Human-Modeling in Sequential Decision-Making: An Analysis through the Lens of Human-Aware AI

Silvia Tulli[1]*, **Stylianos Loukas Vasileiou** [2], **Sarath Sreedharan** [3]

[1]Institute of Intelligent Systems and Robotics (ISIR) - CNRS - INSERM - Sorbonne University
[2]McKelvey School of Engineering at Washington University in St. Louis
[3]Department of Computer Science at Colorado State University
tulli@isir.upmc.fr, v.stylianos@wustl.edu, sarath.sreedharan@colostate.edu

Abstract

"Human-aware" has become a popular keyword used to describe a particular class of AI systems that are designed to work and interact with humans. While there exists a surprising level of consistency among the works that use the label human-aware, the term itself mostly remains poorly understood. In this work, we retroactively try to provide an account of what constitutes a human-aware AI system. We see that human-aware AI is a design oriented paradigm, one that focuses on the need for modeling the humans it may interact with. Additionally, we see that this paradigm offers us intuitive dimensions to understand and categorize the kinds of interactions these systems might have with humans. We show the pedagogical value of these dimensions by using them as a tool to understand and review the current landscape of work related to human-AI systems that purport some form of human modeling. To fit the scope of a workshop paper, we specifically narrowed our review to papers that deal with sequential decision-making and were published in a major AI conference in the last three years. Our analysis helps identify the space of potential research problems that are currently being overlooked. We perform additional analysis on the degree to which these works make explicit reference to results from social science and whether they actually perform user-studies to validate their systems. We also provide an accounting of the various AI methods used by these works.

Introduction

Artificial Intelligence (AI) is currently undergoing a transformational moment, signaling a paradigm shift in its application and perception. There is an escalating optimism surrounding AI-based systems and their potential to significantly enhance the lives of everyday users. This optimism is not just a theoretical construct but has also fostered an intense interest in the development of AI systems that are adept at collaborating with and assisting humans in meaningful ways.

As with any rapidly evolving research domain, this interest has spawned a proliferation of varied research clusters, each with its unique focus. A brief survey of the landscape in human-AI interaction research uncovers a plethora of terms that researchers employ to define their work. Prominent among these are *human-centered AI*, *human-compatible AI*, *human-in-the-loop AI*, and *human-aware AI*. To a novice in the field, these terms might appear bewildering, each suggesting subtly different goals, methodologies, or design principles. Some of these terminologies have become preferred nomenclatures within specific research communities, while others outline distinct research objectives or design paradigms.

Human-centered AI (or human-centric AI) places humans at the core of the design process, primarily focusing on enhancing the human user experience. A system adopting a human-centered approach may not necessarily be human-aware. In this context, being human-centric implies a focus on enhancing user experience by accounting for human factors, such as preferences, needs, and values, in engineering and design, without necessarily incorporating a detailed modeling of human behavior [6]. This contrasts with *human-compatible AI*, which is more about the type of problems being addressed, especially those related to AI safety and ethical considerations [14]. The term *human-in-the-loop AI* has evolved over time; earlier works emphasized more explicit human intervention in the decision-making process [36], but more recent interpretations often relate to machine learning scenarios where humans play a more collaborative role in the learning process of the AI system [29].

Our paper focuses on the last category, namely *human-aware AI*. To the best of our knowledge, this term has been used by multiple research groups and communities in a surprisingly consistent manner, making it a particularly intriguing area of study. The term "human-aware" first entered the AI lexicon consistently with the work of [1]. Even in these early stages, many of the characteristic features of subsequent research using this term were evident. This foundational work demonstrated how an AI agent, such as a robot, needs to model human behaviors and infer their intentions, particularly goals, to facilitate more fluid interactions. The connection between this modeling, mental models, and the theory of mind was noted by Devin and Alami [10], and this concept has since been expanded in later works (cf. [5]) to include a broader range of models. Subsequent research has utilized this premise to offer formal accounts of various phenomena in AI, including explainability [42, 49], trust [56], and value alignment [28], as well as more comprehensive models of human-AI interaction (cf. [19]).

The aim of this paper is not merely to present another for-

*Contact Author

mal account of some aspect of human-AI interaction. Rather, our goal is to step back and provide a general account of what it means for an AI approach to be human-aware. We then intend to apply this framework to analyze a wide array of current AI works that focus on some form of human interaction, assessing their alignment with the human-aware AI paradigm.

We assert that, unlike other terms discussed earlier, human-aware AI systems are characterized by two interrelated yet distinct features:[1]

- *Acknowledgment of Human Interaction* (F_1): This entails an explicit acknowledgement that the AI system will, at some point in its lifecycle, interact with humans.
- *Design Consideration for Human Interaction* (F_2): This feature goes beyond mere acknowledgment, requiring that the AI system's design considers human modeling to account for the anticipated human interaction.

Taking a closer look at F_1, we observe that virtually all AI systems, including those as remote as the Mars rovers, interact with humans in some capacity. However, our focus is on whether this interaction is explicitly acknowledged and integrated into the system's design. Many AI systems are initially conceived as single-agent systems, with human interaction considerations often incorporated as an afterthought. A relevant example is powerful Reinforcement Learning (RL) systems like Alphafold [18], which are designed to solve specific problems (e.g., protein folding) rather than focusing on end-user interaction.

In contrast, F_2 mandates that for an AI system to qualify as human-aware, its design must be influenced by the necessity of human interaction. This feature presupposes F_1, but distinguishing between the two adds clarity. A system designer might be aware that human interaction will occur but may deem explicit human modeling unnecessary for certain use cases. We argue that such a system still qualifies as a human-aware AI system, as the potential for human interaction was considered during its design phase.

It is crucial to note that being a human-aware AI system does not automatically imply effectiveness in this role. Echoing recent discussions in fields like explainable AI (XAI) [12], we propose that the most reliable method to evaluate a human-aware AI system's efficacy is through human subject studies. Thus, an effective human-aware AI system is one that demonstrates practical utility in real-world human interactions.

The rest of the paper is structured as follows: We begin with a discussion of the various roles that humans could play in a human-aware system. Next, we review recent papers from major AI conferences to evaluate whether they align with the features of human-aware AI as outlined above. This analysis will categorize these works based on the roles played by humans, the influence of human interaction on system design, and the degree of utilization of human interactions.

[1]These two features together imply that all human-aware AI systems are inherently multi-agent systems.

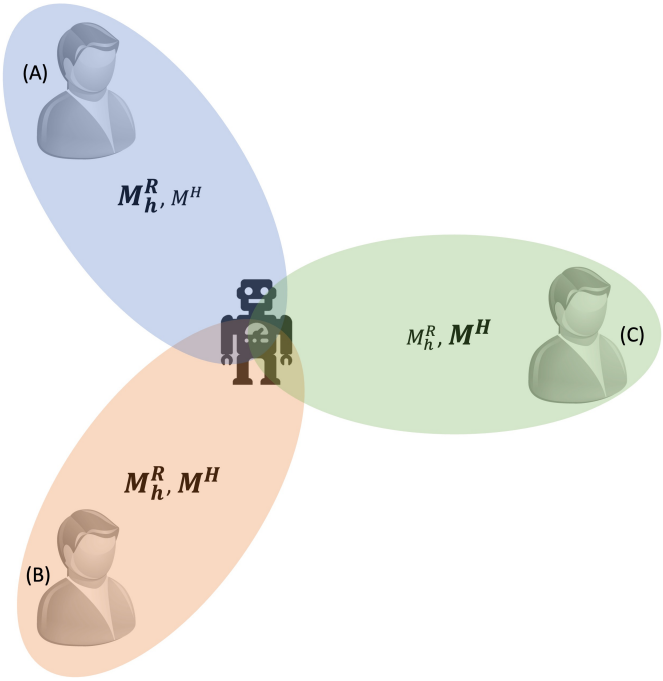

Figure 1: A visualization of the three modes in which the human and AI systems could interact and the models that would come into play in such interactions (the most important one is bolded and put in a slightly larger font). In the context of teaming or collaboration, the three categories correspond to (A) Supervisor/Teacher (B) Teammate (C) End-User.

Humans and Human-Aware AI System

We now examine the specifics of how humans may interact with AI systems. Figure 1 offers an overview of the broad categories into which human roles and interactions with AI might fall. As previously discussed, an explicit model of human behavior is not a prerequisite for a method to be categorized as human-aware AI. Nonetheless, models of human beliefs and capabilities serve as useful tools for understanding and categorizing different modes of interaction between humans and AI.

We approach human-aware AI from the perspective of multi-agent systems, where both humans and AI are considered agents. These agents' actions or decisions are predicated on internal models, which capture their knowledge of the task and decision-making processes. It is important to clarify that our use of "model" here is a heuristic tool for discussing agent states and behaviors, devoid of any assertions about the actual cognitive processes of humans or AI systems. This interpretation accommodates even model-free decision-making paradigms by assuming a basic model comprising reflexive rules.

While there are existing formulations that consider a lot more types of mental models (cf. [55]), for the purposes of this paper, we focus on two primary models:

- $\mathcal{M}^H$: The human model encompasses the human's knowledge about the world state, its capabilities, and cur-

rent goals/preferences.

- $\mathcal{M}_h^R$: The model of the human's perception of the AI agent, detailing the human's knowledge about the agent's understanding of the world, its capabilities, and goals.

These models underpin three principal categories of human roles in human-AI interactions:

1. *Supervisor/Teacher*: In this role, the human oversees the agent's operations, providing feedback or guidance. The critical model here is $\mathcal{M}_h^R$, as the supervisor must have an understanding of the agent's knowledge to offer relevant feedback. However, $\mathcal{M}^H$ also plays a role, as effective supervision requires an awareness of human capabilities and goals.
2. *Teammate*: Here, the human actively collaborates with the agent to achieve a mutual goal. Both models are crucially important, enabling the human to anticipate the agent's decisions ($\mathcal{M}_h^R$) and plan their own actions ($\mathcal{M}^H$). Similarly, the agent requires an understanding of the human's actions and expectations.
3. *End-User*: As an end-user, the human primarily interacts with the AI system as a beneficiary of its services. The predominant model in this scenario is $\mathcal{M}^H$, as the agent's goal is to assist the human. The human's perception of the agent ($\mathcal{M}_h^R$) also bears significance to the extent that it influences the human's expectations and utilization of the system.

It is important to note that these roles are not mutually exclusive; a single individual may assume multiple roles within the same or different interactions. Furthermore, these categories can be extended to multiple user scenarios and even to adversarial contexts, such as:

1. *Attacker*: Corresponding to the Supervisor/Teacher, with a focus on undermining the agent (primary model: $\mathcal{M}_h^R$).
2. *Rival*: Analogous to the Teammate, competing with the agent for resources or goals (both models are significant).
3. *Target*: Similar to the End-User, being the focus of the agent's adversarial actions (primary model: $\mathcal{M}^H$).

This paper, however, will concentrate on collaborative rather than adversarial aspects of human-AI interactions.

The works in this space, tends to characterize the level of human modeling using the following three dimensions [41]:

1. *Knowledge State*: This corresponds to human knowledge or belief, i.e., the contents of the specific models mentioned earlier.
2. *Inferential Capability*: This corresponds to how the human may make use of the given model to come up with plans or decisions. Generally, humans are widely accepted to be bounded rational agents [16], even though they are not always modeled as such. Also, it's worth considering that in the case of $\mathcal{M}_h^R$, the system would need to capture the inferential capability the human ascribes to the system itself.
3. *Vocabulary*: This corresponds to the terms in which the human represents and reasons about the task, which in turn could influence their decisions and interactions. It is worth noting that the same task could in theory be captured using different terms.

Methodology

In this section, we describe our methodology for evaluating recent works with respect to the human-aware AI criteria outlined in the previous section.

We opted for a review of papers published between 2020 and 2023 and targeted major AI conferences[2]. In particular, we opted for a keyword-based search within the Semantic Scholar database, using Semantic Scholar API[3] to gather and analyze literature. Our use of the API and fixed keywords also provide this approach with a level of reproducibility, that is usually missing from most surveys. While this method does not provide us with an exhaustive characterization of the entire landscape of AI methods, it does give us an overview of, at the very least, the recent trends in the field.

We filtered papers by searching for content including the term "human", as well as related terms identified from the search relevance algorithm of Semantic Scholar, and "compatible" or "aware" or "Theory of Mind (ToM)" or "modeling" and "plan". This was done to specifically identify studies that are directly relevant to human users. Additionally, we focused on work considering sequential decision-making processes, a critical aspect of human-AI interaction. Table 1 shows our inclusion criteria and search strings used on the Semantic Scholar API. We narrowed our focus to a more recent three-year period within our initial ten-year range.

Our filtering approach resulted in an initial pool 312 papers. We conducted an exploratory analysis of these papers to gain an overview and identify any general trends. Then, we conducted a manual filtration of this list, further eliminating unrelated work by assessing relevance based on titles, abstracts, and thorough readings of the full texts. We identified 66 relevant papers. To ensure the reliability of our selection, after this initial screening, we redistributed the papers among the authors. This ensured that each paper was deemed relevant by at least two authors. For papers with differing decisions, we discussed their relevance and made a joint decision regarding the paper's inclusion. This left us with a final list of 46 related papers.

Human Assumptions

In our first evaluation, we looked at some of the implicit and explicit assumptions concerning humans. While there is a shared understanding across these papers about the complexity of human behavior and the value of human input in AI systems, they differ in their focus on the nature of human emotions, the specificity of interaction contexts, the degree of human involvement, and the types of contributions humans are expected to make.

Firstly, there is a recognition of *knowledge asymmetry* between humans and AI agents regarding capabilities and

[2]AAAI, IJCAI, ECAI, and ICAPS

[3]https://www.semanticscholar.org/product/api

Topic	Description	Search Term
Human Involvement	Exclude papers without human involvement or considerations (e.g., position papers).	*human* AND *compatible* OR *aware* OR *ToM* OR *modeling*
Sequential Decision Making	Exclude papers that do not use an agent.	*plan*
Recency	Only consider the past 3 years.	*range*(2020, 2023)
Subject Area	Only consider papers from Computer Science.	*Computer Science*

Table 1: Inclusion Criteria and Search Strings used in the Semantics Scholar API.

preferences [44]. Additionally, the research acknowledges that humans often plan individually while also considering parallel planning with other humans [8]. Moreover, human decision-making involves various uncertainties and anxieties about future outcomes [46]. This uncertainty extends to beliefs about AI agents [44].

Human conversations are viewed as goal-oriented and guided by multiple small goals or a global goal [31]. Furthermore, human driving behavior is recognized as diverse and influenced by individual priorities and motivations [37] [38].

The need for *explanations* in human-AI interaction is emphasized in several papers [43, 47, 23, 50, 48, 39]. Additionally, personalized explanations are deemed essential [48]. Effective dialogue with humans also requires topic management [53].

Moreover, the assumption that diverse plans can be used as a proxy to cover unknown human preferences or that human preferences may be private or complex is highlighted [11].

One of the important takeaways from these assumptions is the fact that they act as a way to incorporate information about human models without dealing with the overhead of performing explicit modeling. While baking in fixed assumptions about humans could be limiting from a modeling point of view, we do see them being effective in the scenarios considered by the papers.

Human Models

An overwhelming percentage of papers look at modeling the human's *knowledge state*, with the majority of those papers focusing on using (or learning) $\mathcal{M}_h$. However, many of these works focus on different types of model representation and components. Several papers [31, 57, 20, 37] highlighted the importance of goal-oriented interactions in human-AI systems, which could be formalized using hierarchical or goal-based models. There were also papers that focused on modeling human preferences [53, 11] and discussing preference models, which could be formalized using utility functions and learned using preference elicitation techniques. Wang et al. [2023] and Tuli et al. [2021] explore approaches for learning the human model from human interactions, which could involve techniques such as reinforcement learning with human feedback or imitation learning.

A few papers addressed the modeling of human *inferential capabilities*. Sreedharan et al. [2021] and Amado and Meneguzzi [2020] employ Bayesian models to capture human inference processes, while the work of Zhang, Kemp, and Lipovetzky [2023] and De Peuter and Kaski [2023] address temporal aspects of human behavior and intention recognition.

Finally, the human *vocabulary* modeling was the least represented among the dimensions, with just a few exceptions [48, 23].

Priori/Posteriori Inclusion of the Human

Next, we looked at whether the human considerations were purely taken during the design/decision-making process or whether the system allowed for the humans to directly provide feedback and/or interact with it during the operation of the system. We refer to the former as *priori* inclusion of humans and the latter as *posterior* inclusion.

Examples of priori methods involve those that utilize human input or data as part of the planning or learning process (for example [57]). As a counterpart, examples of posteriori methods include those involve humans after initial planning or decision-making stages, incorporating human feedback or interaction to refine or adjust the system's behavior [3].

While not explicitly mentioned, some approaches may indirectly address value alignment through human-centric design or by considering human feedback in the planning process [60, 53, 11]. These works have discussed the integration of human preferences. Xu et al. [53] incorporates human-inspired strategies to ensure coherent and user-interest-aligned dialogues in open-domain conversation generation. This reflects a priori inclusion of human considerations without direct feedback during operation. Ghasemi et al. [11] introduced a diverse stochastic planning approach to generate varied plans that account for unknown or complex human preferences, considering human factors a priori.

Zheng et al. [60] embedded human roles and preferences into the system's design from the outset, with indirect mechanisms for incorporating human feedback iteratively.

Explanation generation emerges as a prominent theme in several works. While some approaches focus explicitly on generating formal explanations of AI decisions [39] (hence posteriori method), others implicitly aim to make plans or actions more understandable to human end-users through self-explanatory plans or interpretability measures [30] (hence priori methods).

Within explanation generation, one method we found to be quite popular is that of model reconciliation[26, 47, 23, 50, 48].

Role of the Human

The diverse roles humans play in AI systems, from passive data providers to active decision-makers, vary across research work. In many cases, humans serve as *end-users*, benefiting from the outcomes or decisions generated by AI systems without directly influencing any decisions in real-time [7, 17, 13, 27]. In other instances, humans take on more active roles, such as *supervisors* providing guidance or feedback to AI systems [15, 45, 51] or *teammates* collaborating with AI agents [3, 58]. Works [31, 4, 30, 34] have also highlighted the evolving role of humans in AI systems, where they serve both as supervisors and end-users. These examples demonstrate a shift towards more interactive and collaborative AI systems that integrate human feedback and guidance into the decision-making process.

Social Science Theories

In general, it is heartening to see that more works acknowledge the importance of incorporating social science perspectives into AI research, recognizing that human behavior, cognition, and societal dynamics play crucial roles in the development and deployment of AI systems [54, 46, 3, 53, 52, 33, 32, 57, 24, 48, 50].

Among the papers mentioning social sciences theories, there are some based in the concept of *theory of mind* [30, 50, 57, 23]. The theory of mind involves the ability to attribute mental states to oneself and others, enabling individuals to understand and predict behavior based on inferred beliefs, desires, and intentions [35]. Beyond the *theory of mind*, other social science theories mentioned include: *relevance theory* and *population movement and settlement patterns* [54, 48]. *Relevance theory* is a cognitive science theory that seeks to explain how utterances are interpreted [40]. *Population movement and settlement patterns* are focal points in human geography, demography, and urban planning, supported by various social science theories.

It is worth mentioning that while several remaining papers do not explicitly reference or integrate social science theories into their research, this does not necessarily imply a lack of consideration for human factors or societal implications. However, it does raise questions about the depth of understanding of these aspects.

Evaluation Methods and Metrics

We reviewed both *quantitative* and *qualitative* measures employed the relevant papers. Moreover, we aimed to evaluate the importance of incorporating *user studies* and baseline comparisons in the assessment methodologies.

Quantitative measures such as runtime [43], mean reward [8], prediction accuracy [25], precision, and recall [44] offer objective benchmarks for assessing the performance of systems across different tasks and domains. Some work prioritizes robustness metrics, aimed at evaluating the system's resilience against adversarial attacks, noise, or uncertainties in real-world scenarios. For instance, [21] use *max regret* to quantify the worst-case performance deviation from the optimal outcome, providing a measure of robustness against unforeseen circumstances. Similarly, in autonomous driving contexts, evaluation criteria such as success rate and runtime, as seen in [58, 59] reflect the system's ability to adapt and make decisions in dynamic environments.

Only a handful of papers surveyed opt to conduct user studies [31, 30, 23, 48, 50, 57, 53, 32]. Ni et al. [31] focus on the system's ability to facilitate goal-oriented conversations through multi-hierarchy learning. Netanyahu et al. [30] employs evaluation metrics like Average Displacement Error (ADE) and Final Displacement Error (FDE) to measure the accuracy of predictions in physically-grounded abstract social events. These metrics provide insights into the system's performance in understanding and predicting human behavior, useful for applications requiring social interaction and perception. Further, Kumar et al. [23] use metrics such as correction ratio and comprehension score to evaluate the effectiveness of their visualization techniques in conveying explanations to human users. Moreover, user-centric evaluation metrics, such as coherence (intra/inter-topic) and task completion time, are employed by [57] and [48] to assess the human user experience and task efficiency of AI systems.

AI Methods and Learning Paradigms

We finally examined the general AI methods and learning paradigms used across the relevant papers. Many papers use supervised learning for tasks such as understanding natural language [25], recognizing goals [37], and generating human-like dialogues [31]. Some researchers, such as [30], employ mixed methods for identifying and understanding social interactions [30]. They combine supervised learning techniques with reinforcement learning to parameterize learning nodes with learned policies. Additionally, they incorporate imitation learning methods to acquire behavior trees from human demonstration. Planning-based approaches are prevalent, particularly in tasks involving decision-making and action generation. Classical planning techniques are used in various papers [43, 44, 39]. Additionally, more specialized planning methods such as hierarchical planning [30] and behavior tree expansion algorithms [11] are employed in specific domains.

Takeaways

One of our first takeaways from the survey was that human-aware AI proved to be a surprisingly robust tool for analyz-

ing the landscape of papers. At a first glance, human-aware AI might seem like a limited framework to be used as an analysis tool, especially given the fact that most papers do not maintain and manipulate explicit representations of human mental models. On the other hand, works that acknowledge to be human-aware indeed account for explicit human models, with most coming from those related to explanation or related literature [50, 43].

Nevertheless, a closer look at all the papers revealed that many of them are built on top of assumptions that allow for the implicit modeling of humans. With this, in mind, we were able to easily categorize the assumptions into one of the dimensions discussed in the earlier section. We found that a vast majority of works focused on modeling knowledge state (particularly $\mathcal{M}_h$). *This shows a clear lack of work that focuses on $\mathcal{M}_h^R$, and as such are not as adaptive to the human's beliefs.* Moving away from the knowledge state, modeling of inferential capabilities and vocabulary was also considered by less works. The former could be explained by a general lack of robust tools to accurately capture and model human inferential capabilities. The most widely used model, i.e., noisy rational model [16], is known to be insufficient in many cases. In addition, we saw that there is less work overall in capturing vocabulary mismatch. This is particularly surprising given its prevalence within the larger XAI literature [22].

Moreover, we found that most works were focused on cases where the human assumes the role of the end user of the system. There were a few works that looked at humans as supervisors or teammates. However, this might be a result of the venue we chose or the keywords used. We expect to see more work if we had included robotic or multi-agent venues.

In regards to the inclusion of social science concepts and user studies, while there were works that addressed them, it was a clear minority. While most authors in the field publicly acknowledge the importance of both in works related to human-AI interaction, we see that in practice this is usually not the case.

Conclusion and Future Directions

In this survey paper, we hope to both provide a clear and concise description of what it means for an AI system to be considered human-aware. Starting with this description, we provide some characterization and properties of these models and then use it to perform an analysis of some recent works published in different prestigious AI conferences. In our analysis of the paper, we see many glaring omissions in terms of open problems and research opportunities. However, it is still worth noting that our paper focuses on a very small timeframe and only on four conferences. In the future, we hope to perform a more comprehensive survey that considers papers from a number of diverse venues over a larger timeframe.

References

[1] Alami, R.; Clodic, A.; Montreuil, V.; Sisbot, E. A.; and Chatila, R. 2005. Task planning for human-robot interaction. In *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-Aware Services: Usages and Technologies*, sOc-EUSAI '05, 81–85. New York, NY, USA: Association for Computing Machinery. ISBN 1595933042.

[2] Amado, L.; and Meneguzzi, F. 2020. LatRec: Recognizing Goals in Latent Space (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10): 13747–13748.

[3] Bara, C.-P.; Ma, Z.; Yu, Y.; Shah, J. A.; and Chai, J. Y. 2023. Towards Collaborative Plan Acquisition through Theory of Mind Modeling in Situated Dialogue. In *International Joint Conference on Artificial Intelligence*.

[4] Cai, Z.; Li, M.; Huang, W.; and Yang, W. 2021. BT Expansion: a Sound and Complete Algorithm for Behavior Planning of Intelligent Robots with Behavior Trees. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7): 6058–6065.

[5] Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2018. Human-aware planning revisited: A tale of three models. In *IJCAI-ECAI XAI/ICAPS XAIP Workshops*, volume 10.

[6] Chetouani, M.; Dignum, V.; Lukowicz, P.; and Sierra, C., eds. 2023. *Human-Centered Artificial Intelligence: Advanced Lectures*. Lecture Notes in Computer Science. Cham: Springer Cham, 1 edition. ISBN 978-3-031-24348-6.

[7] Czechowski, A.; and Oliehoek, F. A. 2020. Decentralized MCTS via Learned Teammate Models. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 81–88. International Joint Conferences on Artificial Intelligence Organization. Main track.

[8] Czechowski, A.; and Oliehoek, F. A. 2021. Decentralized MCTS via learned teammate models. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20. ISBN 9780999241165.

[9] De Peuter, S.; and Kaski, S. 2023. Zero-Shot Assistance in Sequential Decision Problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10): 11551–11559.

[10] Devin, S.; and Alami, R. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 319–326. IEEE.

[11] Ghasemi, M.; Scope Crafts, E.; Zhao, B.; and Topcu, U. 2021. Multiple Plans are Better than One: Diverse Stochastic Planning. *Proceedings of the International Conference on Automated Planning and Scheduling*, 31(1): 140–148.

[12] Gunning, D.; and Aha, D. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, 40(2): 44–58.

[13] Illanes, L.; Yan, X.; Toro Icarte, R.; and McIlraith, S. A. 2020. Symbolic Plans as High-Level Instructions for Reinforcement Learning. *Proceedings of the International Conference on Automated Planning and Scheduling*, 30(1): 540–550.

[14] Intelligence, H. C. A.; and the Problem of Control. 2019. *Stuart Russell*. Penguin Books.

[15] Jakubik, J.; Hemmer, P.; Vössing, M.; Blumenstiel, B.; Bartos, A.; and Mohr, K. 2022. Designing a Human-in-the-Loop System for Object Detection in Floor Plans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11, 12524–12530.

[16] Jeon, H. J.; Milli, S.; and Dragan, A. D. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 4415–4426. Virtual: Curran Associates, Inc.

[17] Jiang, W.; Zhao, W. X.; Wang, J.; and Jiang, J. 2023. Continuous Trajectory Generation Based on Two-Stage GAN. *ArXiv*, abs/2301.07103.

[18] Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

[19] Kambhampati, S. 2020. Challenges of human-aware ai systems: Aaai presidential address. *AI Magazine*, 41(3): 3–17.

[20] Katz, M.; Ram, P.; Sohrabi, S.; and Udrea, O. 2020. Exploring Context-Free Languages via Planning: The Case for Automating Machine Learning. *Proceedings of the International Conference on Automated Planning and Scheduling*, 30(1): 403–411.

[21] Killian, J.; Xu, L.; Biswas, A.; Verma, S.; Nair, V. J.; Taneja, A.; Hegde, A.; Madhiwalla, N.; Díaz, P. R.; Johnson-Yu, S.; and Tambe, M. 2023. Robust Planning over Restless Groups: Engagement Interventions for a Large-Scale Maternal Telehealth Program. In *AAAI Conference on Artificial Intelligence*.

[22] Kim, B.; Gilmer, J.; Wattenberg, M.; and Viégas, F. 2018. Tcav: Relative concept importance testing with linear concept activation vectors. In *International Conference on Learning Representations*.

[23] Kumar, A.; Vasileiou, S. L.; Bancilhon, M.; Ottley, A.; and Yeoh, W. 2022. VizXP: A Visualization Framework for Conveying Explanations to Users in Model Reconciliation Problems. *Proceedings of the International Conference on Automated Planning and Scheduling*, 32(1): 701–709.

[24] Kumar, A.; Vasileiou, S. L.; Bancilhon, M.; Ottley, A.; and Yeoh, W. 2022. Vizxp: A visualization framework for conveying explanations to users in model reconciliation problems. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 32, 701–709.

[25] Levy, A.; and Karpas, E. 2022. Understanding Natural Language in Context. In *International Conference on Automated Planning and Scheduling*.

[26] Lin, S.; and Bercher, P. 2021. Change the World - How Hard Can that Be? On the Computational Complexity of Fixing Planning Models. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4152–4159. International Joint Conferences on Artificial Intelligence Organization. Main Track.

[27] Liu, L.; Yang, Y.; Yuan, Y.; Shao, T.; Wang, H.; and Zhou, K. 2021. In-game residential home planning via visual context-aware global relation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 336–343.

[28] Mechergui, M.; and Sreedharan, S. 2024. Goal Alignment: Re-analyzing Value Alignment Problems Using Human-Aware AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10110–10118.

[29] Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; and Fernández-Leal, Á. 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4): 3005–3054.

[30] Netanyahu, A.; Shu, T.; Katz, B.; Barbu, A.; and Tenenbaum, J. B. 2021. PHASE: Physically-grounded abstract social events for machine social perception. In *35th AAAI Conference on Artificial Intelligence (AAAI)*.

[31] Ni, J.; Pandelea, V.; Young, T.; Zhou, H.; and Cambria, E. 2022. HiTKG: Towards Goal-Oriented Conversations via Multi-Hierarchy Learning. In *AAAI Conference on Artificial Intelligence*.

[32] Pang, H. N.; Parks, R.; Breazeal, C.; and Abelson, H. 2023. "How Can I Code A.I. Responsibly?": The Effect of Computational Action on K-12 Students Learning and Creating Socially Responsible A.I. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 16017–16024.

[33] Peuter, S. D.; and Kaski, S. 2022. Zero-Shot Assistance in Sequential Decision Problems. In *AAAI Conference on Artificial Intelligence*.

[34] Pingen, G. L. J.; van Ommeren, C. R.; van Leeuwen, C. J.; Fransen, R.; Elfrink, T.; de Vries, Y. C.; Karunakaran, J.; Demirovic, E.; and Yorke-Smith, N. 2022. Talking Trucks: Decentralized Collaborative Multi-Agent Order Scheduling for Self-Organizing Logistics. In *International Conference on Automated Planning and Scheduling*.

[35] Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4): 515–526.

[36] Retzlaff, C.; Das, S.; Wayllace, C.; Mousavi, P.; Afshari, M.; Yang, T.; Saranti, A.; Angerschmid, A.; Taylor, M. E.; and Holzinger, A. 2024. Human-in-the-Loop Reinforcement Learning: A Survey and Position

on Requirements, Challenges, and Opportunities. *J. Artif. Intell. Res.*, 79: 359–415.

[37] Sarkar, A.; and Czarnecki, K. 2021. Solution Concepts in Hierarchical Games Under Bounded Rationality With Applications to Autonomous Driving. In *AAAI Conference on Artificial Intelligence*.

[38] Sarkar, A.; Larson, K.; and Czarnecki, K. 2021. Generalized dynamic cognitive hierarchy models for strategic driving behavior. In *AAAI Conference on Artificial Intelligence*.

[39] Selvey, R.; Grastien, A.; and Thiébaux, S. 2023. Formal Explanations of Neural Network Policies for Planning. In *International Joint Conference on Artificial Intelligence*.

[40] Sperber, D.; and Wilson, D. 1995. *Relevance: Communication and Cognition*. 97355-000. Malden: Blackwell Publishing, 2nd edition. ISBN 0-631-19878-4.

[41] Sreedharan, S. 2023. Human-aware AI—A foundational framework for human–AI interaction. *AI Magazine*, 44(4): 460–466.

[42] Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2021. Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301: 103558.

[43] Sreedharan, S.; Chakraborti, T.; Muise, C.; and Kambhampati, S. 2020. Expectation-aware planning: A unifying framework for synthesizing and executing self-explaining plans for human-aware planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2518–2526.

[44] Sreedharan, S.; Kulkarni, A.; Smith, D. E.; and Kambhampati, S. 2021. A Unifying Bayesian Formulation of Measures of Interpretability in Human-AI. In *International Joint Conference on Artificial Intelligence*.

[45] Tuli, S.; Bansal, R.; Paul, R.; and , M. 2021. TANGO: Commonsense Generalization in Predicting Tool Interactions for Mobile Manipulators. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4197–4205. International Joint Conferences on Artificial Intelligence Organization. Main Track.

[46] Vanhée, L.; Jeanpierre, L.; and Mouaddib, A.-I. 2022. Anxiety-Sensitive Planning: From Formal Foundations to Algorithms and Applications. In *International Conference on Automated Planning and Scheduling*.

[47] Vasileiou, S. L.; Previti, A.; and Yeoh, W. 2021. On exploiting hitting sets for model reconciliation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6514–6521.

[48] Vasileiou, S. L.; and Yeoh, W. 2023. PLEASE: Generating Personalized Explanations in Human-Aware Planning. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*. Proceedings of the European Conference on Artificial Intelligence (ECAI).

[49] Vasileiou, S. L.; Yeoh, W.; Son, T. C.; Kumar, A.; Cashmore, M.; and Magazzeni, D. 2022. A Logic-based Explanation Generation Framework for Classical and Hybrid Planning Problems. *Journal of Artificial Intelligence Research*, 73: 1473–1534.

[50] Vasileiou, S. L.; Yeoh, W.; Tran, S.; Kumar, A.; Cashmore, M.; and Magazzeni, D. 2023. A Logic-based Explanation Generation Framework for Classical and Hybrid Planning Problems (Extended Abstract). In Elkind, E., ed., *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 6985–6989. International Joint Conferences on Artificial Intelligence Organization. Journal Track.

[51] Wang, D.; Wu, L.; Zhang, D.; Zhou, J.; Sun, L.; and Fu, Y. 2023. Human-instructed deep hierarchical generative learning for automated urban planning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press. ISBN 978-1-57735-880-0.

[52] Wang, S.; Hu, L.; Wang, Y.; Sheng, Q. Z.; Orgun, M. A.; and Cao, L. 2020. Intention2Basket: A Neural Intention-driven Approach for Dynamic Next-basket Planning. In *International Joint Conference on Artificial Intelligence*.

[53] Xu, J.; Wang, H.; Niu, Z.; Wu, H.; and Che, W. 2020. Knowledge Graph Grounded Goal Planning for Open-Domain Conversation Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 9338–9345.

[54] Yang, X.; and Liu, W. 2020. Population Location and Movement Estimation through Cross-domain Data Analysis. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 5192–5193. International Joint Conferences on Artificial Intelligence Organization. Doctoral Consortium.

[55] Zahedi, Z.; Sreedharan, S.; and Kambhampati, S. 2022. A mental-model centric landscape of human-ai symbiosis. *arXiv preprint arXiv:2202.09447*.

[56] Zahedi, Z.; Sreedharan, S.; and Kambhampati, S. 2023. A mental model based theory of trust. *arXiv preprint arXiv:2301.12569*.

[57] Zhang, C.; Kemp, C.; and Lipovetzky, N. 2023. Goal Recognition with Timing Information. *Proceedings of the International Conference on Automated Planning and Scheduling*, 33(1): 443–451.

[58] Zhang, Y.; and Williams, B. C. 2023. Adaptation and Communication in Human-Robot Teaming to Handle Discrepancies in Agents' Beliefs about Plans. In *International Conference on Automated Planning and Scheduling*.

[59] Zhang, Y.; Zhang, J.; Zhang, J.; Wang, J.; Lu, K.; and Hong, J. 2020. A Novel Learning Framework for Sampling-Based Motion Planning in Autonomous

Driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01): 1202–1209.

[60] Zheng, G.; Liu, H.; Xu, K.; and Li, Z. J. 2021. Objective-aware Traffic Simulation via Inverse Reinforcement Learning. In *International Joint Conference on Artificial Intelligence*.