

A Mental-Model Centric Landscape of Human-AI Symbiosis

Zahra Zahedi¹ Sarath Sreedharan² * Subbarao Kambhampati¹

¹ Arizona State University ² Colorado State University
 zzahedi@asu.edu, sarath.sreedharan@colostate.edu, rao@asu.edu

Abstract

There has been significant recent interest in developing AI agents capable of effectively interacting and teaming with humans. While each of these works try to tackle a problem quite central to the problem of human-AI interaction, they tend to rely on myopic formulations that obscure the possible inter-relatedness and complementarity of many of these works. The human-aware AI framework was a recent effort to provide a unified account for human-AI interaction by casting them in terms of their relationship to various mental models. Unfortunately, the current accounts of human-aware AI are insufficient to explain the landscape of the work doing in the space of human-AI interaction due to their focus on limited settings. In this paper, we aim to correct this shortcoming by introducing a significantly general version of human-aware AI interaction scheme, called *generalized human-aware interaction (GHAI)*, that talks about (mental) models of six types. Through this paper, we will see how this new framework allows us to capture the various works done in the space of human-AI interaction and identify the fundamental behavioral patterns supported by these works. We will also use this framework to identify potential gaps in the current literature and suggest future research directions to address these shortcomings.

Introduction

As AI systems are becoming an increasing part of our day-to-day lives, there is a growing interest in designing AI systems that can interact with and team with humans fluently and effectively. As such, there is an increasing body of scholarly works that have tried to address various challenges related to human-AI interaction in different settings. The richness and diversity of challenges related to human-AI interaction mean the works tend to use diverse tools and formulations to address the challenges they are trying to address. Unfortunately, this has also meant it is hard to make sense of the overall landscape of the field owing to the lack of a unifying framework, a shortcoming we hope to address in this paper.

Within psychology, mental models and in particular the theory of mind [1] have been proposed as a mechanism to make sense of human-human interaction. One could potentially extend the idea theory of mind and the use of mental

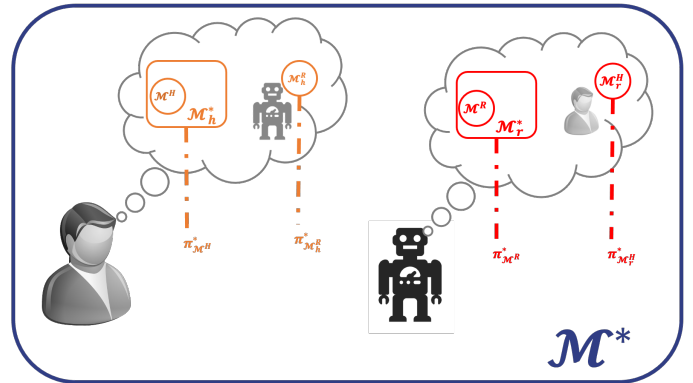


Figure 1: The six models in the GHAI framework. \mathcal{M}^* are the ground truth models of the task; \mathcal{M}^H and \mathcal{M}^R are the task models that the human and the AI agent ascribe to themselves; \mathcal{M}_h^R and \mathcal{M}_r^H are the estimates of the AI agent’s (human’s) model that human (AI agent) has.

models to also ground and categorize interaction between humans and AI systems. The framework of human-aware AI [2] has been proposed to incorporate the intuition of theory of mind into the context of human-AI interaction. Unfortunately, the current versions of the human-aware AI framework are insufficient to adequately explain the various works in the area. This is in part because the original human-aware AI framework was introduced to capture settings where the role of the human is limited to being a passive observer trying to make sense of an AI agent’s decisions. These scenarios only represent a small subset of possible ones that have been considered by the various works studied within the purview of human-AI interaction.

In this paper, we present a more general version of the human-aware AI interaction framework (Figure 1). GHAI will not only allow for scenarios where the human may be an actor but also introduce the notion of a task model \mathcal{M}^* that captures the true joint task specification of both the human and AI agent. Moreover, we will consider each agent’s (i.e. human or the AI agent) perception of the true combined task, which is independent of their beliefs about the other agent’s perception of it. This allows us to capture interaction scenarios where one of the agents may choose to correct the other agent’s beliefs about their task models. With the basic

*Work done while at Arizona State University.

framework in place, we will see how the primary interaction facilitated by the various works is manifestation of either a *model-communication behavior* or a *model-following behavior* made possible by the use of the various mental models that are part of the framework.

This paper brings together works from various communities that are studied under different problem frameworks including human-aware planning, human-AI cohabitation, human-robot interaction, human-in-the-loop ML, and human-AI symbiosis. The works considered here look at facilitating a diverse set of interactions, including explanation, teaming, teaching, and debugging, and have been presented in forums like CHI, HRI, ICAPS, IJCAI, ICML, and HCOMP. By bringing together this rich and varied body of works under a single umbrella, we highlight the centrality of the role played by mental models in the context of human-AI interaction and suggest direction.

Generalized Human-Aware Interaction

In this section, we describe the basic interaction framework that we will be leveraging to analyze the works that we are considering in this paper. The *Generalized Human-Aware Interaction* builds on earlier work done within the context of human-aware planning [2], that was introduced to characterize interpretable behavior generated in the context of a planning agent trying to generate behavior in the presence of a human observer. The framework allowed the agent to leverage human’s expectations about the agent in its decision-making process, by capturing these expectations in the form of a mental model that the human could be thought to ascribe to the agent. Unfortunately, the existing human-aware planning framework is insufficient to explain the wider landscape of works done in the area of human-AI interaction. So in this paper, we introduce *Generalized Human-Aware Interaction (GHAI)* that significantly expands the framework in multiple dimensions to explain and capture the extant literature. Figure 1 presents a diagrammatic overview of GHAI and in the rest of the section, we will provide an overview of the various components of the system.

Agents– Our central modeling objective is to capture the interaction between a human denoted by H and AI agent denoted by R (standing for robot, which we use in a more general sense). While our framework can be extended to cases when we have multiple humans and multiple AI agents, we will focus on this dyadic interaction to allow for more concise exposition. Unlike previous frameworks, we place no restrictions on either the roles played by the agents in the context of the interaction, nor any restrictions on how the objectives of each agent may or may not be related.

Model– The central tool we will be using to capture both the agent decisions and agent’s expectations from others, will be a task model. A task model (usually denoted as \mathcal{M}) is any mathematical model that encodes among other things, an entities beliefs about task objectives, state of the world and how the world may evolve on its own or in response to an agent action. There exist a wide variety of options for what the models might be (for example it may be MDPs [3], differential system equations or symbolic models like PDDL

[4]), but for the purposes of this paper, their exact form is irrelevant. We only require that the models used by an agent is in a form that they can use to derive the required decision or expected decisions. As apparent from Figure 1, GHAI consists of interaction between multiple models. For each individual model, we consider that we have to talk about the entity the model is ascribed to (i.e., the entity that is being described by the model) and the entity that the model belongs to (i.e., the entity that is using the model). In particular, we are interested in three entities, the two agents (denoted as H and R) and the true ground truth task (denoted with $*$). We will be using the superscript to denote the entity described by the model and the subscript will describe the entity that the model belongs to. As additional naming convention, we will use lowercase in the case of subscripts and drop it completely when both superscript and subscript matches. In particular, we only need to consider the following models.

Task models \mathcal{M}^* – This model is meant to capture the entirety of the task. As such it consists of all of the actions, objectives, preferences of both agents and additional facts about the world state that may not fall into the purview of the individual agent models. In this class, we have three specific models to consider. The ground truth model \mathcal{M}^* (same as saying \mathcal{M}_*^*) that is unknown to either of the agents, that captures the true objective and dynamics of the task. This model could be thought of as belonging to nature or to the true task-specifier that is not directly engaging in the interaction. Next we have the task model maintained by each of the agents (\mathcal{M}_r^* and \mathcal{M}_h^*). These models can be thought of as the union of their beliefs about their own model (will be covered later) and their beliefs about what the true model of the other agent should be and any additional task facts.

Model of each agent \mathcal{M}^R and \mathcal{M}^H – This is the model each agent ascribes to themselves. This determines what actions each agent believes they could perform and the objectives and preferences they are trying to satisfy. These models are also part of each agent’s beliefs about what the true task model is, i.e., \mathcal{M}^R is part of \mathcal{M}_r^* and \mathcal{M}^H is part of \mathcal{M}_h^* .

One agent’s model of the other \mathcal{M}_h^R and \mathcal{M}_r^H – These consist of what one agent believes the model of the other agent is, i.e, \mathcal{M}_h^R is the human’s belief about \mathcal{M}^R , and \mathcal{M}_r^H is the AI agent’s belief about \mathcal{M}^H . Note that we don’t expect these models to be same as what each agent thinks the true capabilities and objectives of the other agent may be. For example, from the perspective of the AI agent, what the AI agent thinks the human capabilities and objective should be (as encoded in \mathcal{M}_r^*) need not to be the same as what the AI agent thinks the human ascribes to themselves. These models capture the theories of mind captured by each agent.

These are all models we will need to consider to describe the space of works and various challenges related to human-AI symbiosis. One could of course consider additional models by including more nested beliefs, but in most practical scenarios people generally do not consider too many levels of nesting in their reasoning [5] and as we will see from our survey most existing works tend to avoid such nesting too. Next we will move onto the various decisions that may be

derived from these various models.

Decisions We will generally denote decisions as π . with the subscript corresponding to the model from which it is derived. Again, we will not limit ourselves to a specific decision type and depending on the exact problem the space of decisions could vary from single shot labels one may associate with a classification task to a policy to be carried out by an embodied agent. In some cases, there might also be multiple possible decisions that may be allowed by the model (referred to as valid decisions) and in these cases the model may also enforce an ordering among the possible decisions. We will refer to a decision as being optimal for a given model if there exist no other decision that may be preferred over the given one. The various decisions, we will consider in GHAI setting include,

Decisions made by the actor $\pi_{\mathcal{M}^{R/H}}$ – These are the decisions made by each agent to be carried out in the world (so these come into play when the agent is an actor). These decisions are primarily generated by using $\mathcal{M}^{R/H}$, but depending on the scenario, they may also consider their belief about the other agent’s task model in their decision.

The agent belief about the decision made by the other agent actor $\pi_{\mathcal{M}_{h/r}^{R/H}}$ – These are the decisions that an agent might ascribe to the other agent that is acting. They are primarily derived using their beliefs about the other agent (\mathcal{M}_h^R or \mathcal{M}_r^H).

Landscape of Human-AI Symbiosis

The models described in the previous section could come into play in human-AI interaction in various roles. In particular, one can see the models being used in two critical ways, namely, to drive a *model-communication* behavior wherein one of the agent uses it’s beliefs about the task to mould the belief of the other agent or to drive a *model-following* behavior wherein the agent chooses behaviors that considers the beliefs of the other agent. While previous works in human-aware AI have noted limited forms of such behavior [6], our focus on the larger landscape of human-AI interaction allows us to view these behavioral patterns in a much more general way. Thus we will try to categorize and analyze the works done in this space along these two dimensions. In the case of *model-communication* we will further break the works down based on the source of communication, i.e., is the AI-system trying to communicate to the human (thus the information originates from \mathcal{M}_r^*) or is the human trying to communicate to the AI-system (thus the information originates from \mathcal{M}_h^*). In the case of communication we will consider both explicit communication where the agents are directly providing model information or may engage in implicit communication where the agent tries to influence the other agent’s beliefs through their choice of behavior. Note that these categories aren’t mutually exclusive and as we will see many works in this space support behavior from multiple categories.

Model Information from AI Agent’s End (\mathcal{M}_r^*)

In this section, we have works that mostly use the models from the right side of Figure 1 (AI mental models) to update

models held by the human. For this we start with model information from \mathcal{M}_r^* that includes both their own AI model and their beliefs of the other agent true model. Within this category, we can further categorize works based on the human models that will be updated (i.e., \mathcal{M}_h^R or \mathcal{M}^H) and whether or not the AI agent makes use of a mental model of the human (\mathcal{M}_r^H) to generate the required information (see the summary in Table 1).

Updating \mathcal{M}_h^R The human uses \mathcal{M}_h^R to make sense of the AI’s decisions. So if the AI’s model is different from what the human believes of them, this can result in the human finding the AI behavior incomprehensible or surprising [7]. For this reason, when the AI agent needs to ensure that the human is able to correctly evaluate AI decisions, it could use information from \mathcal{M}_r^* to update \mathcal{M}_h^R , either implicitly or explicitly.

Examples of implicit model communication, include legible behavior generation or transparent planning which enables the human collaborators to infer the goal purely from the agent behavior [8, 9, 10, 11]. Implicit communication can also be used to obfuscate the plans and objectives of the AI agents to protect its privacy from adversarial entities [12] or to synthesize behaviors that are simultaneously legible to friendly entities and obfuscatory to adversarial ones [13]. The AI agent can also update \mathcal{M}_h^R to inform the human about its limitations, by implicitly communicating its incapacibilities through showing how and why it is unable to accomplish its current objectives [14]. It can also use non-verbal cues and feedback to signal how it expects the human to act next to communicate their preferences more effectively [15].

On the other hand, one of the common cases where explicit communication is used to update \mathcal{M}_h^R is in the context of explanations. The explanation may involve analyzing and explaining the details of failures and errors [16]. Such model updates can be good for debugging and troubleshooting purposes. The explanation can also solve the root cause of inexplicable behavior, by reconciling the difference between \mathcal{M}_h^R and AI model \mathcal{M}^R until the behavior becomes explicable to the human [17, 18, 19, 20, 21], or update \mathcal{M}_h^R until it can support aligned behavior (to be covered in detail later) [22, 23]. Also, if the behavior expected by the human is infeasible, the explanation updates the model by expressing incapability. This can be done by presenting unsolvability cores for specific behaviors [24, 25, 26, 27]. In the case of a joint workspace where both the human and robot acts, communicating the policy can update the human model of AI agent \mathcal{M}_h^R for better cooperation [28].

Updating \mathcal{M}^H In this case human is an actor and the robot is trying update the human’s model with what they believe to be information about the true underlying task. This update can again happen implicitly or explicitly. An example of implicit model update is when the robot makes a serendipitous plan to make the task easier for the human [29]. Explicit model updates happen in cases like decision support systems such as RADAR where the human-decision maker may be provided the previously unknown task information to aid their decision-making process of sin-

Table 1: Summary of the various aspects that come into play in cases where model information is coming from \mathcal{M}_r^*

References	Model information from \mathcal{M}_r^*		
	\mathcal{M}_r^H Used	Updating \mathcal{M}^H	Updating \mathcal{M}_h^R
[8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 22, 23, 24, 25, 26, 27, 28, 21]	✗	✗	✓
[35, 2, 20]	✓	✗	✓
[34]	✗	✓	✓
[29, 30, 31, 32, 33]	✓	✓	✗
[37]	✓	✓	✓

gle [30, 31, 32] and multiple [33] humans. It is also possible that the AI agent may be trying to update both \mathcal{M}_h^R and \mathcal{M}^H . This might be the case in joint workspaces where the AI agent needs to both communicate its intentions and actions to update \mathcal{M}_h^R and express what it expects the human to do and thus update \mathcal{M}^H [34].

Use of \mathcal{M}_r^H Here we will look at some works that allow the AI agent to use its mental model of the human when selecting the information to be provided. This includes both cases where the AI system is trying to update \mathcal{M}_h^R or \mathcal{M}^H . The former corresponds to cases when the behavior (implicit update) or explanation (explicit update) is personalized to the human model and preferences. Examples include providing faithful and customized explanations of a black box classifier that accounts for the fidelity to the original model as well as user interest [35], explanations expressed in the human’s vocabulary [36], or choosing explanation based on the human’s preference to the type of information [20].

All works discussed above that allow for the update of \mathcal{M}^H use \mathcal{M}_r^H as basis for the AI agent to decide what information to provide to the human. However it is worth noting that all the works discussed here implicitly assume that the AI agent’s belief about the human model is the same as the human’s own model, i.e., \mathcal{M}_r^H is same as \mathcal{M}^H .

There might also be cases where the AI system may be trying to update both \mathcal{M}^H and \mathcal{M}_h^R while Using \mathcal{M}_r^H . One example is the AI task allocation system AITA [37] that uses a negotiation based algorithm to fairly allocate different tasks to humans and giving contrastive explanation (using \mathcal{M}_r^H) to update both \mathcal{M}^H and \mathcal{M}_h^R , where \mathcal{M}_h^R captures team performance and constraints.

Model Information from Human’s End (\mathcal{M}_h^*)

In this section, we will consider works consider the use of the left side of Figure 1 (Human mental models) to update AI agent models. Thus, we have model information coming from \mathcal{M}_h^* that is used to update \mathcal{M}^R , \mathcal{M}_r^H , or both, and the human might use \mathcal{M}_h^R to select this information (Table 2).

Updating \mathcal{M}^R First we will consider works that use model information and inputs from human models to build, update and improve AI agent model \mathcal{M}^R . The challenge tackled by works that allow for such model updates is not

only to facilitate the update of \mathcal{M}^R but also in deciding how and when to use the model information from \mathcal{M}_h^* [38]. One common example is the problem of crowdsourcing task information from a group of people to create a more accurate task model for a single shot machine learning task (e.g. systems like Crowdsynth and CrowdExplorer) [39, 40, 41], and the challenge here would be to identify the most effective way to collect the information to learn the task [42, 39]. Other works have also looked at using similar settings to collect information about more complex tasks including planning [43]. Model information from, and decisions made using \mathcal{M}_h^* , could also be used indirectly to detect limitations and errors in the AI agent model \mathcal{M}^R , which can then be updated. Thus the information from \mathcal{M}_h^* is used to troubleshoot and then update the model \mathcal{M}^R . For instance, works have looked at analyzing how human and machine decisions differ for the same problem and using that information to update the AI model. [44] looked at using such an approach to improve the effectiveness of recidivism prediction, and [45] also looks at improving the AI systems performance by troubleshooting the systems failure, but performs this analysis and model update iteratively. Furthermore, the model information from \mathcal{M}_h^* can be in the form of an advice or demonstration to the AI agent to update \mathcal{M}^R [46, 47, 15, 48].

Updating \mathcal{M}_r^H The AI agent can also use the input and information from human \mathcal{M}_h^* to make and refine the model they ascribe to the human (\mathcal{M}_r^H) [48, 49]. Examples of this include cases where the AI agent uses the answers to its questions to refine its understanding of the human model \mathcal{M}_r^H [50], or when the AI agent use the human critiques [31, 30, 33] and explanatory queries [32] raised by the human to update and refine \mathcal{M}_r^H .

Use of \mathcal{M}_h^R Now, we will look at works that expects the human to use their mental model of the AI agent when providing the information. This includes both cases where the human information is used to update \mathcal{M}_r^H or \mathcal{M}^R . The latter happens when the human is advising or teaching the AI agent while considering \mathcal{M}_h^R . The human can explicitly advise and give instruction to the AI agent by suggesting actions that the AI agent should take while learning [46, 47], or the human can use \mathcal{M}_h^R to implicitly teach the AI agent through demonstration [15].

There might be cases where the human information is used to update both \mathcal{M}^R and \mathcal{M}_r^H while using \mathcal{M}_h^R . Examples here include cases, where the human demonstration is used to detect blind spots of both the AI agent and the human. Here the demonstration is used by the AI agent to update \mathcal{M}^R and \mathcal{M}_r^H , so it can hand off the task to the most capable agent [48]. Another example includes the AI agent’s use of human instructions to operate in open world scenarios [49].

Model Following Behavior

While the previous section considered how the agents could use communication to address or resolve potential differences between models, in this section we will look

Table 2: Summary of the various aspects that come into play in cases where model information is coming from \mathcal{M}_h^*

References	Model information from \mathcal{M}_h^*		
	\mathcal{M}_h^R Used	Updating \mathcal{M}^R	Updating \mathcal{M}_r^H
[50, 31, 30, 33, 32]	✗	✗	✓
[38, 39, 40, 41, 39, 43, 44, 45]	✗	✓	✗
[46, 47, 15]	✓	✓	✗
[48, 49]	✓	✓	✓

at how agents may choose decisions that take into account multiple models. In particular, we will consider two possible classes of model following behavior. In the first case, called *Multi-Model Alignment* the agent chooses a behavior or decision that is not only valid in one model, but is also valid and even close to optimal in another model. The second case being *response-seeking behavior*, wherein the agent chooses behavior based on their beliefs about the model of the other agent, so as to elicit a specific response or behavior from the other agent. (see Table 3 as summary).

Multi-Model Alignment Current human-AI symbiosis literature have considered multiple models with which the agent could align it’s decisions.

\mathcal{M}_r^* and \mathcal{M}_h^R – An example of such alignment includes the problem studied by [51]. Here the human is final decision-maker and the AI agent is making recommendation to the human. The human is expected to gradually form a mental model of when the AI agent may fail (\mathcal{M}_h^R), and the AI agent can potentially align its recommendation with \mathcal{M}_h^R to generate decisions that are easier for the human to accept and thus optimizing for team performance. Additionally, if \mathcal{M}_r^* is updated overtime and can offer better accuracy, the AI agent can align in recommendation while making a trade off between accuracy and compatibility [52, 53]. Aligning AI agent’s behavior to \mathcal{M}_h^R has also been investigated for sequential decision-making problems under the umbrella of explicable behavior generation [54, 55]. In these works, the AI agent is the actor and when the two models \mathcal{M}_r^* and \mathcal{M}_h^R are different, the AI agent can choose to be explicable to the human by choosing plans that match human’s expectation rather than directly communicating any differences. One can also try to balance between communicating the differences and aligning with the models [22, 23].

\mathcal{M}_r^* and \mathcal{M}_r^H – One example for need for such alignment occurs when the human is the actor and the AI agent tries to come up with policies for the human. Here the AI agent needs to consider potential human limitations including sensor limitations to come up with policies that the human could execute while still maximizing its value in the true model [56].

\mathcal{M}_r^* and \mathcal{M}_h^* – The AI agent and the human can both use their beliefs about the true task model (which may contain different information) to converge on a final decision that agrees with both of these models. We will consider this as a scenario where the decision is being aligned to both \mathcal{M}_r^* and \mathcal{M}_h^* . One example of this is the AI-mix system that

Table 3: Summary of the various aspects that come into play in scenarios involving model following behavior

Model Following Behavior				
Multi-Model Alignment				Response-Seeking Behavior
\mathcal{M}_r^* & \mathcal{M}_h^R	\mathcal{M}_r^* & \mathcal{M}_r^H	\mathcal{M}_r^* & \mathcal{M}_h^*	\mathcal{M}_h^* & \mathcal{M}_h^R	\mathcal{M}_r^H
[51, 52, 53, 54, 55, 22, 23]	[56]	[59]	[57]	[58, 28, 50]

uses the crowd suggestions (i.e. information coming from \mathcal{M}_h^*) and criticizes generated from a partial system model (i.e. information that is part of \mathcal{M}_r^*) to come up with a plan that would be better than one that could have been generated by either parties on their own.

\mathcal{M}_h^* and \mathcal{M}_h^R – This includes scenarios in which the human uses the recommendations made by the AI for their own task. As mentioned earlier, the human makes a mental model of the AI agent’s error boundary that is built over time. Thus the human is effectively aligning the decision they derive using \mathcal{M}_h^* with the model \mathcal{M}_h^R when they accept the AI agent recommendation they believe is more likely to be true [57].

Response-Seeking Behavior One scenario where such behavior is used is in joint workspaces. In such cases, when the AI agent comes up with its decisions, it must consider the model they ascribe to the human \mathcal{M}_r^H to come up with potential coordination strategies. For example, a robot can employ human motion prediction in conjunction with a complete, time-optimal path planner to execute efficient and safe motion in a shared environment [58] or jointly reasoning about policy communication and acting in a collaborative manner [28]. Another work that falls in this category is [50], where the AI agent uses \mathcal{M}_r^H to come up with the right questions to further refine the model they ascribe to the human.

Gaps in Current Literature and Discussion

In this section, we analyze the gaps in the current literature identified by our framework and how the framework connects to some of the other aspects of human-AI interaction including trust, longitudinal interaction and possible differences in inferential capabilities of humans and AI agents. Through this section we hope to provide a brief sketch on some potential future directions for research in human-AI interaction in terms of the characterization of the landscape provided here.

Missing Interactions

First we will consider the parts of the interactions that have received less attention in the current literature and possible opportunities that are being overlooked. One salient observation we can make about the current landscape of work is the focus on settings that only support flow of model information from one side to be used to update the other side. Very few current works (for example [30, 32, 31, 33]) assume the possibility that both agents may have information to provide to the other agent. Arguably, this would be closer to real-life settings and an aspect that may become more ap-

parent as we start allowing for longitudinal and iterated interaction. In fact we would argue that the works that do support bidirectional communication are the ones that consider a form of iterative interaction. We will go over longitudinal and iterated interaction in a separate discussion section.

As we also go over the works in the space, we also see some recurring assumptions about the models. For example, most of the works consider that the model of an agent ascribed to the other agent $\mathcal{M}_{r/h}^{H/R}$ is the same as the true task model they ascribe to the agent (i.e. it coincides with the model information they ascribe to the other agent as part of $\mathcal{M}_{r/h}^*$). Additionally, many of the works assume that the update of one agent’s mental model of the other agent will also cause the agent to update what they believe the true task model of the other agent should be, for example updating \mathcal{M}_r^H causes the robot to update \mathcal{M}_r^* . However, this assumption need not always hold and we could have scenarios that allow for the fact that some agents may believe that the other agent is misinformed about their own task model, an aspect that could also play a role in bidirectional and iterative communication between the agents.

Another aspect of under studied aspect of human-AI interaction relates to the model following behavior. In particular for multi-model alignment, there are no existing works that allow for the possibility that human would try to generate behavior that involves \mathcal{M}^H and \mathcal{M}_r^H . It is easy to see why people may have overlooked this problem as in such cases the human is trying to accommodate for AI agent’s beliefs. However the systems might need to allow for such allowances to accurately capture human behavior in cases where human is aware of some limitation of the AI agent or know that the AI agent isn’t aware of some aspects of the human model.

Bidirectional Communication and Longitudinal Interaction

As mentioned earlier, we can see that most of the works in this area focus on one shot interaction or limit model updates to one direction. So, they assume the implicit or explicit model information from one model can result in the update to the other model. Many of the works that consider bidirectional updates such as [30, 32, 31, 33] focus on decision-support systems where the human is the final decision-maker. Such cases allow for the fact that the human could potentially reject explanations provided by the AI system or could raise explanatory queries that may reveal some previously unspecified preferences. Another work that has bidirectional update is [15] which considers the inverse of the previous setting. In this work, the robot is the final actor and it asks help from the human to update \mathcal{M}_r^R but does so by helping human build a more accurate \mathcal{M}_h^R so the human can provide more helpful information. Both of these works present a limited form of interactive setting, wherein the human and AI agent converge to a final decision over a number of different iterations. Though one would ideally want to see systems that go beyond just these limited iterative setting and consider truly longitudinal interaction, where the human-AI collaboration may extend

over a longer period of time that may cover multiple tasks. Here the results of previous interaction may influence and direct the form of future collaborations. Unfortunately, there aren’t works that can support such forms of interaction. In fact one crucial aspect we will need to better model to allow for such interaction would be an effective way to model trust between the agents [60, 61].

Trust in Human-AI Interaction

In longitudinal interaction, the exact model update to be provided can also depend on the level of trust. For instance, the validity of the assumption that updates to \mathcal{M}_h^R , will get reflected in \mathcal{M}_h^* depends on the human trust in the AI agent and their willingness to comply with the AI agent. Thus, if the human trust is low, then the difference between \mathcal{M}_h^* and \mathcal{M}_h^R may lead the human to think that the AI agent misunderstands the task. As a result, the AI agent’s decision to provide model updates or perform model alignment should also consider the human’s level of trust and it’s effect on reliance and compliance from the human’s end [62]. Furthermore, one could also have cases where if the human’s trust is too high, they may accept information from AI agent and update their model when they shouldn’t [63]. All of these point to the critical role trust plays in the context of longitudinal interaction. Ideally one would want the choice for when the AI agent should provide model information or perform model following behavior and when they shouldn’t, to be driven not only by the human’s trust level but also based on the possible effects the AI agent’s choices may have on the human trust [62]. All of this focuses on the human’s trust and how they determine their reception to the AI agent’s actions, one would also want to develop a similar mechanism at the AI agent’s end to decide when they should accept information or decide to rely on the human.

Knowledge vs. Computational Capability

The majority of the works covered in this paper consider that the potential source of disagreement between the human and the AI agent about any aspect of the problem stems from a purely epistemic or knowledge difference between the corresponding models. However, as one considers more complex problems, an equally relevant aspect becomes the agents ability to effectively reason with the given models. After all, humans are best captured computationally as bounded rational agent and even if they share the same model information the human and AI agent may come to differing decisions. While there are some preliminary works that tries to address this asymmetry most of them are in their infancy. One group of work try to use simplified model information or decisions that are easier for people to follow (examples include [64], [19]). Another set of works tries to build proxy models that tries to predict how human’s may react to particular model update or aligned behavior (examples include [54], [65]).

This brings us to the end of our paper. Through the introduction of the generalized human-aware AI interaction framework, we not only hope to introduce a tool to better understand the current landscape of works dealing with human-AI interaction, but also a way to identify potential limitations of current work and future research directions.

Acknowledgments

This research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1-2442, N00014-18-1-2840, N00014-9-1-2119, AFOSR grant FA9550-18-1-0067, DARPA SAIL-ON grant W911NF-19-2-0006, and a JP Morgan AI Faculty Research grant.

References

- [1] D. Premack and G. Woodruff, “Does the chimpanzee have a theory of mind?” *Behavioral and brain sciences*, vol. 1, no. 4, pp. 515–526, 1978.
- [2] S. Sreedharan, A. Kulkarni, and S. Kambhampati, “Explaining human-ai interaction: A planning perspective,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 16, no. 1, pp. 1–184, 2022.
- [3] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- [4] H. Geffner and B. Bonet, “A concise introduction to models and methods for automated planning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 8, no. 1, pp. 1–141, 2013.
- [5] J. R. Wright and K. Leyton-Brown, “Predicting human behavior in unrepeated, simultaneous-move games,” *Games and Economic Behavior*, vol. 106, pp. 16–37, 2017.
- [6] S. Sreedharan, A. Kulkarni, D. E. Smith, and S. Kambhampati, “A unifying bayesian formulation of measures of interpretability in human-ai interaction,” pp. 4602–4610, 2021.
- [7] T. Chakraborti, A. Kulkarni, S. Sreedharan, D. E. Smith, and S. Kambhampati, “Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior,” in *Proceedings of the international conference on automated planning and scheduling*, vol. 29, 2019, pp. 86–96.
- [8] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, “Legibility and predictability of robot motion,” in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.
- [9] A. Dragan and S. Srinivasa, “Generating legible motion,” *Robotics: Science and Systems*, 2013.
- [10] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, “Effects of robot motion on human-robot collaboration,” in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 51–58.
- [11] A. M. MacNally, N. Lipovetzky, M. Ramirez, and A. R. Pearce, “Action selection for transparent planning,” in *AAMAS*, 2018, pp. 1327–1335.
- [12] A. Kulkarni, S. Srivastava, and S. Kambhampati, “A unified framework for planning in adversarial and cooperative environments,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2479–2487.
- [13] —, “Signaling friends and head-faking enemies simultaneously: Balancing goal obfuscation and goal legibility,” *AAMAS*, pp. 1889–1891, 2020.
- [14] M. Kwon, S. H. Huang, and A. D. Dragan, “Expressing robot incapability,” in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 87–95.
- [15] S. H. Huang, I. Huang, R. Pandya, and A. D. Dragan, “Nonverbal robot feedback for human teachers,” *arXiv preprint arXiv:1911.02320*, 2019.
- [16] B. Nushi, E. Kamar, and E. Horvitz, “Towards accountable ai: Hybrid human-machine analyses for characterizing system failure,” in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 6, no. 1, 2018.
- [17] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, “Plan explanations as model reconciliation: Moving beyond explanation as soliloquy,” *IJCAI*, pp. 156–163, 2017.
- [18] S. Sreedharan, T. Chakraborti, and S. Kambhampati, “Explanations as model reconciliation—a multi-agent perspective.” in *AAAI Fall Symposia*, 2017, pp. 277–283.
- [19] S. Sreedharan, S. Srivastava, and S. Kambhampati, “Hierarchical expertise level modeling for user specific contrastive explanations.” in *IJCAI*, 2018, pp. 4829–4836.
- [20] Z. Zahedi, A. Olmo, T. Chakraborti, S. Sreedharan, and S. Kambhampati, “Towards understanding user preferences for explanation types in model reconciliation,” in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 648–649.
- [21] U. Soni, S. Sreedharan, and S. Kambhampati, “Not all users are the same: Providing personalized explanations for sequential decision making problems,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 6240–6247.
- [22] T. Chakraborti, S. Sreedharan, and S. Kambhampati, “Balancing explicability and explanations in human-aware planning,” *IJCAI*, pp. 1335–1343, 2019.
- [23] S. Sreedharan, T. Chakraborti, C. Muise, and S. Kambhampati, “Expectation-aware planning: A unifying framework for synthesizing and executing self-explaining plans for human-aware planning,” *AAAI Conference on Artificial Intelligence*, 20.
- [24] V. Raman, C. Lignos, C. Finucane, K. C. Lee, M. P. Marcus, and H. Kress-Gazit, “Sorry dave, i’m afraid i can’t do that: Explaining unachievable robot tasks using natural language.” in *Robotics: Science and Systems*, vol. 2, no. 1. Citeseer, 2013, pp. 2–1.
- [25] V. Raman and H. Kress-Gazit, “Towards minimal explanations of unsynthesizability for high-level robot behaviors,” in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 757–762.

- [26] G. M. Briggs and M. Scheutz, “‘sorry, i can’t do that’: Developing mechanisms to appropriately reject directives in human-robot interactions.” in *AAAI Fall Symposia*, 2015, pp. 32–36.
- [27] S. Sreedharan, S. Srivastava, D. Smith, and S. Kambhampati, “Why can’t you do that hal? explaining unsolvability of planning tasks,” in *International Joint Conference on Artificial Intelligence*, 2019.
- [28] V. V. Unhelkar, S. Li, and J. A. Shah, “Decision-making for bidirectional communication in sequential human-robot collaborative tasks,” in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 329–341.
- [29] T. Chakraborti, G. Briggs, K. Talamadupula, Y. Zhang, M. Scheutz, D. Smith, and S. Kambhampati, “Planning for serendipity,” in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 5300–5306.
- [30] S. Sengupta, T. Chakraborti, S. Sreedharan, S. G. Vadlamudi, and S. Kambhampati, “Radar-a proactive decision support system for human-in-the-loop planning,” in *AAAI Fall Symposia*, 2017, pp. 269–276.
- [31] S. Grover, S. Sengupta, T. Chakraborti, A. P. Mishra, and S. Kambhampati, “Radar: automated task planning for proactive decision support,” *Human-Computer Interaction*, vol. 35, no. 5-6, pp. 387–412, 2020.
- [32] K. Valmeekam, S. Sreedharan, S. Sengupta, and S. Kambhampati, “Radar-x: An interactive interface pairing contrastive explanations with revised plan suggestions,” *AAAI’21 Demonstrations Program and ICAPS Workshop on Explainable AI Planning (XAIP)*, 2020.
- [33] S. Sengupta, T. Chakraborti, and S. Kambhampati, “Ma-radar—a mixed-reality interface for collaborative decision making,” *ICAPS UISP*, 2018.
- [34] X. Gao, R. Gong, Y. Zhao, S. Wang, T. Shu, and S.-C. Zhu, “Joint mind modeling for explanation generation in complex human-robot collaborative tasks,” in *2020 29th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, 2020, pp. 1119–1126.
- [35] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, “Faithful and customizable explanations of black box models,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 131–138.
- [36] S. Sreedharan, U. Soni, M. Verma, S. Srivastava, and S. Kambhampati, “Bridging the gap: Providing post-hoc symbolic explanations for sequential decision-making problems with inscrutable representations,” *arXiv preprint arXiv:2002.01080*, 2020.
- [37] Z. Zahedi, S. Sengupta, and S. Kambhampati, “Why not give this work to them? explaining ai-moderated task-allocation outcomes using negotiation trees,” *NeurIPS Workshop on Cooperative AI*, 2020.
- [38] E. Kamar, “Directions in hybrid intelligence: Complementing ai systems with human intelligence.” in *IJCAI*, 2016, pp. 4070–4073.
- [39] E. Kamar and E. Horvitz, “Light at the end of the tunnel: A monte carlo approach to computing value of information,” in *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, 2013, pp. 571–578.
- [40] E. Kamar, A. Kapoor, and E. Horvitz, “Lifelong learning for acquiring the wisdom of the crowd,” in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.
- [41] E. Kamar and E. Horvitz, “Planning for crowdsourcing hierarchical tasks,” in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 2015, pp. 1191–1199.
- [42] B. Wilder, E. Horvitz, and E. Kamar, “Learning to complement humans,” *IJCAI*, pp. 1526–1533, 2020.
- [43] J. Gao, H. H. Zhuo, S. Kambhampati, and L. Li, “Acquiring planning knowledge via crowdsourcing,” in *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.
- [44] S. Tan, J. Adebayo, K. Inkpen, and E. Kamar, “Investigating human+ machine complementarity for recidivism predictions,” *arXiv preprint arXiv:1808.09123*, 2018.
- [45] B. Nushi, E. Kamar, E. Horvitz, and D. Kossmann, “On human intellect and machine failures: Troubleshooting integrative machine learning systems,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [46] O. Amir, E. Kamar, A. Kolobov, and B. Grosz, “Interactive teaching strategies for agent training,” *IJCAI*, 2016.
- [47] L. Torrey and M. Taylor, “Teaching on a budget: Agents advising agents in reinforcement learning,” in *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, 2013, pp. 1053–1060.
- [48] R. Ramakrishnan, E. Kamar, B. Nushi, D. Dey, J. Shah, and E. Horvitz, “Overcoming blind spots in the real world: Leveraging complementary abilities for joint execution,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6137–6145.
- [49] K. Talamadupula, G. Briggs, M. Scheutz, and S. Kambhampati, “Architectural mechanisms for handling human instructions for open-world mixed-initiative team tasks and goals,” *Advances in Cognitive System*, vol. 5, 2017.
- [50] S. Grover, D. Smith, and S. Kambhampati, “Model elicitation through direct questioning,” *ICAPS Workshop on Explainable AI Planning (XAIP)*, 2020.
- [51] G. Bansal, B. Nushi, E. Kamar, E. Horvitz, and D. S. Weld, “Is the most accurate AI the best teammate? optimizing AI for teamwork,” *AAAI*, pp. 11 405–11 414, 2021.

- [52] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz, “Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2429–2437.
- [53] J. Martinez, K. Gal, E. Kamar, and L. H. Lelis, “Improving the performance-compatibility tradeoff with personalized objective functions,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, 2021, pp. 5967–5974.
- [54] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. H. Zhuo, and S. Kambhampati, “Plan explicability and predictability for robot task planning,” in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1313–1320.
- [55] A. Kulkarni, Y. Zha, T. Chakraborti, S. G. Vadlamudi, Y. Zhang, and S. Kambhampati, “Explicable planning as minimizing distance from expected behavior,” in *AAMAS Conference proceedings*, 2019.
- [56] S. Gopalakrishnan, M. Verma, and S. Kambhampati, “Synthesizing policies that account for human execution errors caused by state-aliasing in markov decision processes,” *ICAPS Workshop on Explainable AI Planning (XAIP)*, 2021.
- [57] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, “Beyond accuracy: The role of mental models in human-ai team performance,” in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, 2019, pp. 2–11.
- [58] V. V. Unhelkar, P. A. Lasota, Q. Tyroller, R.-D. Buhai, L. Marceau, B. Deml, and J. A. Shah, “Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2394–2401, 2018.
- [59] L. Manikonda, T. Chakraborti, S. De, K. Talamadupula, and S. Kambhampati, “Ai-mix: using automated planning to steer human workers towards better crowdsourced plans,” in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 2, no. 1, 2014.
- [60] A. Xu and G. Dudek, “Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations,” in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 221–228.
- [61] S. Sengupta, Z. Zahedi, and S. Kambhampati, “To monitor or to trust: observing robot’s behavior based on a game-theoretic model of trust,” in *Proceedings of the Trust Workshop at AAMAS*, 2019.
- [62] Z. Zahedi, M. Verma, S. Sreedharan, and S. Kambhampati, “Trust-aware planning: Modeling trust evolution in longitudinal human-robot interaction,” *XAIP and PlanRob workshop, ICAPS 2021*, 2021.
- [63] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. S. Weld, “Does the whole exceed its parts? the effect of AI explanations on complementary team performance,” *CHI*, pp. 81:1–81:16, 2021.
- [64] S. Gopalakrishnan and S. Kambhampati, “Minimizing robot navigation-graph for position-based predictability by humans,” *AAMAS Conference proceedings*, 2022.
- [65] S. Sreedharan, A. O. Hernandez, A. P. Mishra, and S. Kambhampati, “Model-free model reconciliation,” *IJCAI*, pp. 587–594, 2019.