Preference Proxies: Evaluating Large Language Models in capturing Human Preferences in Human-AI Tasks

Mudit Verma *[1] **Siddhant Bhambri** *[1] **Subbarao Kambhampati** [1]

Abstract

In this work, we investigate the potential of Large Language Models (LLMs) to serve as effective human proxies by capturing human preferences in the context of collaboration with AI agents. Focusing on two key aspects of human preferences - explicability and sub-task specification in team settings - we explore LLMs' ability to not only model mental states but also understand human reasoning processes. By developing scenarios where optimal AI performance relies on modeling human mental states and reasoning, our investigation involving two different preference types and a user study (with 17 participants) contributes valuable insights into the suitability of LLMs as "Preference Proxies" in various human-AI applications, paving the way for future research on the integration of AI agents with human users in Human-Aware AI tasks.

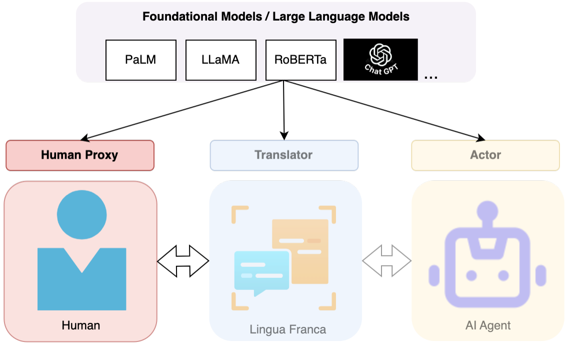

Figure 1: The various roles of Large Language Models in Human Aware AI interaction as a Human Proxy, Translator (common lingua franca), and the Actor. In this work, we investigate the role of LLMs as a Human Proxy (called Preference Proxies) especially when they have to provide answers to queries meant for eliciting human in the loop's preferences.

1. Introduction

As Artificial Intelligence (AI) progresses, the development of the next generation of AI agents requires an enhanced understanding of human thought, processes and behaviors. A vital component of this understanding is the Theory of Mind (ToM), which involves attributing mental states – such as beliefs, intentions, desires, and emotions – to oneself and others, and to understand that these mental states may differ from one's own. Large language models (LLMs) have demonstrated exceptional abilities in various tasks that humans excel at (Hagendorff, 2023; Frieder et al., 2023; Korinek, 2023; Shen et al., 2023; Bubeck et al., 2023), making them suitable candidates for exploring the capabilities of ToM in AI systems (Kosinski, 2023).

Research on LLM's ToM capacities has primarily focused on their ability to model mental states associated with social and emotional reasoning, as well as logical problem-solving (Kosinski, 2023; Baker et al., 2011; Wellman et al., 2001; Astington & Baird, 2005; Cuzzolin et al., 2020; Rescorla, 2015; Çelikok et al., 2019). While LLMs have been used for several tasks like summarization, text generation, comprehension, conversations etc. there is limited literature on testing LLM's ability to predict human preferences. Since these LLMs are infact trained on human generated data available in the wild (Brown et al., 2020) and have been fine-tuned with human feedback on various prompts (Ouyang et al., 2022) a natural question arises :

Can LLMs capture human preferences?

We investigate whether LLMs can serve as human-proxy to the real human in the loop (HiL) and answer queries made by an AI agent meant for the real human. Several prior works in learning human preferences have leveraged human feedbacks of some form, like binary feedback, demonstration, natural language guidance, action guidance, etc. We expect the LLM to work for an AI agent that is acting in the world (powered by an reinforcement learning, planning or other sequential decision-making engines). A common theme across these works has been to model a reward function that

*Equal contribution [1]SCAI, Arizona State University, USA. Correspondence to: Mudit Verma <muditverma@asu.edu>.

Appearing in Proceedings of the First Workshop on Theory of Mind in Communicating Agents, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

captures human's expectations from the agent. Therefore, ToM is an important aspect of such a capability and while prior research has been interested in answering whether LLMs can ascribe correct mental states to the HiL, we go beyond and test whether it can also ascribe correct reasoning mechanisms used by humans. We argue that AI agents (like LLM) must be able to maintain mental states of the HiL and their reasoning process to answer questions that require the agent to know about human's expectations and preferences over the agent. Prior work has already established the potential improvements in team performance when the AI agent's modeling of mental states and reasoning of the HiL is correct (Lim & Klein, 2006; Edwards et al., 2006).

Human preferences over how the agent should behave and what sub-tasks a human-AI team should solve for optimal team performance are important problems being studied by several research communities (Lee et al., 2021; Verma & Metcalf, 2022; Verma et al., 2023; Sreedharan et al., 2020; Guan et al., 2021; Soni et al., 2022; Park et al., 2022; Kambhampati et al., 2022b; Christiano et al., 2017). LLM's ability to correctly identify human's expectations of the agent, or, human's preferences over sub-tasks can be a good direction to study whether LLMs are suitable 'Preference Proxies'. While the term "preferences" has an open-ended definition materializing with respect to the context, we study two important ways in which past research has looked at human preferences over the AI agents when the human is in the loop. First, when the human observer has a preference of the agent acting in the world expecting a degree of explicability. Explicability is the ability to understand human's expectations of the agent and conform to it. Second, the human actor in a human-AI coordination team has a preference for the pursuit of certain sub-tasks by the team among countless possibilities.

When the human in the loop assumes an observer role, our work leverages past research to develop scenarios in which the AI agent's optimal performance depends on its ability to model the human's mental states and reasoning process. This helps the AI to understand the human's expectations of the agent. For instance, in a search and rescue mission where the human serves as an AI robot's commander, it is crucial that the robot can predict how the human would infer and respond to various situations.

In cases where the human user plays an active role in achieving the team's objectives, such as a field commander working alongside a rescue robot, it is vital that the robot identifies the same set of sub-tasks to be accomplished by the team. This necessitates the agent to reason about the human's preferred method for achieving the team's goal, thus going beyond simply ascribing mental states to the human user.

The rest of the paper is structured as follows: we talk about preliminaries for this work in Section 2. we then introduce the readers to our Theory-of-Mind experiments which are divided across Sections 3 and 4 along with their respective results. We describe our user study for understanding the alignment between LLM responses and user responses in Section 5, and finally conclude our investigation of this work in Section 6. An appendix has also been attached at the end. Readers are encouraged to view our additional supplementary material containing prompts and responses from the GPT models at https://tinyurl.com/prefproxiessupp.

2. Theory of Mind, Language Models and Human Preferences

In this section, we will revisit three core concepts essential for our research: Theory of Mind, which facilitates the comprehension and forecasting of human preferences, and the capability of Large Language Models to effectively simulate these aspects.

2.1. What is Theory of Mind?

We follow the definition of "Theory of Mind" from (Sap et al., 2022). Theory of Mind essentially ascribes the ability to ascribe and infer mental states of others. This ability is central to any form of human interactions, communications, empathy, self-consciousness, moral judgment, and even religious beliefs (Albuquerque et al., 2016; Heyes & Frith, 2014; Zhang et al., 2012; Milligan et al., 2007; Seyfarth & Cheney, 2013; Dennett, 1978; Moran et al., 2011). While modeling mental states is a fundamental aspect of theory of mind, it encompasses more than just creating mental models. Modeling the reasoning process over these mental states is an equally important and challenging objective.

2.2. Theory of Mind and Learning Human Preferences

Prior works have tried to advocate how Inverse Reinforcement Learning (IRL) is linked to Theory of Mind and that reward learning mechanisms should take into account several factors like human mental states, their desires, beliefs etc. (Jara-Ettinger, 2019). The field of learning a reward function from human preferences attempts to achieve a similar objective as IRL but assumes access to high-level human feedback (like pairwise comparisons) than explicit demonstrations (Verma & Metcalf, 2022). However, the expectations from the reward function being learned for the case of PbRL is same as that of IRL with respect to Theory of Mind.

Human Preferences can also be defined in various ways like trajectory preferences (Lee et al., 2021), tacit or explicit preferences, goal-oriented preferences (Verma et al., 2023), or more abstract preferences like explicability, predictability, and legibility expectations of the human from the agent

(Chakraborti et al., 2019). The spectrum of human preferences is vast and varied and can touch upon other aspects like levels of autonomy, personalization, and transparency to name a few. While these aforementioned preferences are important in their own right, in this work we focus ourselves on two key preferences prior literature has highlighted, i.e. explicability and sub-task specification.

Under Explicability preference the human expects the agent to behave in a certain way, and the agent proactively attempts to model this expectation and follow it. Hence, by definition, it involves ascribing mental states to the human in the loop and beyond that performing inferences on these states and reasoning about which behaviors would the human prefer. We restrict ourselves to situations where while there may be a human-AI team but the human only observes the agent (and the interaction may involve explanatory dialogues).

Next, we consider a Human-AI teaming scenario where the human plays a more active role and can perform actions in the world alongside our AI agent. On the other hand, sub-task specification preferences involve the agent to come up with the same set of sub-tasks that the human has in mind to achieve the team objective.

2.3. Theory of Mind and LLMs

Large language models have shown great success and exceptional results with many tasks like summarization, conversations, and text generation to name a few. Figure 1 shows the major components of a Human-AI interaction that involves the human user, the AI agent, and a lingua-franca between them (like natural language, formal languages, images, binary feedback etc). Prior literature have tasked LLMs with the roles of Translator (Xie et al., 2023; Kambhampati et al., 2022a), where the LLMs are responsible to ingest natural language inputs from the human user and convert that into a representation that can be easily understood by the AI agent. Additionally, attempts have also been used to utilize LLMs as the actor by asking them to produce actions to be performed (Hu & Sadigh, 2023; Ahn et al., 2022). While the debate regarding the utility of LLMs as a translator and as an actor have not yet settled, we introduce another potential role of LLMs as the “Human Proxy”.

While advances in LLMs-based technology can improve its capabilities as a translator and as an actor, we argue that general-purpose models modeling human preferences can only do so uptill a certain point. This is because human preferences are potentially highly non-stationary, unique to individuals, and at times unknown to the human themselves. Despite this, for several realistic scenarios LLMs can capture reasonable human preferences as shown in later sections 3, 4. Therefore, at best, we are in search for a good human-proxy who can provide general preferences humans may have which can substantially reduce load on the human in the loop.

3. Probing LLMs with Explicability Preferences

In Human-AI scenarios with humans observing AI agents acting in the environment, there is a natural preference, or expectation in particular in this case, that humans may have from the AI agent’s behavior. This expectation is for the AI agent to act such that its actions or plan are explicable to the human. While additional interaction in the form of explanatory dialogue (Chakraborti et al., 2017) can help bridge the gap between the human’s expectations of the agent’s behavior and agent’s final behavior, researchers are also interested in looking for automated ways.

One reasonable approach can be to have these general-purpose large language models (LLMs) reason on behalf of the human in the loop (HiL) who is observing such agents acting in the environment. We test three such scenarios in which the information with the HiL is limited due to varying reasons that may require LLMs to perform ToM and “step into the shoes of the human in the loop” to determine their expectations of the AI agent.

Limited information on agent’s internal workings: We begin with the Rover domain (Zhang et al., 2017), where a rover is navigating in an environment to complete a certain task, while the human observes a top-view of this environment. Note that the human in this case could be an expert of the domain (and possesses knowledge about which actions are possible, effect of those actions etc) but does not know of how the AI agent computes its plan or policy. We test LLMs if they can respond and reason on behalf of this user and can answer questions with respect to explicability. The complete description for this task is given in A.1.1.

Limited information on agent’s actions: Next, we experiment with the Fetch domain (Chakraborti et al., 2017), where the Fetch robot is tasked with picking up a block from one location and transporting it to another location, as given in A.1.2. In this test, the lay user only understands the high level descriptions of the actions the robot can take. However, they are still unaware of the internal workings of the agent, and hence, not understanding the reasons behind its actions. We again probe LLMs for explicability preference in this case.

Limited information due to partial observability: In the third experiment using the Urban Search and Rescue (USAR) domain (Chakraborti et al., 2015; Sreedharan et al., 2017), we have a user who is an expert of the agent and its capabilities, and the task the agent needs to perform. However, this user can only partially observe the dynamics of the environment due to the fact that they only have access to the top-view projection of what is happening on the field, and hence, they may not be completely aware of the other properties the environment like the weight of the medkits. The complete description of this task is given in A.1.3.

3.1. Experiments & Results

We prompt eight LLM models with the same prompts given in A.1.1, A.1.2 and A.1.3, and compare the responses with the ground truth composed from prior works (Zhang et al., 2017; Chakraborti et al., 2017; 2015; Sreedharan et al., 2017). We perform a subjective check: if each LLM model correctly identifies the explicability issue or not, and if so, if the reason provided for the answer is also correct. The results are shown in Table 1.

As part of the prompt, the LLMs are exposed to information available with the AI agent (whom the LLM is trying to assist by modeling the human in the loop), and information available with the human in the loop (like access to only the top-view). It is not, however, explicitly told the impacts of missing information. The objective for LLM is then two-fold, first to correctly identify which is the potential impact of the missing information (for example, the fact that the human is unaware of the 'tuck' motion being part of 'move' in Fetch example) and secondly, to utilize this information to judge whether or not the human in the loop would find the AI agent's actions explicable. We find that newer generations of GPT models perform better than older versions in these tasks. In the cases of Fetch and Rover, the GPT models provide accurate reasoning, but they struggle in the USAR domain. Although the models can correctly predict the explicability label that the human would assign, there is significant room for improvement in their reasoning abilities. While LLMs can offer valuable feedback as preference proxies, their Theory of Mind (ToM) capabilities could be enhanced further. Researchers should use LLMs with caution and continue exploring ways to improve their performance as preference proxies in such settings.

4. Probing LLMs for Sub-Task Preferences

The other set of experiments we perform are based on Human-AI collaborative teaming settings where both, the human and the AI agent are acting in the environment. In this case, we identify at least two categories of human preferences, preferences over the sub-tasks to achieve as a team, and preferences over sub-task assignments between the human and the AI agent. While both of these require extensive modeling of human mental states, we find that preferences over sub-task assignment is usually unique to the human in concern whereas generally, people come up with a finite set of interesting sub-tasks they would want to pursue as a team. Therefore, we restrict the scope of our investigation to sub-task specification based preferences and leave sub-task assignment as a future research objective.

Table 1: Experiments on testing Theory-of-Mind capabilities of LLMs across 3 domains: Rover, Fetch and USAR. **Y**: matches with ground truth, **Y***: matches with ground truth with correct reasoning, **N**: does not match with ground truth, **-**: no response.

	Matches w/ Ground Truth		
Domain/Model	**Rover**	**Fetch**	**USAR**
text-davinci-001	N	Y	Y
text-davinci-002	Y*	N	Y
text-davinci-003	N	Y*	Y
text-ada-001	N	N	N
text-babbage-001	-	-	N
text-curie-001	-	N	Y*
gpt-3.5-turbo	Y*	Y*	Y
gpt4	Y*	Y*	N

We experiment with eight LLM models using the Overcooked domain, a popular 2-player game that has been widely used for training collaborative agents paired with real human partners (Carroll et al., 2019; Yu et al., 2023). We prompt the LLMs with a general description of the game as given in A.2.1, and then also add three specific layout descriptions which have additional specifications on how the two agents can act in the environment, as given in A.2.2, A.2.3, and A.2.4. The objective of the LLM is to respond with a set of seven sub-tasks that the human in the loop would believe as reasonable sub-tasks to be pursued as a team. We use a list of "events" used in prior work (Yu et al., 2023) as the ground truth of what the human would expect.

These layouts are as follows:

Layout 1 - Asymmetric Advantages: This layout tests whether players can choose high-level strategies that play to their strengths. There is a counter in the middle with two stoves that can be accessed from each side. Both players have an onion dispenser, plate dispenser, and serving area on their sides. However, the plates and the serving area are closer to the player on the left, while the onion dispenser is closer to the player on the right.

Layout 2 - Forced Coordination: This layout forces players to develop a high-level joint strategy, since neither player can serve a dish by themselves due to a counter table between them over which the player on the left side can pass over onions and plates, and the right player will take the onions, put them on the cooking stove, plate the cooked soup in a dish, and finally serve them

Layout 3 - Counter Circuit: This layout involves a non-

Table 2: Experiments on testing Theory-of-Mind capabilities of LLMs across 3 Overcooked domain layouts: Asymmetric Advantages, Forced Coordination, and Counter Circuit.

	# Matches (out of 7) w/ Ground Truth		
Layout/ Model	**Asymmetric Advantages**	**Forced Coordination**	**Counter Circuit**
text-davinci-001	1	4	2
text-davinci-002	4	3	5
text-davinci-003	4	2	5
text-ada-001	0	0	0
text-babbage-001	2	2	0
text-curie-001	0	2	3
gpt-3.5-turbo	3	5	4
gpt4	3	3	5

obvious coordination strategy, where onions are passed over a counter in the middle of the kitchen to the pot, rather than being carried around the counter. There is only one path around the counter so the two agents can not cross each other and will collide if they reach the same location in the kitchen.

4.1. Results

For this set of experiments, we check how much overlap exists between the responses of the LLM models and the ground truth as has been described by (Yu et al., 2023) and given in section A.2. This ground truth has been used by (Yu et al., 2023) to account for human preferences when training AI agents to partner with them and accomplish the team goal. Consolidated results can be found in Table 2. We perform a subjective check and report how many, out of the seven predicted 'events' by the LLM, match with the ground truth list. Unlike as in previous section, the performance of GPT4 is at par, if not worse, with an older generation model text-davinci-002 and text-davinci-003 for all the three layouts. The results show that GPT4 model did get several of the expected events, but could not capture all. This furthers our point on using LLMs as a preference proxy (than as a preference substitute).

5. User Study : How aligned are subtask preferences as imagined by LLM with human users?

The experiments and results described in sections 3, 4 use a ground truth human preference either upon agent behavior or sub-task specification obtained from past research works. While these ground-truth preferences are catered by subject matter experts, we also test how well LLMs can act as a human proxy for sub-task specification.

5.1. Setup

We extend experiments of section 4 and designed a user study to answer the following questions :

1. Q1: How well does the LLMs predicted sub-task specification for the Overcooked domain aligns with a lay user's sub-task specification?
2. Q2: Whether a lay user finds LLMs predicted sub-task specification as ”human” generated?

We recruited 17 random participants with varying levels of experience with the Overcooked domain. We described the general theme of the game and showed various layouts of the game and the two agents acting to achieve the task. The layouts were drawn from prior (Carroll et al., 2019) research in multi-agent coordination tasks designed specifically to test certain key ideas like “forced coordination”, “asymmetric advantages” etc. We obtained 60s videos of agents acting in the Overcooked maps using the popular benchmark (Hum, 2023) with Human-Aware PPO agent. The participants were asked to create their own preference list, similar to the expert event-list described in the previous section. After completing their lists, they were shown two lists - one generated by the LLM (referred to them as ”unknown source”) and their own list from the previous step (referred to them as ”List B Created by you in the previous step”). Nex, they were asked to compare the two lists and rate on a Likert scale the degree to which they find the two lists aligned (Question 1 above). Finally, they were asked to rate on a Likert scale the degree to which they believed List A was generated by a human. Please refer to Appendix : figures 6, 7, for details on the study interface, event-lists, and the exact language used to phrase the questions.

Furthermore, as part of the study, we gathered event lists created by actual human users, which could be beneficial for further research.

5.2. Results

For the general Overcooked setting shown in Table 3, we first report the alignment of the responses with our ground truth expert event list. We find that GPT4 does a better job when coming up with event-list for the game in general (as compared to a specific layout). We speculate that this could because of its limited planning capabilities that hinder its ToM abilities to understand nuances of a specific layout. We also collected event-list as responses from our participants in our user study and tested whether the events given by the LLM matches to any of the events given by atleast one participant. While LLM responses may not exactly match an individual participant's response this test allows us to analyse whether there exists atleast one person with similar event-list item as that of the LLM. We find that text-davinci

Table 3: Experiments on testing Theory-of-Mind capabilities of LLMs across the general Overcooked game ground truth as per the domain's provided description, and the user study.

	Overcooked	
Source/ Model	**# Matches (out of 7) w\Ground Truth**	**# Matches (out of 7) w\User Study**
text-davinci-001	4	7
text-davinci-002	4	6
text-davinci-003	3	4
text-ada-001	0	0
text-babbage-001	1	0
text-curie-001	0	0
gpt-3.5-turbo	3	6
gpt4	5	7

models and gpt3.5 and gpt4 models performed exceptionally well achieving close to perfect scores.

The answer to the above-mentioned Q1 tells us that participants also believed that LLMs aligned with their preferences with an average agreement of 3.8/5 ±0.7 on a Likert scale of 1-5. Hence, we infer that LLMs are indeed a reasonable proxy for human preferences, but should not be confused with being a substitute.

For the second question, we note that an agreement of 3.5/5±1.3 from the participants which is indeed borderline, and hence, no real consensus can be drawn on whether the participants really believed that the list was generated from another human. One reason for this could be because typically, the human generated lists were not as structured, lengthy and impressively written.

From the user study, we also note that users, in general, stuck to listing events in their responses which are supposed to be boolean predicates. Very small percentage (17%) of the total participants mentioned strategy-specific responses. Moreover, very few people gave infeasible answers that involved actions or objects not present in the game description. 15% of the people gave more than 2 infeasible answers, while 53% people gave all feasible answers.

6. Conclusion

In this work, we explore the role of large language models to serve as a human proxy for providing answers to preference queries by an AI agent employing LLMs for its Theory of Mind capabilities. Among the several manifestations of human preferences, we explore two key human preferences as explicability preference of a human observer and a sub-task specification preference of a human co-actor in a human-AI team. We borrowed suitable scenarios to probe LLMs for their Theory of Mind abilities to answer whether a human in the loop would find a certain agent behavior explicable, or what sub-tasks would the human in the loop come up with for the team to pursue. We evaluate eight GPT-based models on three explicability preference tasks and three sub-task preference layouts in the Overcooked domain. We also conducted a human user study to confirm that LLMs do show Theory of Mind abilities to be a preference proxy, however, they can provide incorrect reasoning. We also discovered that the study participants generally concurred that there was a substantial correlation between the sub-task list they would have created and what the LLM had provided.

We finally conclude that for these tasks, LLM showed promise to be used as a human proxy. While the earlier LLM models struggled, newer models can perform much better, and real humans agree that it is good enough for these sub-task specification preferences. We hope that future research in learning from and identifying preferences of humans in the loop can utilize our findings and cautiously use LLM for its Theory of Mind capabilities.

References

Overcooked-ai. `https://github.com/HumanCompatibleAI/overcooked_ai`, 2023. Accessed: 2023.

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Albuquerque, N., Guo, K., Wilkinson, A., Savalli, C., Otta, E., and Mills, D. Dogs recognize dog and human emotions. *Biology letters*, 12(1):20150883, 2016.

Astington, J. W. and Baird, J. A. *Why language matters for theory of mind*. Oxford University Press, 2005.

Baker, C., Saxe, R., and Tenenbaum, J. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Çelikok, M. M., Peltola, T., Daee, P., and Kaski, S. Interactive ai with a theory of mind. *arXiv preprint arXiv:1912.05284*, 2019.

Chakraborti, T., Briggs, G., Talamadupula, K., Zhang, Y., Scheutz, M., Smith, D., and Kambhampati, S. Planning for serendipity. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5300–5306. IEEE, 2015.

Chakraborti, T., Sreedharan, S., Zhang, Y., and Kambhampati, S. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317*, 2017.

Chakraborti, T., Kulkarni, A., Sreedharan, S., Smith, D. E., and Kambhampati, S. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In *Proceedings of the international conference on automated planning and scheduling*, volume 29, pp. 86–96, 2019.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Cuzzolin, F., Morelli, A., Cirstea, B., and Sahakian, B. J. Knowing me, knowing you: theory of mind in ai. *Psychological medicine*, 50(7):1057–1061, 2020.

Dennett, D. C. Toward a cognitive theory of consciousness. 1978.

Edwards, B. D., Day, E. A., Arthur Jr, W., and Bell, S. T. Relationships among team ability composition, team mental models, and team performance. *Journal of applied psychology*, 91(3):727, 2006.

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., and Berner, J. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.

Guan, L., Verma, M., Guo, S. S., Zhang, R., and Kambhampati, S. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. *Advances in Neural Information Processing Systems*, 34:21885–21897, 2021.

Hagendorff, T. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*, 2023.

Heyes, C. M. and Frith, C. D. The cultural evolution of mind reading. *Science*, 344(6190):1243091, 2014.

Hu, H. and Sadigh, D. Language instructed reinforcement learning for human-ai coordination. *arXiv preprint arXiv:2304.07297*, 2023.

Jara-Ettinger, J. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29: 105–110, 2019.

Kambhampati, S., Sreedharan, S., Verma, M., Zha, Y., and Guan, L. Symbols as a lingua franca for bridging human-ai chasm for explainable and advisable ai systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12262–12267, 2022a.

Kambhampati, S., Sreedharan, S., Verma, M., Zha, Y., and Guan, L. Symbols as a lingua franca for bridging human-ai chasm for explainable and advisable ai systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12262–12267, 2022b.

Korinek, A. Language models and cognitive automation for economic research. Technical report, National Bureau of Economic Research, 2023.

Kosinski, M. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

Lee, K., Smith, L., and Abbeel, P. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

Lim, B.-C. and Klein, K. J. Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 27(4):403–418, 2006.

Milligan, K., Astington, J. W., and Dack, L. A. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child development*, 78(2):622–646, 2007.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., and Gabrieli, J. D. Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences*, 108 (7):2688–2692, 2011.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Park, J., Seo, Y., Shin, J., Lee, H., Abbeel, P., and Lee, K. Surf: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. *arXiv preprint arXiv:2203.10050*, 2022.

Rescorla, M. The computational theory of mind. 2015.

Sap, M., LeBras, R., Fried, D., and Choi, Y. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*, 2022.

Seyfarth, R. M. and Cheney, D. L. Affiliation, empathy, and the origins of theory of mind. *Proceedings of the National Academy of Sciences*, 110(supplement_2):10349–10356, 2013.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.

Soni, U., Sreedharan, S., Verma, M., Guan, L., Marquez, M., and Kambhampati, S. Towards customizable reinforcement learning agents: Enabling preference specification through online vocabulary expansion. *arXiv preprint arXiv:2210.15096*, 2022.

Sreedharan, S., Kambhampati, S., et al. Balancing explicability and explanation in human-aware planning. In *2017 AAAI Fall Symposium Series*, 2017.

Sreedharan, S., Soni, U., Verma, M., Srivastava, S., and Kambhampati, S. Bridging the gap: Providing post-hoc symbolic explanations for sequential decision-making problems with inscrutable representations. *arXiv preprint arXiv:2002.01080*, 2020.

Verma, M. and Metcalf, K. Symbol guided hindsight priors for reward learning from human preferences. *arXiv preprint arXiv:2210.09151*, 2022.

Verma, M., Bhambri, S., and Kambhampati, S. Exploiting unlabeled data for feedback efficient human preference based reinforcement learning. *arXiv preprint arXiv:2302.08738*, 2023.

Wellman, H. M., Cross, D., and Watson, J. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3):655–684, 2001.

Xie, Y., Yu, C., Zhu, T., Bai, J., Gong, Z., and Soh, H. Translating natural language to planning goals with large-language models. *arXiv preprint arXiv:2302.05128*, 2023.

Yu, C., Gao, J., Liu, W., Xu, B., Tang, H., Yang, J., Wang, Y., and Wu, Y. Learning zero-shot cooperation with humans, assuming humans are biased. *arXiv preprint arXiv:2302.01605*, 2023.

Zhang, J., Hedden, T., and Chia, A. Perspective-taking and depth of theory-of-mind reasoning in sequential-move games. *Cognitive science*, 36(3):560–573, 2012.

Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H. H., and Kambhampati, S. Plan explicability and predictability for robot task planning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 1313–1320. IEEE, 2017.