

Challenges of Human-Aware AI Systems

Subbarao Kambhampati*
Arizona State University
rao@asu.edu

Abstract

From its inception, AI has had a rather ambivalent relationship to humans—swinging between their augmentation and replacement. Now, as AI technologies enter our everyday lives at an ever increasing pace, there is a greater need for AI systems to work synergistically with humans. To do this effectively, AI systems must pay more attention to aspects of intelligence that helped humans work with each other—including social intelligence. I will discuss the research challenges in designing such human-aware AI systems, including modeling the mental states of humans in the loop, recognizing their desires and intentions, providing proactive support, exhibiting explicable behavior, giving cogent explanations on demand, and engendering trust. I will survey the progress made so far on these challenges, and highlight some promising directions. I will also touch on the additional ethical quandaries that such systems pose. I will end by arguing that the quest for human-aware AI systems broadens the scope of AI enterprise, necessitates and facilitates true inter-disciplinary collaborations, and can go a long way towards increasing public acceptance of AI technologies.

*This article is based on the AAAI 2018 Presidential Address that the author had the honor of delivering in New Orleans in February 2018. The video of the talk, along with the slides used, is available at <http://bit.ly/2tHyzAh>

Introduction

Artificial Intelligence, the discipline we all call our intellectual home, is suddenly having a rather huge cultural moment. It is hard to turn anywhere without running into mentions of AI technology and hype about its expected positive and negative societal impacts. AI has been compared to fire *and* electricity, and commercial interest in the AI technologies has sky rocketed. Universities – even high schools – are rushing to start new degree programs or colleges dedicated to AI. Civil society organizations are scrambling to understand the impact of AI technology on humanity, and governments are competing to encourage or regulate AI research and deployment.

There is considerable hand-wringing by pundits of all stripes on whether in the future, AI agents will get along with us or turn on us. Much is being written about the need to make AI technologies safe and delay the “doomsday.” I believe that as AI researchers, we are not (and cannot be) passive observers. It is *our* responsibility to design agents that can and will get along with us. Making such *human-aware* AI agents, however poses several foundational research challenges that go beyond simply adding user interfaces *post facto*. I will argue that addressing these challenges broadens the scope of AI in fundamental ways.

The need for Human-Aware AI Systems

My primary aim in this talk is to call for an increased focus on human-aware AI systems—goal directed autonomous systems that are capable of effectively interacting, collaborating and teaming with humans.¹ Although developing such systems seems like a rather self-evidently fruitful enterprise, and popular imaginations of AI, dating back to HAL, almost always assume we already do have human-aware AI systems technology, little of the actual energies of the AI research community have gone in this direction.

¹In a way, it thus follows in the footsteps of Barbara Grosz’s AAAI Presidential Address [17], which talked about collaborative systems.



Figure 1: *We should build a future where AI systems can be our quotidian partners*

From its inception, AI has had a rather ambivalent relationship to humans—swinging between their augmentation and replacement. Most high profile achievements of AI have either been far away from the humans—think Spirit and Opportunity exploring Mars; or in a decidedly adversarial stance with humans, be it Deep Blue, AlphaGo or Libatus. Research into effective ways of making AI systems *interact, team and collaborate with humans* has received significantly less attention. It is perhaps no wonder that many lay people have fears about AI technology!

This state of affairs is a bit puzzling given the rich history of early connections between AI and psychology. Part of the initial reluctance to work on these issues had to do with the worry that focusing on AI systems working with human might somehow dilute the grand goals of the AI enterprise, and might even lead to temptations of “cheating,” with most of the intelligent work being done by the humans in the loop. After all, prestidigitation has been a concern since the original mechanical turk. Indeed, much of the early

work on human-in-the-loop AI systems mostly focused on using humans as a crutch for making up the limitations of the AI systems [1]. In other words, early AI had humans be “AI-aware” (rather than AI be “human-aware”).

Now, as AI systems are maturing with increasing capabilities, the concerns about them depending on humans as crutches are less severe. I would also argue that focus on humans in the loop doesn’t dilute the goals of AI enterprise, but in fact broadens them in multiple ways. After all, evolutionary theories tell us that humans may have developed the brains they have, not so much to run away from the lions of the savanna or tigers of Bengal but rather to effectively cooperate and compete with each other. Psychological tests such as the Sally Anne Test [30] demonstrate the importance of such social cognitive abilities in the development of collaboration abilities in children.

Some branches of AI, aimed at specific human-centric applications, such as intelligent tutoring systems[29], and social robotics [4, 3, 22], did focus on the challenges of human-aware AI systems for a long time. It is crucial to note however that human-aware AI systems are needed in a much larger class of quotidian applications beyond those. These include human-aware AI assistants for many applications where humans continue to be at the steering wheel, but will need naturalistic assistance from AI systems—akin to what they can expect from a smart human secretary. *Increasingly, as AI systems become common-place, human-AI interaction will be the dominant form of human-computer interaction [2].*

For all these reasons and more, human-aware AI has started coming to the forefront of AI research of late. Recent road maps for AI research, including the 2016 JASON report² and the 2016 White House OSTP report³ emphasize the need for research in human-aware AI systems. The 2019 White House list of strategic R&D priorities for AI lists “developing effective methods for

²<https://fas.org/irp/agency/dod/jason/ai-dod.pdf>

³https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf

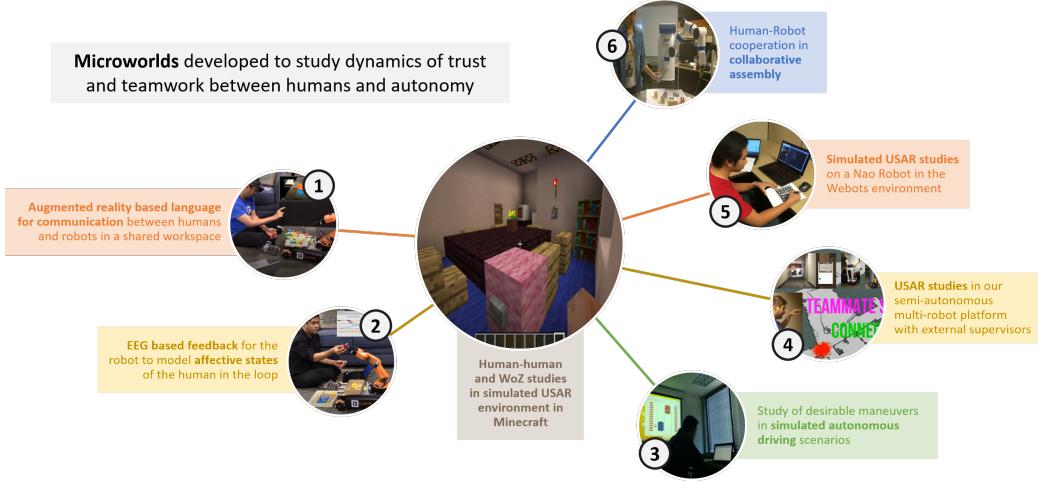


Figure 2: *Test beds developed to study the dynamics of trust and teamwork between autonomous agents and their human teammates.*

human-AI collaboration” at the top of the list of priorities⁴. Human-Aware AI was the special theme for the 2016 International Joint Conference on AI (with the tagline “*why intentionally design a dystopian future and spend time being paranoid about it?*”); it has been a special track at AAAI since 2018.

How do we make AI agents Human-Aware?

When two humans collaborate to solve a task, both of them will develop approximate models of the goals and capabilities of each other (the so called “theory of mind”), and use them to support fluid team performance. AI agents interacting with humans – be they embodied or virtual – will also need to take this implicit mental modeling into account. This certainly poses several research challenges. Indeed, it can be argued that acquiring and reasoning with such models changes almost every aspect of the architecture of an intelligent agent. As an illustration, consider the architecture of an intelligent agent that takes human mental models into account shown in

⁴<https://www.whitehouse.gov/wp-content/uploads/2019/06/National-AI-Research-and-Development-Strategic-Plan-2019-Update-June-2019.pdf>

Figure 3. Clearly most parts of the agent architecture – including state estimation, estimation of the evolution of the world, projection of its own actions, as well as the task of using all this knowledge to decide what course of action the agent should take – are all critically impacted by the need to take human mental models into account. This in turn gives rise to many fundamental research challenges. In [15] we attempt to provide a survey of these challenges. Rather than list the challenges again here, in the rest of this article, I will use the ongoing work in our lab to illustrate some of these challenges as well as our current attempts to address them.⁵ Our work has focused on the challenges of human-aware AI in the context of human-robot interaction scenarios [12], as well as human decision support scenarios [23]. Figure 2 shows some of the test beds and micro-worlds we have used in our ongoing work.

Mental Models in Human-Aware AI

In our ongoing research, we address the following central question in designing human-aware AI systems: “*What does it take for an AI agent to show explainable behavior in the presence of humans?*”. Broadly put, our answer is this: : *To synthesize explainable behavior, AI agents need to go beyond planning with their own models of the world, and take into account the mental model of the human in the loop. The mental model here is not just the goals and capabilities of the human in the loop, but includes the humans model of the AI agents goals/capabilities.*

Let M^R and M^H correspond to the actual goal/capability models of the AI agent and human. To support collaboration, the AI agent needs an approximation of M^H , we will call it \widehat{M}_r^H , to take into account the goals and capabilities of the human. The AI agent also needs to recognize that the human will have a model of its goals/capabilities M_h^R , and needs an approx-

⁵A longer bibliography of work related to human-aware AI from other research groups can be found at <http://rakaposhi.eas.asu.edu/cse591> as part of a graduate seminar at ASU on the topic.

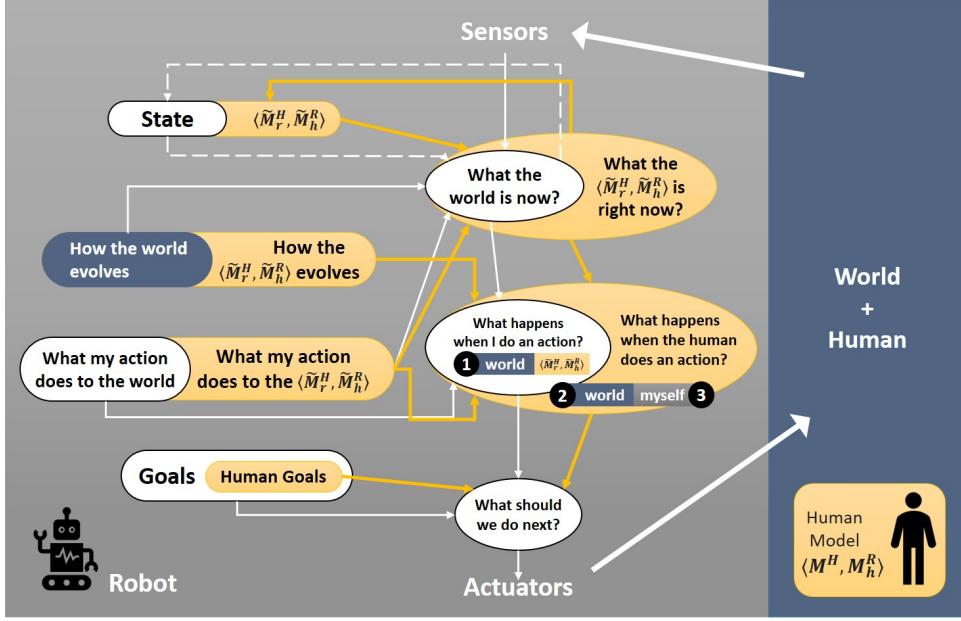


Figure 3: *Architecture of an intelligent agent that takes human mental models into account. All portions in yellow are additions to the standard agent architecture, that are a result of the agent being human-aware. M_h^R is the mental model the human has of the AI agent’s goals and capabilities and M_r^H is the (mental) model the AI agent has of the human’s goal and capabilities (see the section on Mental Models in Human-Aware AI)*

imation of this, denoted \tilde{M}_h^R . All phases of the “sense–plan–act” cycle of an intelligent agent will have to change appropriately to track the impact on these models (as shown in Figure 3). Of particular interest to us in this article is the fact that synthesizing explainable behavior becomes a challenge of supporting planning in the context of these multiple models.

In the following, we will look at some specific issues and capabilities provided by such human-aware AI agents. A note on the model representation: In much of our work, we have used relational precondition-effect models. We believe however that our frameworks can be readily adapted to other model representations; e.g. [26].

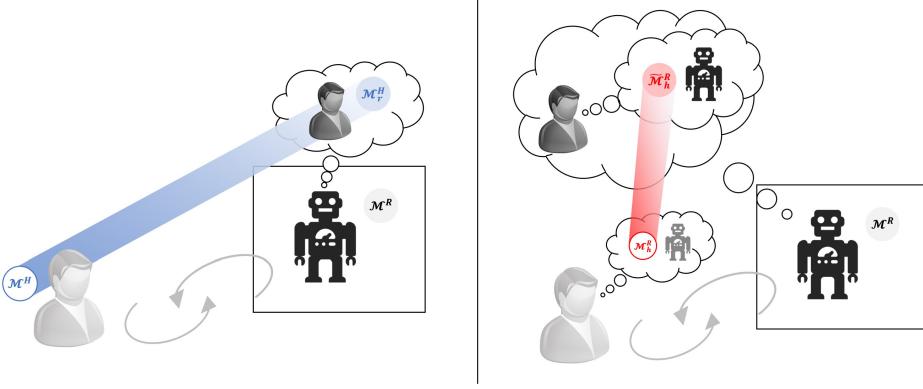


Figure 4: *Use of different mental models in synthesizing explainable behavior.* (Left) The AI system can use its estimation of human’s mental model, M_r^H , to take into account the goals and capabilities of the human thus providing appropriate help to them. (Right) The AI system can use its estimation of human’s mental model of its capabilities M_h^R to exhibit explicable behavior and to provide explanations when needed.

Proactive help: Left to itself, the AI agent will use M^R to synthesize its behavior. When the agent has access to \tilde{M}_r^H , we show how it can use that model to plan behaviors that proactively help the human user—either by helping them complete their goals (c.f. [6]) or avoiding resource contention with them (c.f. [14]).

Explicability: When the agent has access to \tilde{M}_h^R , it can use that model to ensure that its behavior is explainable. We start by looking at generation of *explicable behavior*, which requires the AI agent to not only consider the constraints of its model M^R , but also ensure that its behavior is in line with what is expected by the human. We can formalize this as finding a plan π that trades off the optimality with respect to M^R and “distance” from the plan π' that would be expected according to \tilde{M}_h^R . This optimization can be done either in a model-based fashion, where the distances between π and π' are explicitly estimated (c.f. [19]) or in a model-free fashion, where the

distance is indirectly estimated with the help of a learned “labeling” function that evaluates how far π is from the expected plan/behavior (c.f. [32]). Our notion of explicability here has interesting relations to other notions of interpretable robot behavior considered in AI and robotics communities; we provide a critical comparison of this landscape in [9].

Explanation: In some cases, \widetilde{M}_h^R might be so different from M^R that it will be too costly or infeasible for the AI agent to conform to those expectations. In such cases, the agent needs to provide an *explanation* to the human (with the aim of making its behavior more explicable). We view explanation as a process of “*model reconciliation*,” specifically the process of helping the human bring M_h^R closer to M^R . While a trivial way to accomplish this is to send the whole of M^R as the explanation, in most realistic tasks, this will be both costly for the AI agent to communicate, and more importantly, for the human agent to comprehend. Instead, the explanation should focus on minimal changes \mathcal{E} to M_h^R , such that the robot behavior π is explicable with respect to $M_h^R + \mathcal{E}$, thus in essence making the behavior interpretable to human in light of the explanation. In [13] we show that computing such explanations can be cast as a *meta search* in the space of models spanning M^R and \widetilde{M}_h^R (which is the AI agent’s approximation of M_h^R); see Figure 5. We also provide methods to make this search more efficient, and discuss a spectrum of explanations with differing properties that can all be computed in this framework.

Example: To illustrate the ideas of explicability and explanation in a concrete scenario, consider a simplified “urban search and rescue” scenario depicted in Figure 6. Here the human is in a commander’s role, and is not at the scene of the search and rescue. The robot (AI agent) – which is at the scene – collaborates with the human to search for the injured. Both agents start with the same map of the environment. However, as the robot explores the environment, it might find that some of the pathways are blocked because of fallen debris. In the example here, the robot realizes that the shortest path –

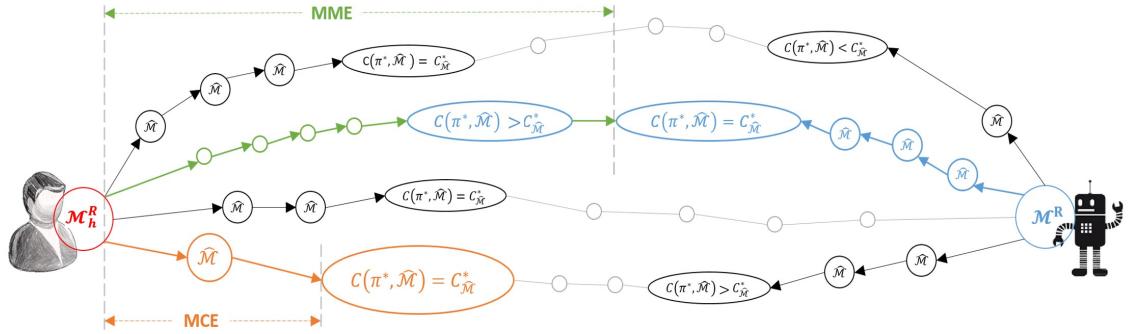


Figure 5: Computing explanations as model reconciliation involves a search in the space of the models. Here the AI agent’s model M^R is on the right end, and the human’s model of the AI agent’s capabilities, M_h^R is on the left. The search transitions corresponds to model changes (for planning models, these might be addition/deletion of preconditions and effects). As discussed in [13], the explanation process involves the AI agent searching for the minimal set of changes to reconcile the human’s model to the actual model of the AI agent in the context of the current problem) for details)

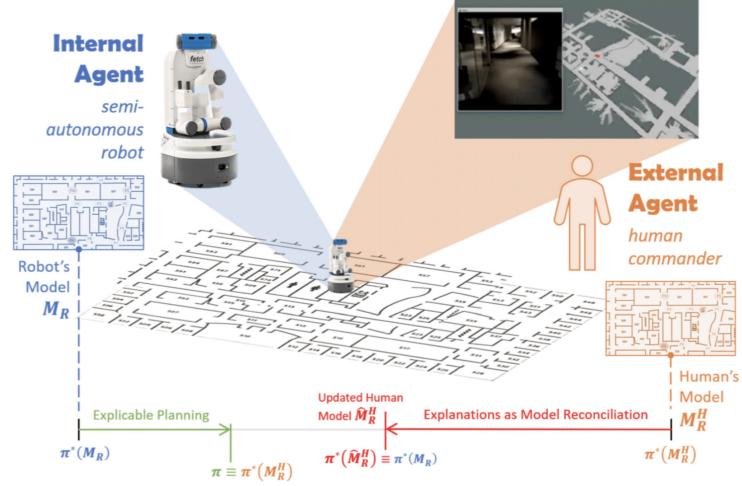


Figure 6: *A simplified urban search and rescue scenario where human and AI agents collaborate.*

as expected by the human – is blocked (see the black “obstacle” on the left in Figure 7). At this point, the robot has two choices. It can be explicable—by going through the path that the human expects. This will however involve the robot clearing the path by removing the obstacle (see Figure 7 right side). Alternately, it can take the path that is optimal to it given the new map. In this case, the robot’s explanation (to the possibly perplexed) human commander involves communicating the salient differences between M_h^R and M^R (see the message on top left in Figure 7).

Balancing Explicability & Explanation: While the foregoing presented showing explicable behavior and giving explanation as two different ways of exhibiting explainable behavior, it is possible to balance the trade-offs between them. In particular, given a scenario where π^* would have been the plan that is optimal with respect to M^R , the AI agent can choose to go with a costlier plan $\tilde{\pi}$ (where $\tilde{\pi}$ is still not explicable with respect to M_h^R), and

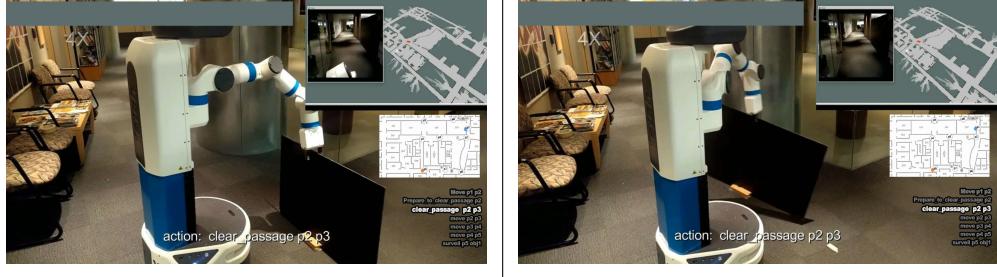


Figure 7: *In the case of explicable behavior, the AI agent behaves in the way the human commander expects it to believe (based on the commander’s model M_h^R). This can be costly (and sometimes even infeasible) for the AI agent—as it is here, for example, where the robot has to remove the obstacle and clear the path so it can navigate it.*





Figure 9: *Assessment of human affective states can be facilitated with brain-computer interface technologies (such as the Emotive helmet used here) that can supplement the normal natural communication modalities*

provide an explanation \mathcal{E}' such that $\tilde{\pi}$ is explicable with respect to $M_h^R + \mathcal{E}'$. In [10], we show how we can synthesize behaviors that have this trade-off.

Model Acquisition: While we focused on the question of reasoning with multiple models to synthesize explainable behavior, a closely related question is that of acquiring the models. In some cases, such as search and rescue scenarios, the human and AI agent may well start with the same shared model of the task. Here the AI agent can assume that as the default mental model. In other cases, the AI agent may have an incomplete model of the human; in [24], we provide an approach to handle the incomplete model, viewing it as a union of complete models. More generally, the AI agent may have to learn the model from the past traces of interaction with the human. Here too, the agent might get by with a spectrum of potential models—starting from fully causal specifications (e.g. PDDL) on one end to correlational/shallow models on the other (see Figure 11) In [28, 31], we discuss some efficient approaches for learning shallow models.

Communicating with Humans: Much of our work focuses on the mechanics of synthesizing explainable behavior assuming the availability of the human mental models. A closely related problem is sensing the affective

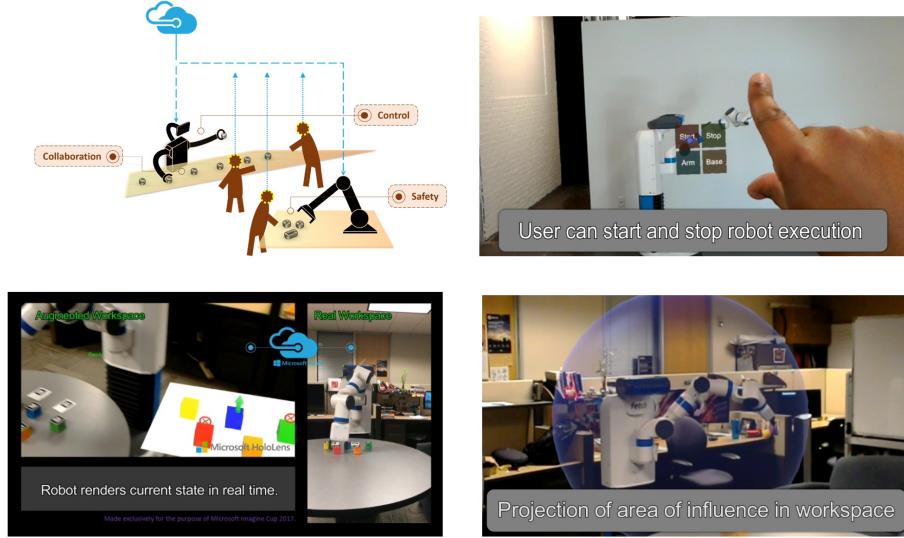


Figure 10: *The AI agent can project its own intentions to the human with the help of augmented reality technologies such as Hololens*

states of human in the loop, and communicating the AI agent’s own intentions to the human. This communication can be done in multiple natural modalities including speech and language and gesture recognition [5]. The human-AI communication can also be supported with the recent technologies such as augmented reality and brain-computer interfaces. Some of our own work looked at the challenges and opportunities provided by these technologies for effective collaboration. Figure 9 shows how off-the-shelf brain computer interfaces supplement natural communication modalities in assessing human affective states. Figure 10 illustrates how the agent can project its intentions with the help of augmented reality technologies such as hololens (that project the agent’s intentions into human visual field). In [25], we look at the challenges involved in deciding when and what intentions to project.

Multiple Humans & Abstraction:: The basic framework above can be generalized in multiple ways. In [27], we show how we can handle situations

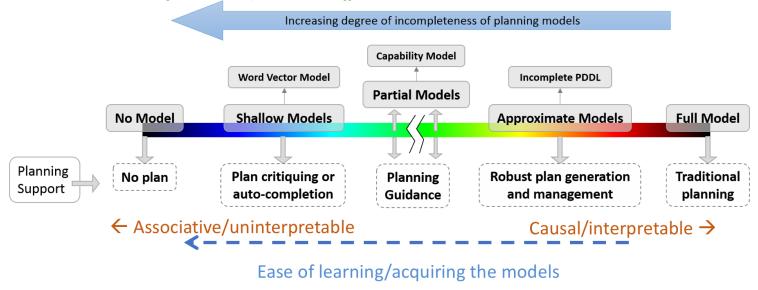


Figure 11: *AI agents can focus on learning a spectrum of human models – starting from fully causal specifications (e.g. PDDL) on one end to correlational/shallow models on the other*

where the human and AI agent have models at different levels of abstraction. In [27] we consider explanations in the context of specific “foils” (e.g. “*why not this other type of behavior?*”) presented by the humans. In [24], we consider how the AI agent can handle multiple humans – obviously with different models (M_{hi}^R) – in the loop, and develop the notions of “conformant” *vs.* “conditional explanations.”

after 1st: We should point out that foils allow for reduction of cognitive load for humans

Self-Explaining Behaviors: While the foregoing considered explanations on demand, it is also possible to directly synthesize *self-explaining* behaviors. In [12], we show how the agent can make its already synthesized behavior more explicable by inserting appropriate “projection” actions to communicate its intentions, and also discuss a framework for synthesizing plans that takes ease of intention projection into account during planning time. In [25], we show how we can synthesize “self-explaining plans,” where the plans contain epistemic actions, which aim to shift M_{hi}^R , followed by domain actions that form an explicable behavior in the shifted model.

Human Subject Evaluations: An important disciplinary challenge posed by research in human-aware AI systems is that of systematic evaluation with

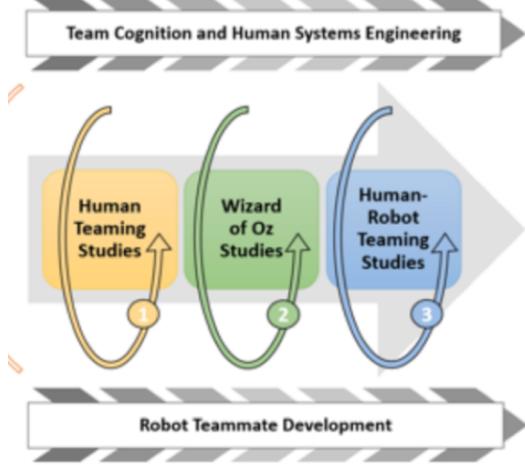


Figure 12: *Evaluation spirals for human-aware AI systems*

human subjects. The temptation of a bunch of engineers unilaterally what sort of support humans will prefer should be resisted. In our own work, we collaborate with researchers in human-factors, and draw on their work on human-human teaming, as well as wizard-of-oz studies [16, 21]. We also evaluate the effectiveness of your systems with systematic human subject studies. Figure 12 shows the evaluation spirals. In [11], we show that people indeed exchange the type of explanations we compute, and that the need for explanations diminishes when the behavior is explicable.

Explanations, Provenance and Explainable AI

Explainable AI (aka XAI) has become quite an active research topic recently. However, much of the work there is concerned with providing “debugging tools” for inscrutable representations (such as those learned by deep networks for perceptual tasks), rather than as a means to human-AI collaboration. A significant part of the work in XAI is concerned with “pointing explanations”—such as pointing the regions of an image that lead to it being classified as an Alaskan Husky or a rare lung disease. Pointing explanations



Figure 13: *A setup for evaluating the effectiveness of explanations produced by the AI agent in a simulated search and rescue scenario. Here the participants assumed the role of external commander and evaluated the plans provided by the AI agent. They could request for plans as well as explanations for those plans, and rate those plans as optimal or suboptimal based on that explanations (from [11])*

are however primitive. Imagine trying to explain/justify a decision that was made by an AI system as part of a sequential decision making scenario. Primitive pointing explanations will have to point to regions of *space-time tubes*. Another thread of research related to “explanations” is providing provenance of decision. Such provenance (or certificate of correctness) is often in terms of the AI agent’s own internal model and is not intended to make sense to the human in the loop. Model reconciliation view, in contrast, can provide explanations in terms of the features of the human and robot models of the task. They thus hew closer to psychological theories of explanation (e.g. [20]).

Ethical Quandaries of Human-Aware AI Systems

Evolutionarily, mental modeling allowed us to both cooperate and compete with each other. After all, lying and deception are possible to a large extent because we can model others' mental states! Thus human-aware AI systems with mental modeling capabilities bring a fresh new set of *ethical quandaries*. We should also be cognizant of the fact that human's anthropomorphizing tendencies are most pronounced for emotional/social agents. After all, no one who saw Shakey for the first time thought it could shoot hoops; yet the first people interacting with Eliza⁶ assumed it is a real doctor and would pour their hearts out to it (prompting Weizenbaum to abort the project!).

Although our primary focus has been on explainable behavior for human-AI collaboration, an understanding of this also helps us solve the opposite problem of generating behavior that is deliberately hard to interpret, something that could be of use in adversarial scenarios. In [18], we present a framework for controlled observability planning, and show how it can be used to synthesize both explicable and obfuscating behavior.

Finally, use of mental models not only helps collaboration but also can open the door for manipulation. In principle, the framework of explanation as model reconciliation allows for the AI agent to tell white lies by bringing M_h^R closer to a model different from M^R . For example, your personal assistant that has a good mental model of you can tell you *white lies* to make you eat healthy. In [8, 7], we explore the question of whether and when it is reasonable for AI agents to lie.

Epilogue

In summary, human-aware AI systems bring in a slew of additional research challenges (as well as a fresh new set of ethical ones). It may seem rather masochistic on our part to focus on these research challenges. As a character from Kurt Vonnegut's Player Piano remarks:

⁶<https://en.wikipedia.org/wiki/ELIZA>

“If only it weren’t for the people, the goddamned people,” said Finnerty, “always getting tangled up in the machinery. If it weren’t for them, earth would be an engineer’s paradise.”

On reflection however, it is easy to see that these are challenges very much worth suiting up for. After all, some of our best friends are human!

Acknowledgments: My views on human-aware AI as well as the specific research described here was carried out in close collaboration with my students and colleagues. Special thanks to my students Tathagata Chakraborti, Sarath Sreedharan, Anagha Kulkarni, Sailik Sengupta, former student Karthik Talamadupula, former post-doc Yu Zhang, and colleagues Nancy Cooke, Matthias Scheutz, David Smith and Hankz Hankui Zhuo. My AAAI address as well as this write-up have benefited from the discussions and encouragement of Dan Weld, Barbara Grosz and Manuela Veloso. Thanks also to Behzad Kamgar-Parsi, Jeffery Morrison, Marc Steinberg and Tom McKenna of the Office of Naval Research for sustained support of our research into human-aware AI systems. Ashok Goel patiently nudged me to complete this write-up for the AI Magazine and provided helpful editorial comments. This research is supported in part by the ONR grants N00014-16-1-2892, N00014-18-1-2442, N00014-18-1-2840, the AFOSR grant FA9550-18-1-0067 and the NASA grant NNX17AD06G. It has been my privilege and singular honor to serve as the president of AAAI at a time of increased public and scientific interest in our field. I sincerely thank the AAAI members for their trust and support.

References

- [1] James F Allen. Mixed initiative planning: Position paper. In *ARPA/Rome Labs Planning Initiative Workshop*, 1994.
- [2] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N.

- Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 3, 2019.
- [3] Cynthia Breazeal. Toward sociable robots. *Robotics and autonomous systems*, 42(3-4):167–175, 2003.
 - [4] Cynthia L Breazeal. *Designing sociable robots*. MIT press, 2004.
 - [5] Rehj Cantrell, Kartik Talamadupula, Paul W. Schermerhorn, J. Benton, Subbarao Kambhampati, and Matthias Scheutz. Tell me when and why to do it!: run-time planner model updates via natural language instruction. In *International Conference on Human-Robot Interaction, HRI’12, Boston, MA, USA - March 05 - 08, 2012*, pages 471–478, 2012.
 - [6] T. Chakraborti, G. Briggs, K. Talamadupula, Y. Zhang, M. Scheutz, D. Smith, and S. Kambhampati. Planning for serendipity. In *IROS*, 2015.
 - [7] T. Chakraborti and S. Kambhampati. (how) can ai bots lie? In *ICAPS Workshop on Explainable Planning (XAIP)*, 2019.
 - [8] T. Chakraborti and S. Kambhampati. (when) can ai bots lie? In *AIES*, 2019.
 - [9] T. Chakraborti, A. Kulkarni, S. Sreedharan, D. Smith, and S. Kambhampati. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In *ICAPS*, 2019.
 - [10] T. Chakraborti, S. Sreedharan, and S. Kambhampati. Explicability versus explanations in human-aware planning. In *AAMAS*, 2018.
 - [11] T. Chakraborti, S. Sreedharan, and S. Kambhampati. Plan explanations as model reconciliation – an empirical study. In *HRI*, 2019.

- [12] T. Chakraborti, S. Sreedharan, A. Kulkarni, and S. Kambhampati. Projection-aware task planning and execution for human-in-the-loop operation of robots in a mixed-reality workspace. In *IROS*, 2018.
- [13] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *IJCAI*, 2017.
- [14] T. Chakraborti, Y. Zhang, D. Smith, and S. Kambhampati. Planning with resource conflicts in human-robot cohabitation. In *AAMAS*, 2016.
- [15] Tathagata Chakraborti, Subbarao Kambhampati, Matthias Scheutz, and Yu Zhang. AI challenges in human-robot cognitive teaming. *CoRR*, abs/1707.04775, 2017.
- [16] Nancy J. Cooke, Jamie C. Gorman, Christopher W. Myers, and Jasmine L. Duran. Interactive team cognition. *Cognitive Science*, 37(2):255–285, 2013.
- [17] Barbara J. Grosz. AAAI-94 presidential address: Collaborative systems. *AI Magazine*, 17(2):67–85, 1996.
- [18] A. Kulkarni, A. Srivastava, and S. Kambhampati. A unified framework for planning in adversarial and cooperative environments. In *AAAI*, 2019.
- [19] A. Kulkarni, Y. Zha, T. Chakraborti, S. Vadlamudi, Y. Zhang, and S. Kambhampati. Explicable planning as minimizing distance from expected behavior. In *AAMAS*, 2019.
- [20] T. Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 2006.
- [21] Nathan J. McNeese, Mustafa Demir, Nancy J. Cooke, and Christopher W. Myers. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human Factors*, 60(2):262–273, 2018.

- [22] Brian Scassellati. Theory of mind for a humanoid robot. *Auton. Robots*, 12(1):13–24, 2002.
- [23] S. Sengupta, T. Chakraborti, S. Sreedharan, S. Vadlamudi, and S. Kambhampati. Radar - a proactive decision support system for human-in-the-loop planning. *AAAI Fall Symposium on Human-Agent Groups*, 2017.
- [24] S. Sreedharan, T. Chakraborti, and S. Kambhampati. Handling model uncertainty and multiplicity in explanations via model reconciliation. In *ICAPS*, 2018.
- [25] S. Sreedharan, T Chakraborti, C Muise, and S. Kambhampati. Planning with Explanatory Actions: A Joint Approach to Plan Explicability and Explanations in Human-Aware Planning . *ArXiv e-prints*, abs/1903.07269, 2019.
- [26] S. Sreedharan, A Olmo, A Mishra, and S. Kambhampati. Model-Free Model Reconciliation. *ArXiv e-prints*, abs/1903.07198, 2019.
- [27] S. Sreedharan, S. Srivastava, and S. Kambhampati. Hierarchical expertise level modeling for user specific contrastive explanations. In *IJCAI*, pages 4829–4836, 2018.
- [28] X. Tian, H. Zhuo, and S. Kambhampati. Discovering underlying plans based on distributed representations of actions. In *AAMAS*, 2016.
- [29] Kurt VanLehn. The behavior of tutoring systems. *I. J. Artificial Intelligence in Education*, 16(3):227–265, 2006.
- [30] H Wimmer and J Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition*, 1983.

- [31] Y. Zha, Y. Li, S. Gopalakrishnan, B. Li, and S. Kambhampati. Recognizing plans by learning embeddings from observed action distributions. In *AAMAS*, 2018.
- [32] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. Zhuo, and S. Kambhampati. Plan explicability and predictability for robot task planning. In *ICRA*, 2017.

Bio: Subbarao Kambhampati (Rao) is a professor of Computer Science at Arizona State University. He received his B.Tech. in Electrical Engineering (Electronics) from Indian Institute of Technology, Madras (1983), and M.S.(1985) and Ph.D.(1989) in Computer Science (1985,1989) from University of Maryland, College Park. Kambhampati studies fundamental problems in planning and decision making, motivated in particular by the challenges of human-aware AI systems. Kambhampati is a fellow of AAAI and AAAS, and was an NSF Young Investigator. He received multiple teaching awards, including a university last lecture recognition. Kambhampati is the past president of AAAI; he served as the president of AAAI during 2016-18, and as a trustee of IJCAI during 2013-18. He was the program chair for IJCAI 2016, ICAPS 2013, AAAI 2005 and AIPS 2000. He served on the founding board of directors of Partnership on AI. Kambhampati's research as well as his views on the progress and societal impacts of AI have been featured in multiple national and international media outlets.

