Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams

BEAU G. SCHELBLE, Clemson University, USA
CHRISTOPHER FLATHMANN, Clemson University, USA
NATHAN J. MCNEESE, Clemson University, USA
GUO FREEMAN, Clemson University, USA
ROHIT MALLICK, Clemson University, USA

An emerging research agenda in Computer-Supported Cooperative Work focuses on human-agent teaming and AI agent's roles and effects in modern teamwork. In particular, one understudied key question centers around the construct of team cognition within human-agent teams. This study explores the unique nature of team dynamics in human-agent teams compared to human-human teams and the impact of team composition on perceived team cognition, team performance, and trust. In doing so, a mixed-method approach, including three team composition conditions (all human, human-human-agent, human-agent-agent), completed the team simulation NeoCITIES and completed shared mental model, trust, and perception measures. Results found that human-agent teams are similar to human-only teams in the iterative development of team cognition and the importance of communication to accelerating its development; however, human-agent teams are different in that action-related communication and explicitly shared goals are beneficial to developing team cognition. Additionally, human-agent teams trusted agent teammates less when working with only agents and no other humans, perceived less team cognition with agent teammates than human ones, and had significantly inconsistent levels of team mental model similarity when compared to human-only teams. This study contributes to Computer-Supported Cooperative Work in three significant ways: 1) advancing the existing research on human-agent teaming by shedding light on the relationship between humans and agents operating in collaborative environments, 2) characterizing team cognition development in human-agent teams; and 3) advancing real-world design recommendations that promote human-centered teaming agents and better integrate the two.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**;

Additional Key Words and Phrases: team cognition, teaming, human-autonomy teaming, artificial intelligence, trust

ACM Reference Format:
Beau G. Schelble, Christopher Flathmann, Nathan J. McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 13 (January 2022), 29 pages. https://doi.org/10.1145/3492832

Authors' addresses: Beau G. Schelble, Clemson University, Clemson, South Carolina, USA, bschelb@g.clemson.edu; Christopher Flathmann, Clemson University, Clemson, South Carolina, USA, cflathm@g.clemson.edu; Nathan J. McNeese, Clemson University, Clemson, South Carolina, USA, mcneese@g.clemson.edu; Guo Freeman, Clemson University, Clemson, South Carolina, USA, guof@g.clemson.edu; Rohit Mallick, Clemson University, Clemson, South Carolina, USA, rmallic@g.clemson.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
2573-0142/2022/1-ART13 $15.00
https://doi.org/10.1145/3492832

1 INTRODUCTION

Teams have consistently leveraged the latest technologies available to them to extend and enhance their operational capabilities, with virtual teams being a relevant and successful example [55]. This intersection between teams and the technologies they leverage has and will continue to exist, serving as a defining feature of the field of Computer-Supported Cooperative Work (CSCW) [30]. More recently, the advent of artificial intelligence (AI) is creating a technological intersection with teamwork that centers around the creation and integration of artificial teammates [64], leading to an emerging CSCW research agenda on human-agent teaming (HAT) [103], and AI agent's roles and effects in modern teamwork [75].

However, while empirical studies on HATs are being published at an ever-increasing rate since the start of the 21st century [75], many key questions regarding the dynamics of teamwork in HATs remain unanswered. One such understudied question centers around the construct of *team cognition* within HATs. *Team cognition* is a construct referencing a group of several related concepts such as team situational awareness, team decision making, shared mental models, and team perceptions [15], each based upon individual team members' perceptions, beliefs, and expectations [47]. The construct is most often identified with shared mental models, which are measured at both the levels of taskwork and teamwork [60]. Important outcomes of team cognition include trust [17, 23], performance [60], and team perceptions of these processes [15]. Team cognition and its outcomes can also be affected by team composition, which refers to the characteristics and attributes of individual team members [27, 89]. In addition, specific to CSCW and the broader HCI community, team cognition is often related to the concept of common ground between teammates [12]. In general, among all teamwork constructs, team cognition is often viewed as the most important and is vital to increasing team performance due to its practical benefits to effective team coordination and communication [8, 13, 68].

While team cognition has been repeatedly quantified and studied in human-human teams [71, 73, 104], how it is formed and perceived in HATs receives limited research attention. The most recent review of existing empirical studies on HATs revealed only a small handful of studies investigating team cognition [75]. Existing research focuses only on perceived team cognition and computational mental models in HATs and fails to address the following research gaps: 1) how to empirically assess the similarities and differences of team cognition development between human-human teams and HATs; and 2) how team composition affects the development of team cognition and its related key outcomes such as perceived team cognition, team performance, and trust in HATS. As an outcome of team cognition, trust is essential to HATs as it is related to team performance, and the mere presence of artificial agents can cause humans to trust all other teammates less, whether they are artificial or not [63]. Further, the existence of a research gap alone is not the only reason to study team cognition in HATs, as there are other relevant real-world implications such as improving artificial agent development, interface design, and AI bias mitigation. The effect of team composition on team cognition and its outcomes also makes studying various HAT compositions (Human-Human-Agent vs. Human-Agent-Agent) vital to address now as HATs begin to enter real-world applications, including manufacturing [83], software development [99], and city management [1]. These industries plan to utilize several different configurations of humans and agents in HATs simultaneously (one HAT with six agents and one human alongside another HAT with six humans and two agents), making team composition another vital research topic.

Therefore, the importance of team cognition to team effectiveness, the unique nature of team dynamics in HATs as compared to human-human teams, and the impact of team composition on team cognition and its outcomes drive this current research study. To address the above mentioned research gaps, this paper explores the following research questions:

RQ1: How is the development of team cognition in human-agent teams similar or different from its development in human-human teams?
RQ2: How does team composition affect the development and outcomes of team cognition in HATs?
RQ2.1: In regard to perceived team cognition?
RQ2.2: In regard to team performance?
RQ2.3: In regard to trust?

The current study addresses the research questions outlined above by conducting a mixed-methods study with three different team composition conditions, which each completed an emergency response management simulation known as NeoCITIES [45]. The study reports on the quantitative and qualitative differences and similarities in team cognition development and outcomes found between human-human teams and two different compositions of HATs. The current study contributes to CSCW research in three main ways. First, by advancing an emerging CSCW research agenda on collaborative activities occurring within human-agent teams and their impacts on humans and agents. Specifically, this work provides the first empirical analysis and comparison of shared mental models in traditional human-human teams and HATs while also investigating the effects of team composition on team cognition and its outcomes in HATs. Second, this study also expands existing research on team cognition in CSCW [31, 44, 71] by shedding light on the nature of the construct in modern HATs and how to improve human experiences of team cognition in such teams. The study also improves the field's understanding of developing common ground in disadvantaged communication environments, a notoriously tricky arena to develop shared understanding in CSCW literature [44]. Third, as HATs are poised to become a significant part of the global workforce in the coming decades [75, 83], this research will help: 1) develop and design more human-centered teaming agents; and 2) integrate humans and agents in teaming environments, leading to a more enjoyable and overall positive experience in future workforces.

2 TEAM COGNITION AND HUMAN-AGENT TEAMS

This section reviews relevant research regarding the definition of team cognition and its relationship with teaming. We also provide context on emerging research covering the role of team cognition in HATs and the particular importance of team composition to team cognition in HATs.

2.1 Team Cognition and CSCW

Team cognition is not a new concept within teamwork research domains, specifically in CSCW. Concepts such as transactive memory systems [96], distributed cognition [41], and common ground [10] all assume that cognitive aspects of a given group or team task can be shared amongst individuals. With this knowledge, it is clear that team cognition is a complex construct with many underlying and related concepts concerning team knowledge, which includes team perceptions, shared mental models, and situational awareness [15]. Since the emergence of such concepts in empirical research, team cognition constructs have been linked to team performance for several decades [60]. The question then becomes, how would effective team cognition be achieved? Research has shown that team cognition can be built through common representations that depict the shared cognitive functions between team members, such as the previously mentioned shared mental model [15]. The concept of shared mental models is one of the most commonly used concepts in team cognition, which is why the current study focuses explicitly on studying shared mental models, which are a component of the larger concept of team cognition [15], in HATs. In a practical sense, shared mental models are meant to measure whether or not team members are "on the same page," in that they share a common understanding of their shared tasks, roles, interdependencies, and strategies [68]. In a technical sense, shared mental models represent organized mental representations of the

various component pieces relevant to a team's overall task [47]. Shared mental models are often broken down into task mental models, covering aspects specific to understanding and completing a shared task, and team mental models, covering aspects specific to cooperation and communication within a team [60]. Measuring team cognition is not limited to only shared mental models but also the theory of interactive team cognition [13]; however, the current study utilizes the shared mental model perspective due to the widespread acceptance and maturity of the technique [68].

The evidence for the relationship of team cognition to a variety of team outcomes has been growing consistently over the years [59, 73]. Shared mental models can be continually developed to be more effective over time and affect various team outcomes, such as objective performance, team viability, member well-being, and strategy. It should also be stated that team cognition concepts like shared mental models are not something to be achieved or not achieved, but that it is an emergent state of teaming, and all teams constantly possess shared, organized, and distributed knowledge among their members [16, 73]. Effective team cognition also promotes the natural development and improvement of other teamwork processes, such as coordination, backup behaviors, and communication [68]. Many of these benefits come from shared mental model's ability to enhance team effectiveness in their action stage [51], which is the task execution phase for teams [58]. The action stage of teaming is explicitly vital to relate to shared mental models as a majority of HAT research focuses on action stage teaming [19, 63, 75].

The advent of technology-supported distributed teaming has also made team cognition a vital concept in CSCW. The field of CSCW was an early adopter to the concept of team cognition, and its sub-constructs of situational awareness, perception, and shared mental models, with early research focusing on the role of these constructs in remote and co-located learning environments [87]. For example, the lens of team cognition and shared mental models has been used in CSCW research to understand and measure the effectiveness of computational systems that attempt to enhance team coordination and communication [93, 100], predict team performance in distributed virtual teams through collective intelligence [46], and understand computer-mediated collaboration within fast-paced virtual environments [71]. Shared mental models have also been used in broader HCI research ranging from understanding the role of non-verbal communication in online multiplayer games [49], and to supporting common ground in computer-supported teamwork [12]. This research thus adds to the existing literature by exemplifying team cognition as a multi-faceted construct in the unique context of HATs. Additionally, the study provides the first empirical measure of shared mental models in HATs and provides comprehensive mixed-methods analysis on their development and outcomes between traditional human-human teams and HATs. This analysis is vital to the CSCW field as HATs are a new domain of collaborative technology unique from human-human teams in several ways.

2.2 Human-Agent Teams, Performance, and Trust

While the term HAT may have only become a fixture in CSCW and related research literature recently [19, 64, 103], the concept has existed since at least the early 1990s [42, 57, 75]. Unfortunately, HAT research was often joined with research focusing specifically on simpler *automated* systems, with no delineation being made between them and the more complex *autonomous* systems. This lack of clarity led to a considerable amount of confusion around what factors truly define HATs, and it was clear that a comprehensive definition was necessary, which was eventually provided in a recent comprehensive review of empirical HAT. O'Neill and colleagues define HAT as: 1) teams where agents are viewed as "agentic" by their human teammates (agents have a significant degree of independent decision making); 2) the agents must have a role interdependent with the roles of their human teammates; and 3) there must be one or more humans and one or more autonomous agents

working towards a common goal [75]. Therefore, when using the term "agent," "AI," or "HAT," the current paper is referring to these definitions laid out by O'Neill and colleagues [75].

ACM communities like CSCW are among the several major fields researching the topic of HATs, contributing significantly to its rapid growth. In particular, existing research on HAT has focused on two main characteristics of HATs: 1) the performance of a HAT; and 2) the relationship between humans and agents in a HAT, including trust.

2.2.1 HAT Performance. HAT performance has been studied in a variety of contexts ranging from medical [95] to military [21], but HATs have oftentimes underperformed or failed to outperform their human-human counterparts [19, 21, 64]. Alternatively, other experiments display incredible HAT performance, outpacing not only human-human teams but even teams consisting of all AI [95]. This disparity can likely be attributed to three major factors: 1) not all studies use true AI, which is capable of expert-level performance when properly trained [25]; 2) the more abstract the team task is, the harder it becomes to train high performing agents [82] (but not impossible [40]); and 3) a distinct lack of proper design and integration within human-agent teams that leads to confusion and poor understanding between the two types of team members. The solution to this discrepancy is to leverage the individual strengths of the AI *and* the human to move team effectiveness beyond what each is capable of achieving alone.

While AI training will always strive for better reliability and individual performance, these performance gains will not mean much if the collective exists in dysfunction, unable to benefit from the unique abilities of each team member. This assertion is backed up by many recent studies which found that while high-performing AI does engender higher trust in human teammates [101], their performance was not a predictor of the teams' performance as a whole [3]. Additionally, improvements to an AI teammate's performance and effectiveness can be offset entirely if those improvements change the user experience and present compatibility problems [4], further emphasizing the need to focus on team-level research and improvement of HATs. Team cognition holds the key to leveraging the unique advantages presented by HATs, as effective shared mental models can allow AI and humans to possess a mutual understanding of what tasks and team functions the other is best suited for, allowing for the quick and efficient allocation of team functions [3].

2.2.2 Human-Agent Relationships and Trust in HATs. Team composition affects characteristics of HATs directly through the social relationship between humans and artificial agents. Similar effects can be seen in human-human teams as team composition is known to affect situational awareness and team cognition [27, 89]. Unfortunately, humans accepting agents as full team members and giving them an equal level of respect as their human counterparts is not nearly as straightforward as it seems, and past CSCW research addresses this. For example, when participants played a video game with AI teammates, they adopted a neo-feudalistic view of the agent teammates that created unequal rights for the agents [97]. Such results are also found in the CSCW domain, where research indicates that humans are more likely to place blame for failures in online cooperative games on AI rather than human teammates, even if that AI teammate was a human pretending to be an AI [67]. Humans were also less likely to save AI teammates than human teammates and significantly misjudged their AI teammate's abilities compared to judging their human teammate's abilities [74].

Such results may be characterized by the consequences of the social identity theory, which posits the existence of "in-group" and "out-group" factions within teams/groups. These factions lead those in the in-group to see others in the group positively and identify with the group's common stereotype [90], while dehumanizing members of the out-group [94]. Recent CSCW research supports the existence of this perspective, as humans were shown to treat AI unfairly and specifically used the terms "I" and "they" to describe humans and AI teammates, respectively [103]. Research on trust in HATs also reflects deficiencies in the relationship between humans and

agents as they make humans trust their teammates less [63], revealing consequences of poor team cognition. Trust was also highly related to team performance [63], which is another outcome of team cognition that further emphasizes the importance of studying the outcomes in concert with the construct itself. Such a dysfunctional relationship between humans and agents may make it exceptionally difficult for HATs to support team cognition. Specifically, human team members may be adversely affected when outnumbered by AI and vice versa, making it essential to understand how team composition affects HATs and if theories like social identity apply to helping those in the field understand the cause of such effects.

These considerations motivate our research questions to highlight: 1) the importance of team cognition to fully leveraging both the human and the agent when designing for HATs; and 2) the significance of team composition to the outcomes of the team cognition being developed and supported by that team design.

2.3 Team Cognition in Human-Agent Teams

While team cognition has received attention in contexts of human-human teaming [68, 73], how it may be fostered and experienced differently in HATs is understudied. The most recent review of existing empirical research of HATs included team cognition as one of several focus areas [75]. The handful of studies collected revealed several insights, despite the overall dearth of literature. For example, it was shown that virtual agents with agreeable personality traits lead to higher perceived team cognition [36]. Additionally, the study's results indicated that agents with personality traits more closely aligned to their human teammates engendered higher perceived team cognition [36]. Perceived team cognition also shared a positive relationship with both verbal and non-verbal communication in HATs and retained its positive relationship with team performance [35]. Unfortunately, other research has indicated roadblocks that may prevent high levels of team cognition compared to human-only teams. Specifically, research shows that HATs may possess more rigid team cognition (inability to adapt to environmental changes rapidly) [18]; however, HATs can overcome this rigidity if they can engage in effective communication and develop accurate situational awareness [20].

Research on the fundamentals of team cognition in HATs is essential to better utilize applied research and interventions that seek to improve team cognition in HATs. For example, studies have focused on applying interventions to HATs to enhance team cognition and team effectiveness. For example, a unique cross-training technique leveraged Markov Decision Chains to represent the autonomous agent's mental model and fine-tune it by training with their human teammate, resulting in significantly improved levels of team performance and trust [72, 86]. Another study deployed a system that shared team members' current cognitive load and beliefs with other human and autonomous agent teammates. The autonomous agent teammates then utilized the information to understand the humans' current status better and develop better shared mental models [22]. These studies demonstrate the potential for developing systems that enhance team cognition. Accordingly, without research into the fundamental nature of team cognition in HATs, such applied efforts are severely limited.

Lastly, research on team cognition in HATs should utilize robust measures of the construct known to capture its content and structure. A majority of the studies measuring team cognition in HATs have used broadly applicable Likert scale questionnaires, which, while more accessible, only capture the content of a shared mental model, not its structure [68]. Therefore, simplified measures of team cognition are only regarded as elicitation tools and not shared mental model measurement techniques [68]. Paired sentence comparison and concept mapping are examples of techniques that measure both content and structure [68], with the current study employing paired sentence comparisons. While a variety of team cognition measurements in HAT research is

positive, given the importance of capturing shared mental model structure and the lack of studies utilizing measures that do so, a significant gap in the literature is exposed.

It is essential for the CSCW community to characterize how team cognition develops in HATs and how team design aspects like team composition affect its outcomes. Understanding team cognition helps to ensure that humans and AI can identify the shortcomings and strengths of each other, leveraging this knowledge to enhance team performance effectiveness. This assertion can be accomplished by addressing the significant gaps that currently exist in HAT team cognition research, such as the lack of empirical research on shared mental model development in HATs vs. human-human teams (RQ1) and the effects of team composition on key outcomes of team cognition including its perception, team performance, and trust (RQ2).

3 METHODS

The current study employs a mixed-methods design to capture and analyze team cognition's formation in teams with varying numbers of agents and humans. The experiment utilized the well-published and validated team research platform known as NeoCITIES [38, 39, 45, 62], which provides an excellent environment to study team cognition and team interaction within both human-human teams [34], and HATs [82]. A 1x3 experimental design was developed as shown in Table 1 to study the effect of various team compositions on the development of team cognition and its related outcomes.

Table 1. Experimental Conditions

Condition Number	Team Composition Pattern
Condition 1 (HHH)	Human-Human-Human
Condition 2 (HHA)	Human-Human-Agent
Condition 3 (HAA)	Human-Agent-Agent

3.1 NeoCITIES Task and Roles

NeoCITIES uses a fictional college town in which three players work together in interdependent roles to respond and complete emergency tasks occurring over time. These roles include Hazmat, Police, and Fire, each with a triad of resources to address events that occur. With three unique interdependent roles, NeoCITIES provides an excellent opportunity to observe the different possible combinations of interactions between humans and agent teammates in the context of emergency response management. The versatility of NeoCITIES is the driving force behind its success as an experimental platform for teams, making it a valuable asset to researching team cognition in HATs. The interface of the simulation can be seen in Figure 1.

The interface of NeoCITIES is designed to simulate an emergency response role as if the user were acting in a supervisory position. As part of their duty in this fictional college town, participants must determine when and where their respective resources need to be assigned based on active events in coordination with other teammates. The interface presents consistent tools amongst all team members regardless of their role to create situational awareness. Tools include a manifest of their resources, active and past events, event descriptions, a chat function to communicate with teammates, and the current objective for all team resources. Participants were also given a spatially accurate map that displayed each teammate's resources, home bases, and currently active events. Accordingly, all team members could establish a level of shared cognition for their responsibilities, resources, strategies, and teammates.

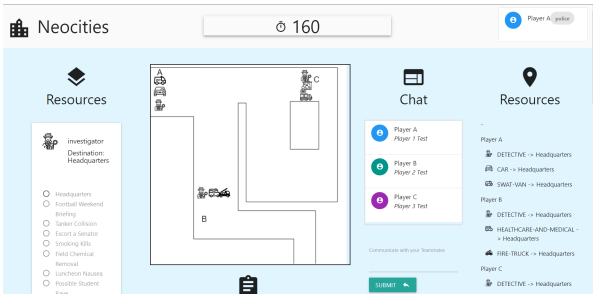


Fig. 1. NeoCITIES Home Screen Interface

During each of the four nine-minute rounds, nine different events occurred that required a response from the team to complete successfully. These nine events and their requirements are shown in Table 2, and each event's location was changed between rounds. Each event must be completed with specific resources, but an additional layer of complexity is introduced as many of the resources themselves differ in speed. For example, in the time limit allowed, some resources could cover more distance compared to other resources; specifically, the slow resources consisted of the investigators, ambulance, and chemical truck. The other resources were equal in speed except for the fire truck, which was the fastest resource of all. Each event in Table 2 can also be categorized into three difficulty ratings (1 = Low, 3 = High), which are based upon the number of resources required, the speed of those resources, and the location of those events. This feature of the simulation made time and distance an additional dimension for the team to consider in decision making, which allows for additional insights into how individuals are cognizant of their respective team members' situations and their approaches to a problem.

Table 2. NeoCITIES Events and Necessary Resources to Complete Them

Order	Emergency Event	Necessary Resources
1.	Football Weekend Briefing	Investigator
2.	Tanker Collision	Squad Car, Fire Truck, Chemical Truck *(In That Order)*
3.	Escort a Senator	SWAT Van
4.	Smoking Kills	Fire Truck
5.	Field Chemical Removal	Chemical Truck
6.	Luncheon Nausea	Ambulance, Investigator
7.	Possible Student Rave	Investigator, Squad Car
8.	Old Main Frame Shoppe Fire	Investigator, Fire Truck
9.	City Hall Bomb Threat	Bomb Squad, Investigator

3.2 Autonomous Agent

This study incorporated an expert system programmed to complete the NeoCITIES task in either the Police and or Hazmat role with high accuracy and flexibility to adapt to needs signaled by its teammates. This system was only applied to team conditions with an agent team member (Condition 2: HHA and Condition 3: HAA; Condition 2 fielded an agent in the Hazmat role only). Expert systems are a branch of applied AI and are designed to represent expert-level human knowledge in a task [53]. In the current study, the expert system continuously managed the allocation of resources

to events based on the simulation state. The expert system was flexible to human teammates because it could make decisions that reacted to the humans' actions or requests to increase the team score. Accordingly, the expert system was developed to recognize its teammates' decisions and plan on the resultant consequences of those decisions. This implementation allowed the expert system to possess a collaborative "mentality" with which they replicate their teammates' level of awareness to assist them better [9].

The chat communication provided by the agent was not a feature of the expert system and was instead accomplished using the Wizard of Oz technique. This technique has a trained experimenter represent a feature of the system (chat communication in this case) to an unknowing participant [61]. The Wizard of OZ technique is often used to simulate capabilities of AI when not fully computationally available [61]. The trained experimenters followed a script developed through multiple iterations of pilot testing. The agent's capabilities were conveyed to participants beforehand to help control for participants expectations of the agent teammate. The agent was described as an expert level player in the role they were assigned (Hazmat in the HHA condition, Hazmat and Fire in the HAA condition). Additionally, the agent was described as having advanced text generation and understanding capabilities similar to Siri or Google Home regarding the NeoCITIES simulation, but no other topics. Thus, the agent could take requests, offer information, and respond coherently as long as the subject involved the NeoCITIES simulation.

3.3 Participants

This study recruited 66 participants (31 Males, 35 Females) from a departmental subject pool at a major university in the USA. The average age of participants was 18.91 (SD = 1.51). The participants were placed into conditions, teams, and NeoCITIES roles at random, and they did not know each other before participation. Participants received course credit for their time as an incentive for their participation.

3.4 Procedure

The novel COVID-19 global pandemic forced in-person research to a standstill due to the highly contagious nature of the coronavirus [80]. Following appropriate social distancing techniques to mitigate the risk of infection, this study was conducted remotely through the high fidelity video-conferencing application Zoom, which is very effective for remote research and was used by multiple researchers in the past year [2, 28]. All Zoom sessions were monitored and conducted by trained experimenters who continuously observed participants, much like a typical in-person experimental setting. Any participants observed within the simulation, survey, and/or Zoom not paying attention or taking the experiment seriously were dismissed. Trained experimenters gave all participants the same information and instructions following a predefined protocol approved by the local Institutional Review Board.

Each condition collected data from 10 teams; however, due to over-scheduling, the HHH condition consisted of 12 teams instead of 10. The experiment was conducted between-subjects where each participant only participated on one team in one condition. Students signed up for a particular testing time and received a Zoom meeting identification and password to enter the secure, virtual environment. Students were instructed through video and audio modalities and interacted with the experimenter in the same fashion. The session began by collecting informed consent from the participants, followed by demographic information.

Afterward, experimenters introduced the study in more detail, providing information on what team cognition is and an overview of the simulation. Participants were assigned their team roles, and participants were informed which role(s) would be taken by an AI teammate, if applicable. Participants were then taken to the simulation training page, where each feature of the simulation

Table 3. Participant Numbers

Overall: **66** (32 Teams)		
HHH: **36** (12 Teams)	HHA: **20** (10 Teams)	HAA: **10** (10 Teams)

was explained in detail alongside video examples. This training page was followed by an in-game training round where all players could ask questions about the interface and the simulation itself. Once the training was complete, and participants agreed to start, the next round began, and they were no longer able to ask any questions.

After completing each round, participants were shown their score and linked to the next round, which began when team members were ready. Each round lasted for nine minutes, and the teams worked together for 36 minutes in total within the NeoCITIES simulation. In congruence with past literature, this amount of time is adequate to develop team cognition [65, 70]. Upon completing the four rounds, the experimenter instructed the team to navigate the survey to complete the post-task measures. The post-task survey collected their team and task mental models, perceived cognition, trust in agent teammates, subjective team performance, and a series of free-response questions. Once completed with the post-task survey, participants were free to leave the Zoom session and were compensated for their participation with course credit.

3.5 Measures

3.5.1 Task and Team Mental Models. Mental Models of the task were measured using paired sentence comparisons [6], a strategy that has long been utilized in the past to measure both the content and structure of shared mental models [48, 60, 68, 91]. Participants were asked to judge the relationship between all significant task attributes on a nine-point Likert scale ranging from -4 to 4 and anchored by "Negatively Related" to "Positively Related" (with 0 representing "Not Related"). Task attributes were identified through comprehensive task analyses with subject matter experts (NeoCITIES simulation designers). The task attributes were as follows: *(1) familiarizing with the simulation layout, (2) determine which resources are at your individual disposal, (3) determine location of event, (4) send resource to event if available, (5) learn what resources your teammates have, (6) recall resources, (7) determine resource allocation based on event importance, (8) send resources in the correct order for critical events.* By assessing how positively related, unrelated, or negatively related each attribute was to the others, a network of relationships can be created, capturing the content and structure of their mental model.

The same methodology was applied to obtain a participant's team mental model, but the collection of teaming attributes was different. The attributes compared are more generalized and were explicitly taken from past shared mental model research [50, 60]: *(1) amount of information, (2) quality of information, (3) role/responsibility, (4) interaction patterns, (5) communication channels, (6) role interdependencies, (7) teammates' skill, (8) teammates' attitudes, (9) teammates' preferences.* Participants were also given definitions of each team attribute listed.

3.5.2 Mental Model Similarity. The Pathfinder network scaling algorithm was used to determine mental model similarity, which is familiar to shared mental model research [14, 60, 68, 69]. This algorithm inputs the participant's pairwise comparisons of the predefined attributes to create graphical representations of their mental models [85]. Each attribute represents a node in the graph, and the assessed relationships between attributes are the links between nodes. A similarity metric is produced by comparing two networks to provide a similarity rating between zero (no similarity) and one (perfect similarity) of the two. Pathfinder can only provide a similarity metric for two human team members at a time and therefore the HHH team had their three possible pairings

averaged together for a single team similarity metric, which is standard practice [54, 81]. AI agents cannot provide any ratings, so the HHA condition produced a single comparison, while the HAA condition could not produce any comparison. This method was the same for the task and team mental models.

3.5.3 Perceived Team Cognition. Perceived team cognition was measured using the Teamwork Schema Questionnaire [76, 78]. Participants were asked to rate the importance of a series of statements to their idea of teamwork. The participants were then presented with the same statements, but this time were asked to rate how important they believed each statement was to their human teammate(s) idea of teamwork (one assessment for both human teammates). If the participant had AI teammate(s), they also completed an assessment for them (one assessment for both AI teammates). Together these questions created a measure of congruence representative of total perceived team cognition. Perceived team cognition was calculated by taking the absolute difference between the participant's teamwork ratings and the teamwork ratings they chose for their human and AI teammates. The scores were then scaled by the number of comparisons made on the team. Scores ranged between 0 and 84, with lower scores indicating higher perceived team cognition.

3.5.4 Objective Team Performance.

$$EventScore = \frac{(end - start)}{(limit - start) * difficulty} \tag{1}$$

$$TeamScore = \frac{100 * [(worstScore - rawScore)}{(worstScore - bestScore + 1)]} \tag{2}$$

NeoCITIES calculates team performance using Equations 1 and 2, which have been utilized in past HAT research [65]. The variables within Equation 1 like "(end - start)" refer to the duration it took the team to complete the event successfully, from the moment the event became active, while "limit" referred to the time limit associated with that event. The "difficulty" variable referred to that event's particular difficulty rating. The variables in Equation 2 like "Raw Score" represent the cumulative sum of the actual earned "Event Scores," while "Worst Score" represents the cumulative sum of the theoretical worst "Event Scores." Similarly, "Best Score" is the cumulative sum of the theoretical best "Event Scores." Accordingly, the equation produced objective team performance scores ranging from 0 to 100 based upon the team's accuracy, speed, and ability to complete events, with higher scores indicating higher objective team performance. Additionally, the team scoring equations punished for wasting valuable resources on events that were not actually resolved.

3.5.5 Perceived Team Performance. Perceived team performance was measured using the Team Effectiveness Scale [79]. Subjects were asked to respond to questions that gauge how well they believe their team performed in the task on a five-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree." The resulting scores ranged from 8 to 40, with higher scores indicating higher perceived team performance.

3.5.6 AI Trust. Participants' trust in the agent teammate they worked with in NeoCITIES was measured using statements derived from the principles of trust and distrust as defined in recent literature [56]. An example statement was, "Did you feel confident in the AI you just worked with?" which was rated on a five-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree." Scores ranged from 6 to 30, with higher scores indicating higher levels of trust in the agent.

3.5.7 Qualitative Questions. Participants were presented with an opportunity to provide more details about their experiences and opinions after the experiment within the post-task survey. The open-ended questions were designed to extract participants' experiences working with the agent

and how team cognition developed within their team. Responses underwent thematic analysis [7, 26, 92], where participants' statements were analyzed for themes that related to the study's research questions. After the analysis was completed, quotes were extracted to concisely illustrate those themes.

Thematic analysis offers insight into how individuals construct their perceptions, understandings, and accounts of the felt experiences in teamwork [88]. The analysis used the following phases: 1) two of the authors read through all of the narrative information provided by participants to achieve an understanding of how team cognition developed; 2) two of the authors then combed through the narratives to identify thematic topics based upon participants descriptions of how team cognition did or did not develop within their team, how, and why, as stated in RQ1; 3) all authors reviewed and debated the themes and sub-themes identified in Phase 2; 4) the first author identified strong example quotes that best represented the themes and sub-themes identified in Phase 3; 5) all authors again reviewed and debated the refined themes and sub-themes, using the quotes identified in Phase 4 to synthesize the similarities and differences in team cognition development between traditional human-human teams and HATs. Differences between the two authors identifying themes in Phase 2 were resolved through open discussion and, if necessary, discussed and resolved with all authors in Phase 3.

4 RESULTS

To answer our RQs, we present our findings in two parts. Both sections report on data addressing the two stated research questions with the quantitative results reporting on the analysis of the empirical measures of performance, shared mental models, and the associated perceptions of team cognition, performance, and trust in AI. The qualitative section focuses on the similarities and differences of developing team cognition in human-human teams vs. HATs. The quantitative analysis section is organized by dependent variable, while the qualitative results section is organized by each major theme identified.

4.1 Quantitative Results

The quantitative results are divided into three major sections to address the study's research questions, and each is explicitly focused on one of the three categories of dependent variables measured. Each dependent variable and its mean and standard deviation can be seen in Table 4. Average score and perceived team performance are covered first, followed by the team cognition variables of team and task mental model similarity and perceived team cognition, and lastly, trust in AI is covered as the final component of the team perceptions that make up team cognition. Additionally, all statistical assumptions for tests (i.e., normality, homoscedasticity) were met for all analyses unless otherwise stated.

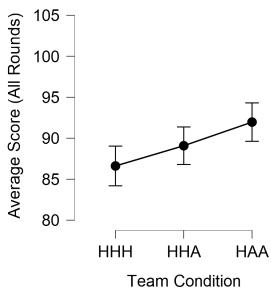
4.1.1 Objective and Perceptive Team Performance. The current study took two measures of team performance in all three conditions, the first being an *objective* measure of the teams' collective performance in the NeoCITIES simulation, and the second was a measure of how well the team *perceived* their collective performance to be. Analyzing differences in the three conditions' perceived and objective performance contributes to answering RQ2.2, shedding light on how team composition affects the outcomes of team cognition.

To determine whether team performance changes as a function of team composition, a one-way ANVOA was conducted. The main effect of team composition on objective team performance was statistically significant, $F\ (2, 27) = 6.07$, $p = .007$. Also, the effect size ($\eta^2 = .31$) indicated that 31% of the variance in objective performance could be explained by team composition, which is a large effect size [11].

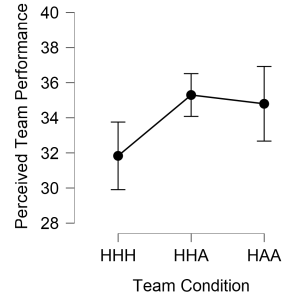
Table 4. Mean and Standard Deviations for Dependent Variables

Measure	HHH		HHA		HAA	
	Mean (*N*)	**SD**	**Mean** (*N*)	**SD**	**Mean** (*N*)	**SD**
Team Performance	86.62 (12)	3.81	89.08 (10)	3.20	91.97 (8)	2.80
Trust in AI	*N/A* (0)	*N/A*	26.55 (10)	2.33	24.30 (10)	2.26
Perceived Team Performance	31.83 (12)	3.03	35.30 (10)	1.70	34.80 (10)	2.97
Perceived Team Cognition	8.19 (12)	6.18	11.28 (10)	3.28	12.10 (10)	7.74
Team Mental Model Similarity	0.30 (12)	0.06	0.28 (10)	0.12	*N/A* (0)	*N/A*
Task Mental Model Similarity	0.31 (12)	0.07	0.35 (10)	0.08	*N/A* (0)	*N/A*

Because team composition was found to be significantly related to objective team performance, post hoc analyses were conducted using Tukey's HSD. This analysis revealed that the HHH condition (M = 86.62, SD = 3.81) did not have significantly different objective performance from the HHA condition (M = 89.08, SD = 3.20). The HAA condition (M = 91.97, SD = 2.80), however, did have significantly higher objective performance than the HHH condition, but not the HHA condition. These differences can be seen in Figure 2a.



(a) Objective Team Performance (Includes Training)



(b) Perceived Team Performance

Fig. 2. Measures of objective and perceived team performance

An additional one-way ANOVA was conducted to assess whether perceived team performance changes as a function of team composition. This analysis revealed that the effect of team composition on perceived team performance was statistically significant, F (2, 29) = 5.53, p = .009. In addition, the effect size (η^2 = .28) indicated that 27.6% of the variance in perceived team performance could be explained by team composition, which is a large effect size.

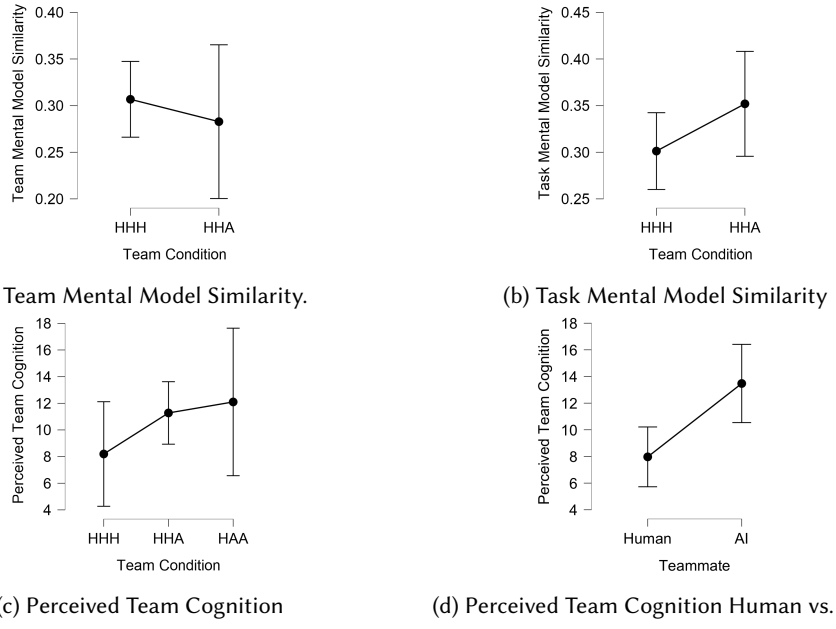
Since the ANOVA revealed significant differences in perceived team performance as a function of team composition, Tukey's HSD post hoc analyses revealed that the HHH condition (M = 31.83, SD = 3.03) reported significantly lower perceived team performance when compared to the HHA condition (M = 35.30, SD = *1.70*). Additionally, the HHH condition had significantly lower perceived team performance when compared to the HAA condition (M = 34.80, SD = 2.97), and there were no significant differences between the HHA and HAA conditions. The results of these analyses can be seen in Figure 2b.

While the difference is not significant, it is interesting to point out that the trend seen in Figure 2a is not maintained in Figure 2b. This trend points to an apparent misconception of how humans

perceived their team's performance when they were the only human on the team compared to their team's objective performance.

4.1.2 Team Cognition. Three different measures of team cognition were collected, but all three did not apply to all conditions. Team and task mental model similarity applied only to the HHH and HHA conditions and measures how similar the content and structure of the individual team members' mental models were. The last measure of team cognition, perceived team cognition, was measured in all three conditions and measures only the perception of team cognition within each team. Finally, the HHA condition was in a unique position to measure perceived team cognition with human teammates and perceived team cognition with agent teammates within the same team. This subset of analyses contributes to answering RQ1 and RQ2.1, which investigate the similarities and differences in team cognition's development and perception across team compositions.

A check of statistical assumptions revealed significant heteroscedasticity between the HHH and HHA conditions in team mental model similarity. Because of this a Mann-Whitney U test was used to determine the effect of team composition on team mental model similarity. The average team mental model similarity for those in the HHH condition (M = .31, SD = .06) was higher than those in the HHA condition (M = .28, SD = .12), but this difference was not statistically significant $U($ $N_{(HHH)}$ = 12, $N_{(HHA)}$ = 10) = 70, p = .54, rb = .17. These results can be seen in Figure 3a. Finding significant heteroscedasticity is noteworthy as this result can be more than the violation of a statistical assumption and can instead contribute meaningfully to understanding how various individual differences may contribute to teams and teamwork [84].



(a) Team Mental Model Similarity.

(b) Task Mental Model Similarity

(c) Perceived Team Cognition

(d) Perceived Team Cognition Human vs. Agent

Fig. 3. Measures of team and task mental model similarity and perceived team cognition. All error bars represent bootstrapped 95% confidence intervals

To assess the effect of team composition on task mental model similarity, an independent samples t test was conducted between the HHH and HHA conditions. Teams in the HHH condition averaged a lower task mental model similarity (M = .30, SD = .07) than those in the HHA condition (M =.35,

SD = .08), but this difference was not statistically significant, $t(20) = 1.66$, $p = .113$, $d = .71$, with the estimated Cohen's D indicating a medium effect size [11]. These results can be seen in Figure 3b.

To assess whether perceived team cognition changes as a function of team composition a one-way ANOVA was conducted, which indicated that the differences seen in Figure 3c were not significant, $F(2, 29) = 1.30$, $p = .287$. Additionally, the effect size ($\eta^2 = .08$) indicated that 8.3% of the variance in perceived team cognition could be attributed to team composition, which is a medium effect size.

Lastly, since participants in the HHA condition provided a perceived cognition score for both their agent teammate as well as their human teammates, the two values can be compared to determine if humans perceive different levels of team cognition with human and agent teammates. An independent samples *t* test revealed that teams in the HHA condition perceived lower levels of team cognition with their agent teammate ($M = 13.48$, $SD = 6.27$) than with their human teammate ($M = 7.97$, $SD = 5.06$), and this difference was significant, $t(40) = 3.14$, $p = .003$, $d = .97$, with the estimated Cohen's D indicating a large effect size. This difference can be seen in Figure 3d.

In summary, while shared mental model results indicated no *significant* differences between human-human and human-agent teams, there is value in insignificant results [43], and there are also essential trends to identify here. HATs had lower *team* mental model similarity, and their similarity levels were significantly more varied than human-human teams, but HAT's had greater *task* mental model similarity than human-human teams. This trend reveals that even though HAT's team mental models suffer, the agent teammate is capable of setting an example for their teammates, and in doing so, they enhance the team's shared understanding of the task, as posited in prior HAT research [65]. Finally, human team members perceived significantly less team cognition with agent teammates than human teammates, as shown in the independent samples *t* test, and this trend was reflected in the ANOVA of the three conditions.

4.1.3 AI Trust. The final quantitative analysis is focused on differences in AI trust or how much trust team members had for their agent teammates. This measure applied only to the HHA and HAA conditions. This analysis supports RQ2.3, which explores how trust in AI is affected by team composition as an outcome of team cognition.

To determine whether AI trust was affected by manipulations in team composition, an independent samples *t* test was conducted. HHA teams reported higher levels of AI trust ($M = 26.55$, $SD = 2.33$) than HAA teams ($M = 24.30$, $SD = 2.26$), and this difference was significant, $t(18) = 2.19$, $p = .042$, $d = .98$, with the estimated Cohen's D indicating a large effect size. This difference is shown in Figure 4.

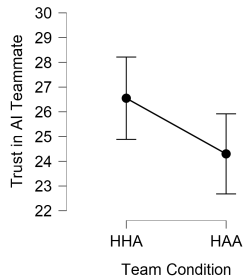


Fig. 4. Trust in AI Teammate(s). Error bars represent bootstrapped 95% confidence intervals.

4.2 Contextualizing the Development of Team Cognition with Human vs. Agent Teammates

The qualitative findings provide additional detail and context to RQ1 by directly revealing participants' relevant opinions and lived experiences throughout the collaborative simulation. In concert with the quantitative results, this analysis provides a holistic and detailed picture of team cognition development in HATs compared to human-human teams. Each quote is followed by a participant's identifier and their assigned condition. Additionally, the context of each quote given by participants is supplemented by words taken directly from the question they were answering at the time, indicated by the square brackets.

4.2.1 Similarities in Team Cognition Development Between Human-Human Teams and HATs. The findings revealed two clear similarities in team cognition development between human-human teams and HATs. The first was the iterative nature of team cognition, and the shifting focus human teammates have as they gain experience with their teammates (agent or human). The other similarity was how effective communication rapidly accelerates the formation of team cognition and how essential that accelerant is to many teams that cannot suffice on implicit communication alone.

4.2.1.1 Developing team cognition is an iterative process. Much like human-human teaming, the development of team cognition in HATs appears to be an iterative process, taking place throughout a series of shared experiences with teammates. Teams begin with very little if any shared experience apart from the instructional video, and the process of becoming accustomed to the game together presents an excellent opportunity to develop team cognition. This iterative process represents the natural progression of team cognition developing due to collective successes and failures over time. P206 and others explicitly noted how they were aware of their team cognition steadily growing throughout each game:

> *"It [team cognition] happened in the later games. Personally, I did not know what I was doing in the first game, but then got an understanding of it as the game went on." (P206-HHA)*
> *"Yes [I feel team cognition was established], through each game, team cognition grew and grew." (P218-HHA)*
> *"It [team cognition] happened in the later games, because we got more comfortable with the tasks as we moved up levels." (P121-HHH)*

According to these participants, the initial games saw them learning more about the task and the various events they needed to respond to at an individual level, while in the later games, they become familiarized with it collectively. For example, P121 highlighted their team's collective comfort with their task. P206 exemplifies this same iterative process, clearly demonstrating how similar the iterative process of team cognition is as team member focus shifts from individual to collective familiarization.

The iterative process exemplified above can be broken down further as the additional rounds present an opportunity to continue to learn the intricacies of the simulation and their teammates' tendencies. Specifically, the later rounds show teammates what to expect from certain roles within the team and their interdependencies with the other two NeoCITIES roles. These later rounds present an opportunity for teams to take advantage of this acclimation and leverage a collective strategy:

> *"Later games went much better than the earlier ones as we got a feel for the strategy." (P223-HHA)*

"Yes definitely [team cognition developed], as the game progressed I think we all developed an understanding of the game and of each member's responsibilities." (P110-HHH)

According to these participants, when it comes to the transformation of team cognition over time, HATs are similar to human-human teams in that adding high-performing teammates to teams does not simply speed the process up. As P110 suggested, it took time for human teammates to familiarize themselves with the shared task and the specific responsibilities of team members. Agents appear to be along for this ride and play a significant role in cultivating and sharing this iterative experience as the following theme's showcase.

4.2.1.2 Communication is a rapid accelerant of team cognition development. The importance of communication extends to both human-human teams and HATs. Participants indicated that the chat was the main focus while completing the simulation with their team, and they heavily associated communication with the establishment of shared cognition:

"[Everyone thought about cooperating and responding to events the same because] When we told each other what would be quickest, we listened and the outcome was better than what it would have been." (P132-HHH)
"[Everyone thought about cooperating and responding to events the same because] Each member started to say where they were sending their resources and were asking others to send resources based on closeness to the event." (P219-HHA)

The above quotes showcase how vital the chat communication feature was to HATs and human-human teams alike. As P132 specifically highlighted, communication was instrumental to their team developing a shared understanding and subsequently improving their teaming outcomes. P219 also identified how the chat supported a shared strategy that their team developed over time, accelerated by communication.

On the other hand, some teams reported that the lack of communication specifically indicated their team lacked team cognition:

"[Team cognition was] Not at all [established] there was not a lot of communication and in the end we still failed [the] task." (P135-HHH)

P135 is a clear example of teams that, for one reason or another, are unable to develop any form of shared understanding through implicit communication and specifically require the acceleration and support that explicit communication provides for the development of team cognition.

HAA teams echoed this sentiment. These teams stated that they did not communicate with the agent teammates much or outright stated that they would have communicated more if working with humans instead of agents:

"No [I don't think team cognition was established], I think there would have been more discussing if it was with other humans." (P308-HAA)
"I didn't really communicate [with the agent teammates] that much." (P303-HAA)

These quotes should signal the growing need for more discussion-driven features to be included in some agents to overcome the barrier put in place by some human teammates.

Surprisingly, a handful of teams felt capable of developing and establishing team cognition without the need for explicit communication. These teams utilized spatial information to implicitly coordinate themselves in response to each other's actions, intentions, and implied strategy, perfected over time. This sentiment is, of course, not shared across all teams with little communication. Nevertheless, it does reveal that some team tasks may support the development of team cognition through implicit communication and spatial information:

"[Teammates] Partially [thought about cooperating and responding to events the same], it seemed like we anticipated each others movements and responded to each other somewhat"

(P136-HHH)
"[Team cognition was] Somewhat [established], we did not quite communicate but we ended up doing pretty well" (P205-HHA)

While explicit communication may not be necessary to develop team cognition in all cases, it is clearly a significant driver and accelerant of team cognition. This assertion makes it essential that task-related spatial information be included whenever possible for teams, especially for HATs, as many agents have significant limitations to their communicative abilities.

4.2.2 Differences in Team Cognition Development Between Human-Human Teams and HATs. The findings also revealed two significant differences in team cognition development between human-human teams and HATs. Each center around communication and design. The first actionable difference between human-human teams and HATs was the importance of action-based communication from the artificial agent to the human teammates when developing and supporting team cognition. The second was how foundational shared goals were to convey through communication and design to help connect human teammates to agents.

4.2.2.1 Building team cognition in HATs centers around actionable communication from the agent. Not all forms of communication are equally important to developing team cognition in HATs. Action-related communication events were consistently identified as fundamental to building team cognition in HATs. Action-related communication is of specific importance as it is the most task-related and is typically time-sensitive, meaning these communication events have significantly more emphasis by their very nature. As a result, human teammates place particular importance on communication events relating to task actions, with P208 and P219's quotes illustrating such importance:

"Yes [I trusted my agent teammate] because I could ask them to do certain things" (P208-HHA)
"Yes [everyone thought about cooperating and responding to events the same], [because] each member started to say where they were sending their resources and were asking others to send resources based on closeness to the event." (P219-HHA)

According to P208 and P219, team cognition is developed between team members as the event-related communications with the agents help human teammates better understand facets of the simulation like task events and their resources. Such action-related communication also helps put those clearly defined roles into actual practice, focusing on the more nuanced developments of team cognition found in the later rounds. Specifically, P308 and P206 responded positively to agents putting their role and strategies into practice by making requests and conveying intent and strategy:

"[An example of team cognition in my team was when] The other teammates would write their next steps and discuss how they were going to move forward" (P308-HAA)
"[An example of team cognition in my team was] When a team member (the AI) would ask one of us to do something." (P206-HHA)

Additionally, HATs identified when the agent began communication within their team or the agent communicating an actionable request as specific examples of team cognition:

"[A specific example of team cognition in my team was] At the beginning when the AI communicated." (P212-HHA)
"[A specific example of team cognition in my team was] When a team member (the AI) would ask one of us to do something." (P206-HHA)

The quotes above indicate that agents can bear the burden of initiating communication within a team and help jump-start the accelerated development of team cognition. They also show that human teammates see cooperative actionable communication from agents as especially valuable to shared understanding.

Based on these findings, human members of HATs seem to specifically value action-related communication from their agent teammates over other forms of communication. This form of communication is significantly more effective in helping develop team cognition throughout the entirety of the iterative process. Communication within HATs should focus on perfecting these aspects of communication to allow HATs to develop high levels of team cognition.

4.2.2.2 Explicit shared goals in HATs are foundational for building team cognition in HATs. The starting point for cultivating team cognition is a complex process that generally begins with teammates familiarizing themselves with one another and the task. Agents do not share this process with humans, making it difficult for many HATs to begin developing team cognition. Instead, humans seem to rely on beginning the process of developing team cognition from the goal shared between themselves and the agent, using it as a launching point for iteration through discussion and shared learning. P306 and P218 specifically noted shared goals as being a foundation to shared team cognition:

> *"Yes [everyone thought about cooperating and responding to events the same], because everyone had the same goal in mind." (P306-HAA)*
> *"Yes [everyone thought about cooperating and responding to events the same], [because] it seemed that all team members cared about the goal of the game and cooperated together to try to achieve it." (P218-HHA)*

HATs are in a unique position to utilize shared goals to launch the development of team cognition as AI-powered agent teammates can be very high-performing team members. Agent teammates are so high performing in some cases that human teammates look to them as an exemplar of how to develop their strategy within the simulation:

> *"They were probably the best member on the team. They were able to get all of their tasks done on time." (P223-HHA)*
> *"I liked it [experience with the agent teammate], and trusted it more than myself and my human teammate." (P206-HHA)*
> *"It [the agent] displayed helpful abilities to the team." (P204-HHA)*

From these quotes, it is clear that human teammates are willing and actively looking to their agent teammates for examples of effective taskwork when their shared goals are clearly defined. Therefore, agents should be designed to set examples for human teammates in various facets of taskwork and even communication. If such features are deployed, it will help unify the team's collective strategy and speed up team effectiveness, giving HATs the ability to form similar shared mental models in task spaces rapidly. These quotes also show that participants had positive perceptions of the agent's abilities in communication and task performance, which had a positive effect on their overall experience.

Unfortunately, facilitating team cognition from the start with clearly defined shared goals may not always be enough for all individuals or teams as individual differences may lead some teammates to doubt the agent in some way. This doubt or lack of understanding may lead human teammates to ignore a dialogue with the agent despite its repeated communications, leading such teams to perceive their agent teammate as a black box entity:

> *"Can't really say [that the human team members paid attention to the agent teammate]. You can't see what the AI is doing behind the scenes." (P222-HHA)*

"No [everyone did not think about cooperating and responding to events the same because] everyone had their own ways of thinking about the events." (P210-HHA)
"I didn't really communicate that much [with my agent teammates]" (P303-HAA)

The quotes above illustrate that even if agents can provide the action-based communication utilized in the current study, there are still certain teams that cannot develop a shared understanding with their agent teammates. P210 illustrated that they felt every teammate had individual strategies to complete the simulation, while P303 indicated that they did not even communicate with their agent teammates (despite the agent teammates communicating with them).

While clearly defined shared goals between humans and agents represent an excellent starting point for team cognition to begin conceptualizing, this is not enough for some HATs. This problem may be associated with certain individual differences and is an issue for certain HATs, as seen in the following quotes:

"I feel that since I was the only teammate that wasn't an AI, I had to think harder and more about the task." (P308-HAA)
"I feel the human players acted on their own for a large part of the experiment, so while we were working towards the same goal, without much experience in the game, it is hard to say cooperation was very high." (P223-HHA)
"It seemed like myself and my human teammate responded similarly, but the AI was much more confident in its actions." (P205-HHA)

These quotes display the effect that individual differences, such as biases and assumptions regarding AI, impact team cognition development. Because of this shortcoming, clearly defined shared goals should be coupled with high levels of transparency separate from the communication of intent and action-related communications used in this study.

5 DISCUSSION

By manipulating team composition in teams completing the NeoCITIES task simulation and collecting data on their shared mental models, performance, and trust, a holistic and detailed picture of team cognition development can be gathered. Our findings have the following highlights in response to the research questions: RQ1) Team mental model similarity levels did not differ significantly between the HHH and HHA conditions; however, the HHA condition had significantly higher variance than the HHH condition indicating a greater inconsistency in the HHA teams ability to develop team mental models. Qualitative results indicated that team cognition is a highly iterative process greatly accelerated by communication for traditional human-human teams and HATs alike, however, action-based communication and explicitly shared goals were of much greater importance to HATs than traditional human-human teams when developing team cognition; RQ2.1) Objective performance results saw teams perform incrementally better with the addition of more autonomous agent teammates, while the trend of perceived performance was not as consistent; RQ2.2) Human teammates perceived significantly more team cognition with their human teammates than their agent teammates; RQ2.3) Human teammates trusted their agent teammates significantly more when they had one human and one agent teammate than when they had only two agent teammates and this result suggests that the addition of a second agent teammate lowered their trust in the agent despite the teams achieving objectively higher performance. The following discussion highlights the implications of these findings to advance existing CSCW research on HATs and team cognition while also identifying design recommendations for agents to enhance HAT team cognition development. It also explains the limitations of the current research, which identify and inform areas for future CSCW research.

5.1 New Perspectives of Human-Agent Team Cognition through the Lens of Team Composition

While society has already begun integrating HATs into the workforce, their effectiveness will be severely limited without considering the human element within teaming. The current study results indicate that several novel aspects of the human experience within teams play a significant role in forming team cognition in HATs, and the formation of team cognition subsequently impacts the human experience. These insights thus add to the current CSCW knowledge on HATs and team cognition by providing new perspectives of human-agent team cognition through the lens of team composition, including: 1) how team composition may have adverse effects on team cognition outcomes; and 2) the crucial role of individual differences in humans in the formation of HATs shared mental models.

5.1.1 Team Composition Can Have Negative Effects on Team Cognition Outcomes. Specifically, the quantitative results of the study identify a disconnect between objective and perceived team performance trends when comparing the two. While the difference in average perceived performance between the HHA and HAA conditions was not significant, HAA teams perceived their performance much more inconsistently than HHA teams, showing significant heteroscedasticity between the two. It is possible that being a minority member of a team led the participants in the HAA condition to misjudge their teammates' performance, which echos findings in human-human teaming where teams with minority members had lower performance ratings by the internal members despite external observers noting no such differences [5]. These results provide further evidence that the inclusion of AI teammates may lead humans to create negative in-group, out-group dynamics in HATs. This assertion posits that human team members have a discriminatory bias against AI, which is supported by past research [77], but this bias does not extend to judgments of its ability [37]. This negative effect from bias may be especially prevalent in HATs with humans in the minority or have only a single human member, as seen in the current study's trends.

Additionally, the HHA teams reported significantly poorer perceived team cognition with their agent teammate than their human teammate. This result is notable as it helps explain the significant variance shown in the HHA condition's team mental models. This assertion is bolstered further by the finding that trust in AI teammate(s), which is a byproduct of positive team interaction and team cognition [17, 23], was significantly lower in HAA teams than in HHA teams. The qualitative data provided additional insight by revealing that many HAT members reported a significant disconnect between the two human teammates and the agent teammate. Practically, such results mean that HATs could suffer from dissatisfaction with their team and teammates, reduced effectiveness, and a lack of shared understanding, making it difficult for team cognition to manifest. Applied HATs in manufacturing roles could be facing a significantly uphill battle as many factories seek to pair a single human with multiple agent teammates [24, 83]. Choosing how, where, and when to make humans the minority team member with AI teammates should be a careful practice coupled with adherence to the most effective interventions identified here and in the literature to help counter the adverse effects identified while also enhancing trust and communication and coordination. However, it has been shown in prior research that positive prior experiences with agents can increase humans trust in the agent [32, 33, 83], as such, this finding may change if participants were to go through multiple teaming experiences with the agent like many real world HATs do.

5.1.2 Individual Differences in Humans Play a Key Role in the Formation of Team Cognition in HATs. Results directly addressing team cognition through shared mental models showed no significant differences between the HHH and HHA conditions in similarity levels; however, HHA teams did have significantly higher variance (more inconsistent) in team mental model similarity compared

to human-human teams. While it is a positive result that HHA teams could develop shared mental models to the same level as HHH teams, it is concerning that their team mental models were less consistent. HATs in practice may have unpredictable teamwork efficacy because of this inconsistency [60], and this result indicates that there are likely additional mechanisms that affect how HATs develop their team mental models.

The qualitative data sheds light on some potential factors at play as several HHA teams reported instances where human teammates did not utilize explicit communication, did not clearly understand the agent's goals or perceived the agent as separate from themselves and their human teammate. Alternatively, other teams reported that the communication of their agent sparked helpful communication amongst the entire team, that they trusted their agent teammate the most, or that they used the agent's actions as guidance for themselves. These diametric results display that the importance of individual differences, which have been a vital topic in CSCW [29], play a pivotal role in the efficacy of HATs, especially in regard to the formation of team cognition. This study demonstrates how these individual differences may lead to contradictory perspectives from humans regarding the HATs they operate within. These insights can aid in the deployment and development of future HATs by informing CSCW researchers and practitioners how team composition in HATs can both negatively and positively impact the human element of HATs based on the specific perspectives of the prospective human teammates. As measured by mental model similarity, the inclusion of team cognition becomes all the more important in HATs as it allows for an empirical quantification of the variation that may exist between teammates due to individual differences. Therefore, CSCW researchers and practitioners should consistently consider team cognition and shared mental models to build HATs that can overcome individual differences to build a more cohesive and effective team.

5.2 Design Recommendations for Agents to Enhance HAT Team Cognition Development

Grounded in our findings, we propose three design recommendations that both CSCW researchers and practitioners can use as leverage to produce more effective HATs and overcome some of the adverse effects of team composition on team cognition. These design recommendations are centered around communication, which is not surprising given how important it is to developing team cognition [13]. Artificial agents also face significant challenges to effective communication given the current struggles of natural language processing [102], meaning HATs struggle in this area without effective design. The current design recommendations are essential and timely to the current literature as they serve to enhance a critical facet of team cognition development (communication) within an environment that historically faces extraordinary challenges to effective communication.

5.2.1 Agent Teammates Should Point Out Exemplar Behavior to Accelerate The Development of Team Cognition. Our study's quantitative and qualitative findings support the assertion that agents working in HATs can enhance their team's effectiveness and team cognition by being an exemplar for their teammates and explicitly stating this feature. Direct quotes reveal human members of HATs indicating that they trusted their agent teammate more than anyone else on the team (including themselves), considered them the best team member, and displayed beneficial abilities to their team. Agents in HATs should capitalize on this sentiment and leverage their strengths to initiate the formation of team cognition. In agent design, this would be accomplished by designing the agent to explicitly state that they can be seen as an exemplar for learning to complete the task effectively and do so early on in the task. It should be stated, however, that the current study did not utilize any natural language processing technology, and all communication was conducted using a script and the Wizard of Oz methodology. As such, the following design recommendation considers the

difficulty and practicality of natural language processing and only suggests predefined automated communication snippets. Regardless, by providing human teammates with an exemplar of practical strategies and tactics and calling attention to them via predefined communication snippets, the team can develop faster, become more effective team members, and coalesce towards a more robust shared task mental model. At the same time, this design feature would serve to break the ice in team communication and may potentially lead to more communication overall. The current study initiated communication from the agent in a similar manner by calling attention to the agent's decision to send a resource to an event, "Sending investigator to the Football Weekend Briefing."

In practice, this would involve designing agents to adhere to high levels of performance and effective strategy (as they would typically), but also by pointing out to human teammates when they are engaging in these behaviors by saying, "I just sent my investigator to the Smoking Kills event because my resource was the closest." This way, teammates know why a decision was made and the strategy behind it, providing examples of how to operate for their human teammates who may still need help while also enhancing the explainability of the AI. The content of these predefined communication snippets and what action would trigger them would need to be determined by a collaboration between the developers, project managers, and users to ensure the design feature is practical and feasible.

5.2.2 Agent Teammate Communication Should Center Around Needed, In Progress, and Completed Actions. Because team cognition was identified as an iterative process in HATs, communication is a critical factor in accelerating its formation in HATs. The results overwhelmingly indicated that communication rapidly accelerated team cognition development within teams and indicated that communication related to action events was incredibly beneficial. These findings advance team cognition literature by identifying a specific type of communication that significantly contributes to team cognition development in HATs. From a design perspective, agents should be designed to provide short, action-related communication that updates humans on actions that need to be done, actions that are currently in progress, and completed actions. This dialogue should be associated with spatial and temporal information while also happening in concert with implicit communication done through task-related actions [52]. The frequency and timing of these communications are left to the designer as these decisions should be made based on the specific task; however, their utilization should still begin early in a team's lifespan to ensure the requirements of the first design recommendation are still met. Implementing these design recommendations allows humans working within HATs to understand their agent teammates better, develop more effective team cognition, and enhance trust in their agent teammates through the cross-validation of actions agents take and their clear communication regarding teaming actions. Agent communication should be designed to be effective and action-based, "I am sending my Ambulance to the Luncheon Nausea event." This type of communication should change based on the task but and should not over-saturate the communication feed.

5.2.3 AI Teammates Should Explicitly Utilize Shared Goals During Communication to Accelerate Team Cognition's Formation. Clearly defined individual expectations and shared goals also represent a significant leverage point for HATs attempting to develop team cognition. The teams in the current study identified with the agent(s) when they understood its expectations and saw its goals as being aligned with their own. Specifically, teams reported that when they felt the agent shared their own goals, they identified it as a contributing factor to the development of team cognition. As previously stated, human teammates find agent teammates fundamentally different from themselves, which is backed up in the current study's data. Therefore, designing to emphasize shared goals should encompass the following aspects: 1) the agent should convey its individual goals to the team clearly and concisely; 2) the agent should emphasize how its individual goals integrate with other team

goals; and 3) these details should only be communicated at the beginning of team formation unless explicitly asked for again. If done correctly, this design should act as a type of agent equivalent of the "norming" stage seen in traditional human-human teams [98]. Through practice, the team will better understand what the agent is working towards and how it contributes to their shared goal, connecting two fundamentally different types of teammates through their shared tasks. For example, the agent in NeoCITIES could be improved by clearly stating its goals and how they overlapped, "My goals are to send resources to events as efficiently as possible to complete as many events successfully as we can. I cannot do this without everyone's help and we must work together to complete joint events as they occur."

5.3 Future Research and Limitations

One limitation of this study was that only a limited amount of qualitative data was collected in the HAA condition compared to the other two, as only 10 participants operated from this condition (a consequence of the experimental design). More qualitative data would likely uncover additional themes relating to the differences between the HHA and HAA conditions and should be explored in future research. Additionally, the current study was unable to collect structural information on the shared mental models of the HAA condition because there were no other human teammates to compare. This limitation is a consequence of both the current study and the measurement methods selected for the shared mental model, even if it is the most robust measurement. Also, due to this limitation, the current study was unable to measure the mental models of agent teammates and was only able to characterize the team cognition of human team members. While perceived team cognition can help make up for this limitation, measuring shared mental model content is no substitute for proper measures of shared mental models [68]. The finding that the agent initiating communication helps develop team cognition should also be interpreted with the limitation that the current study was designed to have the agent communicate first, potentially leading towards this finding. As alluded to in the above discussion, research has been conducted showcasing that perceptions of agent teammates can be improved through positive prior experiences with agents. As such, the current results focus on teams that only interacted for a single teamwork session, and participant perceptions could change if given multiple teaming sessions. Participants also appeared to perceive the NeoCITIES simulation as a game, which could alter their perceptions of the agent and team when compared to real-world HATs, but this is a common limitation and trade-off of simulated task environments with high internal validity.

Finally, the current study presents a host of exciting avenues for additional research to investigate. Reliability in agents operating as full team members is an area of the team cognition literature with very little prior research. While the current study indicated that reliability was critical to team cognition, it is possible that with adequate transparency, any adverse effects of poor reliability could be mitigated, just as past research on decision aid agents has found [66]. The effects of agent teammates' role in communication and communication development as it relates to team cognition development should also be a topic of future research as the current study found evidence for its importance. Additional future research should investigate if, as stated above, participants perceptions change when given the opportunity to complete multiple teaming sessions and if perceptions change when participants are given the opportunity to complete teaming sessions in different HAT compositions (HHA and HAA). Finally, future research should also seek to develop a validated methodology for measuring the mental models of agent teammates, or at the least, disambiguation of the construct that agents possess that can be compared to its human teammates.

6 CONCLUSION

The burgeoning CSCW literature on HATs motivating this research has yet to empirically characterize team cognition, let alone compare the similarities and differences in its development between traditional human-human teams and HATs. The current study was conducted to address this significant gap and explore the effects of team composition in HATs on various outcomes of team cognition like trust and perception. The study found that team cognition in HATs develops similarly in its iterative quality and the accelerative effects of communication, but HATs valued action-related communication and explicit shared goals much more than human-human teams. HATs with only one human (HAA) had inconsistent judgements of their performance and trusted their agent teammates less than HATs with two humans (HHA), while HATs with both types of teammates (HHA) perceived less team cognition with their agent teammate than their human teammate. These findings are essential to the CSCW literature as they: 1) provide a characterization of team cognition and its development in HATs for the first time; 2) outline the effects of team composition in HATs on team cognition's outcomes; and 3) provide specific design recommendations to applied stakeholders directed at mitigating a weakness inherent to HATs (communication). The study leaves future research directions for the field in further contextualizing the nuanced effects of team composition on HATs and the various complexities of communication, both explicit and implicit.

7 ACKNOWLEDGEMENTS

This work was supported by ONR Award N000142112336 and AFOSR Award FA9550-20-1-0342 (Program Manager: Laura Steckman). Thank you to Julie Schelble for her valuable feedback on the paper.

REFERENCES

[1] Zaheer Allam and Zaynah A Dhunny. 2019. On big data, artificial intelligence and smart cities. *Cities* 89 (2019), 80–91.

[2] Mandy M Archibald, Rachel C Ambagtsheer, Mavourneen G Casey, and Michael Lawless. 2019. Using zoom videoconferencing for qualitative data collection: perceptions and experiences of researchers and participants. *International Journal of Qualitative Methods* 18 (2019), 1609406919874596.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[5] S Gayle Baugh and George B Graen. 1997. Effects of team gender and racial composition on perceptions of team performance in cross-functional teams. *Group & Organization Management* 22, 3 (1997), 366–383.

[6] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[7] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012).

[8] Janis A Cannon-Bowers and Eduardo Salas. 2001. Reflections on shared cognition. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 22, 2 (2001), 195–202.

[9] Lorenzo Barberis Canonico, Christopher Flathmann, and Dr. Nathan McNeese. 2019. Collectively Intelligent Teams: Integrating Team Cognition, Collective Intelligence, and AI for Future Teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63, 1 (2019), 1466–1470. https://doi.org/10.1177/1071181319631278

[10] Herbert H Clark and Edward F Schaefer. 1987. Collaborating on contributions to conversations. *Language and cognitive processes* 2, 1 (1987), 19–41.

[11] Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Academic press.

[12] Gregorio Convertino, Helena M Mentis, Mary Beth Rosson, Aleksandra Slavkovic, and John M Carroll. 2009. Supporting content and process common ground in computer-supported teamwork. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2339–2348.

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

[14] Nancy J Cooke, Preston A Kiekel, Eduardo Salas, Renée Stout, Clint Bowers, and Janis Cannon-Bowers. 2003. Measuring team knowledge: A window to the cognitive underpinnings of team performance. *Group Dynamics: Theory, Research, and Practice* 7, 3 (2003), 179.

[15] Nancy J Cooke, Eduardo Salas, Janis A Cannon-Bowers, and Renee J Stout. 2000. Measuring team knowledge. *Human factors* 42, 1 (2000), 151–173.

[16] Chris W Coultas, Tripp Driskell, C Shawn Burke, and Eduardo Salas. 2014. A conceptual review of emergent state measurement: Current problems, future solutions. *Small Group Research* 45, 6 (2014), 671–703.

[17] William DeLone, J Alberto Espinosa, Gwanhoo Lee, and Erran Carmel. 2005. Bridging global boundaries for IS project success. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE, 48b–48b.

[18] Mustafa Demir, Nancy J Cooke, and Polemnia G Amazeen. 2018. A conceptual model of team dynamical behaviors and performance in human-autonomy teaming. *Cognitive Systems Research* 52 (2018), 497–507.

[19] Mustafa Demir, Nathan J McNeese, and Nancy J Cooke. 2017. Team situation awareness within the context of human-autonomy teaming. *Cognitive Systems Research* 46 (2017), 3–12.

[20] Mustafa Demir, Nathan J McNeese, and Nancy J Cooke. 2020. Understanding human-robot teams in light of all-human teams: Aspects of team interaction and shared cognition. *International Journal of Human-Computer Studies* 140 (2020), 102436.

[21] Xiaocong Fan, Shuang Sun, Michale McNeese, and John Yen. 2005. Extending the recognition-primed decision model to support human-agent collaboration. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. 945–952.

[22] Xiaocong Fan and John Yen. 2010. Modeling cognitive loads for evolving shared mental models in human–agent collaboration. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41, 2 (2010), 354–367.

[23] Rosemarie Fernandez, Sachita Shah, Elizabeth D Rosenman, Steve WJ Kozlowski, Sarah Henrickson Parker, and James A Grand. 2017. Developing team cognition: a role for simulation. *Simulation in healthcare: journal of the Society for Simulation in Healthcare* 12, 2 (2017), 96.

[24] Christopher Flathmann, Nathan McNeese, and Lorenzo Barberis Canonico. 2019. Using Human-Agent Teams to Purposefully Design Multi-Agent Systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 1425–1429.

[25] Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 122–130.

[26] Helen Gavin. 2008. Thematic analysis. *Understanding research methods and statistics in psychology* (2008), 273–282.

[27] Jamie C. Gorman and Nancy J. Cooke. 2011. Changes in team cognition after a retention interval: the benefits of mixing it up. *Journal of Experimental Psychology: Applied* 17, 4 (2011), 303.

[28] Lia M Gray, Gina Wong-Wylie, Gwen R Rempel, and Karen Cook. 2020. Expanding qualitative research interviewing strategies: Zoom video communications. *The Qualitative Report* 25, 5 (2020), 1292–1301.

[29] Jonathan Grudin. 1988. Why CSCW applications fail: problems in the design and evaluationof organizational interfaces. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*. 85–93.

[30] Jonathan Grudin. 1994. Computer-supported cooperative work: History and focus. *Computer* 27, 5 (1994), 19–26.

[31] Pranav Gupta and Anita Williams Woolley. 2018. Productivity in an era of multi-teaming: The role of information dashboards and shared cognition in team performance. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.

[32] Feyza Merve Haf? ızoğlu and Sandip Sen. 2018. The Effects of Past Experience on Trust in Repeated Human-Agent Teamwork. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 514–522.

[33] Feyza Merve Hafizoglu and Sandip Sen. 2018. Reputation Based Trust In Human-Agent Teamwork Without Explicit Coordination. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. 238–245.

[34] Katherine Hamilton, Vincent Mancuso, Dev Minotra, Rachel Hoult, Susan Mohammed, Alissa Parr, Gaurav Dubey, Eric McMillan, and Michael McNeese. 2010. Using the NeoCITIES 3.1 simulation to study and measure team cognition. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 54. SAGE Publications Sage CA: Los Angeles, CA, 433–437.

[35] Nader Hanna and Deborah Richards. 2014. The impact of communication on a human-agent shared mental model and team performance. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1485–1486.

[36] Nader Hanna and Deborah Richards. 2015. The impact of virtual agent personality on a shared mental model with humans during collaboration. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. 1777–1778.

[37] Yugo Hayashi and Kosuke Wakabayashi. 2017. Can AI become reliable source to support human decision making in a court scene?. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 195–198.
[38] D Benjamin Hellar and David L Hall. 2009. NeoCITIES: an experimental test-bed for quantifying the effects of cognitive aids on team performance in C2 situations. In *Modeling and Simulation for Military Operations IV*, Vol. 7348. International Society for Optics and Photonics, 73480K.
[39] D Benjamin Hellar and Michael McNeese. 2010. NeoCITIES: A simulated command and control task environment for experimental research. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 54. SAGE Publications Sage CA: Los Angeles, CA, 1027–1031.
[40] Sean D Holcomb, William K Porter, Shaun V Ault, Guifen Mao, and Jin Wang. 2018. Overview on deepmind and its alphago zero ai. In *Proceedings of the 2018 international conference on big data and education*. 67–71.
[41] Edwin Hutchins. 1991. The social organization of distributed cognition. (1991).
[42] Leila J Johannesen, Richard I Cook, and David D Woods. 1994. Cooperative communications in dynamic fault management. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 38. SAGE Publications Sage CA: Los Angeles, CA, 225–229.
[43] Douglas H Johnson. 1999. The insignificance of statistical significance testing. *The journal of wildlife management* (1999), 763–772.
[44] Brennan Jones, Anthony Tang, and Carman Neustaedter. 2020. Remote communication in wilderness search and rescue: implications for the design of emergency distributed-collaboration tools for network-sparse environments. *Proceedings of the ACM on human-computer interaction* 4, GROUP (2020), 1–26.
[45] Rashaad ET Jones, Michael D McNeese, Erik S Connors, Tyrone Jefferson Jr, and David L Hall Jr. 2004. A distributed cognition simulation involving homeland security and defense: The development of NeoCITIES. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 48. SAGE Publications Sage CA: Los Angeles, CA, 631–634.
[46] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W Malone. 2017. What makes a strong team? Using collective intelligence to predict team performance in League of Legends. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2316–2329.
[47] Richard Klimoski and Susan Mohammed. 1994. Team mental model: Construct or metaphor? *Journal of management* 20, 2 (1994), 403–437.
[48] Kurt Kraiger and Lucy H Wenzel. 1997. Conceptual development and empirical evaluation of measures of shared mental models as indicators of team effectiveness. *Team performance assessment and measurement: Theory, methods, and applications* 63 (1997), 84.
[49] Alex Leavitt, Brian C Keegan, and Joshua Clark. 2016. Ping to win? non-verbal communication and team performance in competitive online multiplayer games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4337–4350.
[50] Miyoung Lee and Tristan E Johnson. 2008. Understanding the effects of team cognition associated with complex engineering tasks: Dynamics of shared mental models, Task-SMM, and Team-SMM. *Performance Improvement Quarterly* 21, 3 (2008), 73–95.
[51] Jeffery A LePine, Ronald F Piccolo, Christine L Jackson, John E Mathieu, and Jessica R Saul. 2008. A meta-analysis of teamwork processes: tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel psychology* 61, 2 (2008), 273–307.
[52] Claire Liang, Julia Proft, Erik Andersen, and Ross A Knepper. 2019. Implicit communication of actionable information in human-ai teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[53] Shu-Hsien Liao. 2005. Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert systems with applications* 28, 1 (2005), 93–103.
[54] Beng-Chong Lim and Katherine J Klein. 2006. Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 27, 4 (2006), 403–418.
[55] Jessica Lipnack and Jeffrey Stamps. 1999. Virtual teams: The new way to work. *Strategy & Leadership* (1999).
[56] Fabrice Lumineau. 2017. How contracts influence trust and distrust. *Journal of management* 43, 5 (2017), 1553–1577.
[57] Jane T Malin, Debra L Schreckenghost, David D Woods, Scott S Potter, Leila Johannesen, Matthew Holloway, and Kenneth D Forbus. 1991. Making intelligent systems team players: Case studies and design issues. Volume 1: Human-computer interaction design. (1991).
[58] Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. 2001. A temporally based framework and taxonomy of team processes. *Academy of management review* 26, 3 (2001), 356–376.
[59] John Mathieu, M Travis Maynard, Tammy Rapp, and Lucy Gilson. 2008. Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of management* 34, 3 (2008), 410–476.

[60] 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.

[61] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 277–284.

[62] Michael D McNeese, Vincent F Mancuso, Nathan J McNeese, Tristan Endsley, and Pete Forster. 2014. An integrative simulation to study team cognition in emergency crisis management. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58. SAGE Publications Sage CA: Los Angeles, CA, 285–289.

[63] Nathan McNeese, Mustafa Demir, Erin Chiou, Nancy Cooke, and Giovanni Yanikian. 2019. Understanding the role of trust in human-autonomy teaming. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

[64] Nathan J McNeese, Mustafa Demir, Nancy J Cooke, and Christopher Myers. 2018. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors* 60, 2 (2018), 262–273.

[65] Nathan J. McNeese, Beau G. Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. 2021. Who/What Is My Teammate? Team Composition Considerations in Human–AI Teaming. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 288–299. https://doi.org/10.1109/THMS.2021.3086018

[66] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human Factors* 58, 3 (2016), 401–415.

[67] Tim R Merritt, Kian Boon Tan, Christopher Ong, Aswin Thomas, Teong Leong Chuah, and Kevin McGee. 2011. Are artificial team-mates scapegoats in computer games. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 685–688.

[68] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton. 2010. Metaphor no more: A 15-year review of the team mental model construct. *Journal of management* 36, 4 (2010), 876–910.

[69] Susan Mohammed, Richard Klimoski, and Joan R Rentsch. 2000. The measurement of team mental models: We have no shared schema. *Organizational Research Methods* 3, 2 (2000), 123–165.

[70] Geoff Musick, Thomas A O'Neill, Beau G Schelble, Nathan J McNeese, and John B Henke. in press. What Happens When Humans Believe Their Teammate is an AI? An Investigation into Humans Teaming with Autonomy. *Computers in Human Behavior* (in press).

[71] Geoff Musick, Rui Zhang, Nathan J McNeese, Guo Freeman, and Anurata Prabha Hridi. 2021. Leveling Up Teamwork in Esports: Understanding Team Cognition in a Dynamic Virtual Environment. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–30.

[72] Stefanos Nikolaidis and Julie Shah. 2013. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 33–40.

[73] Ashley A Niler, Jessica R Mesmer-Magnus, Lindsay E Larson, Gabriel Plummer, Leslie A DeChurch, and Noshir S Contractor. 2020. Conditioning team cognition: A meta-analysis. *Organizational Psychology Review* (2020), 2041386620972112.

[74] Christopher Ong, Kevin McGee, and Teong Leong Chuah. 2012. Closing the human-AI team-mate gap: how changes to displayed information impact player behavior towards computer teammates. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*. 433–439.

[75] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2020. Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors* (2020), 0018720820960865.

[76] Laura J Pape. 1998. *The effect of personality characteristics on team member schema similarity*. Ph.D. Dissertation. Wright State University.

[77] Martin Ragot, Nicolas Martin, and Salomé Cojean. 2020. AI-generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence?. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–10.

[78] Joan Rentsch, Michael D McNeese, Laura J Pape, Dawn D Burnett, and Darcy M Menard. 1998. *Testing the effects of team processes on team member schema similarity and team performance: examination of the Team Member Schema Similarity model*. Technical Report. WRIGHT STATE UNIV DAYTON OH DEPT OF PSYCHOLOGY.

[79] Joan R Rentsch and Richard J Klimoski. 2001. Why do 'great minds' think alike?: Antecedents of team member schema agreement. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 22, 2 (2001), 107–120.

[80] Parya Saberi. 2020. Research in the time of coronavirus: continuing ongoing studies in the midst of the COVID-19 pandemic. *AIDS and Behavior* 24, 8 (2020), 2232–2235.

[81] Catarina Marques Santos, Sjir Uitdewilligen, and Ana Margarida Passos. 2015. A temporal common ground for learning: The moderating effect of shared mental models on the relation between team learning behaviours and performance improvement. *European Journal of Work and Organizational Psychology* 24, 5 (2015), 710–725.

[82] Beau Schelble, Lorenzo-Barberis Canonico, Nathan McNeese, Jack Carroll, and Casey Hird. 2020. Designing Human-Autonomy Teaming Experiments Through Reinforcement Learning. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 64. SAGE Publications Sage CA: Los Angeles, CA, 1426–1430.

[83] Beau G Schelble, Christopher Flathmann, and Nathan McNeese. 2020. Towards Meaningfully Integrating Human-Autonomy Teaming in Applied Settings. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. 149–156.

[84] Amber N Schroeder, Patrick J Rosopa, Julia H Whitaker, Ian N Fairbanks, and Phoebe Xoxakos. 2020. HETEROSCEDASTICITY IN ORGANIZATIONAL RESEARCH. *Research Methods in Human Research Management: Toward Valid Research-Based Inferences* (2020), 67.

[85] Roger W Schvaneveldt. 1990. *Pathfinder associative networks: Studies in knowledge organization*. Ablex Publishing.

[86] Julie Shah, Been Kim, and Stefanos Nikolaidis. 2012. Human-inspired techniques for human-machine team planning. In *2012 AAAI Fall Symposium Series*.

[87] Mark K Singley, Moninder Singh, Peter Fairweather, Robert Farrell, and Steven Swerling. 2000. Algebra jam: supporting teamwork and managing roles in a collaborative learning environment. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 145–154.

[88] JA Smith and M Osborn. 2003. Chapter 4: Interpretive phenomenological analysis. *Qualitative psychology: A practical guide to methods* (2003), 53–80.

[89] Harlan E. Spotts and Anthony F. Chelte. 2005. Evaluating the Effects of Team Composition and Performance Environment on Team Performance. *Journal of Behavioral & Applied Management* 6, 2 (2005).

[90] Jan E Stets and Peter J Burke. 2000. Identity theory and social identity theory. *Social psychology quarterly* (2000), 224–237.

[91] Renée J Stout, Janis A Cannon-Bowers, Eduardo Salas, and Dana M Milanovich. 1999. Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors* 41, 1 (1999), 61–71.

[92] Anselm L Strauss. 1987. *Qualitative analysis for social scientists*. Cambridge university press.

[93] Zachary O Toups, Andruid Kerne, William Hamilton, and Alan Blevins. 2009. Emergent team coordination: From fire emergency response practice to a non-mimetic simulation game. In *Proceedings of the ACM 2009 international conference on Supporting group work*. 341–350.

[94] Jeroen Vaes, Maria Paola Paladino, Luigi Castelli, Jacques-Philippe Leyens, and Anna Giovanazzi. 2003. On the behavioral consequences of infrahumanization: The implicit role of uniquely human emotions in intergroup relations. *Journal of personality and social psychology* 85, 6 (2003), 1016.

[95] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016).

[96] Daniel M Wegner. 1987. Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior*. Springer, 185–208.

[97] Rina R Wehbe, Edward Lank, and Lennart E Nacke. 2017. Left them 4 dead: Perception of humans versus non-player character teammates in cooperative gameplay. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 403–415.

[98] Susan A Wheelan. 1994. *Group processes: A developmental perspective*. Allyn & Bacon.

[99] H James Wilson and Paul R Daugherty. 2018. Collaborative intelligence: humans and AI are joining forces. *Harvard Business Review* 96, 4 (2018), 114–123.

[100] Anna Wu, Xiaolong Zhang, Gregorio Convertino, and John M Carroll. 2009. CIVIL: support geo-collaboration with information visualization. In *Proceedings of the ACM 2009 international conference on Supporting group work*. 273–276.

[101] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[102] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine* 13, 3 (2018), 55–75.

[103] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An Ideal Human" Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

[104] Bin Zheng, Nasim Hajari, and M Stella Atkins. 2016. Revealing team cognition from dual eye-tracking in the surgical setting. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. 321–322.

Received July 2021; revised September 2021; accepted October 2021