Sequential Theory of Mind Modeling in Team Search and Rescue Tasks

Huao Li, [1] Long Le, [2] Max Chis, [1] Keyang Zheng, [1]
Dana Hughes, [3] Michael Lewis, [1] Katia Sycara [3]
[1]School of Computing and Information Science, University of Pittsburgh
[2]College of Information and Computer Sciences, University of Massachusetts Amherst
[3]The Robotics Institute, Carnegie Mellon University

Abstract

The ability to make inferences about other's mental states is referred to as having a Theory of Mind (ToM). Such ability is fundamental for human social activities 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. In this paper, we propose a computational ToM model which observes team behaviors and infer their mental states in a simulated search and rescue task. The model structure consists of a transformer-based language module and an RNN-based sequential mental state module in order to capture both team communication and behaviors for the ToM inference. To provide a feasible baseline for our ToM model, we present the same inference task to human observers recruited from Amazon MTurk. Results show that our proposed computational model achieves a comparable performance with human observers in the ToM inference task.

Introduction

The famous Sally–Anne test has been widely used in developmental psychology to test if children are able to distinguish their own mental states and others. 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, and Tenenbaum 2011), or affect and empathy (Baron-Cohen, Leslie, and Frith 1985) have also been linked 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 and Ravenscroft 1992).

In this paper, we propose a sequential Theory of Mind model that reasons about human dynamic beliefs in 3-people urban-search-and-rescue (USAR) teams. The USAR rescuer teams have to navigate through an environment and clear obstacles to locate and triage victims. Multimodal observations, including both team communications and individual actions, are used to infer hidden beliefs of rescuers regarding the meaning of different markers. A team communication module is implemented based on a pre-trained language model to identify marker-related utterances in rescuers' verbal communication. Then those detected timestamps are used as potential transition points of human beliefs about marker meanings. The dynamic belief module is based on recurrent neural networks and capable of mapping observable action sequences to belief sequences. The overall framework of marker ToM inference model is shown in Fig. 1. Each component of the framework is explained in subsequent sections of this paper.

Human ToM inference

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 mental entities of one human are not known to other humans, an observer must infer them on the basis of very little evidence. Humans do this readily (Wimmer and Perner 1983) albeit often in error (Samson and Apperly 2010; Birch 2005).

In this paper, we introduce a computational inference model 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 a computational ToM model would be a human observer. However, the human observer's accuracy in ToM inference is not necessarily expected to be high. Despite mastery of ToM reasoning in everyday life, people often fail to employ it, in taking directions (Samson and Apperly 2010), for example, or fail in reasoning about content of others' minds due to biases toward their own perspectives and knowledge (Birch 2005). Nonetheless, it would still be useful to assign the same ToM inference tasks to human observers to provide a performance baseline for our proposed model.

Computational ToM Models

Several authors have proposed a computational framework to model human goal inference as a Bayesian inverse planning process, i.e. the Bayesian Theory of Mind (BToM) framework (Baker, Saxe, and Tenenbaum 2009). BToM models have enjoyed success in explaining human ToM inference in goals, desires (Liu et al. 2017), and (false) beliefs (Baker et al. 2017). However, most of the Bayesian-based methods are applied in simple environments with small state space (e.g. grid world) and yet to be tested with real human data in complex task scenarios. Computational ToM models based on neural networks have been shown to successfully reason about both machine agents internals (Rabinowitz et al. 2018) and human mental states (Jain et al. 2020; Oguntola, Hughes, and Sycara 2021; Guo et al. 2021). In this paper, we continue to explore the possibilities of incorporating ToM models with state-of-the-art deep learning techniques.

In the literature, little attention has been paid to modeling human mental states in team settings. When multiple humans form teams, the complexity of their joint mental model increases tremendously. Team members have to intensively communicate about task information to maintain a shared situational awareness. Thus humans in a team have even more dynamic desires, beliefs, and intentions that change over time when compared with humans operating in isolation. In this paper, we take a first step in this direction by investigating human beliefs in 3-person urban-search-and-rescue (USAR) teams.

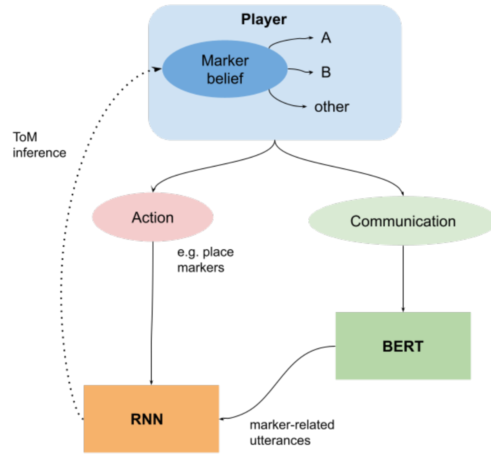

Figure 1: Overall Framework for sequential ToM Inference.

Decision Points

In our work, we analyze three types of human data: verbal communications between team rescuers, rescuer actions, and action prediction, that were collected from human observers to assess our ToM model. ToM inferences involve at least two entities; one team player presumed to have mental states which may on occasion lead to observable actions or communications and an observer who attributes mental states and transitions between them to be the cause of observed actions by the first party. Because human observers are well practiced at making such inferences and making 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 our proposed model 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 marked by others that could be entered. The other kind of decision points are when team members explicitly communicate about their mental state, for example reporting change of marker meanings. These decision points provide stopping points during a USAR mission, where our ToM computational model can be updated, and inference tasks can be posed to the model and human observers.

Simulated Search and Rescue Task

Task scenario

(Fiore et al. 2020) describe a Minecraft environment designed to reproduce the uncertainties and hazards of a collapsed building for use in the study of urban search and rescue teams. Using data shared with us by Arizona State University from this task and environment, we built our ToM inference model and collected action predictions at decision points made by Mechanical Turk workers. The search and rescue map developed for this task is shown in Figure 2. The scenario portrays a structurally damaged office building after an unspecified incident. It contains 54 rooms and multiple connecting corridors. The building layout and connectivity may be changed by perturbations such as rubble. The 3-person team needs to search the building and rescue as many victims as possible within 15 minutes. Their performance is measured by points earned from saving victims. There are 55 injured victims inside the building. Out of the 55 injured, five are critical victims with severe injures and others are regular victims. Critical victims are worth more points but can not be rescued until all team members are present at the victim's location. This encourages communication, for example about a critical victim's location, and coordination between teammates. The three rescuers in a team are named Red, Blue, Green for easy identification.

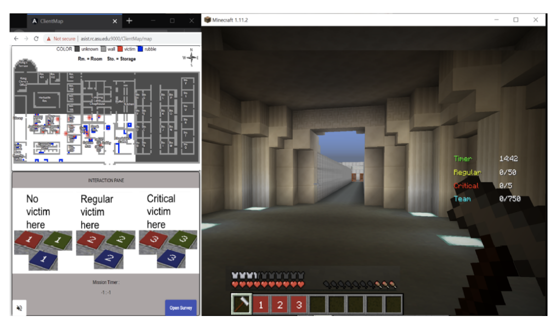

Figure 2: Human participants serving as rescuers see the first-person view on the right and client map and marker block legend on the left.

Team roles

Rescuers can choose from three interdependent roles, each of which have different capabilities and limitations. Teams can choose to have any team composition they would like (e.g., one of each role, all three at the same role). Each role's tools have limited uses before they need to be replenished. Rescuers can change their roles or replenish their tools at any time by returning to the base.

- Heavy Equipment Specialist (a.k.a, engineer): Can remove rubble with a hammer, and move at a slow speed.
- Medical Specialist (a.k.a., medic): Can rescue victims and travel at a medium speed.
- Searching Specialist (a.k.a, searcher): Can move regular victims to a different location, making it easier for the medic to rescue multiple victims at one spot, and move at a fast speed, allowing for scouting.

Information Asymmetry

To provide heterogeneous mental models associated with task knowledge, the participants receive different information when the mission starts. On the client map interface, each rescuer is provided with unique information about a certain map regions. The information includes the room name and locations of rubble and victims divided by regions. Each participant receive a map that provides details in two regions and no information in four regions. In addition, three rescuers each has a set of marker blocks (numbered 1, 2, and 3) that they can put on the ground to indicate the state of a room. Each marker block also indicates the identity of its author (Red, Green, or Blue). In the mission, one rescuer will receive a legend to the meanings of numbered blocks that conflicts with the legend given to the other two rescuers. Two possible semantics are given in Table 1. As such, a rescuer with the marker semantic A would most likely place a marker 1 in front of a room that she believes to contain no victim.

Semantics	Marker 1	Marker 2	Marker 3
A	No victim	Regular v.	Critical v.
B	Regular victim	No v.	Critical v.

Table 1: Marker semantics

An example of asymmetrical information received by rescuers is shown in Fig. 3. Note that those rescuers are not aware of this manipulation among team members and will need to communicate to share information.

In the present study, we develop a ToM model to predict the marker semantic belief of each rescuer. Note that this belief is inherently dynamic i.e. it can change over time. For example, a rescuer might be initially assigned marker semantic A but later switch to semantic B after the team realize, and communicate about the information asymmetry, and agree to all switch to semantic B.

Data set

19 groups, each of 3 participants, took a series of training sessions and surveys before entering the actual search and

Figure 3: Asymmetrical information on client map and marker block legend received by each rescuer in the team.

rescue task. Each team completed two consecutive 15-min missions on two different map configurations. Game state was recorded for in-game events (e.g. rescue victims, switch roles, and place marker blocks) at 30Hz. Rescuer screen recordings and verbal communication audio were saved for post processing.

Team communication model

Given the experimental setting, rescuers in the team need to intensively communicate about task information to maintain a shared situational awareness among members. Especially for the manipulated marker legends, such information asymmetry can lead rescuers to have different mental models about the meaning of maker blocks, then form false beliefs when seeing makers placed by other rescuers. If such team mental state misalignment is not resolved by communication early in the mission, team performance is harmed severely. Therefore, we propose a natural language processing method to identify team communication entries related to marker blocks. The model takes in a single communication transcription and outputs a binary decision: whether this communication is related to marker blocks or not.

Model structure

Our proposed model is based on general language models pre-trained on a large corpus and fine-tuned on our data set. An illustration of the model structure is shown in Fig. 4. Communication transcripts are prepossessed and tokenized then fed into a pre-trained Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018). BERT embeddings with a 0.3 dropout rate go through a neural network with two fully connected layers with 64 and 1 hidden neurons each. The ReLU activation function is used between two FC layers. The single output value from neural network is used to determine the binary prediction result.

Training details

The human data set is divided into two parts, where the training data set consists of 2725 entries from 10 teams and the test dataset consists of 3434 entries from 9 teams. All transcriptions are manually coded by experimenters to serve as ground truth labels of each communication entry. There are in total 67 positive cases out of 2725 entries in the training data. To overcome imbalanced data labels, we oversample positive entries by 20 times to create an augmented training data set. We use binary cross-entropy with logit loss as

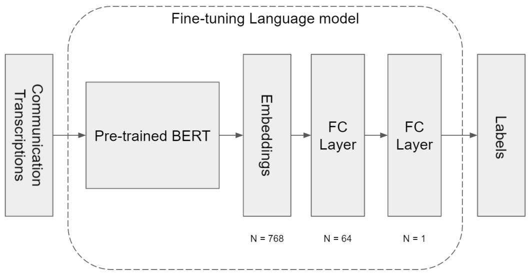

Figure 4: Communication model structure.

the loss function and Adam optimizer to regulate the training (Kingma and Ba 2014). Training batch size is 16 and learning rate is set to 1e-05. The whole model including both BERT and FC modules are fine-tuned together for 1 epoch.

Experimental results

Cross-validation is used on training set to explore different model structures and hyper-parameters. Specifically, training data from 1 team is held out for validation, and the other 8 teams are used for training. This process repeats 9 times and the average validation result is used to evaluate model performance. In addition, since the aim of this model is to identify potential mental state transition points, we care more about how relevant returned results are instead of how complete they are. Therefore, precision is used as the performance metric in model evaluation. The average precision of cross-validation on training set is **98.6%** and the test precision is **81.0%**, indicating a good model performance and a reasonable generalization between teams.

Experimental results show that our proposed model is capable of identifying marker-related entries from team communication transcriptions. Those communication points will be used in dynamic belief modeling and human observation experiment as key decision points where rescuers' mental state are highly likely to change.

Dynamic belief model

The goal of our ToM model is to infer human's dynamic beliefs based on observable behaviors. Here we again concentrate on inferring what marker block meanings each rescuer was using during the mission. Because one rescuer in the team was initially assigned with a different marker legend than other team members and the team may realize and resolve this manipulation at any time during the mission, this ToM model is trying to infer dynamic beliefs of humans that may change over time. By observing an individual rescuer's behaviors for a certain time interval, the model should be able to estimate the most likely marker legend that the rescuer is using. Since we have identified time points where the team communicates about marker blocks, we can assume that the mental state of rescuers remains the same during intervals between those communication points. We make such an assumption based on the observation that it is unlikely for a certain rescuer to suddenly change her marker block meaning without informing other team members.

Data processing

In total, 16 trials of game data from 9 teams are used in model training and evaluation. We first slice the game log of each trial by marker-related communication points generated by the BERT communication model, then calculate each individual rescuer's actions within those N intervals. Specifically, we count the number of action sequences that potentially reveal the rescuer's belief about marker blocks, in each interval. For example, if a rescuer sees a regular victim and then immediately places a number 2 marker block, it is more likely for her to hold semantics A (the complete action list is shown in Table 2). With a 12-dimension action vector, each time a listed action is observed within a certain time interval following its prerequisite, the count of corresponding dimension adds one. Some actions share the same dimension as they refer to the same semantic meanings in different forms, e.g. perception (marker in FOV – rescue victim) and intention (victim in FOV – place marker). By counting those actions in each observation window, we have an input observation sequence with the shape of $(N, 12)$ per trial per rescuer.

Prerequisite	Action	Dimension
Other	Place marker 1/2/3	1,2,3
Marker 1/2/3 in FOV	Other	1,2,3
Regular v. in FOV	Place marker 1/2/3	4,5,6
Marker 1/2/3 in FOV	Rescue regular v.	4,5,6
Critical v. in FOV	Place marker 1/2/3	7,8,9
Marker 1/2/3 in FOV	Rescue critical v.	7,8,9
Marker 1/2/3 in FOV	Clean rubble	10,11,12

Table 2: Action sequences list.

In addition, the actual marker semantics rescuers were using during the mission are manually coded by experimenters into three categories: semantics A, semantics B, and other. These label sequences are used as the supervised ground truth for model training and evaluation.

Model structure and Experimental results

Fig. 5 shows the structure of our dynamic belief model. It uses Gated recurrent units (GRUs) to process input observation sequence and transit hidden state through timestamps (Cho et al. 2014). At each timestamp, a fully connected neural network takes in the hidden state and outputs prediction results. Both the GRU and FC module share weights along the timeline. Hyper-parameters used in training is as follows: dropout rate = 0.3, learning rate = 0.001, batch size = 1, loss function = cross entropy loss, optimizer = Adam.

The model is trained on 12 trials of data and tested on 4 held-out data. We run 20 epochs of training 10 times to balance the influence of random initialization. The average prediction accuracy on test data set is **70.5%**.

Human observation experiments

Because the initial goal of our computational ToM model is to replicate human performance at ToM tasks, we have col-

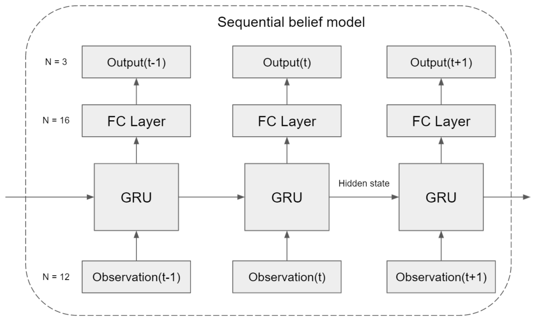

Figure 5: Belief model structure.

Figure 6: Game state video presented to human observers

lected predictions from human observers to provide a baseline with which to compare the model performance. Note that our human observer experiment goes beyond just inferring the marker block semantics to include other aspects such as team scores and map. Thus, this data can be used in future studies for other ToM tasks.

We devise a team-level and an individual-level experiment. In the team-level experiment, the observer is asked to predict the final team performance and make inferences about the team mental model such as which rescuer was assigned with different marker semantics than other team members. In the individual-level experiment, the observer is asked to predict individual rescuer's next action based on potential false belief about marker block semantics.

Materials

For human experiment materials, we use the same 4 team trials as in the test data set for belief model evaluation. Based on the collected team trajectories, we generated the following materials: game screen video recordings and dynamic mini-map videos. Human observers can watch the first person screen recordings to understand what the rescuers were doing, and refer to the dynamic maps to locate the rescuers' current location and navigation path. The individual-level experiment has a similar setup to the team-level one except the observer can have access to only the field of view of one player and the minimap does not include the locations of other players (unless their locations are within the field of view). Video materials were segmented by 'decision points' at which behaviors occur such as spotting a victim or placing a marker block. The specific types of the decision points are explained below.

Table 3 gives the "decision points" for each experiment. For the team performance prediction task, the video is paused at regular intervals, i.e. 4,9,14 minutes. All other decision points are triggered by events. In particular, decision points in the team-level experiment are triggered by communication between team members in the video. For example, a team-level map prediction question is triggered whenever a rescuer mentions the map in verbal communication. For the purpose of detecting communication about objects of interest (blocks, maps, ect...), we used the team communication model introduced earlier. For the individual-level decision points, we use similar method as in dynamic belief modeling to detect when a rescuer places or sees a marker block.

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 a subset of decision points for each type and generated 15 video segments with corresponding prediction questions for each trajectory.

Procedure

102 human observers were recruited from Amazon Mechanical Turk in which 42 participated in the team-level experiment and 60 participated in the individual level experiment. 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 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 with one trajectory from a rescue team. Therefore, each trial of team trajectories were annotated by at least 5 independent observers. In each of the decision points from the trajectory, human observers were presented a video clip and the corresponding prediction questions. The length of this human observation experiment was around 45 minutes.

Results

To ensure quality control, we removed data with missing answer entries or extreme completion times (i.e. it took too long for the observers to answer). For team performance prediction, observer's prediction is compared with the discrete actual team performance, thus the average RMSE of observers is calculated as the metric. For other predictions, since the result is categorical, the majority result of observers is used. For example, if 3 out of 5 observers predict the player will enter the room, this is the result that is compared with ground truth.

Table 4 show the prediction accuracy of human observers. In addition to the majority voting results, we also calculated the 90th percentile of individual observers indicating the performance of proficient humans in ToM inference tasks.

Experiment	Prediction tasks	Decision points	Ground Truth
Team	Team performance	Fixed intervals	Final team performance
Team	Map information	Map-related comm.	Initial map assignment
Team	Marker semantics	Marker-related comm./ Place markers	Initial semantics assignment
Individual	Reaction to markers	Markers in FOV	Actual next action

Table 3: Decision points list.

Prediction	Majority voting	90th percentile
Performance	125.2	48.6
Map	20.29%	44.44%
Marker	**58.52%**	**77.78%**
Reaction	32.65%	55.56%

Table 4: Human observation accuracy.

Discussion

In this paper, we propose a sequential Theory of Mind model that reasons about human dynamic beliefs in a team task. Specifically, the model observes team communication and individual actions to infer marker meanings each rescuer used during the search and rescue mission. A team communication module is implemented based on a pre-trained language model to identify marker-related utterances in rescuers' verbal communication. Then those detected timestamps are used as potential transition points of human beliefs about marker meanings. The dynamic belief module is based on recurrent neural networks and capable of mapping observable action sequences (e.g. see a regular victim then place a marker 2) with belief sequences (e.g. hold semantics A). The overall framework of marker ToM inference model is shown in Fig. 1. Both communication and belief models are trained and tested on data previously collected from human teams. Test set results show that the communication model achieves **81.0%** precision in identifying marker-related utterances, and the belief model achieves **70.5%** accuracy in inferring rescuer's marker semantics. To provide a feasible baseline to evaluate our model performance, we assign the same test materials and ToM inference tasks to human observers recruited from Amazon MTurk. Results show that majority populations only achieve **58.52%** in inferring marker semantics and even worse in other ToM inference tasks. Even for the more competent human observers, i.e. the 90th percentile, the inference accuracy is only **77.78%**. This aligns with previous findings in the literature that inferring other humans' mental states in complicated task scenarios is challenging for human observers (Li et al. 2021). We can tell from the above comparison that our proposed computational ToM model achieves a human-level performance in inferring dynamic beliefs in human teams.

This research bears its own limitations that we would like to improve in future steps. First, the current model structure deals with the mental state of each individual in the team separately. Although the communication model considers information shared among all team members, the dynamic belief model only takes in action sequence from one rescuer when inferring her mental state. However, in such a team task, the mental states of three members are dependent on each other and conditioned on team roles, which might lead to more complicated belief structure such as nested second-order beliefs (rescuer Red thinks that rescuer Blue has marker semantic A). A more reasonable method to model team ToM is to incorporate action sequences of all individuals in the team and infer the joint team mental state at once. In addition, the current model is trained and tested on a relatively small dataset and limited to a narrow belief regarding marker meanings. Further experiments on larger dataset (collected from humans or data augmentation) and general mental beliefs are needed to test the model effectiveness and robustness.

Acknowledgement

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

References

Baker, C.; Saxe, R.; and Tenenbaum, J. 2011. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* .

Baker, C. L.; Jara-Ettinger, J.; Saxe, R.; and Tenenbaum, J. B. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1(4): 1–10.

Baker, C. L.; Saxe, R.; and Tenenbaum, J. B. 2009. Action understanding as inverse planning. *Cognition* 113(3): 329–349.

Baron-Cohen, S.; Leslie, A.; and 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.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Fiore, S.; Bracken, B.; Demir, M.; Freeman, J.; Lewis, M.; and 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* .

Guo, Y.; Jena, R.; Hughes, D.; Lewis, M.; and 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)*, 784–791. IEEE.

Jain, V.; Jena, R.; Li, H.; Gupta, T.; Hughes, D.; Lewis, M.; and 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)* .

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Li, H.; Zheng, K.; Sycara, K.; and Lewis, M. 2021. Human Theory of Mind Inference in Search and Rescue Tasks. In *Proceedings of the 65th Annual Meeting of the Human Factors and Ergonomics Society*. Human Factors and Ergonomics Society.

Liu, S.; Ullman, T. D.; Tenenbaum, J. B.; and Spelke, E. S. 2017. Ten-month-old infants infer the value of goals from the costs of actions. *Science* 358(6366): 1038–1041.

Oguntola, I.; Hughes, D.; and Sycara, K. 2021. Deep Interpretable Models of Theory of Mind. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 657–664. IEEE.

Rabinowitz, N.; Perbet, F.; Song, F.; Zhang, C.; Eslami, S. A.; and Botvinick, M. 2018. Machine theory of mind. In *International conference on machine learning*, 4218–4227. PMLR.

Samson, D.; and 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.; and Ravenscroft, I. 1992. What is Folk Psychology? *Cognition* 447–468.

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