TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 00 (2023) 1–25

ISSN: 1756-8765 online
DOI: 10.1111/tops.12648

This article is part of the topic "Building the Socio-Cognitive Architecture of COHUMAIN: Collective Human-Machine Intelligence," Cleotilde Gonzalez, Henny Admoni, Anita W. Woolley and Scott Brown (Topic Editors).

Establishing Human Observer Criterion in Evaluating Artificial Social Intelligence Agents in a Search and Rescue Task

Lixiao Huang,[a] Jared Freeman,[b] Nancy J. Cooke,[a] Myke C. Cohen,[a] Xiaoyun Yin,[a] Jeska Clark,[a] Matt Wood,[b] Verica Buchanan,[a] Christopher Corral,[a] Federico Scholcover,[a] Anagha Mudigonda,[a] Lovein Thomas,[a] Aaron Teo,[a] John Colonna-Romano[b]

[a]*Center for Human, Artificial Intelligence, and Robot Teaming, Arizona State University*
[b]*Aptima, Inc*

Received 1 June 2022; received in revised form 18 March 2023; accepted 20 March 2023

Abstract

Artificial social intelligence (ASI) agents have great potential to aid the success of individuals, human–human teams, and human–artificial intelligence teams. To develop helpful ASI agents, we created an urban search and rescue task environment in Minecraft to evaluate ASI agents' ability to infer participants' knowledge training conditions and predict participants' next victim type to be rescued. We evaluated ASI agents' capabilities in three ways: (a) comparison to ground truth—the actual knowledge training condition and participant actions; (b) comparison among different ASI agents; and (c) comparison to a human observer criterion, whose accuracy served as a reference point. The human observers and the ASI agents used video data and timestamped event messages from the testbed, respectively, to make inferences about the same participants and topic (knowledge training condition) and the same instances of participant actions (rescue of victims). Overall, ASI agents performed better than human observers in inferring knowledge training conditions and predicting actions. Refining the human

Correspondence should be sent to Lixiao Huang, Center for Human, Artificial Intelligence, and Robot Teaming, Arizona State University, Room 167A, ISTB3, 7418 E. Innovation Way South, Mesa, AZ 85212, USA. E-mail: Lixiao.Huang@asu.edu

criterion can guide the design and evaluation of ASI agents for complex task environments and team composition.

Keywords: Artificial social intelligence; Theory of mind; Evaluation; Human observer criterion; Baseline; Minecraft; Search and rescue

1. Introduction

In the two tragic airline accidents involving the Boeing 737 Max 8, it seems likely that advanced automation or artificial intelligence (AI) system failed human teammates. News reports stated that automation on each aircraft (Maneuvering Characteristics Augmentation System or MCAS) incorrectly predicted an engine stall, took control of flight surfaces, and pitched the aircraft into the earth; 346 people died (Hamblen, 2020). Imagine if a more advanced AI system could infer the pilots' beliefs about the system status and its cause, predict what pilots would do to restore stable flight, and give pilots instructions on how to diagnose and solve the problems. Flight operations would be much safer, and some disasters would likely be avoided.

The ability to infer a human's states (e.g., beliefs, emotions, knowledge) relates to the level two cognitive process in Endsley's (1995) situation awareness framework—comprehension, and the ability to predict their future state associates with the level three cognitive process—prediction; both influence individual performance. Related laboratory research found precursors to disasters in the brittle interactions between a synthetic pilot and the crew of a simulated uninhabited aerial vehicle (UAV; McNeese et al., 2018). In the McNeese et al. (2018) study, a communicative synthetic pilot failed to anticipate and adapt to the information needs of human teammates, and team performance suffered. In these real-world and laboratory cases, intelligent systems lacked the social intelligence to recognize the beliefs of the operators and failed to assist the human team members in achieving joint goals. For AI agents to assist humans in succeeding in critical missions, they should first be able to infer the state of humans and predict their actions.

There is limited literature on ASI agents' dynamic inferences and predictions. This study aims to understand humans and ASI agents' strengths and weaknesses through different ways of accuracy evaluations in a search and rescue task environment. As the first attempt to develop and evaluate ASI agents' capabilities in complex task environments, this research provides empirical data to the literature on ASI agent development and evaluation withone human player in a dynamic task environment.

The following sections include a brief review of the theoretical background of artificial social intelligence (ASI), a description of the experiment, the process of establishing the human observer criterion of inferring players' states and predicting their next actions, and the evaluation of ASI agents and human observers' accuracy in inferences and predictions.

1.1. Social intelligence and Artificial Social Intelligence

Thorndike (1920, p. 228) originally defined humansocial intelligence as "… the ability to understand and manage men and women, boys and girls—to act wisely in human relations." This definition includes the cognitive understanding of other individuals and behaviorally interacting with them appropriately—at the right time in the right way for a good outcome. Likewise, many researchers defined social intelligence in terms of behavioral effectiveness in social situations (Ford & Tisak, 1983; Walker & Foley, 1973). Barnes & Sternberg (1989, p. 263) defined social intelligence partly as inferring human relationships through "the ability to accurately decode social information." The goal of accurately decoding social information is to behave effectively in social situations, so both understanding and actingare important.

Appropriate social tasks are required to develop effective ASI agents to test their capabilities. Thus, we reviewed tasks that exercise varying aspects of human social intelligence. The first aspect is the ability to infer a person's habitual dispositions based on the facts of a certain group of people to whom the person belongs (Frith and Frith, 2006). People may judge on the impression of warmth, competence, and morality (Fiske et al., 2007, Wojciszke, 1994). One study used telephone surveys, describing a typical group of people and asking for participants' responses (Cuddy et al., 2007). Another study used pictures to ask participants to judge (a) whether a couple in a picture was in a real relationship or just two strangers and (b) which of the two people in a picture was the supervisor (Barnes & Sternberg, 1989). The relationship judgment study carefully examined the accuracy of participants' inferences and rationales. However, describing the impression of an individual based on a written description of a group or judging relationships in pictures cannot capture changes over a series of tasks.

The second aspect of social intelligence is the ability to recognize others' emotions and share their feelings, or empathy, via facial expressions or bodily behaviors (Frith and Frith, 2006). The Reading-the-Mind-in-the-Eyes test (RMIE; Baron-Cohen et al., 2001) has been widely used to test adults' ability to recognize complex emotions (e.g., shame, curiosity) through viewing 36 images of eyes. This task is context-free and was found to be a good predictor of humans' theory of mind (ToM; Leslie, 1987) under constrained communication (Engel et al., 2014). However, the test differs from people's tasks in real social interactions.

The third aspect of social intelligence is the ability to infer others' beliefs and knowledge—understanding another entity's point of view (Frith and Frith, 2006). This concept is the same as developing a ToM, which can be achieved through (a) spatial perspective taking (Creem-Regehr et al., 2013) or (b) mental perspective taking or mentalizing (Frith and Frith, 2006). In a study exploring spatial perspective taking, participants were asked to report what someone else would see from a different position (Vogeley et al., 2004). Spatial perspective-taking tasks have been used to study the development of ToM in children (Viana et al., 2016). In mental perspective-taking research, participants were given verbal or pictorial instructions to interpret a character's thinking and predict their behaviors (Wolpert et al., 2003). In these cases, participants normally execute these tasks offline (i.e., looking at static pictures without direct interactions), and the findings bear the same inability to reflect changes over time. Rabinowitz et al. (2018) proposed the novel concept of Machine Theory of Mind (MToM), which uses meta-learning to build models of other agents' mental states, including inferring false beliefs. The concept of MToM has not been empirically tested.

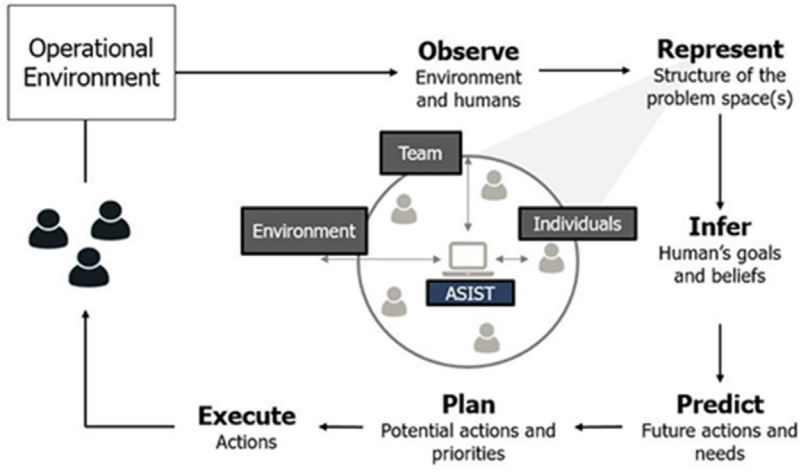

Fig. 1. ASIST model (adopted from OUTREACH@DARPA.MIL, 2019).

The fourth aspect of social intelligence is the ability to predict a person's next possible actions, intentions, and bodily and mental states on a movement-to-movement basis (Frith and Frith, 2006). This process considers the possible dynamics of individual actions. For example, a robotic agent can observe humans' actions, infer latent human mental states, and adjust its goals (Hoffman and Breazeal, 2004). It is worth noting that an observer's domain expertise influences the ability to predict another person's course of action. For example, a high-performing basketball player can outperform a novice basketball player in predicting another basketball player's actions (Frith and Frith, 2006).

Social intelligence literature inspires us to define ASI as the software agents' capabilities to infer humans' mental states, predict their next actions, and act appropriately. Testing these capabilities through descriptive survey questions (e.g., Cuddy et al., 2007) and isolated images (e.g., Barnes & Sternberg, 1989, Baron-Cohen et al., 2001, Viana et al., 2016) is insufficient for continuous real-world tasks, especially due to the emerging need for real-time detections (DeCostanza et al., 2018). When AI and robotic agents work together with humans toward a joint goal, the ability to infer individuals' and teams' mental states (e.g., knowledge, beliefs) in realtime and predict their next actions in an interactive and ongoing task scenario is critical for AI to provide helpful and timely interventions. Fig. 1 illustrates the general processes of human–AI interaction and the goal of generating effective human–AI collaborations (OUTREACH@DARPA.MIL, 2019). ASI agents would proactively engage people in a social manner to achieve a goal. These intelligent agents will play an increasingly important role in advising and executing tasks previously handled solely by humans.

1.2. Evaluating ASI agents and establishing human observer criterion

Evaluating ASI agents' capabilities is an important step in guiding the direction of development, identifying weak points for improvement, and providing evidence for deployment

decisions. Evaluating the underlying ASI agents' capabilities is crucial throughout the life cycle of developing and deploying such ASI systems.

Using ground truth (i.e., participants' actual knowledge and actions) to evaluate the ASI agents' accuracy in inferences of human states (e.g., emotions, knowledge, and goals) and predictions about human actions is the most objective approach when ground truth is available. We could also compare the performance of multiple agents and human capability on the same inference and prediction tasks, then use human performance as a criterion (or reference point) for comparing ASI agent capabilities. Human observers could provide their rationales concerning their inferences on prediction tasks. This is to set up what is reasonable to expect from ASI agents. Some rationales and insights may help ASI agent developers improve ASI agents' ability to understand human players, though not all developers care about human rationales, and humans' rationales are not always reliable. The capabilities of ASI agents depend on designers' understanding of social intelligence, as well as the programming skills of their developers. This comparison approach could help determine whether ASI agents are significantly better (or worse) than humans at assisting other humans or executing specific human tasks. That is, it enables us to identify areas in which ASI or humans excel, based on differing strengths of humans and ASI agents. Fitts (1951) listed humans' strengths in detection, perception, judgment, induction, improvisation, and long-term memory, as well as machines' strength in speed, power, computation, replication, simultaneous operations, and short-term memory. Recent AI transformer models, such as Chat Generative Pre-trained Transformer (ChatGPT), suggest that AI may be proficient at generating coherent descriptions and explanations that are nonetheless factually inaccurate (Open AI, 2022). There have been very effective human–AI teams that take advantage of the strengths and weaknesses of each entity and that work together in a complementary approach (Muller, 2022). This study focuses on ASI agents and humans' certain inference and prediction capabilities through different ways of accuracy evaluations in a search and rescue task environment.

2. Methods

Researchers from six research institutions in the United States developed six ASI agents (#3, #4, #5, #6, #7, and #8) to run in a customized urban search and rescue task environment in Minecraft. The agents were allowed to infer some aspects of human mental states and predict certain actions of the researchers' choices. Only three agents (#3, #6, and #7) demonstrated capabilities of inferring specifically knowledge training conditions and predicting the next victim type to be saved, which were comparable to human observers' tasks. The following sections describe the relevant components of the experiment, the process of establishing the human observer performance criterion, and the relative inferential and predictive accuracy of these three ASI agents and human observers. The details of this experiment (a.k.a., ASIST Study 1) and survey items are accessible on the Open Science Framework website (see Huang et al., 2020).

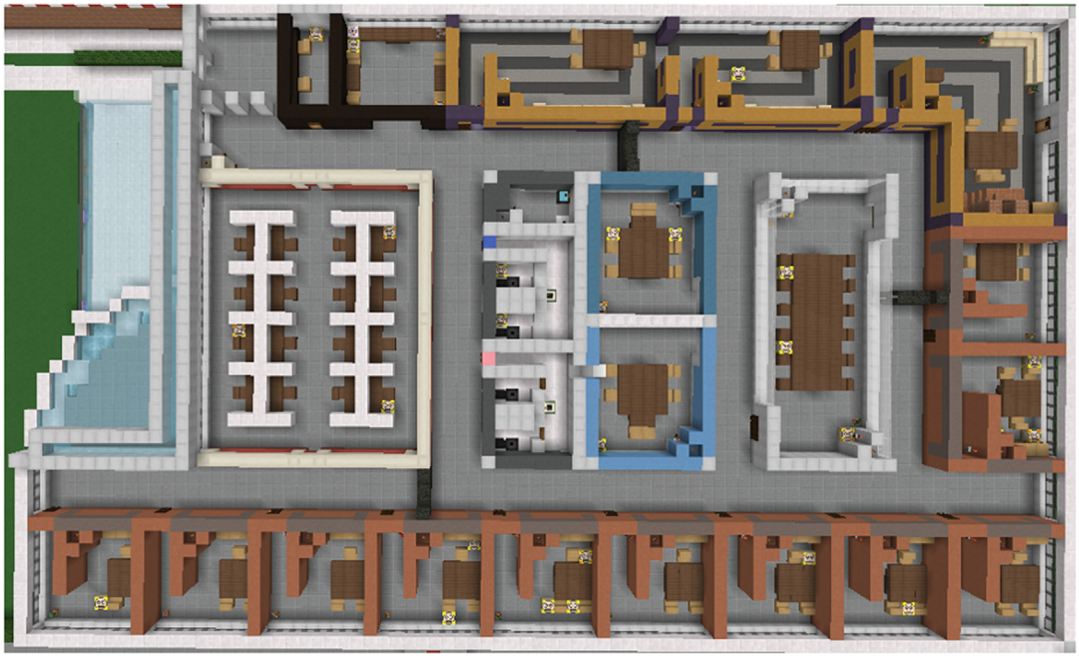

Fig. 2. The Bird's-eye view of the Minecraft building structure (black blocks in the hallways are blockages).

2.1. Experiment environment

Minecraft has many benefits as a research platform (Bartlett and Cooke, 2015, Corral et al., 2021), so we developed a search and rescue task environment in Minecraft (see Figs. 2 and 3). The mission goal was to maximize points by rescuing critical and noncritical victims, represented by yellow and green Minecraft blocks with a face on the surface. Each participant completed three 10-min missions at varying complexity levels (i.e., easy, medium, and difficult) in a counterbalanced order. Three mission maps had the same number of victims (with a different distribution based on the distance from the start point) and varying blockages. The participant's computer interface (see Fig. 3) showed a Minecraft environment, a 2D static mission map, a notepad, the number of points accumulated, and the mission countdown timer throughout the mission.

2.2. Experimental design

To evalua the accuracy of human and agents' inference and prediction, we manipulated two types of knowledge trainin addition to the basic training. The first type was the knowledge of a victim detection device to detect victim types at the door without entering a room, regardless of whether the room was blocked by rubble or not. The text "beep" appeared in the chat window to indicate only low-risk victims (LRVs) in a room (see Fig. 3), and "beep beep" indicated at least one high-risk victim (HRV) in a room, regardless of whether there are LRVs

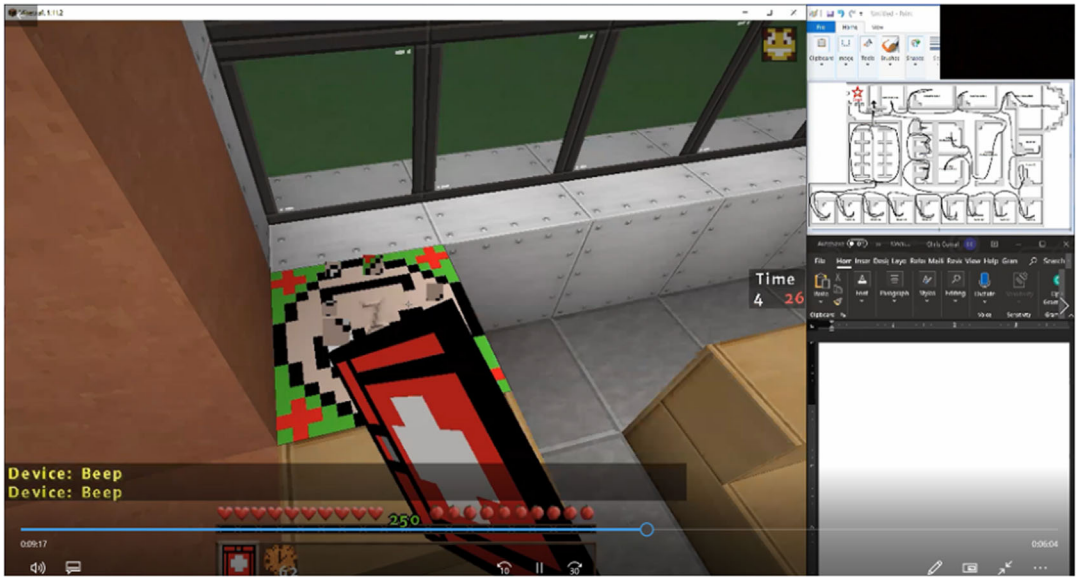

Fig. 3. Participant interface. Note: Left $\frac{3}{4}$ window = Minecraft with a first-person view rescuing a low-risk victim in green; bottom left text = beep signal for low-risk victims; bottom center in the Minecraft window = points accumulated; top-right black square = participant face video (covered to protect identifiable info); mid-right = screenshots of their planned route on the building map;bottom right = notepad for o.

also present. The victim detection device did not differentiate the number of victims of each type in a room and did not beep for any other reason than victim presence.

The second type of knowledge training concerned the trade-off between the different point values for the two types of victims and the time required to rescue each victim. Two types of victims awaited rescue: 24 low-risk victims (LRVs; colored green, each was worth 10 points and took 7.5 s to rescue) and 10 high-risk victims (HRVs; colored yellow, each was worth 30 points and took 15 s to rescue). LRVs were always rescuable during the entire mission, whereas HRVs would die within 5 min after the mission started, making them no longer rescuable.

The knowledge manipulation created three between-subject conditions: 1 (training for both knowledge: victim detection device signals + victim trade-off), 2 (trade-off training), and 3 (neither training). We did not includthe trainiconditdetection devicealonbecause justdetecting the types of victims would not lead to rescue prioritization behaviors witho. Both the victim detection device signals and victim value trade-off were functional in all three conditions, with the only difference being participants in Conditions 1 and 2 were explicitly trained on both or one type of knowledge. Participants in Condition 3 (neither training) would not know the meaning of the device signals (i.e., the "beep" and "beep beep" messages) nor the higher number of points assigned to the HRVs. Participants in all three conditions receive common training on basic game rules except for these two types of knowledge.

Participants completed a training knowledge test via Qualtrics to ensure they understood the training knowledge in their conditions (Appendix B). If one question was answered

incorrectly, participants would review the training slide, repeating it four times until their answer was correct. Most people could answer all questions correctly within two tries. However, one participant tried four times and partially failed one question: he picked only one of the two correct options from a multiple-choice question: "The device beeps twice when (Mark all that apply)." The correct answers should be "There are only yellow victims in the room" and "There are both yellow and green victims in the room." We did not exclude this participant because (a) his error was discovered after the experiment was completed, (b) in his first two tries, he did select the two correct answers separately, and (c) he answered all other questions correctly.

2.3. Participants

An a priori power analysis was conducted using G*Power3 (Faul et al., 2007) to test the difference between the means of three knowledge training conditions by three counterbalanced missions using an F-test with an effect size ($f = 0.25$), an alpha of 0.05, and a power of 0.95. We selected the F-tests family with the option of "ANOVA: Repeated measures, within-between interaction" which resulted in the requirement of 54 participants. To prevent data loss due to technical issues, we ran 57 participants in total, with 19 participants in each condition.

Participants were recruited through the online forums of a large southwest university and Minecraft player communities. Participants were required to be fluent in English, have normal color vision, have experience playing Minecraft with a computer mouse and keyboard, and be at least 18 years old. The 57 participants were 18–39 years old ($mean = 21.19$, $SD = 3.80$), with 42 males, 14 females, and one who preferred not to report gender. The participants in the three conditions had similar gender distribution. All participants had at least some college-level education; 47 of them were currently enrolled college students. A total of 27 participants identified as White, 16 as Asian, five as Hispanic or Latino, one as Native American, two as Middle Eastern, and six as mixed other.

We asked 12 gaming proficiency questions (see Appendix A). An ANOVA analysis showed that the gaming proficiency score ranged from 7 to 24.20, but there was no significant difference between participants in the three conditions ($mean_{both} = 18.62$, $SD = 2.69$; $mean_{trade\text{-}off} = 19.85$, $SD = 2.20$; $mean_{neither} = 17.95$, $SD = 3.62$, $F\ (2,54) = 2.11$, $p = 1.31$). Combining all conditions, the two-tail Pearson correlation between gaming proficiency and an average score of three trials was low and not significant ($r = .19$, $p = .15$, $n = 57$, two missing data). In our customized Minecraft Competency Test that had 14 action-based subtasks, an ANOVA analysis also showed a lack of significant differences among participants in the three conditions ($F\ (2,54) = .26$, $p = .77$).

The three human observers were graduate students in Human System Engineering (two females and one male; ages ranged from 23 to 30). They had been experimenters in the research program for over six months. They watched more than 20 videos of the pilot testing of the experiment, knew the experimental design of the three conditions, and played the game themselves. However, they were blind to the knowledge training condition of each participant's video before viewing it.

2.4. Dataset

This experiment produced a rich dataset, including surveys (pre-trial & post-trial), video recordings of screen activities per mission, and timestamped event messages (e.g., location, actions) on the testbed per trial (see Huang et al., 2020 for all measures).[1] ASI agents' developers had access to these data to design and train the ASI agents. ASI agents conducted inference and prediction tasks using survey data plus participants' action timestamps and events. The human observers generated their criterion inferences and predictions over the same events (i.e., rescue victims), but those events were represented to the human observers in video recordings of screen activities from each trial. Human observers were not able to process all other data (e.g., timestamped event messages generated by the Minecraft testbed and Qualtrics surveys) within a short time. Here, we present only the measures that are relevant to the analysis of this paper.

Participants were asked to answer five questions (victim priority, self-location, self-efficacy, anxiety, and effort) orally during three pauses in the game (at countdown minute 9, minute 6.5, and minute 4). The question related to victim priority was, "Which type of victim will you save next—Yellow, green, or whoever comes first?" Human observers could hear participants' answers in the video recordings.

2.5. Experimental procedure

Participants joined the experiment remotely via Zoom. After consenting to participate, participants reviewed a voice-over PowerPoint presentation of instructions concerning basic Minecraft operations. They then went through hands-on practice of a short version of the game in Minecraft. Next, participants took a Minecraft competency test. Participants were then randomly assigned to one of three knowledge training conditions and went through corresponding game rule training slides with or without the explicit information of the victim detection device signals and victim trade-off. Afterward, participants completed a training knowledge test. Then, participants completed another customized hands-on practice in Minecraft with more details of the game. Before each mission began, participants spent additional 3 min planning their search and rescue strategy on the 2D static building map, using Zoom annotation to draw the paths. The game was paused three times during a mission for the participants to respond verbally to the incident commander. After the third mission, participants answered additional survey questions and were briefed to leave.

2.6. Inference and prediction tasks

In evaluating ASI agents' capabilities, we set two testing tasks: inference and prediction. The first task was inferring which knowledge training condition each participant was in. The inference was to be made four times: at the start of every pause (i.e., at 9, 6.5, and 4 min remaining in the mission) and at the end of every mission. Human observers did the same. The second task was predicting whether a victim in the field of view (FOV) would be rescued next. We chose this task because this task was related to our manipulation of knowledge training conditions. This created the same prediction instances for both ASI agents and

human observers. Thus, the accuracy rates of human observers and ASI agents were directly comparable.

2.7. Human observer annotation procedure

When establishing the human observer criterion, human observers watched only video recordings of participants' screen activities because human observers were not able to process other types of given data (e.g., surveys hosted on Qualtrics, JSON format testbed messages) rapidly to make their inferences and predictions. Thus, a total of 57 participants completing three missions each were subject to human observer annotation for a total of 57*3 = 171 video clips. We had three observers available, so we divided the workload of coding among the three. On the assumption that humans have varying levels of social intelligence and that this variance was inherently important, we chose not to train the human observers to an agreed-upon standard but let them learn by themselves to apply their unique social intelligence and develop by themselves a suitable level of the social model for the participant and task at hand. Each observer had some practice with the pilot study data and then observed the first seven participants' videos (a total of 21 video clips) to serve as training data. Afterward, each participant's videos were assigned to two observers, and their inter-rater reliability of inference and prediction was calculated. So, each human observer annotated a total of 19 participants' videos (57 video clips).

Human observers could see the condition of the participant only after finishing annotating the participant's video. Their task was to predict, when encountering every victim (either an LRV or rescue an HRV) in the FOV, the participants' next action (which victim type would be rescued next, in other words, whether to rescue or skip that victim) and infer, at four timepoints, the training condition (both detection device and victim trade-off training, trade-off training, and neither training). Observers did this while watching the video clips in the sequence of a participant's three missions. Human observers did the annotation in real time, occasionally paused to edit an entry, but were not allowed to replay the video. The observers also provided corresponding rationales.

Categorical predictions were annotated using BORIS (Friard and Gamba, 2016), open-source software designed primarily to record human subject behavior observations in prerecorded experimental video footage (see Fig. 4). For this task, human observers took note of (a) their inference of a player's experimental condition and (b) their prediction of the victim type to be rescued next, operationalized as whether a victim in the player's line of sight would be saved next. These were precisely the same tasks ASI performed, and the instances of rescue predictions were the same between ASI and human observers. Table 1 lists the ethogram codes.

A codebook of topics and phrases for Zoom annotations is in Table 2. These phrases were designed to facilitate extracting word tags from Zoom annotations using text analysis but were sufficiently conducive for human observers to speak in natural language when they annotated the contents. In addition, observers provided qualitative descriptions of behaviors that they perceived as meaningful.

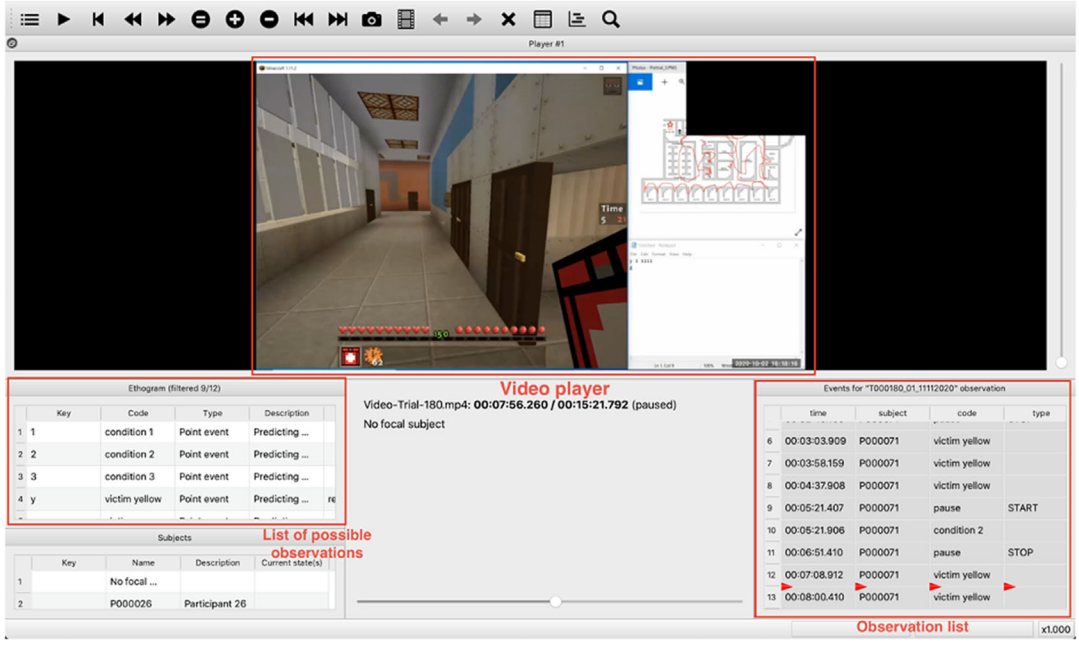

Fig. 4. BORIS annotation interface. Note: The layout of the BORIS interface is adjustable. On this layout, top half = video of screen recording, bottom left up= ethogram codes, bottom left down= subjects with participant ID, bottom middle = progress barbottom right = annotation events.

Table 1
BORIS ethogram (categorical inference and prediction codes)

Code	Key	Description	Status
victim green	g	Predicting next victim to save is green	point
victim yellow	y	Predicting next victim to save is yellow	point
condition 1	1	Inferring training condition 1	point
condition 2	2	Inferring training condition 2	point
condition 3	3	Inferring training condition 3	point
minute 0	0	Mission start	point
minute 5	5	Approaching minute 5	point
pause	p	Video paused to ask incident commander inquiries	state

Note: yellow = high-risk victims (HRVs); green = low-risk victims (LRVs); status = nature of the prediction (point = one-time event; state = a lasting event with a start point and an endpoint).

Accompanying the BORIS categorical predictions were observer prediction rationales recorded through semi-formatted thought vocalizations extracted from automated transcriptions by Zoom. In addition to the two precoded prediction categories to be noted in BORIS, a third category, whether a participant has gained the knowledge of important features (e.g., the meaning of the victim detection device signals "beep" and "beep beep," the different values and rescue times of victim types) was also recorded.

Table 2
Zoom rationale annotation codes

Zoom code	BORIS code(s)	Description
Minute zero	NA	Record the start
Predict <color> (optional: because <reason>)	victim first, victim green, victim yellow	Prediction and explanation of next victim type
Condition <number>1<reason>	condition 1, condition 2, condition 3	Running inference of condition followed by explanation
Participant <has/has not> learned <behavior> because <reason>	NA	Observation of participant behavior acquisition and explanation for noting such. <behavior> can refer to listening for beeps, prioritizing victim colors, and other strategies

Predictions about the type of next-to-be-saved victim were made by each observer as soon as a victim appeared in the player's line of sight. Human observers made a training condition inference at the start of every pause (i.e., at 9, 6.5, and 4 min remaining in the mission) and at the end of every mission. ASI agents did precisely the same tasks at the same time points for comparison with the human observers. Training conditions of teams were revealed to human observers after they finished annotating a participant's three missions to provide feedback and aid observer learning. To facilitate comparisons between observer predictions done in BORIS and their annotations of the reasons for that prediction in Zoom, we combined each observer's data from both sources for each trial using the zoom2boris package developed in R 4.0.3 (R: The R Project for Statistical Computing, 2021) using the tidyverse suite of packages (Wickham et al., 2019).

Cohen's kappa was used to make pairwise comparisons between human observers to evaluate the extent to which raters were making similar judgments for the same participants and trials. The agreement was fair to good between pairs of observers (observers 1 and 2, Cohen's kappa = 0.52, z-score = 5.53; observers 1 and 3, Cohen's kappa = 0.53, z-score = 6.02; observers 2 and 3, Cohen's kappa = 0.62, z-score = 5.52, all p-values < .01), suggesting they were using similar indicators from player behaviors in the videos to identify the condition to which they were trained.

3. Results

3.1. Did the three training conditions contribute to different performance scores?

A UNIANOVA analysis of the three trials' average score showed that the knowledge training condition was a significant factor for performance score (F (2, 54) = 5.37, $p < .01$, $\omega^2 = .16$). Post hoc analysis showed that the average trial score in Condition 1 (both training: victim detection device and trade-off; *mean* = 398.42, *SD* = 34.41) was significantly higher than Condition 3 (neither training; *mean* = 342.37, *SD* = 67.47); but Condition 3 (neither training) significantly lower than Condition 2 (trade-off training), *mean* = 366.49, *SD* = 45.06).

Table 3
Percent correct inferences of the condition by human observers

Observer ID	Correct inferences	Total inferences	Percent correct
1	79	120	65.8%
2	63	107	58.9%
3	71	118	60.2%

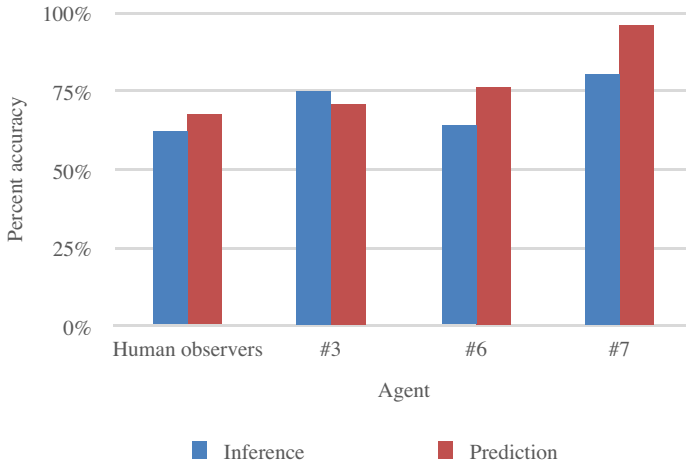

Fig. 5. Accuracy comparison for inferring knowledge condition and predicting next victim type.

3.2. Human observers inferring knowledge training conditions

Regarding the human observers' ability to correctly infer training conditions, results showed that observers were able to correctly infer the participant training condition in an average percentage of 61.7% of trials (see Table 3 and Fig. 5).

Cohen's kappa was calculated to assess the accuracy of participant conditions while controlling for correctness expected by chance (∼33%) under conditions of random guessing, given the equal likelihood of the three experimental condition classes in the data. These statistics were calculated using the *irr* package for R (Gamer et al., 2019). Cohen's kappa values were calculated by comparing observer assessments of condition to ground truth: observer 1 = 0.48 (z-score = 7.66); observer 2 = 0.39 (z-score = 6.13); observer 3 = 0.40 (z-score = 6.22); all p-values $< .01$. The agreement between each observer and ground truth was rated fair (Fleiss et al., 2013, Landis and Koch, 1977).

3.3. Human observers predicting the next victim type to be rescued

To assess the accuracy of the observers' ability to predict the next type of victim to be rescued, we aligned observer data in time with actual rescue events in each mission. For each victim rescued by a participant, the observer prediction reported before the victim rescue was

Table 4
Human observer prediction correctness by condition

Condition	Count of victims	Count of correct prediction	Correctness percent
1	3700	2542	69%
2	3436	2554	74%
3	3393	2197	65%
Grand total	10,529	7293	69%

used to assess accuracy. If the human observer missed one victim, that was counted as an incorrect prediction. Three observers were able to predict the next type of victim that would be rescued with an average of 67.3% accuracy (ranging from 66% to 69%).

Then, we broke down the correctness by conditions and found that Condition 2 (trade-off) resulted in the highest correctness percentage (see Table 4). It seems that Conditions with some training (detection device and trade-off, or just trade-off) resulted in higher prediction accuracy than Condition 3 (neither training). It is possible that participants in Condition 3 (neither training) did not show any priority in entering certain rooms and rescuing certain victims.

3.4. Human observers' rationales for their inferences and predictions

When human observers made predictions via BORIS, they also used Zoom to narrate their reasons for each prediction and later cleaned the Zoom auto-transcript. Table 5 shows a list of reasons human observers used to support their inferences and predictions.

Developing the list of rationales was a learning process for human observers. Each human observer contributed a subset of all final rationales and used some rationales more often than others. Humans continued to learn in the process of doing inference and prediction tasks. Some of the rationales did not emerge until after watching many participants' videos and were modified in the process. Human observers also reported that they inferred more accurately and felt more confident at the end of a trial because more information had been gathered while watching for a few minutes or sometimes after a whole mission. For example, a participant in Condition 1 (both training) was inferred as Condition 3 (neither training) because the participant was saving whatever victims came first during the first mission, though the participant started displaying knowledge of the "beep" signals and prioritizing yellow victims during the second mission through the third mission. In other words, human observers learned between trials and between participants. In contrast, ASI agents were restricted from learning from a participant's three missions but not between participants. These ensured that the same agent completed the task (rather than an agent in successive states of learning).

Using one single rationale did not guarantee correctness; sometimes, multicriteria had to be considered. For example, a human observer made an incorrect inference about a participant in Condition 2 (trade-off training) as Condition 3 (neither training) because the participant was saving whatever victim they saw first without any prioritization. In another case, for example, a participant in Condition 3 (neither training) was incorrectly inferred as Condition 2 (trade-

Table 5
Reasons for predicting next victim type and inference of conditions

Prediction	Reasons
Predicting HRV	1. The participant saw an HRV in the room.2. The participant saw "Beep Beep" and then entered the room to rescue HRVs.3. Prioritizing yellow victim before the 5-min marker.4. More likely to predict HRV for participants in Condition 1 (both).
Predicting LRV	1. The participant saw an LRV in the room.2. The participant saw a "Beep," and then entered the room to rescue LRVs.3. The participant is approaching the 5-min mark.4. It is past the 5-min mark.5. The participant has not found any HRVs for a while and has started approaching an LRV skipped previously.
Inferring Condition 1 with both knowledge	1. The participant seemed aware of signals for victims when the participant stopped in front of the door after "Beep" or "Beep Beep" appeared, then the participant entered the room to rescue the victim associated with "Beep."2. The participant would prioritize HRVs before the 5-min mark.
Inferring Condition 2 with trade-off only	1. The participant has learned the different values of the victims because the participant did not focus on saving the HRV in the first 5 min.2. The participant did not use the device to detect the room, and the participant had just ignored the LRV before the 5 min mark.
Inferring Condition 3 with neither knowledge	1. The participant has saved an LRV which means the participant has not learned the meaning of the victims.2. The participant easily alternates between saving LRVs and HRVs, even while prioritizing HRVs, suggesting that they only know that HRVs die at the 5-min mark.

off training) because the participant was saving exclusively yellow victims before the 5-min mark throughout all missions, to the point that they did not save any green victims in the first mission (so they possibly knew the different values of victim types, or maybe they just focused on the fact that the yellow victim would expire soon). The participant never displayed knowledge about the meaning of the "beep" signals (so they are certainly not Condition 1).

3.5. ASI agents inferring knowledge training conditions

Three research institutes completed the inference tasks. The accuracy of the ASI agents inferring training conditions ranged from 64% to 80% (see Fig. 5). These inferences were made using varied methods: ASI agent #3 used a Modular ToM Architecture-Neural Network and reported 74.7% accuracy (Sycara et al., 2021); ASI agent #6 reported an accuracy of 64% in an internal report without specifying the name of the method; ASI agent #7 used a Dynamic Bayes Network and reported 80% accuracy (Pyarelal, 2022). The detailed methods they used to build the ASI agents will be introduced in other publications.[2]

3.6. ASI agents predicting the next victim type to be rescued

Three research teams reported their ASI agents' prediction accuracy on whether the victim in FOV would be saved next. ASI agent #3 used long short-term memory neural networks (70.44% accuracy; Guo et al., 2021), ASI agent #6 used sequential machine learning (76% accuracy), and ASI agent #7 performed 10-fold cross-validation (94% accuracy for LRV and 98% accuracy for HRV; overall average 96%). The rationales and analyses of the researcher teams' agents were conducted separately from this paper and will be shared in future publications.

4. Discussion

4.1. Comparison of human observers versus ASI agent prediction capabilities

As the first empirical study that examines ASI inference and prediction capabilities for participants in a continuous and dynamic task scenario, the results showed that human observers did not do as well as most ASI agents in inferring training conditions and predicting participants' next actions. A possible reason for the performance difference in inferring training conditions might be that the training conditions of knowledge about the trade-off and knowledge about the victim detection device can be demonstrated in multiple undefined ways in the Minecraft task scenario. Humans may not be able to consistently track all the behavioral features. As Frith and Frith (2006) mentioned, many studies about social intelligence used offline tasks with static pictures (e.g., Barnes & Sternberg, 1989) without interacting with individuals. There is limited literature on dynamic inference studies to compare this work with.

ASI agents may have predicted actions better than the human observers because ASI agents' computational power to estimate action probability was better than human heuristics (see Table 5) in this fast-paced task environment. More specifically, human observers may have relied on participants' oral statements about their intent about which type of victims to rescue next, whereas ASI mainly considered participant behaviors. Two examples illustrate situations in which the human observer and ASI might make different predictions. In the first example, the game was paused, and the experimenter asked, "Which type of victim will you save next—Yellow, green, or whoever comes first?" The participant answered "yellow" (HRV) and stated the reason as yellow HRVs will expire within 5 min. The participant then saw and saved a green LRV. At the experiment debrief after completing the experiment tasks, some participants explained that if they passed this victim, it would be hard to find this one later, especially at locations with long detours due to hallway blockages. In a second example, a participant stated that they would save an HRV next, but after checking several empty rooms in a row, they rescued the next LRV they saw. The human observers could have, but may not have, considered that when the ideal option is not present, the alternative becomes satisfactory. The ASI agents may have followed one single rule or adapted to this flexible principle. In sum, participants' vocalized intent was not always matching their actions, which

might be misleading for human observers. ASI agents' calculation of probabilities based on participants' action patterns may have been more robust than human observer heuristics.

The human observers' limited attention span could also be a cause of prediction inaccuracy. The fast-paced game required human observers to be sensitive to subtle strategies, remember the participants' previous actions, and process new information rapidly to predict the next actions. Predicting the next action involves tracking and remembering a lot of information, as well as learning and detecting patterns for players, a computational process that humans would struggle with compared to computers. This intensive task took hours during the data processing window and may have caused fatigue, so the observers may have missed little clues in the video and made random guesses. ASI agents would not tire of their high-speed processing. The distinct capabilities of humans and AI could be exploited for collective human–machine intelligence. ASI agents could possibly process the data and make recommendations to help humans with decision-making.

4.2. The human observer criterion method

Human observers inferred the training conditions and predicted whether a victim in FOV would be rescued or skipped, and then used the participants' actual rescue action to verify or adjust their inferences.

Human observers used only video recordings as the data source. They could not make use of other available information because they did not have time to do so or were incapable of reading the testbed messages easily. ASI agents used some surveys and timestamped event messages on the testbed to infer participants' mental states and predict their actions. We presented human observers with only video screen recordings because this representation was accessible, rich, and relatively easier to process than other types of data that require processing voluminous text data. Other research task scenarios may lead to different judgments about what data human observers should use and analyze in prediction tasks to establish human observer criteria.

Regarding the criteria for selecting (or testing) human observers, the human observers were experimenters who worked on the project over a period of months. We could use the codebook to train coders that do not have background knowledge about the experiment to compare their results with experimenter observers. Other researchers recruited observers from Amazon Mechanical Turk to do similar tasks (Li et al., 2021).

4.3. Limitations

Zoom transcription has a low quality of accuracy in transcribing human observers' narration of rationales. So, it took observers a very long time to double-check the accuracy of the transcription. Also, using unrestricted natural language made it difficult to calculate the frequency of the reasons for each prediction; it required an additional layer of coding and matching with the prediction in the BORIS system. Using BORIS alone is more efficient for making both the inference/prediction and rationales and thus is highly recommended.

Human observer criteria have great value but are also costly to establish, so we only chose one inference task and one action prediction task, not all other inference and prediction tasks the ASI agents did.

4.4. Future directions

This paper evaluated the accuracy of three ASI agents on inference and prediction tasks in the context of continuous action data and compared the accuracy to that of human observers (our criterion as a reference point), with all accuracy based on ground truth. This new research topic leads to several research directions.

First, the mental state inference and prediction rationales of human observers are potentially useful to ASI agent developers, even though human observers did not infer players' states and predict their actions as accurately as ASI agents did. Human observers can discern emergent tactics that ASI designers may have failed to represent in their models because they did not discover the tactics in analyses of training data, which used a limited number of participants and confederates. ASI agents could build a computational model to infer participants' training conditions based on rationales in Table 5. Human observers' rationales may be complementary to ASI agent developers' rules and help the developers improve the accuracy of inferences and predictions by the next generation of ASI agents.

Second, rather than processing the continuous events in the 10-min task scenario, further research could explore humans' and ASI agents' capabilities in treating discrete events (e.g., presenting each rescue event in a mini clip as a separate event). This will provide better control of prediction time. It may be less ecologically valid because it eliminates the context before each event, but it could be effective for training purposes for beginning observers. Other researchers (Li et al., 2021) selected 10 out of 300 decision points and made mini clips for novice observers to annotate and found the task was still challenging for humans. Both continuous action annotation and discrete event annotation methods can be used to establish a human observer criterion.

Third, it is also worth exploring the choices of how to establish a human criterion: (a) train the human observers to their highest accuracy at inference and prediction using the same heuristics across observers, or (b) allow each participant to learn at their own pace, to represent the variety of human capability levels. For a given research study or application, we may not want variance between observers but the best human practices, which requires training. On the other hand, individual differences among human observers do exist, even at the expert level.

Fourth, Fitts' (1951) list of MABA-HABA (machines are better at vs. humans are better at) stated that humans and machines have different strengths in conducting different types of tasks. Fitts's list needs to be operationalized in the development of ASI. For example, perhaps humans are good at abstract information perception and inferring mental states at slower speeds or longer time horizons but bad at highly frequent predictions of actions. Further research needs to be conducted to identify the boundaries of human and ASI agents' performance on such tasks, or maybe humans are bad at all rapid information processing tasks compared to machines.

Fifth, ASI agents could potentially use the rich data (i.e., all the surveys, testbed messages, and videos) to generate participant profiles to help infer their mental states and actions. The participant profiles may correlate with participants' performance. If displayed appropriately, the participant profiles may inform the human observers to predict participants' actions better. Other advanced modeling approaches (e.g., DiDonato et al., 2013) could be used to explore the dynamical patterns of participant actions that provide more holistic or abstract cues to inform human observers and agents' inference of participant states or prediction of actions.

Sixth, because we, by design, allowed human observers to learn by themselves and use different heuristics, the three human observers performed at varying levels. To ease the calculation of inter-rater reliability, we recommend training the observers with a set of agreed-upon criteria and having two observers code all data. If three observers observed the same participants' data, the Hallgren (2012) method is recommended to calculate inter-rater reliability.

5. Conclusion

Studying ASI requires interdisciplinary effort to program the agents, propose underlying social theories and hypotheses, design human subject experiments to collect data, evaluate agent performance, and test hypotheses. This pioneering study examined ASI inference and prediction capabilities for participants in a continuous and dynamic task scenario. The presented method of establishing a human observer criterion is valuable in evaluating any human–AI team. The collective human–machine intelligence of a human–AI dyad paves the road for the later phase of teamwork involving multiple human team members (Freeman et al., 2023; Huang et al., 2021; Huang, Freeman, Cooke, Colonna-Romano, et al., 2022; Huang, Freeman, Cooke, Dubrow et al., 2022).

The distinctions between the human observer and ASI agent data sources offer an opportunity to investigate ways of diversifying the data sources that humans use within the limits of comprehension, given the scarcity of cognitive resources. It also suggests opportunities to investigate what specific inference and prediction tasks human observers perform better than ASI, which tasks ASI perform better than humans, and how to manage human–ASI collaboration over those tasks.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001119C0130. All performer teams contributed to the study design. Robert Hoffman provided the gaming proficiency items and scoring sheet. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

Open Research Badges

 This article has earned Open Materials badge. Materials are available at https://doi.org/10.17605/OSF.IO/ZWAU9.

Notes

1 The data will become publicly available in 2023 by agreement with the subjects in informed consent and the Institutional Review Board that approved the human subject research protocol.
2 Study 1 publications and data will be announced on this website: https://artificialsocialintelligence.org/

References

Barnes, M. L., & Sternberg, R. J. (1989). Social intelligence and decoding of nonverbal cues. *Intelligence*, *13*(3), 263–287.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251.

Bartlett, C. E., & Cooke, N. J. (2015). Human–robot teaming in urban search and rescue. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 250–254.

Corral, C. C., Tatapudi, K. S., Buchanan, V., Huang, L., & Cooke, N. J. (2021). Building a synthetic task environment to support artificial social intelligence research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 660–664.

Creem-Regehr, S. H., Gagnon, K. T., Geuss, M. N., & Stefanucci, J. K. (2013). Relating spatial perspective taking to the perception of other's affordances: Providing a foundation for predicting the future behavior of others. *Frontiers in Human Neuroscience*, *7*, 596.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). Behaviors from intergroup affect and stereotypes: The BIAS Map. *Journal of Personality and Social Psychology*, *92*, 631–648.

DeCostanza, A. H., Marathe, A. R., Bohannon, A., Evans, A. W., Palazzolo, E. T., Metcalfe, J. S., & McDowell, K. (2018). *Enhancing human agent teaming with individualized, adaptive technologies: A discussion of critical scientific questions*. US Army Research Laboratory Aberdeen Proving Ground United States.

DiDonato, M. D., England, D., Martin, C. L., & Amazeen, P. G. (2013). Dynamical analyses for developmental science: A primer for intrigued scientists. *Human Development*, *56*(1), 59–75.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65–84.

Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014). Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face. *PLoS One*, *9*(12), e115212.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Fitts, P. M. (1951). *Human engineering for an effective air-navigation and traffic-control system*. National Research Council.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.

Ford, M. E., & Tisak, M. S. (1983). A further search for social intelligence. *Journal of Educational Psychology*, *75*(2), 196.

Freeman, J., Huang, L., Wood, M., & Cauffman, S. J. (2023). Evaluating artificial social intelligence in an urban search and rescue task environment. In *Computational Theory of Mind for Human–Machine Teams: First International Symposium, ToM for Teams2021, Virtual Event, November 4–6, 2021, Revised Selected Papers* (pp. 72–84). Cham: Springer Nature Switzerland.

Friard, O., & Gamba, M. (2016). BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, *7*(11), 1325–1330.

Frith, C. D., & Frith, U. (2006). How we predict what other people are going to do. *Brain Research*, *1079*(1), 36–46.

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement* (0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

Guo, Y., Jena, R., Hughes, D., Lewis, M., & Sycara, K. (2021). Transfer learning for human navigation and triage strategies prediction in a simulated urban search and rescue task. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (pp. 784–791). IEEE.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23–34.

Hoffman, G., & Breazeal, C. (2004). Collaboration in human–robot teams. *AIAA 1st Intelligent Systems Technical Conference*.

Huang, L., Freeman, J., Cooke, N., Buchanan, V., Wood, M., Freiman, M., Colonna-Romano, J., & Demir, M. (2020). *Experiment 1 study preregistration*. https://doi.org/10.17605/OSF.IO/ZWAU9

Huang, L., Freeman, J., Cooke, N., Colonna-Romano, J., Wood, M., Buchanan, V., & Cauffman, S. J. (2022). *Exercises for artificial social intelligence in Minecraft search and rescue for teams*. https://doi.org/10.17605/OSF.IO/JWYVF

Huang, L., Freeman, J., Cooke, N., Dubrow, S., Colonna-Romano, J., Wood, M., Buchanan, V., & Cauffman, S. J. (2021). *ASIST Study 2 June 2021 exercises for artificial social intelligence in Minecraft search and rescue for teams*. https://doi.org/10.17605/OSF.IO/GXPQ5

Huang, L., Freeman, J., Cooke, N., Dubrow, S., Colonna-Romano, J., Wood, M., Buchanan, V., Cauffman, S., & Yin, X. (2022). Artificial Social Intelligence for Successful Teams (ASIST) Study 2 *[Data set]*. ASU Library Research Data Repository. https://doi.org/10.48349/ASU/BZUZDE

Hamblen, M. (2020). Killer software: 4 lessons from the deadly 737 MAX crashes. *Fierce Electronics*. https://www.fierceelectronics.com/electronics/killer-software-4-lessons-from-deadly-737-max-crashes

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1) 159–174. https://doi.org/10.2307/2529310

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review*, *94*(4), 412.

Li, H., Zheng, K., Lewis, M., Hughes, D., & Sycara, K. (2021). Human theory of mind inference in search and rescue tasks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. *65*, No. 1, pp. 648–652). Los Angeles, CA: SAGE Publications.

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: Insights into human–autonomy teaming. *Human Factors*, *60*(2), 262–273.

Muller, E. (2022). How AI–human symbiotes may reinvent innovation and what the new centaurs will mean for cities. *Technology and Investment*, *13*(1), 1–19.

Open AI. (2022). Introducing ChatGPT. https://openai.com/blog/chatgpt

OUTREACH@DARPA.MIL. (2019). *Using AI to build better human–machine teams*. Defense Advanced Research Projects Agency.

Pyarelal, A. (2022). Experiment 1 preregistration. https://doi.org/10.17605/OSF.IO/C926K

R: The R Project for Statistical Computing. (2021). Retrieved from https://www.r-project.org/

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In *International Conference on Machine Learning* (pp. 4218–4227). PMLR.

Sycara, K., Lewis, M., & Hughes, D. (2021). Experiment 1 results. https://doi.org/10.17605/OSF.IO/VA2CR
Thorndike, E. L. (1920). Intelligence and its uses. *Harper's Magazine*, *140*, 227–235.
Viana, K. M. P., Zambrana, I. M., Karevold, E. B., & Pons, F. (2016). Beyond conceptual knowledge: The impact of children's theory-of-mind on dyadic spatial tasks. *Frontiers in Psychology*, *7*, 1635.
Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K., & Fink, G. R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of Cognitive Neuroscience*, *16*(5), 817–827.
Walker, R. E., & Foley, J. M. (1973). Social intelligence: Its history and measurement. *Psychological Reports*, *33*(3), 839–864.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.
Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology*, *67*(2), 222–232.
Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 593–602.

Appendix A: Gaming Proficiency Scale with scoring key (adopted from Robert Hoffman)

Question	Response options	Points
1. Please rate your comfort level with playing computer games.	Not comfortable	1
	Somewhat comfortable	2
	Comfortable	3
	Very comfortable	4
2. If you play computer games, when did you first start playing First-Person Action or Real-time Strategy games (e.g., Minecraft)?	Never played first-person action game	0
	Less than 5 years ago	1
	About 5 years ago	2
	More than 5 years ago	3
3. In a typical month, how often do you play First-Person Action or Real-time Strategy games?	Rarely	1
	A few times a month	2
	Once a week	3
	A few times a week	4
	Daily	5

(*Continued*)

4. How long do you play First-Person Action or Real-time Strategy games each time when you play?	1–2 h per session 3–4 h per session 5–6 h per session	1 2 3
5. Of all the computer games you have played, please check all that apply. (sports, action/adventure, third-person shooter, educational, survival horror, puzzle, role-playing, real-time strategy, beat em ups, and other type) (please specify).	Free form response	1 point for every five games played
6. Of all the games you have played, which ones are you particularly good at?	Free form response	1 point for every game listed
7. What game are you currently playing the most?	Free form response	1 point if it is one of: First Person Shooter, First Person Action, Real-time Strategy
8. Think for a moment about the people with whom you have played video games the most. Please check off one of these.	Some of them are very good Some of them are good	Not scored
9. Referring to those people who are very good or good please check off which of these applies.	None of them would say that I am better Some of them would say that I am better All of them would say that I am better	1 2 3
10. Have you ever participated in tournaments or competitions in 11. First-Person Action and Real-time Strategy games?	If you have, please list them	1 point for every tournament
11. In those competitions, how many have you ever "won" or scored the top 10 logged point value?	Free form response	I point for every tournament

The data collected from participants will present a range of scores: participants with considerable game experience and those with little or no game experience. A participant who does not play video games at all would receive a score of zero. The table that follows presents the scoring for a hypothetical highly proficient gamer.

Measurement	Points
Has played for over 5 years	3
Plays daily, for 4–6 h	6
Has played 10 different games	10
Is self-rated as good at all 10	10
Currently plays an Action or Strategy game the most	1
Has team played in Action and Strategy games	2
Plays those games often	1
All peers would say the player is better than they	3
Has played in five tournaments	5
Won in three of them	3
Total	**44**

Appendix B: Preknowledge Survey

This survey assesses whether the participants understand the training they received. It is intended in part as a manipulation check.

Common knowledge section:

The Common Knowledge section is presented to all participants in all training conditions.

1. What is your primary goal for the mission?
   - Find as many victims as I can
   - Locate the fire extinguisher
   - Finish the tasks as soon as I can
   - Maximize my points through rescuing victims
2. What are the two types of victims you need to rescue?
   - Green and Yellow
   - Green and Red
   - Yellow and Red
   - Red and Safe block
3. How long will the GREEN victims survive in each mission?
   - 5 min
   - 7 min
   - Green victims do not die in the mission
   - Not sure
4. How long will the YELLOW victims survive until they die?
   - 5 min
   - 7 min
   - Yellow victims do not die in the mission
   - Not sure
5. Which of the following statements is FALSE?
   - When yellow victims die, they turn red.
   - When green victims are rescued, they show a safe sign.

- When yellow victims are rescued, they show a safe sign.
- When red victims die, they show a safe sign.

6. How much time do you have for each mission?
   - 10 min
   - 15 min
   - 20 min
   - 30 min
7. How many victims in total are in the building?
   - 8
   - 10
   - 24
   - 34

Condition training section:

Only participants in training conditions 1 (both training) and 2 (trade-off training) will be asked the following questions:

1. How many YELLOW victims are in the building?
   - 8
   - 10
   - 24
   - 34
2. How many points are the GREEN victims worth?
   - 10
   - 25
   - 30
   - Unsure
3. How many points are the YELLOW victims worth?
   - 10
   - 25
   - 30
   - Unsure

Only participants in training condition 1 (both training) will be asked the following questions:

1. The device beeps once when:
   - There are only yellow victims in the room.
   - There are only green victims in the room.
   - There are both yellow and green victims in the room.
   - There are no victims in the room.
2. The device beeps twice when (Mark all that apply):
   - There are only yellow victims in the room.
   - There are only green victims in the room.
   - There are both yellow and green victims in the room.
   - There are no victims in the room.