Human Theory of Mind Inference in Search and Rescue Tasks

Huao Li, Keyang Zheng, Michael Lewis
School of Computing and Information
University of Pittsburgh
Pittsburgh, PA, 15260

Dana Hughes, Katia Sycara
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

The ability to make inferences about other's mental state is referred to as having a Theory of Mind (ToM). Such ability is the foundation of many human social interactions such as empathy, teamwork, and communication. As intelligent agents being involved in diverse human-agent teams, they are also expected to be socially intelligent to become effective teammates. To provide a feasible baseline for future social intelligent agents, this paper presents a experimental study on the process of human ToM reference. Human observers' inferences are compared with participants' verbally reported mental state in a simulated search and rescue task. Results show that ToM inference is a challenging task even for experienced human observers.

INTRODUCTION

The Sally–Anne test is a psychological test, used in developmental psychology to measure a person's social cognitive development. 3-4 year old (Wimmer & Perner, 1983) and autistic children of all ages (Baron-Cohen, Leslie, & Frith, 1985) almost always fail the Sally-Anne test while 6-9 year olds almost always pass. In the test a child is asked the control question of recalling the actor's names (the Naming Question). A short skit is then enacted; Sally takes a marble and hides it in her basket. She then "leaves" the room and goes for a walk. While she is away, Anne takes the marble out of Sally's basket and puts it in her own basket. Sally is then reintroduced and the child is asked the key question, the Belief Question: "Where will Sally look for her marble?" To pass, the child must recognize that Sally has not seen Anne move the marble, although the child herself has, and therefore should predict that Anne will look for it in the box where Anne left it. This ability to make inferences about another's mental state is referred to as having a Theory of Mind (ToM). While reasoning about false beliefs is the capability most commonly associated with ToM, other inferences such as preference orderings (Baker, Saxe, & Tenenbaum, 2011), or affect and empathy (Baron-Cohen et al., 1985) have also been associated with ToM along with other explanatory concepts involving mental states such as desires and intentions (Bratman, 1987) which have been referred to inclusively as Folk Psychology (Stich & Ravenscroft, 1992).

A panel (Fiore et al., 2020) at last year's HFES meeting introduced ASIST (Artificial Social Intelligence for Successful Teams), a DARPA program to develop AI agents capable of employing ToM reasoning. Because ToM is defined through its role in folk psychology and human commonsense reasoning the appropriate baseline for guiding development and evaluating an agent employing ToM would necessarily be the naïve human observer. However, despite mastery of ToM reasoning in everyday life, people often fail to employ it, in taking directions (Samson & Apperly, 2010), for example, or fail in reasoning about content of others' minds due to biases toward their own perspectives and knowledge (Birch, 2005). Therefore, it would be useful to collect direct accounts of mental states as well as attributions of observers.

ToM Inference

In this paper we describe and analyze three types of human data: think aloud verbal protocols, action prediction, and explanations of action, that were collected to assess our agent's ToM. ToM inference involves at least two entities; one presumed to have mental states which may on occasion lead to observable actions and an observer who attributes mental states and transitions between them to be the cause of observed actions by the first party. One formalized version of folk psychology, the Belief-Desire-Intention model (Bratman, 1987), holds that agents form intentions to act in order to bring about desired states, with beliefs describing the allowable states and transitions. Because these entities cannot be observed, the observer must infer them on the basis of very little evidence. Humans do this readily (Wimmer & Perner, 1983) albeit often in error (Birch, 2005; Samson & Apperly, 2010).

In an experimental setting, however, it may sometimes be possible to gain access to mental states by requesting the performer to report them. In a small 8 person study reported here, participants were instructed to 'think aloud' as they performed the search task. Following (A. Ericsson & Simon, 1993)'s advice for preserving validity reporting was unconstrained, not requiring a distinct format or content. These data while presenting ground truth, varied widely in specificity and content providing insight in how the task was being performed but were too incomplete to fully validate the agent's ToM inference.

Decision Points

Comparing to a human reference provides a second avenue to validating agent performance that hews more closely to the original objective of replicating human ToM. Because human observers are well practiced at making such inferences and make them on the basis of incomplete evidence, human inferences are likely to vary in confidence and accuracy with the ambiguity of observations. The accuracy of an agent should vary in a similar way with ambiguity in observations, which makes the comparisons with human 'experts' a good test of inference capabilities.

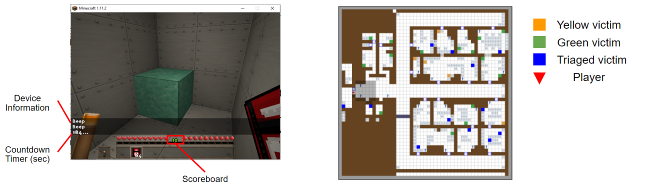
Because a ToM model is expected to evolve over time but only reveals itself intermittently through observed actions, it needs to be maintained and updated in order to converge to a more accurate model. To choose the decision points for making these updates, it is necessary to consider whether an action is taken or not when an opportunity occurs, for example encountering a door that could be entered, and whether new information potentially initiating a current or future action is encountered, for example sensing an unexpected obstacle. Based on these decision points, a player's trajectory through the game was segmented stopping just before each opportunity for action. This segmentation supports two forms of judgments: 1-prediction of actions player will/will not take and 2-explanation of why that action will be taken. For predictions, there are three values to compare: action taken by a human player, action predicted by an agent, and action predicted by a human observer. Because so little is known about the human player's actual mental state, many possible ToM models may predict the same (phenotypic) action, making the accuracy of prediction high but not the accuracy of the ToM. For explanations, there are eight possible outcomes depending on whether an agent and/or a human correctly predicts the action and whether or not their explanations agree. The extent to which agents and human observers predict the same action and attribute to the same cause measures the extent to which the agent has replicated human ToM reasoning regardless of the accuracy of the action prediction.

SIMULATED SEARCH AND RESCUE TASK

Task scenario

Task scenario. Since the Kobe earthquake and the twin towers disaster a few years later, urban search and rescue has grown to be a signature research area for human-robot interaction (Casper & Murphy, 2003; Nourbakhsh et al., 2005). In the current study (Fiore et al., 2020) is collecting data from human participants searching for and triaging victims in a Minecraft environment reproducing the uncertainties and hazards of a collapsed building. In this paper we describe the analysis of data collected at the University of Pittsburgh including think aloud verbal protocols from local participants and action predictions and explanations at decision points made by Mechanical Turk workers. The search and rescue map developed for (Fiore et al., 2020) is shown in Figure 1. The scenario portrays a structurally damaged office building after an unspecified incident. It contains 26 discrete areas consisting of corridors, rooms, and elevators. The building layout and connectivity may be changed by perturbations such as collapses, wall openings, and sporadic fires. There are 20 injured victims inside the building in need of rescue. Out of the 20 injured victims, five are high-risk victims with severe injures (denoted in yellow) and would die if not treated in time (after 7 minutes). Others are low-risk victims (denoted in green) and stay alive throughout the mission. The rescuer needs to search the building and rescue all victims within 15 minutes. Their performance is measured by the number of saved victims, they are also encouraged to completed the task as fast as possible.

Rescuers were given a building map and a victim detection device to help them with the task. The map on the right contains the static building layout (e.g. room numbers and connectivity) before the disaster, which might help rescuers plan their search strategy. Perturbations are not shown on the map, thus rescuers need to handle unexpected environmental changes such as blockages and holes. The victim detection device is capable of detecting victims in a room when the rescuer approaches its entrance. Participants were assigned to one of three knowledge conditions: naive, knowledge of scoring, or knowledge of device and scoring mirroring the design of the larger experiment (Fiore et al., 2020).



(a) 3D Minecraft view to the human participants

(b) 2D map layout of the environment

Figure 1. Human participants serving as rescuers see the egocentric view in (a). Human observers viewing replayed segments of the trajectory see both (a) and the floor plan view of the environment shown in (b).

Data collection

To gain greater insight into the task, 8 local participants were recruited for closer scrutiny and assigned to the three knowledge conditions used in the larger experiment (Fiore et al., 2020) being conducted at Arizona State University.

Participants were instructed to think aloud as they performed the task. Specifically, they were instructed to "try to report any thoughts that come to your mind, including what you are doing, what you are thinking about, and what you are trying to do next" (K. A. Ericsson & Simon, 1984). After a training trial on a smaller map to practice the task and verbalization process, participants completed three trials of up to 15 minutes followed by a debrief. Three map variants with different perturbations and victim locations were used in the trials.

Our experimental testbed recorded logs of the rescuer's in-game behavior, first-person game screen videos and audio recordings of their think aloud protocols. One subject did not finish the last trial due to technical issues and 5 trials were removed from analysis due to missing data or low quality recording. In total 18 trials were retained for analysis.

Thinking aloud protocol

Method. Eighteen think aloud videos were analyzed from 8 participants with an average length of 10 minutes 30 seconds. Two independent raters coded transcriptions auto-generated by Otter.ai. Recording were assigned 63.95 codes on average. Two raters first discussed the standards and principles of coding and then each independently coded half of the recordings. One trial was coded by both raters to evaluate the inter-rater reliability. Since the speech-to-text transcription did not generate semantically meaningful sentence segmentation for oral reports, raters are allowed to assign multiple codes to a single sentence. Thus we used the percentage agreement instead of Cohen's Kappa to measure the inter-rater reliability (McHugh, 2012). Result shows that two raters have 64.4% agreement in the 27-category coding task, which indicates a moderate level of consistency. Most of the disagreement occur when more codes were assigned to the same part of the protocol by one coder in addition to the agreed code both has assigned to.

Coding schema. The coding schema is developed based upon the search and rescue mission itself and the key principles in Belief-Desire-Intention model (Bratman, 1987), and later optimized through a pilot trial. The coding schema consists of four categories: *Goals*, *Information*, *Behaviors*, *Emotions* which cover the main topics of the participants' think aloud protocols. There are 27 detailed codes within those categories to better capture the content of verbal reports.

1. Goal: rescuer's reports about future planing.
2. Information: participants reporting new information they observed about the task.
3. Behavior: participants describing their current actions.
4. Emotion: emotional reports such as the feeling of stress.

In addition, a special type of code, *Explanation*, was used when the participant explicitly explained the rationale behind their behavior. Judgments in this category can then be compared with human observations to reveal differences between introspective explanations and ToM attributions.

Results. Figure 2 shows the distribution of the top ten codes among the think aloud protocols. *Information* and *Goals* are the two most popular categories, especially for location-related information and planning. This shows that the rescuers were most frequently focused on navigation involving their current location and plans to move to the next location during the mission.

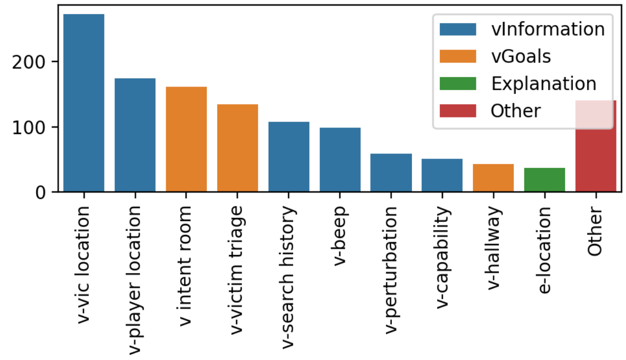


Figure 2. Distribution of think aloud transcription coding.

HUMAN OBSERVATION EXPERIMENTS

Because the initial goal of our artificial socially intelligent agent is to replicate human performance at ToM tasks we have collected predictions and causal attributions from human observers to provide a baseline with which to compare agent performance.

Materials

For the 18 rescuer trajectories, depending on the performance of the rescuer, the length of each trajectory range from 8 minutes to 15 minutes. Based on the collected human trajectories, we generated the following materials: game screen video recordings, dynamic mini-map videos and a static building layout image. Human observers can watch the first person screen recordings to understand what the rescuers were doing, and refer to the dynamic/static maps to locate the rescuers' current location and navigation path. Note, human observers have no access to rescuers' think aloud protocols. Video materials were were segmented by 'decision points' at which behaviors occur such as spotting a victim or leaving a room. The specific types of the decision points are explained below.

- Triage decision points: When a victim block enters rescuer's field of the view (FOV), the observer is asked whether the rescuer will triage the victim.
- Navigation decision points: When a room entrance (door/hole) enters rescuer's FOV, the observer is asked whether the rescuer will enter the room?
- General decision points: when the rescuer finishes triaging a victim or leaving a room, the observer will need to answer three general questions, including predicting the next location rescuer will go, as well as the rescuer's triage strategy and knowledge condition.

At each of the different decision points, human observers were given different prediction tasks including predicting next room and triage strategy of the rescuer etc. They were asked to choose among alternative locations or strategies from menus. In addition, observers were asked to provide explanations from a pre-designed list of options.

Video segments were presented in chronological order so that prior segments can inform judgments. The actual action taken by the rescuer in the video segments was then presented at the start of the following sequence providing knowledge of results. The total number of decision points in one trajectory is around 300, which is too demanding for human observers to annotate. Thus we sampled 10 decision points for each type and generated 30 video segments with corresponding prediction questions for each trajectory.

Procedure

50 human observers were recruited from Amazon Mechanical Turk. Participant accessed the online survey on their own computer. Detailed instruction were given to the observers about the search and rescue environment and the prediction and explanation tasks they were to complete. Then the observers were required to pass a quiz about basic knowledge of the experiment in order to proceed to the experimental task. Each observer was assigned one trajectory from a human rescuer. In each of the 30 decision points from the trajectory, human observers were presented a video clip and the corresponding prediction/explanation questions. The length of this human observation experiment was around 45 minutes.

Results: Prediction

Human observers' predictions were compared with the ground truth of rescuers' behavior to calculate the prediction

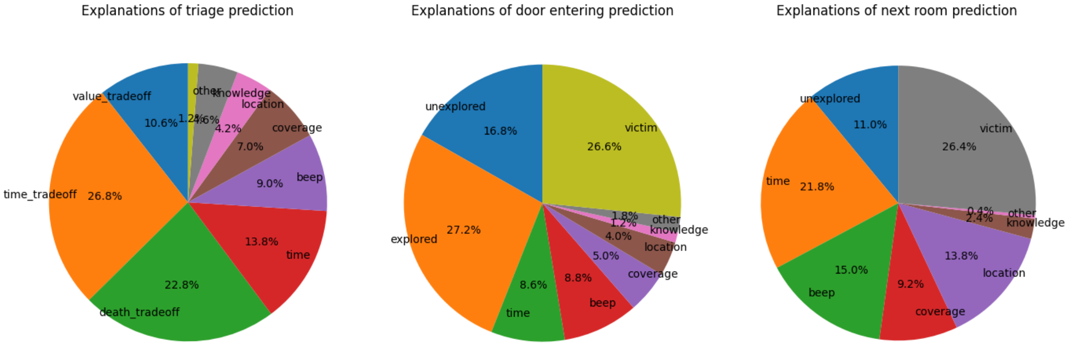


Figure 3. Distribution of human observation explanations.

accuracy shown in Table 1. A recent work reports the performance of social intelligence agents in similar ToM inference tasks (Jain et al., 2020). As a reference, agents achieve 98.8% accuracy in predicting triage strategy and 68.5% in predicting next location.

Table 1. *Prediction accuracy of human observers*

Prediction task	Task type	Human accuracy
Triage action	Binary	56.0%
Door entering action	Binary	31.2%
Next location	Multi-class	58.2%
Triage strategy	Three-class	65.5%
Knowledge condition (Beep)	Binary	46.2%
Knowledge condition (Victim)	Binary	74.0%

From these results it is evident that the ToM inference task is challenging even for human observers. The prediction accuracy in most tasks is around random guess levels except for next location prediction. This might be explained by the systematic searching routes the rescuers were likely to adopt such as checking adjunct rooms one by one. Such searching patterns are relatively noticeable for human observers to capture, and thus could be leveraged in their inference. This also aligns with our finding in think aloud protocols that rescuers' current location and next intended location are the two highest frequency codes which contribute to 26.2% of the reports. On the other hand, behaviors depending on rescuer's current mental state (belief) such as whether to enter a room are harder for human observers to predict because the rescuer's current mental state may not always be inferable by the observer. For instance, the rescuer might revisit an explored room by mistake because he holds a false belief about his search history in which he has forgotten that he has already been to the room. Because evidence of this false belief is not available to the observer prior to the revisit, the observer can not predict the revisit.

Additionally, if we plot the prediction accuracy over time as in Figure 4, we can see that navigation and triage prediction accuracy slightly decreases over the progress of experiment. This is counter intuitive considering human observers are continuously learning by watching episodes from the rescuer's trajectory. This might be due to the fact that as rescuers explore more rooms, their memory load may increase leading to increased forgetting and more false beliefs. Based on evidence provided by a trajectory human observers under most conditions can not correctly predict probable rescuer errors such as revisiting cleared rooms or missing victims in their field of view because these actions run counter to those commonly taken.

This argument can be supported by the pattern we observed in rescuers' thinking aloud reports. By comparing the reports in first and second half of the task, we found that the frequency of protocols that might contain false beliefs (i.e. player location, search history, capability) increases from 17.0% to 34.8%. A typical quote from a rescuer is as follows: “I don't remember if there were any in 207. So I'm gonna have to run by and check just to be safe. So now I believe I have cleared all the rooms or maybe not Janitor's room." This happened when the participant had already searched all rooms but forgot where were the last few untreated victims. So he had to re-enter multiple rooms to verify if his memory was correct. Such situations are challenging for either humans or agents to infer without privileged knowledge about player's mental state.

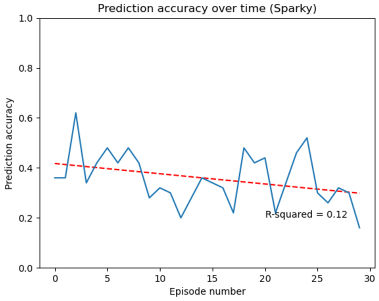
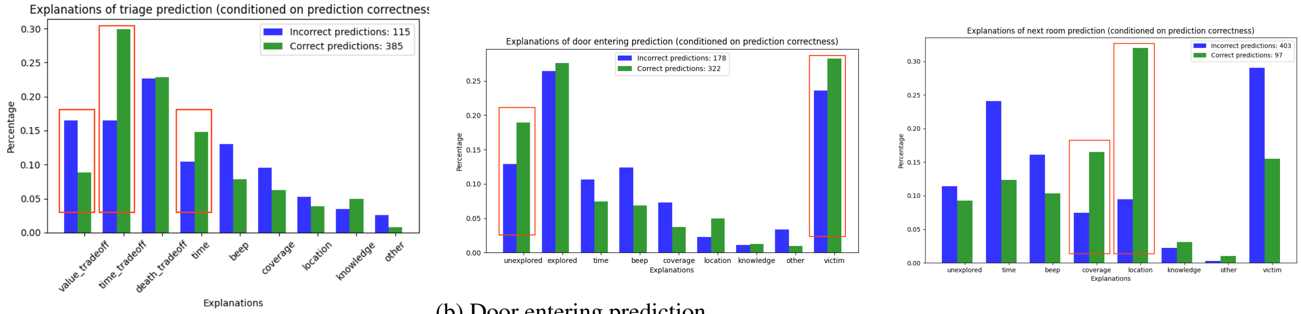


Figure 4. Prediction accuracy of triage and navigation decision points over time.

Results: Explanation

We summarized the explanations provided by human observers for each types of decision points. The distribution of human observation explanation is shown in Figure 3. Another per-



(a) Triage prediction

(b) Door entering prediction

(c) Next room prediction

Figure 5. Human observation explanation conditioned on prediction correctness.

spective to evaluate human observation explanation is to condition it on the prediction correctness as shown in Figure 5. Since observers gave their prediction and explanation at the same time, those who correctly predicted the rescuer's behaviors can be presumed to show greater insight into the rescuer's mental state.

For triage decisions, most of the explanations fall on the cost/benefit trade-off between two kinds of victims. There are a few considerations when deciding whether to rescue a victim in front of you. One is the fact that high-risk victims take a longer time to triage (i.e. time trade-off), but also are worth more points (i.e. value trade-off). The other concern is that all high-risk victims will die if not get treated by 7 minutes (i.e. death trade-off). Those trade-offs plus the overall task objective of saving all victims cause the rescuer to take different triage strategies. One typical strategy is to prioritize high-risk victims first while ignoring low-risk victims until all high-risk victims are either saved or expired. Another strategy is to save any victims encountered without considering the type. These strategies as well as mixture of them were observed in the human trajectories. To better understand if human observers had the correct insights when giving those explanations, we plot the distribution conditioned on prediction correctness as in Figure 5(a). It shows that observers who made correct predictions on victim triage decisions focus more on the time trade-off and current task time. While incorrect predictions were influenced more by the value trade-off. Similar reports are found in rescuers' think aloud records about the ratio between time spent on locating victims by exploring rooms and the time spent on rescuing victims, as well as the efficiency of the different strategies.

Figure 5(c) shows an example pattern of relative percentages of explanations in the next-location decision points. When asked about which room the player will go next, observers who made correct predictions focus more on players' current location and map area coverage so far. This aligns with our finding in the think aloud protocols that rescuers' current location and next intended location are the most frequent factors which contribute to 26.2% of the reports, while incorrect predictions were influenced more by other factors, like beliefs about victim location or mission time. Human observers could have a wrong assumption about a rescuer's belief about those factors. For example, the rescuers might not remember the victim location that the observer thought they would have remembered, a typical failure in ToM inference.

ACKNOWLEDGMENTS

This research was supported by DARPA award HR001120C0035 and AFOSR/AFRL award FA9550-18-1-0251.

REFERENCES

Baker, C., Saxe, R., & Tenenbaum, J. (2011, 01). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*.

Baron-Cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a "theory of mind. *Cognition*, 37-46.

Birch, S. (2005). When knowledge is a curse: children's and adults reasoning about mental states. *Current Directions in Psychological Science*, 1-5.

Bratman, M. (1987). *Intentions, plans, and practical reason.* Harvard University Press.

Casper, J., & Murphy, R. R. (2003). Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *33*(3), 367-385. doi: 10.1109/TSMCB.2003.811794

Ericsson, A., & Simon, H. (1993). *Protocol analysis: Verbal reports as data, revised edition.* MIT Press.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* the MIT Press.

Fiore, S., Bracken, B., Demir, M., Freeman, J., Lewis, M., & Huang, L. (2020). Transdisciplinary team research to develop theory of mind in human-ai teams. *Proceedings of the Sixty-fourth Annual Conference of the Human Factors and Ergonomics Society*.

Jain, V., Jena, R., Li, H., Gupta, T., Hughes, D., Lewis, M., & Sycara, K. (2020). Predicting human strategies in simulated search and rescue task. *Accepted at NeurIPS 2020; Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response (AI+HADR 2020)*.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Nourbakhsh, I. R., Sycara, K., Koes, M., Yong, M., Lewis, M., & Burion, S. (2005). Human-robot teaming for search and rescue. *IEEE Pervasive Computing*, *4*(1), 72-79. doi: 10.1109/MPRV.2005.13

Samson, D., & Apperly, I. (2010). There is more to mind reading than having theory of mind concepts: new directions in theory of mind research. *Infant Child Development*, 443-454.

Stich, S., & Ravenscroft, I. (1992). What is folk psychology? *Cognition*, 447-468.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 103-128.