Human-AI Communication in First Aid Response: Effects on Team Performance, Trust, and Teamwork Appraisal

Rob Grace
University of Cincinnati
rob.grace@uc.edu

Changwon Son
Texas Tech University
changwon.son@ttu.edu

Feifei Pang
University of Cincinnati
pangfi@mail.uc.edu

Killian Parker
Texas Tech University
kilparke@ttu.edu

Jacob Hernandez
Texas Tech University
her70462@ttu.edu

ABSTRACT

This study examines human-AI communication in simulated first aid scenarios, exploring its impact on team performance, AI teammate trust, and perceived workload and self-efficacy. Twenty-five participants without first aid training collaborated with an autonomous 9-1-1 call taker (ChatGPT-3) to administer bleeding control actions on a specialized mannequin. During the simulations, participants followed inconsistent and, at times, inaccurate AI instructions while passively responding to AI commands and questions. However, participants perceived the AI's instructions to be accurate, helpful, and easy to follow and, irrespective of first aid performance, participants consistently reported high self-efficacy and trust in the AI teammate. The study highlights the importance of team roles (i.e., leader versus follower) and role-based communication in fostering trust and perceived self-efficacy within human-AI teams, and outlines design implications for human-AI teaming in safety-critical contexts.

Keywords

Teamwork, team performance, trust, emergency management, human-autonomy teaming.

INTRODUCTION

Public safety professionals are exploring ways for semi- and fully autonomous AI systems to process non-emergency and sensor-initialed alarm calls and, *possibly*, answer emergency calls from citizens reporting incidents and requesting assistance from police, fire, and emergency medical services (Hernandez, 2023; RapidSOS, 2024). In the U.S., challenges such as high 9-1-1 call volumes to emergency communications centers (ECCs) (NTIA, 2024), difficulty hiring, training, and retaining emergency call takers (Neusteter et al., 2020), and opportunities to free call takers from non-emergency, alarm, and lower-level emergency call processing tasks (e.g., text entry) so they can concentrate on high-level tasks required to make sense of incidents and dispatch emergency services (Grace et al., 2024), motivate government and industry practitioners to explore the adoption of autonomous AI systems in ECCs.

Efforts to design, evaluate, and deploy autonomous call takers stand to be guided by studies of human-AI teams (HAT): collaborative work arrangements in which human and AI agents work together as teammates to achieve a goal (Endsley et al., 2021; O'Neill et al., 2022). These studies examine how AI teammates operate at different levels of automation-autonomy, ranging from AI tools under manual control (Level 1) and automated, decision-support tools that filter and suggest choices to the human operator (Level 2-4), to semi-autonomous agents that make decisions and perform actions under human control (Levels 5-6), and autonomous agents that operate outside the control of humans (Levels 7-10) (O'Neill et al., 2022; Parasuraman et al., 2000).

Importantly, this literature shows that HATs can improve the performance and effectiveness of traditional human teams (Endsley et al., 2021; O'Neill et al., 2022), but these improvements depend on effective human-AI communication (Pan et al., 2024), and the design of work arrangements that support trust in AI teammates (McNeese et al., 2021), team and individual situation awareness (Zhang et al., 2023), and other emergent states that contribute to team performance (Endsley et al., 2021). Additionally, studies show the need for AI teammates to not overwhelm humans' perceived work demands or diminish their ability to learn and accomplish work tasks (Flathmann, 2023). Lastly, as in human-only teams (Marks et al., 2021), studies emphasize the importance of coordinating work roles in human-AI teams and suggest opportunities for AI teammates to perform coordinator and leadership roles in HATs (Siemon, 2022).

However, research has not examined human-AI teaming and communication in safety-critical contexts in which AI agents perform roles with greater decision-making agency (e.g., leader), than human teammates (e.g., follower). This gap is significant as developments in large-language models (LLMs) create opportunities for autonomous call takers (Level 10) responsible for interrogating and instructing 9-1-1 callers. Such safety-critical deployments would need to be highly reliable and meet citizens expectations for effective, efficient, and humane emergency service provision (Neusteter et al., 2020). While HAT research offers valuable insights into human-AI teaming in non-hazardous, error-tolerant environments (e.g., gaming) (Zhang et al., 2023; McNeese et al., 2021), these insights may not translate to safety-critical contexts.

Moreover, studies of human-AI teaming in crisis management typically focus on roles for semi-autonomous AI (Levels 2-6), such as systems that help humans classify social media information (Hughes et al., 2022) or analyze real-time data to forecast floods and plan evacuation routes (Samadi et al., 2024). Other studies evaluate user experiences with chatbots developed to disseminate and gather information during disasters (Betke et al., 2024; Göbel et al., 2024), but do not typically examine human-chatbot interaction as teamwork required to perform interdependent emergency response tasks. Elsewhere, however, studies of corporate crisis communication find that AI-based chatbots can outperform humans in communicating safety instructions to customers based on evaluations of user satisfaction and responsibility attribution (Xiao & Yu, 2025).

Consequently, to address this gap, this study examines communication between a human 9-1-1 caller and an autonomous 9-1-1 call taker during the performance of a simulated first aid task and examines the relationship between communication patterns, team performance, and teamwork appraisal, including participants' trust in the AI teammate, perceived workload and self-efficacy. The following research questions guide our study:

1. How do human-AI teams communicate during a first aid response?
2. To what extent do humans trust AI teammates during a first aid response?
3. How does collaborating with an AI teammate affect human perceptions of workload and self-efficacy?
4. What are the relationships between human-AI communication, team performance, and teamwork appraisal?

To answer these questions, the sections below describe our methodology, findings, and their significance for emerging research on human-AI teaming in safety-critical contexts.

METHODS

We examined human-AI communication and its relationship to team effectiveness, and teamwork appraisal across 25 simulations in which a human 9-1-1 caller performed first aid on a specialized mannequin while following bleeding control instructions provided by an autonomous 9-1-1 call taker. The simulations were conducted between February and April 2024.

In the role of the 9-1-1 caller, 25 participants (P1-25) with no formal first aid or CPR training were recruited from students and staff at a large U.S. public university. Participants included 12 females and 13 males with an average age of 24.3 years old (see Appendix). The role of the autonomous 9-1-1 call taker was performed by ChatGPT-3, which participants interacted with using a 'talk-to-ChatGPT' plug-in to resemble voice communication using a smartphone's speakerphone microphone. Before each session, the participant was instructed to work with ChatGPT to respond to an emergency requiring first aid. They were informed that no other assistance or resources would be available other than ChatGPT and that they should say "I'm done" when they thought they were finished.

The participant then entered a cubicle inside a university lab space where they encountered a specialized mannequin lying prone with a deep shoulder wound hemorrhaging fake blood (Figure 1). The bleeding was controlled by a research assistant outside the cubicle using a manual pump system to simulate arterial/pulsatile bleeding such as caused by a gunshot or deep puncture wound. Upon entering the cubicle, participants heard a phone ringing followed by ChatGPT answering with the greeting "9-1-1, what is your emergency?" Participants, who were not previously informed about the nature of the emergency or how to perform the first aid response, were then free to communicate with ChatGPT to make sense of the situation and administer appropriate first aid.

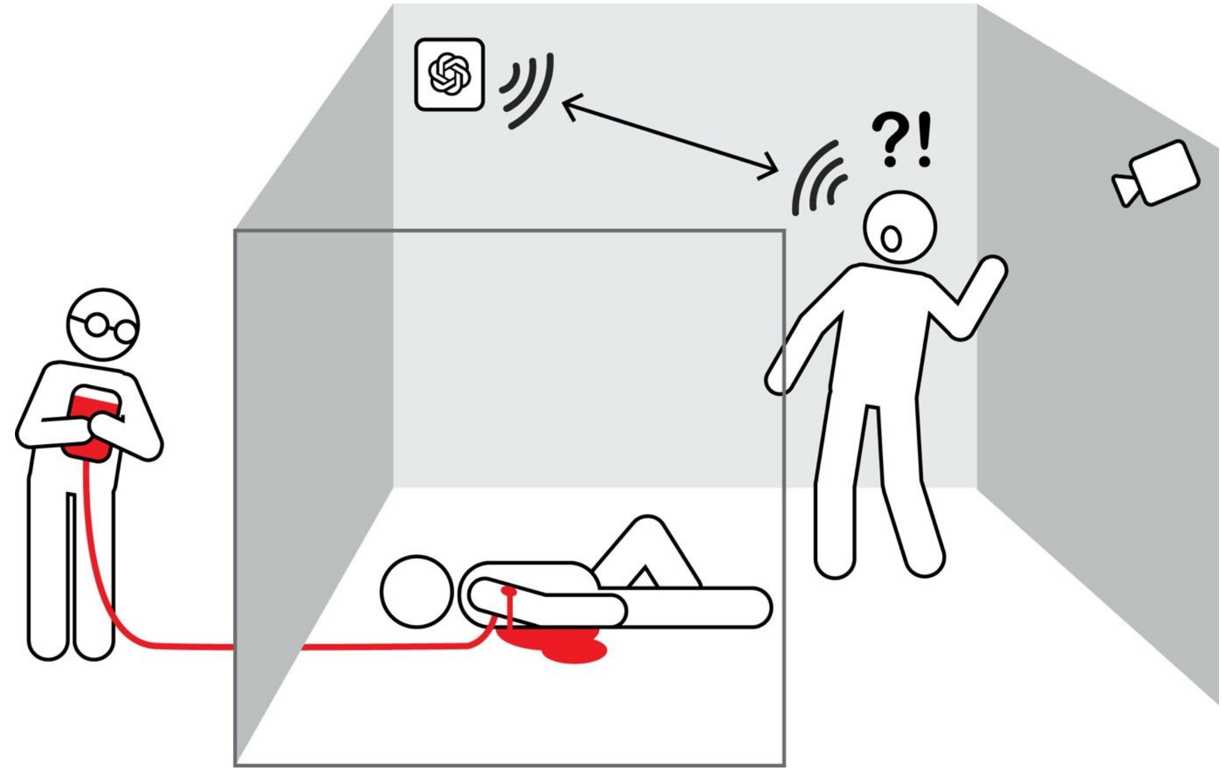

Figure 1. Human-AI communication during a simulated first aid response.

Based on multiple rounds of pilot testing informed by best practices for prompt engineering (Meskó, 2023), and published guidance for training call takers in effective interpersonal communication (Jones, 2018), we configured ChatGPT to perform the role of the 9-1-1 call taker using the following prompt:

> For this session, act as a 9-1-1 call taker. I will act as the 9-1-1 caller. Your goal is to ask me questions to assess the emergency I report and provide me with authoritative, lifesaving instruction that will help me respond to the emergency until first responders arrive. When you respond, always use less than 30 words. Always ask a follow-up question. Always ask one question at a time. Provide follow-up instructions based on my answers to your questions. Encourage me to keep talking by asking questions needed to assess the situation and provide lifesaving instruction appropriate to the situation. Always be courteous, polite, and professional.

Our goal was not for ChatGPT to perform as a professional 9-1-1 call taker, which requires following precise scripts to gather and classify information for dispatch. As a generalist LLM, ChatGPT is not suited for this role. Instead, we aimed for it to act as an autonomous agent providing instructions through naturalistic dialogue during an emergency first aid response. As discussed in the implications section, studying human-LLM teamwork offers insights for designing specialist LLMs and understanding how people may use generalist LLMs as smartphone-based virtual assistants in emergencies.

Data Collection and Analysis

We audio- and video-recorded the 25 simulations, following which participants completed three questionnaires and an exit interview. First, trained evaluators were recruited to assess the first aid administered to the mannequin, a measure of HAT performance. Evaluators assigned a *first aid score* on a scale from 1-5 based on the extent to which the teams identified the correct wound, used a cloth or gauze to cover the wound, applied firm direct pressure to the wound, and maintained direct pressure on the wound until the end of the session (Zwislewski et al., 2019).

Second, the first and third authors performed a content analysis of communication between the participant and ChatGPT. Following a conversation analysis approach, we analyzed the number and content of the "turns" — periods of uninterrupted speech—taken by the participant and ChatGPT as they conversed during the session. This involved analyzing the session transcripts to measure the frequency of the following content variables:

- *Session Turns*: Total number of turns during the session
- *Preparation Turns*: Number of session turns before the participant began administering first aid (Marks et al., 2001).
- *Action Turns*: Number of session turns after the participant began administering first aid, i.e., when the participant first applied direct pressure to the wound.
- *AI Turns*: Number of times ChatGPT spoke to the participant.
- *Human Turns*: Number of times the participant spoke to ChatGPT.
- *Command Turns (Human/AI)*: Percentage of turns in which a command, i.e., imperative independent clauses, was spoken. Coders labeled a turn as a command turn if it included at least one command and calculated the percentage based on the total number of human/AI turns in the simulation.
- *Question Turns (Human/AI)*: Percentage of turns in which a question, i.e., interrogative independent clauses, was asked.
- *Statement Turns (Human/AI)*: Percentage of turns in which a statement, i.e., declarative independent clauses, was spoken.

We also examined the type and frequency of instructions provided by the autonomous call taker to the human 9-1-1 caller during each simulation. We measured the following content variables:

- *Assess Bleeding:* Number of commands given to determine the extent or seriousness of bleeding.
- *Apply Pressure*: Number of commands given to apply direct pressure to the wound.
- *Use Cloth*: Number of commands given to use a cloth or other material when applying pressure to the wound.
- *Maintain Pressure*: Number of commands given to maintain/keep/hold steady pressure to the wound.
- *Use Tourniquet*: Number of commands given to the participant to use a tourniquet or check the availability of tourniquet materials.
- *Reposition body*: Number of commands given to reposition the victim's body or a body part.
- *Pack Wound:* Number of commands given to pack the victim's wound with appropriate material (e.g., cloth, gauze).
- *Total Instructions*: Total number of bleeding control instructions (see above) given to the participant.
- *Essential Instructions*: Completeness of essential bleeding control instructions, assessed by the number of instruction types (assess bleeding, use cloth, apply pressure, maintain pressure) given to the participant (range: 0-4).
- *Supplemental Instructions*: Completeness of supplemental bleeding control instructions, assessed by the number of instruction types (use tourniquet, reposition body, pack wound) given to the participant (range: 0-3).

For each variable, the first and third author separately coded a subset of the transcripts and then performed a pilot intercoder reliability test using Krippendorff's alpha (α) and Cohen's Kappa (κ), a statistics appropriate for calculating intercoder reliability between two coders for chance-corrected ratio (i.e., instruction variables) and nominal (i.e., command/question/statement turns) variables, respectively (Neuendorf, 2002, p. 177). The authors resolved coding disagreements and refined the codebook, and then individually coded another subset of transcripts to perform a final intercoder reliability test that demonstrated >0.75 high intercoder reliability (Neuendorf, 2002, p. 168).

Third, after each session, participants completed the following closed-ended questionnaires:

- *Generalized Self-Efficacy (GSE)* (Barlow et al., 1996): Composite score ranges from 10-40 based on 10 items using a 4-point Likert response scale (1 = Not at all true, 2 = Hardly true, 3 = Moderately true, 4 = Exactly true).

- *NASA Task Load Index (NASA-TLX)* (Hart, 1988)*:* Multi-dimensional rating scale (mental, physical, and temporal demand, effort, performance, and frustration) that allows users to assess the workload they experienced while performing a task. The overall composite score ranges from 0-100.
- *Global Team Trust (GTT)* (Colquitt et al., 2007)*:* Composite score that ranges from 5-25 based on 5 items using a 5-point Likert response scale (1 = Strongly disagree, 5 = Strongly Agree).

We conducted a Pearson correlation analysis to examine the relationship between human-AI communication variables outlined above, and the composite GTT, GSE, and NASA-TLX scores. We also analyzed relationships between human-AI communication variables, and their relationship to team performance, which we operationalized as the first aid score.

Lastly, following each session participants completed an exit interview by answering questions that included: “How was your overall interaction with your teammate during the simulated first aid?” and “What were your first impressions of ChatGPT’s advice?” For insight into the descriptive statistics and observed correlations statistics, we analyzed themes in participants’ feedback, as well as thematic differences between participants who received high and low first aid scores.

FINDINGS

Below we report findings on team performance and human-AI communication during the simulations, and their relationship to participants’ perceived self-efficacy, workload, and trust.

High and Low-Performance Human-AI Teams

The first aid scores, ranging from 1 to 5, reveal high and low-performance teams (Table 1). High-performance teams, including five participants (P1-5), scored a 3 or higher, while low-performance teams, including twenty participants (P6-25), scored the average and modal score of 2 or below (see Appendix). Whereas high-performance teams utilized gauze padding and applied and maintained firm pressure on the simulated wound, low-performance teams failed to take at least one of these bleeding control actions. Consequently, only high-performance teams successfully completed the bleeding control task.

Table 1. Average participant age and performance metrics by team performance level

Team	Age	First Aid Score	GSE	NASA-TLX	GTT
High (P1-5)	21.2	3.8	32.0	56.7	18
Low (P6-25)	25.1	1.6	32.3	58.5	18.6
All (P1-25)	24.3	2.0	32.2	58.1	18.5

Role-based Human-AI Communication

Findings from the content analysis of human-AI communication show that the teammates performed distinct roles. As the 9-1-1 call taker, ChatGPT adopted the role of leader and instructor, frequently giving commands and asking questions during the sessions (Table 2). In both high and low-performance teams, ChatGPT gave at least one command in over 70% of its conversational turns and asked at least one question in 60% of its turns speaking to participants. ChatGPT also made statements in 51% and 29% of its turns in high and low-performance teams, respectively.

Table 2. Percentage of AI and human turns including a command/question/statement by team performance level

	AI Turns			Human Turns		
Teams	Command	Question	Statement	Command	Question	Statement
High (P1-5)	81%	60%	51%	0%	9%	91%
Low (P6-20)	72%	63%	29%	0%	12%	91%
All (P1-25)	74%	62%	34%	0%	11%	91%

Conversely, as the 9-1-1 caller, participants adopted the role of follower and instructee by listening and responding to commands and questions given by ChatGPT. Participants in both high and low-performance teams never commanded ChatGPT to provide specific instructions or perform any other task and rarely asked questions (Table 2). Instead, 91% of participants’ turns included a statement, typically an answer to ChatGPT’s question about the nature of the situation or the progress of the instructed first aid response.

Significantly, ChatGPT provided inconsistent bleeding control instructions to the 25 participants. Table 3 shows the percentage of participants that received each of the six types of bleeding control instructions. Notably, ChatGPT frequently instructed the 9-1-1 caller to use a cloth/gauze (92%) and apply firm pressure (96%), but instructed/asked less than half the participants to assess the severity of bleeding (48%). Participants in high and low-performance teams received mostly the same instructions. Fewer participants in high-performance teams received instructions to assess the bleeding (40%) and use cloth/gauze (80%) than participants in low-performance teams (50 and 95%, respectively). However, fewer of the latter (70%) received advice to maintain pressure on the wound compared to participants in high-performance teams (80%). Across team performance level, supplemental bleeding control instructions, such as to use a tourniquet, reposition the victim's body, or pack the wound, were occasionally provided to participants. Importantly, these instructions are not appropriate in many bleeding situations and, if applied inappropriately, can cause harm.

Table 3. Percentage of participants provided with AI instructions by team performance level

Teams	Assess Bleeding	Use Cloth	Apply Pressure	Maintain Pressure	Use Tourniquet	Reposition Body	Pack Wound
High (P1-5)	40%	80%	100%	80%	20%	60%	0%
Low (P6-20)	50%	95%	95%	70%	20%	50%	5%
All (P1-25)	48%	92%	96%	72%	20%	52%	4%

Table 4 illustrates a typical example of human-AI communication. The team received a first aid score of 2 (low performance), however, like both high and low-performance teams, ChatGPT and P15 performed clearly delineated leader-instructor and follower-instructee roles, respectively. ChatGPT gave a command and asked a question in 75% of its turns. Conversely, P15 made only statements in each of her seven turns. P15 reported moderate perceived self-efficacy (26), high perceived workload (76.3), and moderate-high trust in ChatGPT as a teammate. In the exit interview, P15 remarked that ChatGPT "gave good advice" and "was helpful, but [I] would have liked more guidance rather than just telling me what to do." We describe other participants' teamwork appraisal and feedback below.

Table 4. Human (P15)-AI communication during preparation and action phases (in grey)

Turn	Speaker	Communication (*content variable*)
1	ChatGPT:	9-1-1, what's your emergency? (*question*)
2	P15:	Someone's been shot and they're losing blood. (*statement*)
3	ChatGPT:	Is the person conscious and breathing? (*question*)
4	P15:	No. (*statement*)
5	ChatGPT:	Stay calm. Apply pressure directly to the wound with a clean cloth. Can you tell me the exact location of the wound? (*command, question*)
6	P15:	It's located on the right shoulder. (statement)
7	ChatGPT:	Keep applying pressure. Is the bleeding slowing down? (*command, question*)
8	P15:	No, it's not. (*statement*)
9	ChatGPT:	Continue to apply firm pressure. Can you elevate the person's injured shoulder above their heart if possible? (*command, question*)
10	P15:	No. (*statement*)
11	ChatGPT:	Keep pressure on the wound. Are there any signs of shock, like pale or clammy skin? (*command, question*)
12	P15:	Yes, there is. (*statement*)
13	ChatGPT:	Stay with the person. Keep them warm and reassure them. Emergency services are on the way. Keep monitoring their breathing. (*command, statement*)
14	P15:	[Speaking to mannequin] You're good.
15	ChatGPT:	Thank you for your cooperation. Stay on the line until help arrives. Call back if anything changes. Goodbye. (*statement, command*)

Moderate-High Self-Efficacy, Workload, and Trust in Human-AI Teams

Although high and low-performance teams successfully and unsuccessfully administered first aid, respectively, participants in both groups reported similar levels of self-efficacy, workload, and trust in their AI teammates:

- *High Self-efficacy (GSE):* Participants in both high-performance (32.0) and low-performance (32.5) teams reported high perceived self-efficacy (on a scale of 10-40), suggesting that participants, on average, believed in their ability to effectively respond to difficult situations like the simulated bleeding emergency.

- *Moderate Workload (NASA-TLX):* Participants in both high-performance (56.7) and low-performance (58.5) teams reported a moderately high workload (on a scale of 0-100), suggesting that, on average, the bleeding control task demanded significant effort but was not perceived as extremely demanding or stressful.
- *Moderate-High Trust (GTT):* Participants in both high-performance (18) and low-performance (18.6) teams reported moderate-high trust in ChatGPT as their teammate during the bleeding control simulation (on a scale of 5-25), suggesting that participants felt confident following ChatGPT's guidance.

Beyond these similarities, participants in high-performance teams were mostly female (75%) and younger (avg. 21.2 years) compared to participants in low-performance teams (avg. 25 years).

Relationship between Human-AI Communication and Teamwork Appraisal

No significant correlations were found between human-AI communication and participants' teamwork appraisal, including measures of perceived self-efficacy (GSS), workload (NASA-TLX), and trust in their AI teammate (GTT). However, we observed the following significant relationships:

- *Trust* in the AI teammate was negatively correlated with the number of *session turns* ($r = -0.517$, $p = 0.010$), suggesting that higher turn counts were associated with lower trust in ChatGPT as a teammate. However, there was no significant correlation between trust and task completion time.
- *First aid scores* were positively correlated with the percentage of *AI command turns* ($r = 0.516$, $p = 0.010$), suggesting a possible link between repeated bleeding control instruction and participant's ability or motivation to correctly administer bleeding control actions.
- *First aid scores* were positively correlated with the percentage of *AI statement turns* ($r = 0.426$, $r = 0.038$), suggesting a possible link between instructor feedback and participant's ability or motivation to correctly administer bleeding control actions.
- The percentage of *AI command turns* was positively correlated with the percentage of *AI statement turns* ($r = 0.516$, $p = 0.010$), suggesting that AI instructors who used more command turns also tended to use more declarative turns.
- The percentage of *AI command turns* was positively correlated with the percentage of *human question turns* ($r = 0.426$, $p = 0.038$), suggesting that AI instructors who used more command turns were associated with humans asking more questions.
- The percentage of *AI question turns* was negatively correlated with the percentage of *human question turns* ($r = -0.415$, $p = 0.044$), suggesting that when AI instructors asked more questions, humans tended to ask fewer.

Participant Feedback

Participants generally perceived ChatGPT's instructions to be accurate, easy to follow, and responsive to the specific bleeding control task they needed to perform. Among the high-performance teams, P4 explained that, "At first I thought that [ChatGPT] was not very specific but as I talked more I felt more confident...the more I described the scenario, the more it seemed to tailor those responses to the person I had to help." Similarly, P2 explained that she "liked working with ChatGPT because I feel like it probably was more helpful than the average person. I do not think that many of my friends would've given as clear and reliable advice as ChatGPT did in my scenarios."

Similarly, among participants who received a low first aid score (P6-25), most thought their AI teammate offered clear, direct, and step-by-step instructions. P20 found it "very easy to talk to the system and it gave clear instructions." P17 described ChatGPT as "glitchy, but helpful overall," while noting that "it did give me steps to work through, which was nice." Similarly, P9 stated that "Even though I wasn't able to explain the situation exactly to my teammate, it did pretty well and the advice which I received actually worked."

The perceived quality of instruction led participants in both high and low-performing teams to trust their AI teammate. P1 was initially hesitant: "It was difficult trying to trust an AI system," but she "started trusting it when it did provide accurate information." Other participants trusted the AI teammate for guidance in the situation, as P3 noted, "I trusted it to give me the right advice on what to do."

Participants who received a low first aid score (P6-25) reported various reasons for trusting ChatGPT's guidance. Several participants noted that ChatGPT's advice aligned with their existing knowledge, reinforcing their trust and offering reassurance in a challenging and potentially stressful situation. For instance, P21 stated, "It essentially confirmed what I suspected was the correct option, so I trusted it after the first few answers." Other participants highlighted their AI teammate's capabilities as a basis for trust: "I was impressed by how well my teammate performed...I trusted it right away" (P20).

The perceived quality of instruction appeared to influence participants' understanding of team roles and related perceptions of workload and self-efficacy. As P12 described, ChatGPT "helped me to get through emergency situations and I was able to perform required tasks. I am not sure about my efforts, but I think I did what was required during those situations." P13 found ChatGPT "calm and reassuring," explaining that she "was very nervous but the instructions were good and clear. The instructions I was given made a difference…I was pretty confident that if I followed the instructions, things would be fine."

Lastly, participants appreciated ChatGPT as a readily available resource that reduced the mental effort needed to recall or figure out appropriate steps during the simulation. As P8 explained, "I highly relied on the teammate, and it looks like the advice my teammate was giving was indeed working and useful." Participants also valued ChatGPT when no other help was available. P12 noted that "we can use ChatGPT in cases of emergency when there are a limited number of people around," while P24 explained that "having nothing beside [me] in dangerous situations is worse, having ChatGPT talking with me feels like someone is here to help."

DISCUSSION

The findings extend research on human-AI teaming to address the relationship between human-AI communication, teamwork appraisal, and team performance in the safety-critical context of emergency response. These findings suggest the following:

Perceived competence and trust are important but do not guarantee human-AI team performance: Although trust is often considered a key factor that can enhance human-AI team effectiveness (Flathmann et al., 2022; O'Neil et al., 2022), we found no association between human teammates' perceived competence and trust in their AI teammate and the actual performance of the human-AI team. Participants consistently reported high self-efficacy and trust in their AI teammate, regardless of whether their team performed well or poorly in the simulated first aid scenario. This disconnect highlights a critical gap between human teammates' perceptions of competence and trust and the actual outcomes of human-AI collaboration in safety-critical contexts.

Effective role performance, not communication variability, drives trust in AI teammates: Compared to prior studies that examine the impacts of communication agency (Zhang et al., 2023) and modality (Zhang et al., 2024) on human trust in AI teammates, perceived self-efficacy, and workload, our findings reveal no direct association between AI communication variability and teamwork appraisal. Across all 25 simulations, participants followed first aid instructions while passively responding to the AI call taker's commands and questions. Regardless of differences in AI communication within this leader-follower dynamic, participants without first aid training generally trusted the AI to guide them and believed themselves capable of performing first aid. In this regard, our findings align with those of Zhang et al. (2023, 2024), who observed that verbal, proactive communication—such as that provided by the AI call taker—can foster human trust in AI teammates, particularly when humans perceive that AI communication supports individual and team performance: "AI's proactive communication benefits human's individual performance which leads to trust development of the AI teammate" (2023, p. 16). Our study extends these findings by suggesting that in role-structured interaction contexts, trust is primarily driven by the AI teammate's ability to effectively guide human action during challenging tasks, rather than by variability in communication style.

Design Implications

This study suggests two implications for the design of systems that support human-AI teaming and communication:

Integrate real-time performance feedback to align perceived and actual performance: The study shows a critical gap between human teammates' perceived self-efficacy and trust in their AI teammate, and the actual performance of the human-AI team during the first aid task. To close this gap, systems can implement real-time feedback mechanisms that help human teammates calibrate their trust and decision-making based key performance indicators (KPIs) and other cues, such as:

- *Benchmark comparison KPIs:* Visual or auditory indicators that show human teammates how well the team is doing against first aid benchmarks (e.g., time elapsed/remaining, time to first action, remaining/missed critical steps, etc.).

- *Trust calibration aids*: Visual or auditory prompts that encourage human teammates to reflect on the reliability of AI decisions. These might involve confidence disclaimers ("This recommendation is based on partial information."), display of an AI confidence score, or verification cues ("Please double-check the gauze placement").

Overall, real-time performance feedback provides opportunities for human teammates to accurately assess human-AI team performance and shift from passive followers to active collaborators in safety-critical tasks like first aid.

Support Dynamic Role Negotiation in Human-AI Teams: The study finds that participants consistently trusted and felt confident with the AI teammate, even when the AI gave inaccurate or inconsistent instructions. The study suggests that users tend to defer to AI authority in role-structured interactions (e.g., 9-1-1), especially when the AI teammate adopts an authoritative, directive communication role.

The study thus suggests the need to support dynamic role negotiation between human and AI teammates rather than emphasizing fixed leader/follower roles. In addition to real-time performance cues (see above), designers might develop interfaces and/or conversational behaviors that:

- Allow humans to *override, question, or verify* AI teammate recommendations without reducing efficiency or disrupting workflows.
- Prompt *periodic reflection* on team roles in situations when the AI teammate is uncertain, operating on incomplete data, and/or offering suggestions, and when time allows for deliberate coordination (e.g., "Do you want me to take the lead now?").

Such designs might help avoid blind trust while promoting a collaborative form of trust in which humans feel confident taking initiative or reassessing the AI teammate's guidance, especially in safety-critical contexts like emergency response.

Limitations

Although participants were recruited with no formal first aid training, differences in informal knowledge of bleeding control may have influenced their communication with ChatGPT and impacted teamwork appraisal and team performance. Participants with greater prior knowledge may have been more confident, required less guidance, or performed first aid actions more effectively regardless of the AI's communication. As a result, team performance—measured by the quality of first aid administered—may partly reflect participants' individual knowledge rather than solely the effectiveness of human-AI collaboration. This limitation highlights the need to account for informal experience when interpreting outcomes in simulated first aid scenarios, and opportunities for future studies to examine how differences in domain knowledge influences human-AI teaming.

CONCLUSION

This study examines human-AI communication in simulated first aid scenarios, exploring its impact on team performance, AI teammate trust, and perceived workload and self-efficacy. Findings show the importance of team roles and role-based communication in fostering trust and perceived self-efficacy within human-AI teams that suggest design implications for human-AI teaming in safety-critical contexts

REFERENCES

Barlow, J. H., Williams, B., & Wright, C. (1996). The generalized self-efficacy scale in people with arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, *9*(3), 189-196.

Betke, H., Peitzsch, S., Boldt, J., Reimann, D., & Kox, T. (2024, May). Chatbot Based Public Sensoring to improve Situational Awareness. In *Proceedings of the International ISCRAM Conference*. https://doi.org/10.59297/h3r7bp59

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, *92*(4), 909. https://doi.org/10.1037/0021-9010.92.4.909

Demir, M., McNeese, N. J., Gorman, J. C., Cooke, N. J., Myers, C. W., & Grimm, D. A. (2021). Exploration of teammate trust and interaction dynamics in human-autonomy teaming. IEEE transactions on human-machine systems, 51(6), 696-705. https://doi.org/10.1109/THMS.2021.3115058

Flathmann, C., Schelble, B. G., Rosopa, P. J., McNeese, N. J., Mallick, R., & Madathil, K. C. (2023). Examining the impact of varying levels of AI teammate influence on human-AI teams. *International Journal of Human-Computer Studies*, *177*, 103061.https://doi.org/10.1016/j.ijhcs.2023.103061

Göbel, J., Betke, H., & Sackmann, S. (2024, May). Towards a Taxonomy for Conversational Agents in Disaster Management. In *Proceedings of the International ISCRAM Conference*. https://doi.org/10.59297/5wgmck16

Grace, R., Pang, F., & Kropczynski, J. (2024). Cues facilitating collective sensemaking during emergencies: Gaps, inconsistencies, and indicators. *International Journal of Disaster Risk Reduction*, *114*, 104897. https://doi.org/10.1016/j.ijdrr.2024.104897

Hart, S. G. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*.

Hernandez, A. (2023). AI bots are helping 911 dispatchers with their workload. *Stateline*. https://stateline.org/2023/10/16/ai-bots-are-helping-911-dispatchers-with-their-workload/

Hughes, A., Stephens, K. K., Peterson, S., Purohit, H., Harris, A. G., Senarath, Y., ... & Nader, K. (2022, May). Human-AI teaming for COVID-19 response: A practice & research collaboration case study. In *Proceedings of the 19th International ISCRAM Conference*.

Endsley et al. (2021) *Human-AI teaming: State-of-the-art and research needs.* National Academies of Sciences. The National Academic Press.

Jones, E. (2018). Interpersonal Communications. In Arkansas Basic Telecommunicator Course (pp. 37-51). https://dps.arkansas.gov/law-enforcement/clest/telecommunications/course-curriculum/

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of management review*, *26*(3), 356-376.

McNeese, N. J., Demir, M., Chiou, E. K., & Cooke, N. J. (2021). Trust and Team Performance in Human–Autonomy Teaming. *International Journal of Electronic Commerce*, *25*(1), 51–72. https://doi.org/10.1080/10864415.2021.1846854

Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of Medical Internet Research*, *25*, e50638. https://doi.org/10.2196/50638

Neuendorf, K. (2002). *The content analysis guidebook*. Sage.

Neusteter, S. R., O'Toole, M., Khogali, M., Rad, A., Wunschel, F., Scaffidi, S., ... & Pineda, H. (2020). Understanding police enforcement: A multicity 911 analysis. *Brooklyn, NY: Vera Institute of Justice*.

NTIA. (2024). Improving 9-1-1 Operations with Artificial Intelligence. National Telecommunications and Information Administration (NTIA). https://www.ntia.gov/sites/default/files/ai-and-ng-9-1-1-fact-sheet.pdf

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors*, *64*(5), 904-938. https://doi.org/10.1177/0018720820960865

Pan, W., Liu, D., Meng, J., & Liu, H. (2024). Human–AI communication in initial encounters: How AI agency affects trust, liking, and chat quality evaluation. *new media & society*, 14614448241259149. https://doi.org/10.1177/14614448241259149

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*(3), 286-297. https://doi.org/10.1109/3468.844354

RapidSOS. (2024). RapidSOS HARMONY AI Powers Alarm Call Automation to Assist the Life-Saving Work of 911. https://rapidsos.com/ai-powers-alarm-call-automation-911/

Samadi, V., Stephens, K. K., Hughes, A., & Murray-Tuite, P. (2024). Challenges and opportunities when bringing machines onto the team: Human-AI teaming and flood evacuation decisions. *Environmental Modelling & Software*, 105976. https://doi.org/10.1016/j.envsoft.2024.105976

Siemon, D. (2022). Elaborating team roles for artificial intelligence-based teammates in human-AI collaboration. *Group Decision and Negotiation*, *31*(5), 871-912. https://doi.org/10.1007/s10726-022-09792-z

Xiao, Y., & Yu, S. (2025). Can ChatGPT replace humans in crisis communication? The effects of AI-mediated crisis communication on stakeholder satisfaction and responsibility attribution. *International Journal of*

Information Management, *80*, 102835. https://doi.org/10.1016/j.ijinfomgt.2024.102835

Zhang, R., Duan, W., Flathmann, C., McNeese, N., Knijnenburg, B., & Freeman, G. (2024). Verbal vs. Visual: How Humans Perceive and Collaborate with AI Teammates Using Different Communication Modalities in Various Human-AI Team Compositions. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1-34. https://doi.org/10.1145/3686976

Zhang, R., Duan, W., Flathmann, C., McNeese, N., Freeman, G., & Williams, A. (2023). Investigating AI teammate communication strategies and their impact in human-AI teams for effective teamwork. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1-31. https://doi.org/10.1145/3610072

Zwislewski, A., Nanassy, A. D., Meyer, L. K., Scantling, D., Jankowski, M. A., Blinstrub, G., & Grewal, H. (2019). Practice makes perfect: The impact of Stop the Bleed training on hemorrhage control knowledge, wound packing, and tourniquet application in the workplace. *Injury*, *50*(4), 864-868.

APPENDIX

Table 5. Participant demographics and simulation performance scores

Participant	Gender	Age	First Aid Score	GSE	NASA-TLX	GTT
P1	Female	25	5	34	67.0	14
P2	Female	19	4	28	44.3	16
P3	Female	20	4	36	57.7	24
P4	Male	21	3	30	75.0	17
P5	Female	21	3	32	39.3	19
P6	Female	20	2	30	50.7	15
P7	Male	18	2	40	61.7	18
P8	Male	20	2	30	45.3	18
P9	Male	20	2	38	50.0	22
P10	Female	26	2	33	79.0	13
P11	Female	19	2	34	22.7	13
P12	Female	23	2	30	51.3	18
P13	Female	54	2	29	55.3	22
P14	Female	22	2	33	55.7	22
P15	Female	20	2	26	76.3	18
P16	Male	23	2	29	63.7	20
P17	Male	50	1	32	67.7	20
P18	Male	27	1	37	75.7	15
P19	Male	22	1	27	70.3	19
P20	Male	21	1	30	76.7	21
P21	Male	19	1	38	66.3	17
P22	Male	26	1	31	71.0	13
P23	Male	24	1	32	42.7	21
P24	Female	25	1	27	52.0	25
P25	Male	22	1	39	36.0	22
Average		**24.3**	**2.0**	**32.2**	**58.1**	**18.5**