLAS with those advised by a human.

II. Related Work

Intelligent agents designed to interact with humans have matured to the point that distinct categories have emerged.

*This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0036. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). This work was also supported by AFOSR/AFRL Award FA9550-18-1-0251.

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA {danahugh,ioguntol,sstepput,jacampbe,sycara}@andrew.cmu.edu

[2]School of Computing and Information, University of Pittsburgh, Pittsburgh, PA 15260, USA {mac372,huao.li,kez20,cmlewis}@pitt.edu

[1]https://artificialsocialintelligence.org/

Intelligent tutoring systems (ITS) are systems aimed at improving a human learner's knowledge of a domain by maintaining a model of various student factors—such as knowledge, learning styles, etc.—through repeated and multiple interaction with the student; identifying knowledge deficiencies; and providing hints, feedback, or recommendations to compensate for the identified knowledge deficiency [5]. Intelligent tutors maintain accurate models of student knowledge and skills by operating within a closed domain.

Recommender systems [8], takes advantage of similarity of likes and dislikes among individuals (collaborative filtering) and user profiles (content matching) to model user preferences in a shallower but less constrained way. By refining preference models, recommenders can capture subtle and complex relations among recommendations through their reflection in user responses without addressing them directly.

Intelligent assistants such as Apple's Siri, Microsoft's Cortana, and Amazon's Alexa have become ubiquitous largely through offering a hands-free speech interface replacing the GUI. These descendants of DARPA's Personalized Assistant that Learns (PAL) program [9] learn user preferences, as do recommender systems, but assistants also learn to act in response to explicit user commands, e.g., play music.

Creating agents capable of effective social interactions with individuals and groups is substantially more difficult. Socially intelligent agents are expected to possess a theory of mind [10] allowing them to reason about others' beliefs and intentions, an understanding of individual differences, and recognition and participation in social exchanges. Despite these difficulties we demonstrated more than twenty years ago that interventions targeting team communications and processes may be more beneficial than targeting individual tasks alone [11].

Teamwork is defined as a set of interrelated reasoning, actions and behaviors of team members that combine to fulfill team objectives [12]. The ability to adapt is believed to lie at the heart of team effectiveness [13]. Because of the complex interdependence and dynamics within teams, members need to account for their teammates' behaviors in order to be effective. Teamwork theory and experiments have identified a set of states and processes contributing to team effectiveness [14], [15] where cognitive processes and states comprise constructs, such as shared mental models, macrocognition, situation awareness, transactive memory while affective team processes comprise team cohesion and intragroup conflict [16], [17] which according to [14] are all grounded in the basic capabilities of communication and coordination. A recent compilation of these processes [18] served as a guide for targeted interventions by ATLAS.

A. Challenges

While ATLAS shares characteristics with earlier systems, challenging factors not previously addressed together remain.

- **Multiple Actors.** Unlike systems designed to interact with individuals, ATLAS needs to interact not only with individual team members but the *team as a whole* which entails that it has to *understand the interaction between team members*, which considerably complicates inferring intent or mental state of individuals and team state.
- **Lack of Privileged Knowledge.** In the search and rescue task described, both team members and ATLAS are given a minimal amount of prior knowledge about the environment. The implication of this is that ATLAS not only needs to maintain and update its own knowledge of the environment, but also infer the knowledge of the team members, and predict their behavior *under uncertainty* in order to provide best advice possible.
- **On-Line Critical Timing of Interventions.** ATLAS is required to give intervention advice on line and at *at the right time*. This presents challenges in computational efficiency and reasoning algorithms for detection of team suboptimal behavior, advice calculation, and advice time determination as the players rapidly navigate.
- **Lack of Authority.** In contrast to intelligent assistants that obey user commands, ATLAS's advice aims to change user cognition and behavior via its advice. However, unlike game coaches ATLAS has no authority over user behavior, and so must persuade the user to accept its advice e.g., through phrasing or providing convincing reasoning and explanations, which present additional challenges in their own right.
- **Lack of Predefined Workflow(s).** In contrast to intelligent assistants, which can leverage predefined workflows in response to user requests, ATLAS cannot utilize predefined workflows, since search is by definition nondeterministic and, in addition the search and rescue domain, as most real world domains, is dynamic.

III. Intervention Framework

While monitoring team behavior, ATLAS performs several tasks, including maintaining a model of the beliefs about the environment and predicting intended behavior for all participants [10], estimating player knowledge from team discourse, and managing team interventions and presenting advice. This paper focuses on describing and assessing the intervention framework of ATLAS.

A. Intervention Lifecycle

For the purpose of this paper, *Interventions* denote the core components used by ATLAS for monitoring individual and team behavior, and *Followups* denote related components used by ATLAS to monitor the effect of presented advice or other interactions with the team by ATLAS based on an Intervention. Interventions are self-contained specifying triggers, targeted team processes, and Followups. In the present study Interventions were presented in text through chat. While monitoring a team, ATLAS maintains the state of multiple Intervention instances, and updates the state of each instance based on observed events. Intervention categories are defined by the behavior they aim to influence. Finally, Interventions are designed to be anticipatory, and require that certain observed behaviors be satisfied prior to ATLAS interacting with the team.

Interventions and Followups are designed to be atomic, in that the evolution of one intervention instance does not influence the evolution of another. This structuring enables the rapid development of scenario-specific Intervention categories in a common framework that leverages scenario-agnostic team monitoring and performance analysis. Additionally, this structuring allows individual Interventions to be developed independently and rapidly, and allows ATLAS to be easily be tailored to perform specific Interventions / attend to specific deficiencies in the team.

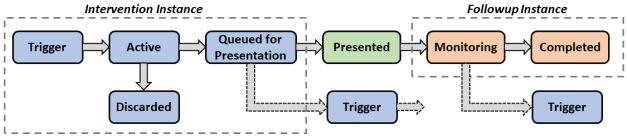

Fig. 1. Lifecycle of interventions and followups

Figure 1 summarizes the lifecycle of Intervention instances; each stage of this lifecycle is described below:

- **Triggered.** An Intervention instance is *spawned* when one of a pre-defined set of triggering events is observed by ATLAS. As Interventions are designed to be anticipatory, the trigger for an Intervention is assumed to be stateless (i.e., any required state should be encapsulated by the spawned Intervention instance itself). Intuitively, triggers anticipate and initialize potential interventions pro-actively; however, not every triggered intervention will lead to advice being presented.
- **Active.** Once spawned, the Intervention instance actively monitors individual and team behaviors. Once a sufficient amount of behavior is observed, the Intervention is either *Queued for Presentation*, if providing advice is deemed suitable for the observed behaviors, or *Discarded*, if providing advice is no longer relevant.
- **Queued for Presentation.** Once an Intervention is queued for presentation, monitors indicate to ATLAS that advice should be presented. When queued for presentation, the Intervention will generate an advisory message to present, and determine which participants on the team the message should be presented to. Additionally, in this state, the Intervention may spawn further interventions, in cases where the players' actions may necessitate further Interventions.
- **Discarded.** If an Intervention is deemed no longer relevant, it is discarded. ATLAS maintains the collection of discarded Interventions, as discarded Interventions may be relevant for determining if advice should be withheld or modified, when encountering similar triggering events in the future.

When an Intervention is queued for presentation, ATLAS determines whether the advice generated by the Intervention should be presented or withheld, based on the current state of the team or their performance, the history of participant reactions to advice, and/or the recent rate of presented advice. Furthermore, ATLAS may opt to reroute advisory messages to different participants (e.g., the advice may be routed to a team leader, who would presumably inform the team) and/or modify the contents of the message (e.g., providing additional explanation of the provided advice).

If an Intervention is presented by ATLAS, one or more *Followup* instances may be created to monitor individual and/or team response to the provided advice. As with Interventions, a Followup will observe behaviors until relevant conditions are met. A Followup instance may spawn further Interventions if relevant conditions are met. The primary class of Followups involve monitoring for *compliance* to provided advice, by an individual and/or the team as a whole. Individuals can be deemed *compliant* (i.e., they follow the advice), *non-compliant* (i.e., they do not follow the advice), or *non-applicable* (i.e., compliance to the advice becomes irrelevant). Team compliance is determined from the compliance of each advised team member.

B. Characteristics

The framework described above enables several desirable characteristics, as described below.

1) Anticipatory Intervention: A key characteristic is that the triggering of an Intervention implies that ATLAS may issue advice in the *future* to address behavior associated with the triggering event. In contrast, a *reactive* Intervention would present advice in immediate response to an observed event. Treating Interventions as anticipatory enables ATLAS to continue monitoring individual and/or team behavior once triggered—this provides an opportunity for ATLAS to determine the applicability of potential advice, or the team to self-correct deficient behavior before advice is presented.

As an example, when ATLAS observes a player entering a room, it may trigger multiple interventions based on *possible* undesirable behaviors—e.g., the team member may attempt a risky task without asking for assistance from teammates, they may deviate from a pre-established protocol, and/or they may fail to provide critical updates to teammates. These active Interventions may become irrelevant based on the team member performing desired behaviors, or changes to the mission context. The Intervention lifecycle explicitly captures these conditions, allowing ATLAS to both allow the team to self-correct behavior, as well as providing statistics on the rate that each of the Interventions are discarded or result in presented advice for each player.

2) Explainability: While *Active*, each Intervention monitors the players and environment for the occurrence of specific events. The Intervention can store the sequence of relevant events. If advice is presented, an explanation of *why* the advice was presented can be constructed from the stored sequence of events. This can include, for instance, specific actions that the player took, or environment context that led to the advice being presented.

3) Assessment of Presented Advice: Followups are designed as part of the intervention lifecycle to provide ATLAS with feedback on player or team responses to presented advice. This stage of the lifecycle can be used to provide feedback to ATLAS on the influence an Intervention has on

Fig. 2. Example image of the USAR scenario used for evaluation.

team behavior, and can be leveraged to adapt future advice. For example, a Followup may observe a player to see if they comply with a given piece of advice within a certain timeframe; if the player repeatedly fails to comply with the advice from the same type of Intervention, ATLAS may opt to modify or withhold future advice due to its ineffectiveness on altering the target player's behavior.

IV. Evaluation

A. Scenario

Our agent was evaluated in a simulated Urban Search and Rescue (USAR) environment[1] developed using Minecraft[2], as shown in Figure 2. In this scenario, teams of three humans were tasked with searching a collapsed building for victims, and stabilizing and evacuating victims [19]. Each victim had a single injury type—abrasion (A), bone damage (B), or critical (C). Points were awarded for each victim successfully stabilized and evacuated to one of two zones corresponding to their injury type. Teams were awarded 50 points for evacuating stabilized critical victims, and 10 points each for evacuating stabilized victims with the other injury types. In each trial, 15 critical victims and 20 victims with abrasion or bone damage were present in the building for each trial, for a maximum score of 950 points. Multiple "threat rooms" are present in the building; when a player enters a threat room, rubble collapses in front of the entry to the room, trapping the player inside. Finally, a perturbation event occurs during each mission, which either removes communication ability among the team (including the dynamic map described below) or blocks off an evacuation zone with rubble.

Each participant was assigned fixed roles, namely a Medic, an Engineer, and a Transporter. Medics are able to identify the injury type of each victim, and can stabilize victims with abrasion or bone damage injuries alone. Stabilizing a critical victim requires another teammate to be present. Engineers can remove rubble, and are provided with knowledge of the locations of threat rooms at the beginning of each trial. Transporters have a signal tool allowing for the detection of the presence of victims in rooms and injury severity (critical or non-critical) of the victim without entering the room. Each role navigates at a different speed: Transporters move 20% faster than Medics, and Engineers move 15% slower than Medics. Each participant can pick up, carry, and place stabilized victims, and can only carry a single victim at a time; while carrying victims, movement speed is reduced.

Each player can arbitrarily place and remove marker blocks as the environment is being searched. Marker blocks indicate injury types of non-critical victims, the presence of critical, non-critical or no victims in a room, indication of threat rooms, location of rubble, and blocks for requesting help. Each player has a dynamic 2-D map of the building prior to collapse, which provides players with their location in the building (but not the location of their team mates), and the location and type of marker blocks in the environment. Players can communicate verbally throughout the mission.

Teams completed a pair of two 17-minute missions, and received advice from one of six artificial advisor agents (including ATLAS), a human advisor, or no advisor for both trials. Prior to the start of the missions, each player completed a hands-on training mission, consisting of training for role-specific subtasks, training for team communication and coordination, and a competency test on the role-specific subtasks. At the start of each mission, each player receives unique information regarding meetings where potential victim may be located: Medics are provided with meeting location, Transporters are provided with the number of attendees for each meeting, and Engineers are provided with the severity of rooms where meetings took place. During the first two minutes of each mission, teams are instructed to discuss the mission, and have the potential for integrating unique provided knowledge—denoted as the Knowledge Integration Task (KIT)—to gain insight into the number of victims in some rooms. Teams perform the search and rescue task during remaining 15 minutes of each mission.

B. Interventions

For the described scenario, we developed Interventions aimed at addressing deficiencies in and improving various team processes, as defined in [1], [18]. Specifically, we developed Interventions targeting the following first-order constructs: Mission Analysis, Goal Specification, Strategy Formulation and Planning, Monitoring Progress towards Goals, System Monitoring, Team Monitoring and Backup Responses, Coordination, Motivation and Confidence Building. Table I provides a list of developed interventions, including descriptions and team processes targeted by each. Evolution of interventions were based on observed behaviors of individual players, such as victim stabilization and evacuation, marker placement, etc. ATLAS did not utilize the verbal communication between players, as a well-trained NLP system was unavailable for understanding utterances in the USAR domain. Additional details are available in [7]. As training on both *taskwork* and *teamwork* has a positive influence on team performance [3], we developed Interventions with both types of content.

1) Presentation: For the USAR scenario, ATLAS withholds advice for the reasons below. Threshold levels were determined empirically through small pilot trials.

1) Advice was withheld from the team if ATLAS predicted the team performance, as measured by final

[2]https://www.minecraft.net

TABLE I

DESCRIPTION OF INTERVENTIONS DEVELOPED FOR THE DESCRIBED USAR SCENARIO. INTERVENTION IDENTIFIERS PROVDED IN ITALICS.

Targeted Team Process	**Description of Illustrative Interventions**	**RI-role independent, RD-role dependent, TS-team specific**
System Monitoring	*C8 Passage.* ATLAS suggests player share presence of hard to see passage through room C8 when discovered.	
	Evacuation Zone Distance. ATLAS alerts player to presence of closer evacuation zone if they appear to travel to further zone.	RI
	Remind Rubble Perturbation. If a player discovers an evacuation zone is blocked due to a perturbation event, ATLAS suggests that the player communicate this with the rest of the team (if not Engineer) or remove the rubble (if Engineer).	RI
Coordination	*Encourage Proximity to Medic.* ATLAS suggests to a player to remain the medic's proximity, to assist with critical victims.	RI,TS
	Evacuate Critical Victims. ATLAS encourages the team to evacuate stabilized high-value critical victims, if they appear to ignore these victims.	RI
	Inform About Stabilized Victims. Alert the Engineer or Transporter of nearby stabilized victims that may be evacuated.	RI,RD
	Stabilize Critical Victims. ATLAS encourages the Medic to stabilize high-value critical victims, if they appear to ignore these victims.	RD
Strategy Formulation and Planning	*Encourage or Discourage Proximity.* ATLAS suggests that the team either group together (to improve taskwork) or spread out (to improve searching), based on context.	RI
	Synchronize Stabilization and Evacuation. ATLAS suggests a strategy to synchronize evacuation of victims with stabilization.	TS
	Transporter Initial Strategy. ATLAS suggests to the Transporter to initially mark rooms with victim contents based on signal tool, to maximize team knowledge.	RD
Monitoring Progress towards Goals	*Clean Markers.* ATLAS suggests that one player should remove stale marker blocks, to reduce cognitive burden on the team.	RI
	Remind Transporter of Signal. ATLAS suggests to the Transporter to mark the outside of rooms based on victim presence indicated by the signal tool.	RD
	Start Evacuation. ATLAS suggests to the team to begin evacuation of victims if there is no benefit to continuing search or stabilization.	TS
	Remind to Place Victim Marker. ATLAS reminds the Transporter or Engineer to place a marker block if they see a victim and fail to place one.	RD
	Time Elapsed. ATLAS reminds the team when three minutes remain in the trial.	RI
	KIT Recommendation. ATLAS provides suggestions to discuss high-value rooms which the team did not discuss during the KIT.	TS
	Team Encouragement. ATLAS provides a message to high-performing teams as encouragement.	RI,TS
Team Monitoring and Backup Responses	*Remind Change Marker.* ATLAS reminds the Medic to change marker blocks to reflect victim stabilization if they fail to do so.	RD
	Remind Medic to Inform About Stabilized Victim. If the medic fails to inform the team about stabilized victims ATLAS reminds the Medic to share this information.	RD
	Remind to Place Stabilized Victim Marker. ATLAS reminds the Medic to place a marker indicating victim type if they fail to do so after stabilizing.	RD
	Correct Victim Misidentification. ATLAS alerts the Medic to correct victims that they misidentified through incorrect marker block placement.	RD
Mission Analysis	*Team Welcome Message.* Introduces ATLAS and provides a brief explanation of its expected behavior.	RI
	Reminder To Share KIT Information. If a player does not participate in the KIT, ATLAS reminds the player to discuss their unique information.	RI,TS

score, to be over a threshold of 780;

2) Advice for a given class of Intervention was limited to a maximum number of presentations during a mission;
3) Advice was withheld from an individual if the rate of presentation exceeded two messages per minute;
4) Advice was withheld from the team until a minimum of two minutes of time has elapsed during the search and rescue phase of the mission, in order to collect sufficient data to predict team performance

C. Dataset

In this paper, we evaluate ATLAS using data collected in [20], publicly available at [21]. This data consists of 15 trials of two missions (denoted Mission 1 and Mission 2 by sequential order) each using ATLAS as an artificial advising agent, as well as 14 trials using a human advisor. Unique teams were used for each trial. We omit the no advisor condition from our analysis, as participants in this condition were significantly more proficient in video game playing, and were more experience in Minecraft, than the participants advised by ATLAS or the human [20]. As described in [22], more proficient participants in the no advisor condition would be able to more effectively learn the experimental task than those advised by ATLAS or the human, which would mask the influence of the effect of advisor interventions.

In addition to assessing team performance based on final score, each player was asked two pairs of questions to rate their perceived utility of the advisor on team performance and coordination (denoted "Utility") and the trustworthiness of the advisor (denoted "Trust") on a 7-point Likert scale; the questions asked are provided in Table II and ratings in Table V. Finally, each player's competency was calculated as the time to complete the pre-mission competency test.

TABLE II

SURVEY QUESTIONS ASKED TO ASSESS PERCEPTION OF ADVISOR.

Identifier	Question Text
Utility-1	The advisor's recommendations improved our team score.
Utility-2	The advisor's recommendations improved our team coordination.
Trust-1	I felt comfortable depending on the ASIST agent.
Trust-2	I understand why the advisor made its recommendations.

D. Results

To assess the influence of our agent interventions, we consider the final team performance (i.e., final score), rate of interventions presented in each mission, the distribution of intervention lifecycle state, and the subjective player assessment of perceived utility and trustworthiness of the agent advice. Where applicable, we compare teams advised by ATLAS to those advised by a human advisor.

1) Team Performance: Teams advised by ATLAS had an average final score of 534.7 after the Mission 1, and 616.0 after Mission 2. These compare well with human-advised teams, which achieved average scores of 542.7 and 620.0 for Missions 1 and 2, respectively.

We also considered the influence of advisors based on the initial team competence. Teams were separated into two categories based on the average time for participants to complete the competency mission: Teams with an average competency mission time of below the median (i.e. 100 seconds) were labeled as "High Competency," while others were labeled as "Low Competency."

Figure 3 shows the average improvement of team performance between Missions 1 and 2 based on advisor and team competence. Two-way ANOVA shows a significant interaction effect between advisor type and team competency, $F(1, 80) = 5.87, p = .018, \eta^2 = 0.07$. To reveal the effect of ATLAS on teams with different competence levels, we conducted pairwise t-tests. Low Competency teams advised by ATLAS showed medium improvement than Low Competency teams advised by the human ($t(38.9) = 2.56, p = 0.014, \eta^2 = 0.12$): 31.2% improvement for teams advised by ATLAS vs. 17.4% for teams advised by the human. Results show that ATLAS has a strong effect at improving Low Competency teams compared to High Competency teams ($t(28.3) = 2.94, p = 0.06, \eta^2 = 0.19$): 31.2% improvement for Low Competency teams vs. 6.7% improvement for High Competency teams. High Competency teams advised by the human had an improvement of 18.2%, however, this improvement is not significantly different from improvements of High Competency teams advised by ATLAS or Low Competency teams advised by the human ($ps > .05$). In summary, ATLAS was more effective at improving performance of Low Competency teams than High Competency teams. And for Low Competency teams, ATLAS is more helpful compared to human advisors.

2) Intervention Influence Across Missions: The average number of times each Intervention category was presented

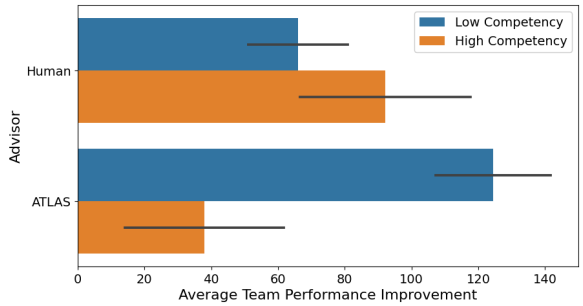

Fig. 3. Comparison between human advisor and ATLAS influence on Team Performance. Teams with low competency demonstrated greater improvement when advised by ATLAS, while teams with high competency demonstrated greater improvement when advised by the human advisor. Error bars indicate standard error.

(presentation rate) in each trial is summarized in Table III. Of the 23 Interventions categories provided, the presentation rate for relevant Interventions was lower in the second mission for 12 of the intervention categories, and higher for 6 intervention categories. We omit the "Remind Rubble Perturbation" and "Team Encouragement" Interventions, as the rubble perturbation was present only for Mission 1, and team encouragement was not intended to modify a specific behavior, respectively. The total number of interventions presented in Missions 1 and 2 for all 15 trials (30 missions total) was 314 and 292, respectively, corresponding to a decrease of 7% in intervention rate.

TABLE III

AVERAGE INTERVENTION PRESENTATION RATES, FOR EACH MISSION. STANDARD DEVIATION IS GIVEN IN PARENTHESES.

Intervention Identifier	Mission 1	Mission 2
C8 Passage	**1.00 (1.03)**	1.53 (1.09)
Encourage Proximity to Medic	2.87 (1.59)	**2.40 (1.62)**
Encourage or Discourage Proximity	0.07 (0.25)	**0.00 (—)**
Evacuate Critical Victims	0.87 (0.34)	**0.80 (0.34)**
Evacuation Zone Distance	0.80 (0.54)	**0.60 (0.61)**
Clean Markers	**0.00 (—)**	0.13 (0.34)
Inform About Stabilized Victims	**1.33 (1.19)**	1.53 (0.72)
KIT Recommendation	0.07 (0.25)	**0.00 (—)**
Remind Change Marker	1.80 (1.47)	**1.20 (0.98)**
Remind Medic to Inform About Stabilized Victim	0.47 (0.81)	**0.20 (0.54)**
Reminder To Share KIT Information	0.13 (0.50)	0.13 (0.50)
Remind Rubble Perturbation	1.13 (0.72)	**N/A**
Remind Transporter of Signal	**1.87 (1.15)**	2.13 (1.45)
Start Evacuation	0.93 (0.25)	**0.87 (0.34)**
Synchronize Stabilization and Evacuation	0.27 (0.44)	**0.13 (0.34)**
Team Encouragement	**0.00 (—)**	0.13 (0.34)
Remind to Place Victim Marker	3.20 (0.54)	**2.80 (0.75)**
Team Welcome Message	1.00 (0.00)	1.00 (0.00)
Time Elapsed	1.00 (0.00)	1.00 (0.00)
Transporter Initial Strategy	0.80 (0.40)	**0.73 (0.44)**
Stabilize Critical Victims	0.87 (0.34)	**0.60 (0.49)**
Remind to Place Stabilized Victim Marker	**0.47 (0.50)**	0.60 (0.71)
Correct Victim Misidentification	**0.53 (0.88)**	0.67 (0.87)

To determine whether teams learn from the advice pro-

vided by ATLAS, we calculate the change in average number of times each intervention was presented during each mission, as summarized in Table III. A decrease in intervention presentation rate (bold values in the table indicate which Mission had lower presentation rates) implies improvement in team processes from interventions provided by ATLAS in Mission 1, however, none of these decreases were found to be significant ($p > 0.05$).

3) Intervention Lifecycle State: For each mission, we calculated the number of Interventions that were in each terminal state (*Queued for Presentation* and *Discarded*) at the end of each mission, and the number of Interventions whose presentation was withheld. Table IV summarizes the percentages of these states.

TABLE IV

FINAL STATE OF INTERVENTIONS TRIGGERED BY ATLAS.

Intervention State	Mean (Std-Dev.)	Proportion
Presented	20.7 (3.8)	7.4%
Discarded	215.0 (53.4)	77.2%
Withheld—Total	42.9 (11.5)	15.4%

The table demonstrates that a very large percentages of interventions were discarded. An ablation study showed that 65% of this reduction is attributable to beliefs maintained by the ToM model. Another consequence of the anticipatory nature of the developed interventions is that rather than intervening immediately, ATLAS waits to observe if the player will take the desired action unprompted before intervening.

4) Player Assessment: Players' subjective evaluation of advisor utility and trustworthiness, as defined in Table II, is summarized in Table V. As can be seen from the table, the human advisor scored higher on these measures ($F(1, 80) = 31.8, p < .001; \eta^2 = 0.28$ for Trust; $F(1, 80) = 34.8, p < .001; \eta^2 = 0.29$ for Utility). A possible reason for this is that human advisors were able to (1) better understand what players were saying to one another, (2) verbally communicate advice to participants, and (3) answer questions or engage is dialog with players, capabilities that ATLAS did not possess.

In addition, players assessment is also influenced by the initial team competency. Figure 4 shows the average player assessment of advisor Utility and Trust, for High Competency and Low Competency teams. Although not significant, Human advisors receive higher trust and utility ratings from High Competency teams compared to Low Competency teams, while ATLAS receive higher subjective ratings from Low Competency teams compared to High Competency teams. These results align with above findings in team performance that ATLAS is more effective for participants with less experience.

V. DISCUSSION

The results described in Section IV-D highlight the benefit to our proposed framework, and illustrate how the characteristics of the framework address many of the challenges described in the introduction. The developed Intervention lifecycle and framework helps to generalize ATLAS to novel

TABLE V

SUBJECTIVE ASSESSMENT MEASURES IN EACH MISSION BY ADVISOR. STANDARD DEVIATIONS IN PARENTHESES.

Measure	ATLAS		Human Advisor	
	Mission 1	Mission 2	Mission 1	Mission 2
Utility-1	4.7 (1.3)	4.7 (1.8)	6.1 (1.0)	6.2 (0.9)
Utility-2	4.5 (1.6)	4.8 (1.4)	6.1 (1.1)	6.1 (1.0)
Trust-1	4.3 (1.6)	4.5 (1.8)	5.8 (1.2)	6.1 (1.2)
Trust-2	5.8 (1.3)	5.8 (1.4)	6.6 (0.6)	6.6 (0.7)

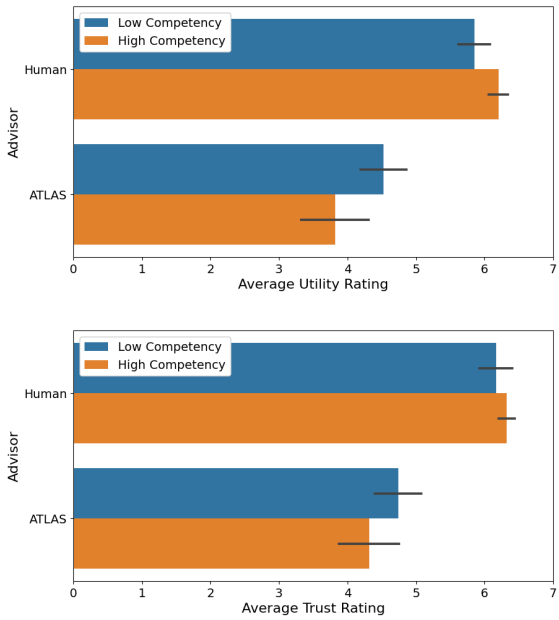

Fig. 4. Average Utility and Trust rating of ATLAS and the human advisor, separated by team competency. Error bars indicate standard error.

scenarios. Interventions listed were rapidly developed and integrated into the framework for the USAR scenario described. ATLAS can be readily adapted to novel scenarios by simply reusing these existing Interventions, where suitable, and defining additional scenario-specific Interventions.

The ToM model and anticipatory nature of developed interventions played a significant role in the behavior of ATLAS. Of the interventions triggered, roughly 77% were discarded due high confidence in predicted beliefs or prior occurrence of behaviors that the Interventions were intended to address. Had the Interventions been purely reactive, then players would have been given roughly four times the amount of advice—an undesirable scenario, as the quantity of interventions might be distracting or significantly increase the cognitive burden players have attending to the advice. Moreover, giving advice when the team would have performed correctly would probably cause decreased assessment of utility and trust in ATLAS, and hence reduced compliance. Finally, this percentage of discarded interventions highlights a fundamental difference between human advisors and ATLAS: ATLAS can attend to a much larger range of potential behavioral deficiencies players exhibit than a human advisor, due to cognitive limits of the human. ATLAS may,

for instance, more quickly notice repeated deficiencies and provide correcting advice quicker than a human.

When comparing ATLAS with the human advisor, the competency of the team was a determining factor in which advisor was more suitable, with ATLAS showing greater impact on Low Competency teams. This observation points to a significant role ATLAS can play during training—ATLAS can be used as an advisor to train teams to a given competency level, after which more costly human advisors can be used to further advise teams once that level is achieved. Additionally, in the case where human advisors are not available, ATLAS can still effectively advise High Competency teams to increase team performance, though to a lesser extent than the human advisor would provide.

VI. Conclusion and Future Work

We presented a framework used by ATLAS for managing team interventions. The framework treats interventions as atomic components, and manages the lifecycle of each intervention through presentation, as well as followups to interventions once presented. The key benefit of this framework is that it allows for rapid development of scenario-specific Interventions that leverage scenario-agnostic team models.

We demonstrated the effectiveness of our intervention framework in a simulated USAR task. Specifically, we show that Low Competency teams advised by ATLAS had higher improvement in performance across the two missions than teams advised by a human. For a majority of the designed Interventions, ATLAS advised teams less on the Intervention in the second mission, implying that teams learn from the provided advise. Finally, we show that measures of team performance are comparable between ATLAS and a human advisor. Subjective assessment of ATLAS's utility and trustworthiness were positive, despite lacking the interaction modalities that the human advisor possessed.

The described framework offers several avenues for further analysis and development. Further analysis of individual interventions, such as the relative impact of each, would assist in determining which interventions to present more prominently. Modeling the response of individual interventions over time, using time-series analysis [23] or sequence-to-sequence [24] speech models, would enable the influence of interventions to be quantified over time, potentially providing an additional mechanism for determining when an intervention is relevant. Finally, we aim to demonstrate the generalizability of the framework to novel domains, such as bomb disposal scenarios.

References

[1] M. A. Marks, J. E. Mathieu, and S. J. Zaccaro, "A temporally based framework and taxonomy of team processes," Academy of management review, vol. 26, no. 3, pp. 356–376, 2001.

[2] L. A. Delise, C. Allen Gorman, A. M. Brooks, J. R. Rentsch, and D. Steele-Johnson, "The effects of team training on team outcomes: A meta-analysis," Performance Improvement Quarterly, vol. 22, no. 4, pp. 53–80, 2010.

[3] E. Salas, D. DiazGranados, C. Klein, C. S. Burke, K. C. Stagl, G. F. Goodwin, and S. M. Halpin, "Does team training improve team performance? a meta-analysis," Human factors, vol. 50, no. 6, pp. 903–933, 2008.

[4] J. Yin, M. S. Miller, T. R. Ioerger, J. Yen, and R. A. Volz, "A knowledge-based approach for designing intelligent team training systems," in Proceedings of the fourth international conference on Autonomous agents, 2000, pp. 427–434.

[5] E. Mousavinasab, N. Zarifsanaiey, S. R. Niakan Kalhori, M. Rakhshan, L. Keikha, and M. Ghazi Saeedi, "Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods," Interactive Learning Environments, vol. 29, no. 1, pp. 142–163, 2021.

[6] C. Myers, J. Ball, N. Cooke, M. Freiman, M. Caisse, S. Rodgers, M. Demir, and N. McNeese, "Autonomous intelligent agents for team training," IEEE Intelligent Systems, vol. 34, no. 2, pp. 3–14, 2018.

[7] K. Sycara, M. Lewis, and D. Hughes, "Cmu-ri study 3 preregistration," https://osf.io/yj52e, May 2022, [Online; accessed 10-April-2023].

[8] J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, and B. Gipp, "A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation," in Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, ser. RepSys '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 7–14. [Online]. Available: https://doi.org/10.1145/2532508.2532511

[9] N. G. Canbek and M. E. Mutlu, "On the track of artificial intelligence: Learning with intelligent personal assistants," Journal of Human Sciences, vol. 13, no. 1, pp. 592–601, 2016.

[10] I. Oguntola, D. Hughes, and K. Sycara, "Deep interpretable models of theory of mind," in 2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN), 2021, pp. 657–664.

[11] K. Sycara and M. Lewis, "Integrating intelligent agents into human teams," in Team cognition: understanding the factors that drive process and performance, E. Salas and S. Fiore, Eds. Washington, DC: American Psychological Association, 2004, pp. 203–231.

[12] B. B. Morgan, A. S. Glickman, E. A. Woodard, A. S. Blaiwes, E. Salas, W. J. Campbell, D. L. Miller, R. C. Montero, and S. Zimmer, Measurement of team behaviors in a Navy training environment. Old Dominion University Research Foundation, 1986.

[13] C. S. Burke, K. C. Stagl, E. Salas, L. Pierce, and D. Kendall, "Understanding team adaptation: A conceptual analysis and model." Journal of Applied Psychology, vol. 91, no. 6, p. 1189, 2006.

[14] E. Salas, D. E. Sims, and C. S. Burke, "Is there a "big five" in teamwork?" Small group research, vol. 36, no. 5, pp. 555–599, 2005.

[15] E. Salas, N. J. Cooke, and M. A. Rosen, "On teams, teamwork, and team performance: Discoveries and developments," Human factors, vol. 50, no. 3, pp. 540–547, 2008.

[16] N. Cooke, J. Gorman, W. JL, and F. Durso, "Team cognition," Handbook of Applied Cognition, vol. 2, no. 1, pp. 239–268, 2007.

[17] N. Cooke, J. Gorman, C. Myers, and J. Duran, "Interactive team cognition," Cognitive Science, vol. 37, no. 1, pp. 255–285, 2013.

[18] J. E. Mathieu, M. M. Luciano, L. D'Innocenzo, E. A. Klock, and J. A. LePine, "The development and construct validity of a team processes survey measure," Organizational Research Methods, vol. 23, no. 3, pp. 399–431, 2020.

[19] L. Huang, J. Freeman, N. Cooke, J. Colonna-Romano, M. D. Wood, V. Buchanan, and S. J. Caufman, "Exercises for artificial social intelligence in minecraft search and rescue teams," https://osf.io/jwyvf, April 2022, [Online; accessed 06-March-2023].

[20] J. Freeman, L. Huang, M. Demir, L. Markham, N. Cooke, Z. Klinefelter, A. Fouse, M. Hidalgo, M. Wood, C. Corral, X. Yin, M. Cohen, J. C. Clark, V. Buchanan, M. Willett, X. He, A. Chethikattil, J. Lee, J. Perry, K. Bronzi, and R. Gundala, "Asist study 3 evaluator results," https://osf.io/djeuf, January 2023, [Online; accessed 06-March-2023].

[21] L. Huang, J. Freeman, N. Cooke, J. olonna-Romano, M. Wood, V. Buchanan, and S. Caufman, "Artificial Social Intelligence for Successful Teams (ASIST) Study 3," 2022. [Online]. Available: https://doi.org/10.48349/ASU/QDQ4MH

[22] K. A. Orvis, D. B. Horn, and J. Belanich, "The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames," Computers in Human behavior, vol. 24, no. 5, pp. 2415–2433, 2008.

[23] G. Kirchgässner, J. Wolters, and U. Hassler, Introduction to modern time series analysis. Springer Science & Business Media, 2012.

[24] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina et al., "State-of-the-art speech recognition with sequence-to-sequence models," in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018, pp. 4774–4778.