Challenges of Human-Aware AI Systems

Subbarao Kambhampati

■ *From its inception, artificial intelligence (AI) has had a rather ambivalent relationship to humans — swinging between their augmentation and their 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 help 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 and 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 interdisciplinary collaborations; and can go a long way toward increasing public acceptance of AI technologies.*

Artificial intelligence (AI), the discipline we all call our intellectual home, is suddenly having a rather large 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 with fire and electricity in its overall importance to humanity, 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

Figure 1. We Should Build a Future where AI Systems Can Be Our Quotidian Partners.

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 (figure 1). 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 article 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 Arthur C. Clarke's *HAL 9000*, 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.

From its inception, humans have had a rather ambivalent relationship with AI — swinging between their augmentation and their replacement. Most high-profile achievements of AI have either been far away from humans — think of the rovers *Spirit* and *Opportunity* exploring Mars; or in a decidedly adversarial stance with humans, chess-playing programs such as IBM's Deep Blue and DeepMind's AlphaGo Zero, or Carnegie Mellon University's poker-playing program, Libratus. 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 humans 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 18th century's 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 for the limitations of the AI systems (Allen 1994). 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

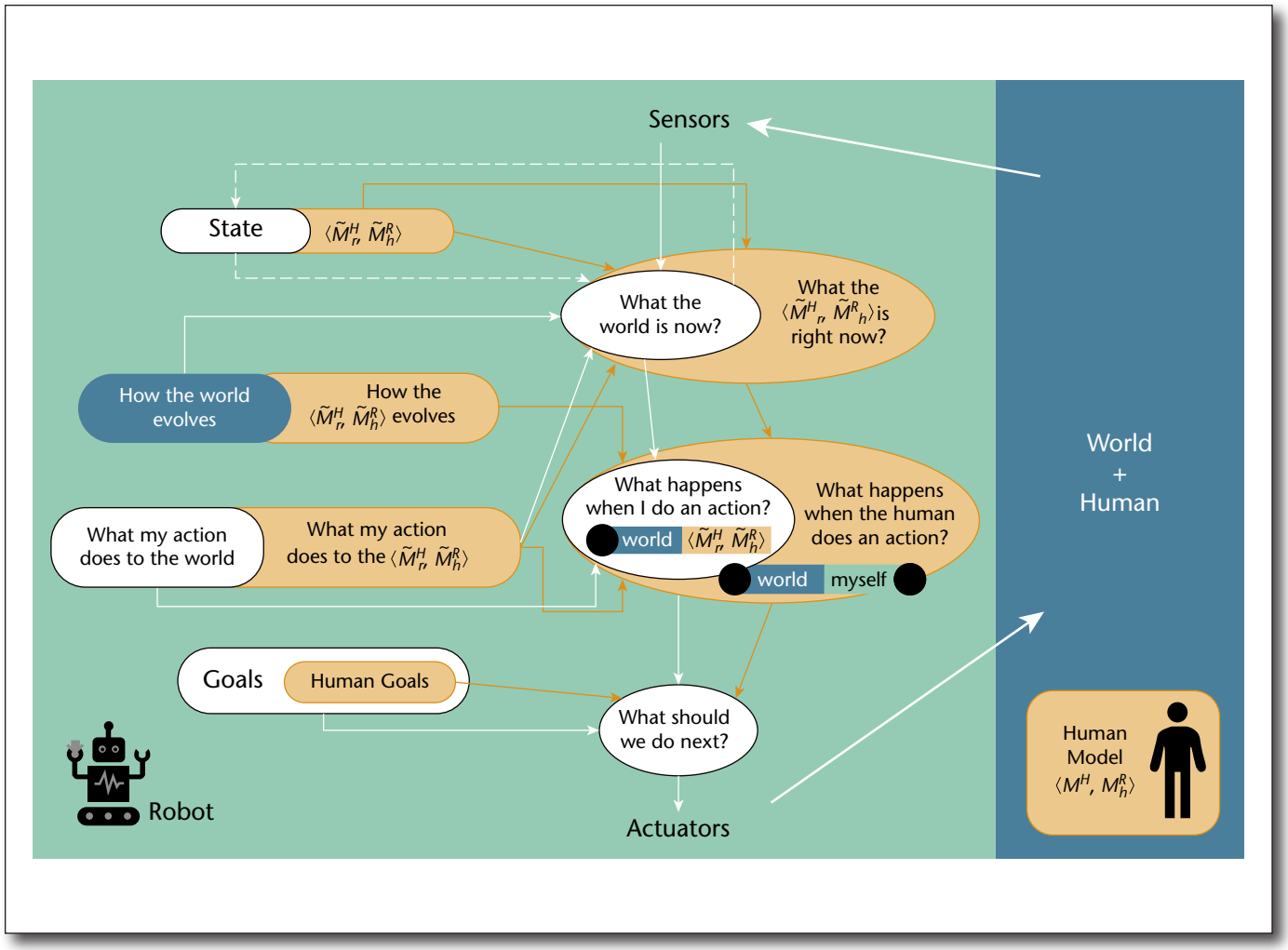

Figure 2. Architecture of an Intelligent Agent that Takes Human Mental Models into Account.

All portions in yellow are additions to the standard agent architecture, which are a result of the agent being human-aware. M^R_h is the mental model the human has of the AI agent's goals and capabilities, and M^H_r is the (mental) model the AI agent has of the human's goal and capabilities (see "Mental Models in Human-Aware AI").

have not so much to run away from the lions of the savanna or the tigers of Bengal, but, instead, to effectively cooperate and compete with each other. Psychological tests such as the Sally Anne Test (Wimmer and Perner 1983) demonstrate the importance of such social cognitive abilities in the development of collaborative abilities in children.

Some branches of AI, aimed at specific human-centric applications such as intelligent tutoring systems (VanLehn 2006) and social robotics (Breazeal 2004, 2003; Scassellati 2002), 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 commonplace, human-AI interaction will be the dominant form of human–computer interaction (Amershi et al. 2019).

For all of 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 report[2] and the 2016 White House OSTP report,[3] emphasize the need for research in human-aware AI systems. The 2019 White House list of strategic research and development priorities for AI places developing effective methods for human-AI collaboration at the top of it list.[4] 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 the Association for the Advancement of Artificial Intelligence (AAAI) since 2018.

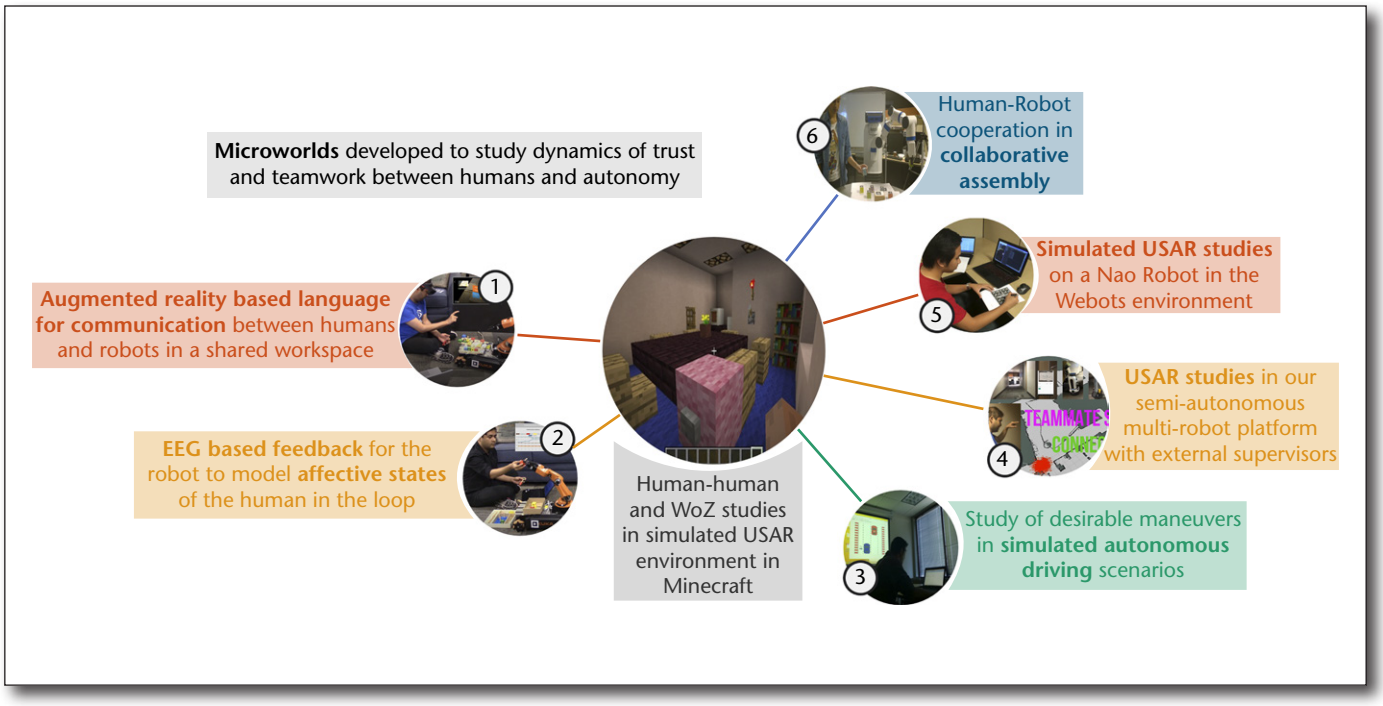

Figure 3. Test Beds Developed to Study the Dynamics of Trust and Teamwork between Autonomous Agents and Their Human Teammates.

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 (see figure 2). Clearly most parts of the agent architecture — including state estimation, estimation of the evolution of the world, projection of its own actions, and 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 a 2017 article (Chakraborti, Kambhampati, Scheutz, and Zhang 2017), we attempted to provide a survey of these challenges. Rather than list the challenges again here, in the remainder of this article, I will use the ongoing work in our laboratory 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 (Chakraborti, Sreedharan, Kulkarni, and Kambhampati 2018), as well as human decision support scenarios (Sengupta et al. 2017). Figure 3 shows some of the test beds and microworlds 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 human's model of the AI agent's goals and capabilities.

Let M^R and M^H correspond to the actual goal or capability models of the AI agent and human. To support collaboration, the AI agent needs an approximation of M^H, which we will call it $\tilde{M}^H_r$, 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^R_h, and needs an approximation of this, denoted $\tilde{M}^R_h$. All phases of the sense–plan–act cycle of an intelligent agent will have to change appropriately to track the impact on these models (figure 2). 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, as illustrated in figure 4.

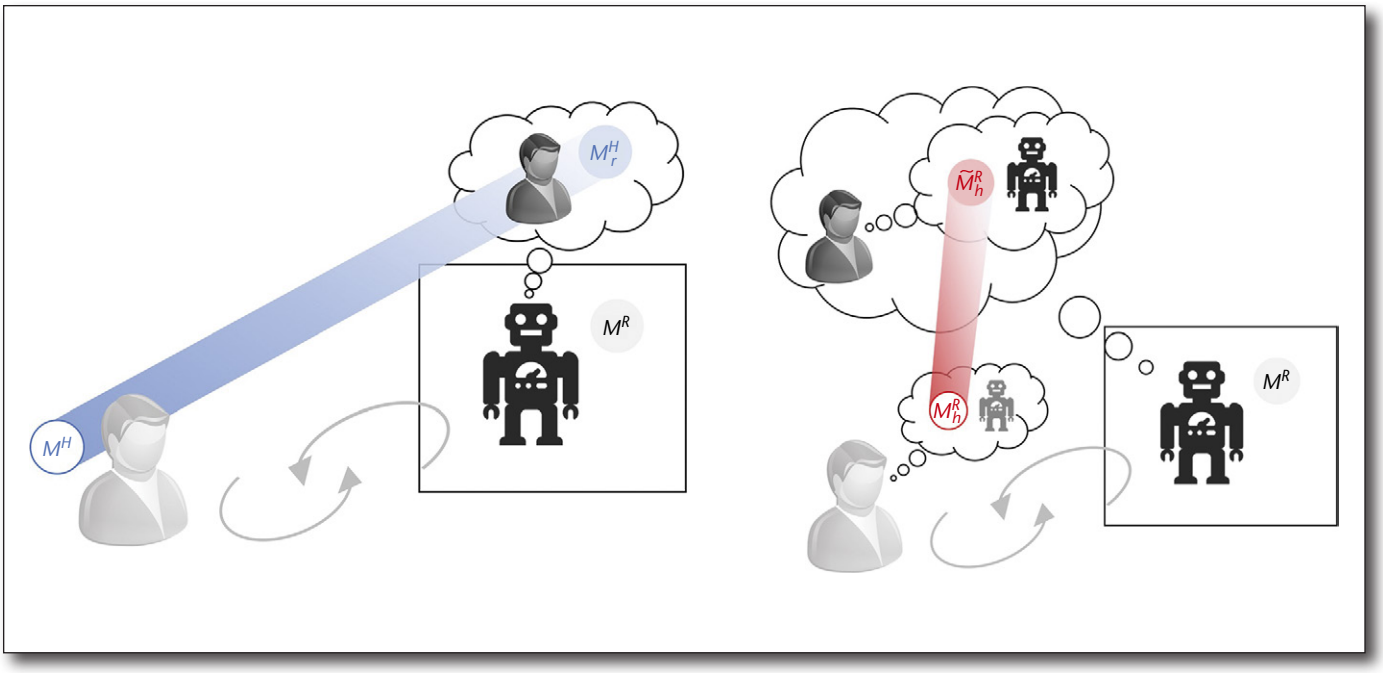

Figure 4. Use of Different Mental Models in Synthesizing Explainable Behavior.

(Left) The AI system can use its estimation of human's mental model, M^H_r, 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 a human's mental model of its capabilities M^R_h to exhibit explicable behavior and provide explanations when needed.

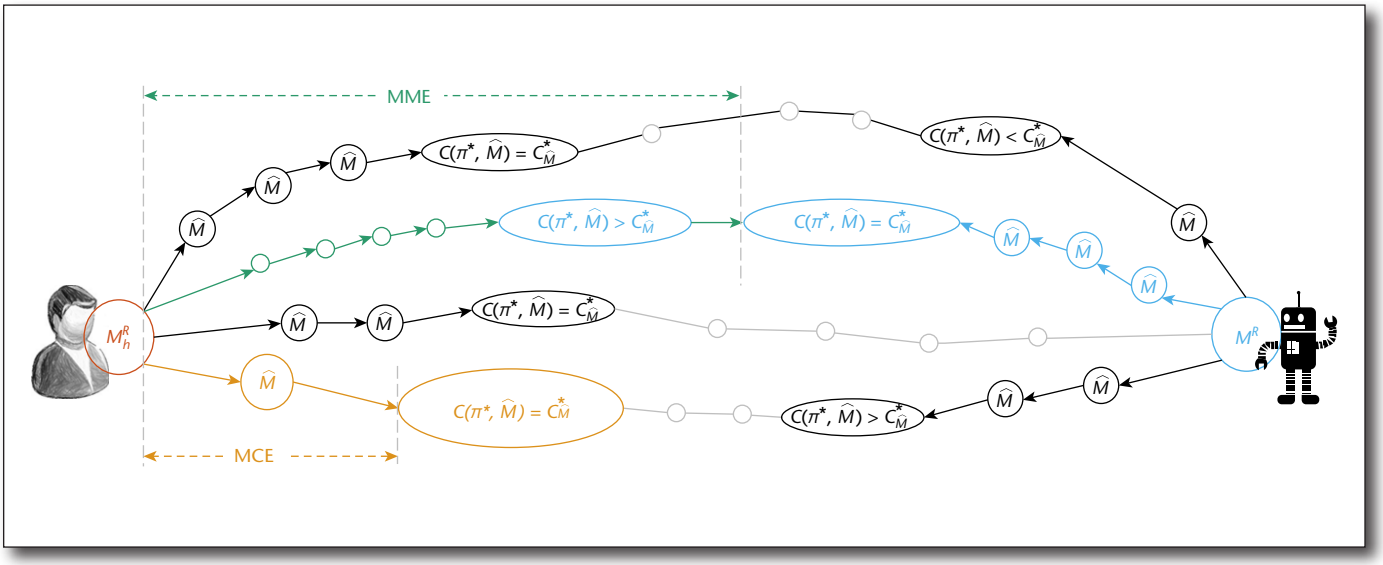

Figure 5. Computing Explanations as Model Reconciliation Involves a Search in the Space of the Models.

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^R_h, is on the left. The search transitions correspond to model changes (for planning models, these might be addition or deletion of preconditions and effects). As is discussed in Chakraborti et al. (2017a), 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.

In the following subsections, 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 (for example, see Sreedharan, Olmo, Mishra, and Kambhampati, 2019).

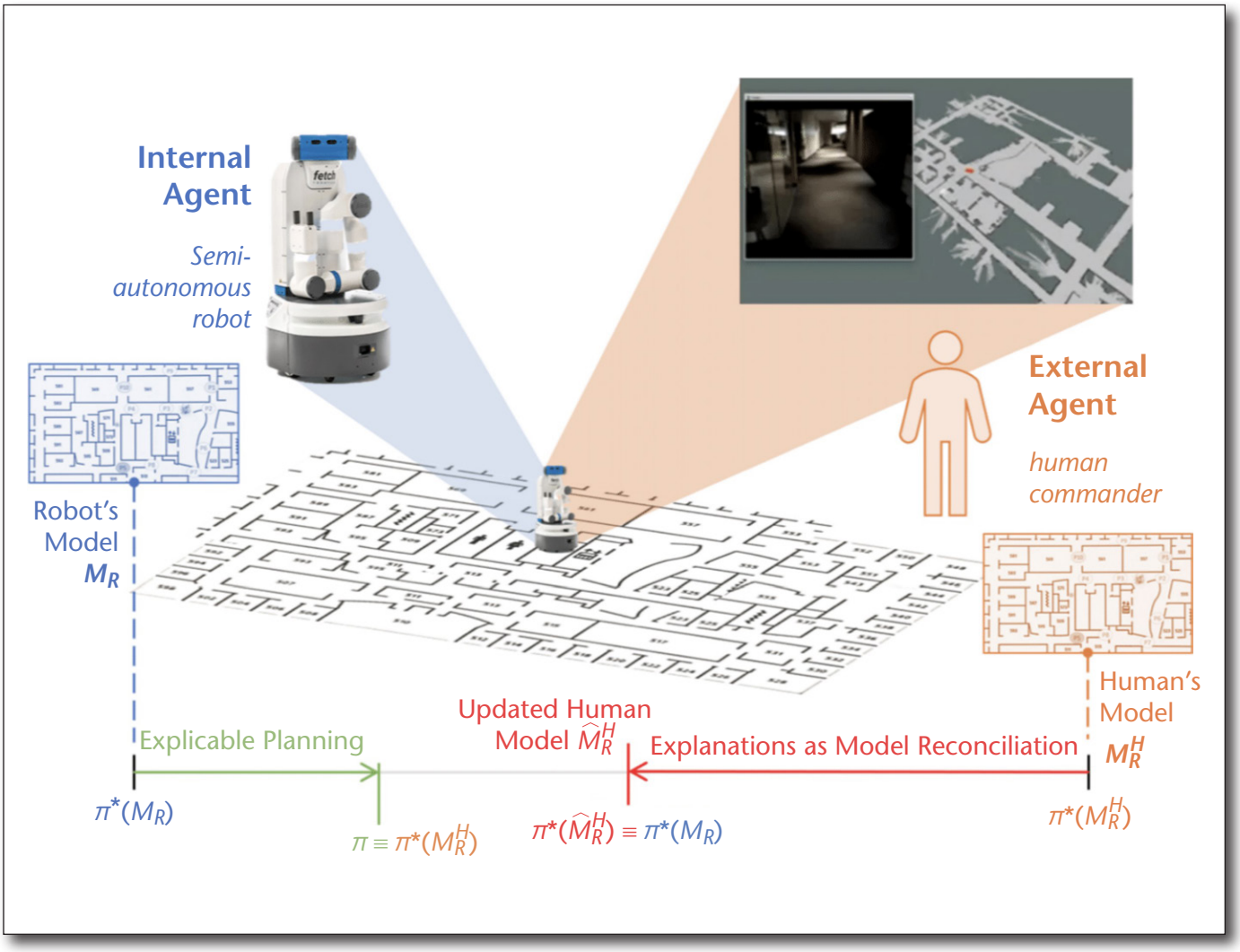

Figure 6. A Simplified Urban Search and Rescue Scenario Where Human and AI Agents Collaborate.

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 (Chakraborti et al. 2015) or avoiding resource contention with them (Chakraborti, Zhang, Smith, and Kambhampati 2016).

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 (Kulkarni, Zha et al. 2019), 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 or behavior (Zhang et al. 2017). 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 the article by Chakraborti, Kulkarni, et al. (2019).

Explanation

In some cases, $\tilde{M}_h^R$ 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

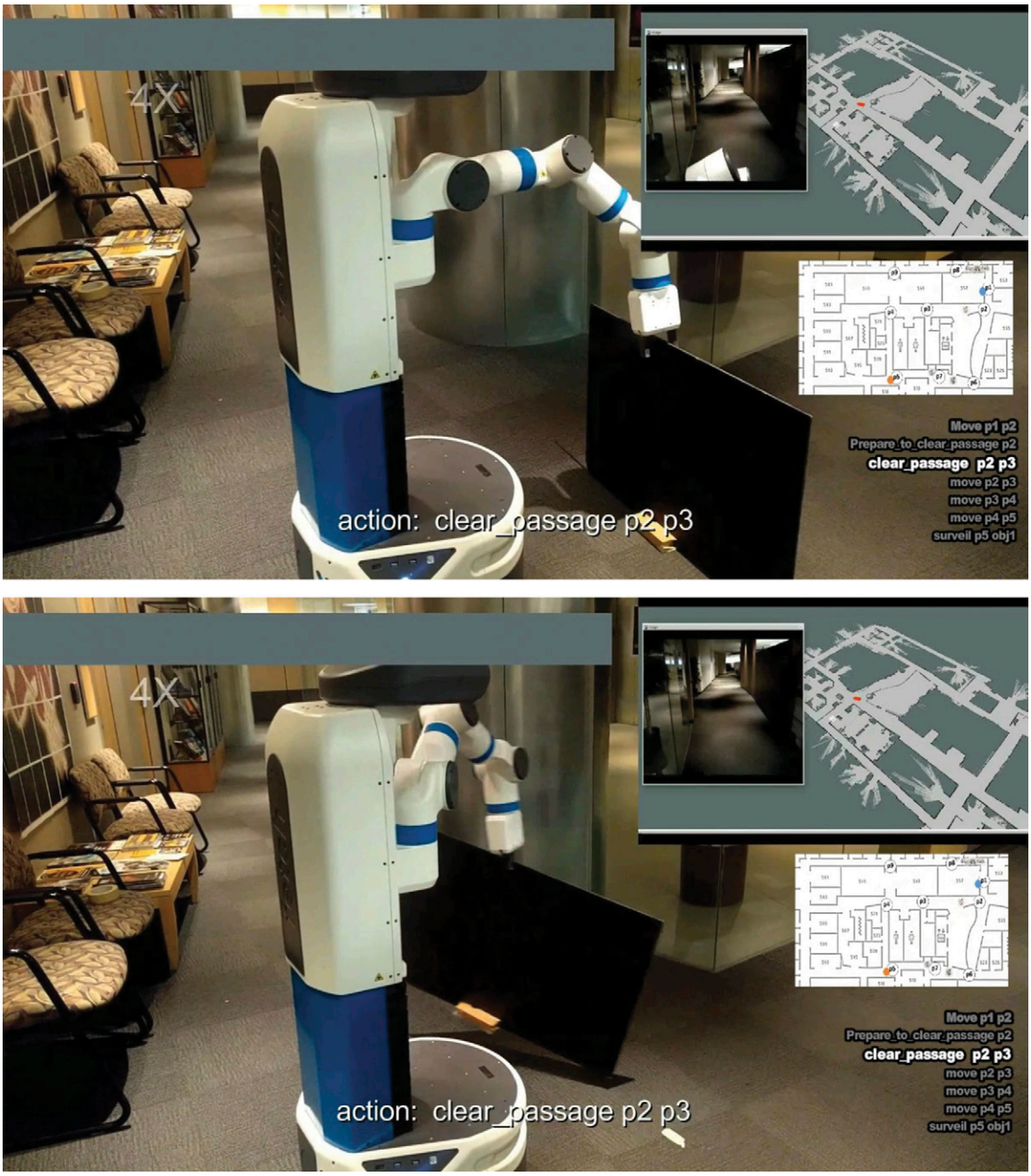

Figure 7. The Explicable Choice.

In the case of explicable behavior, the AI agent behaves in the way the human commander expects it to, 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.

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.

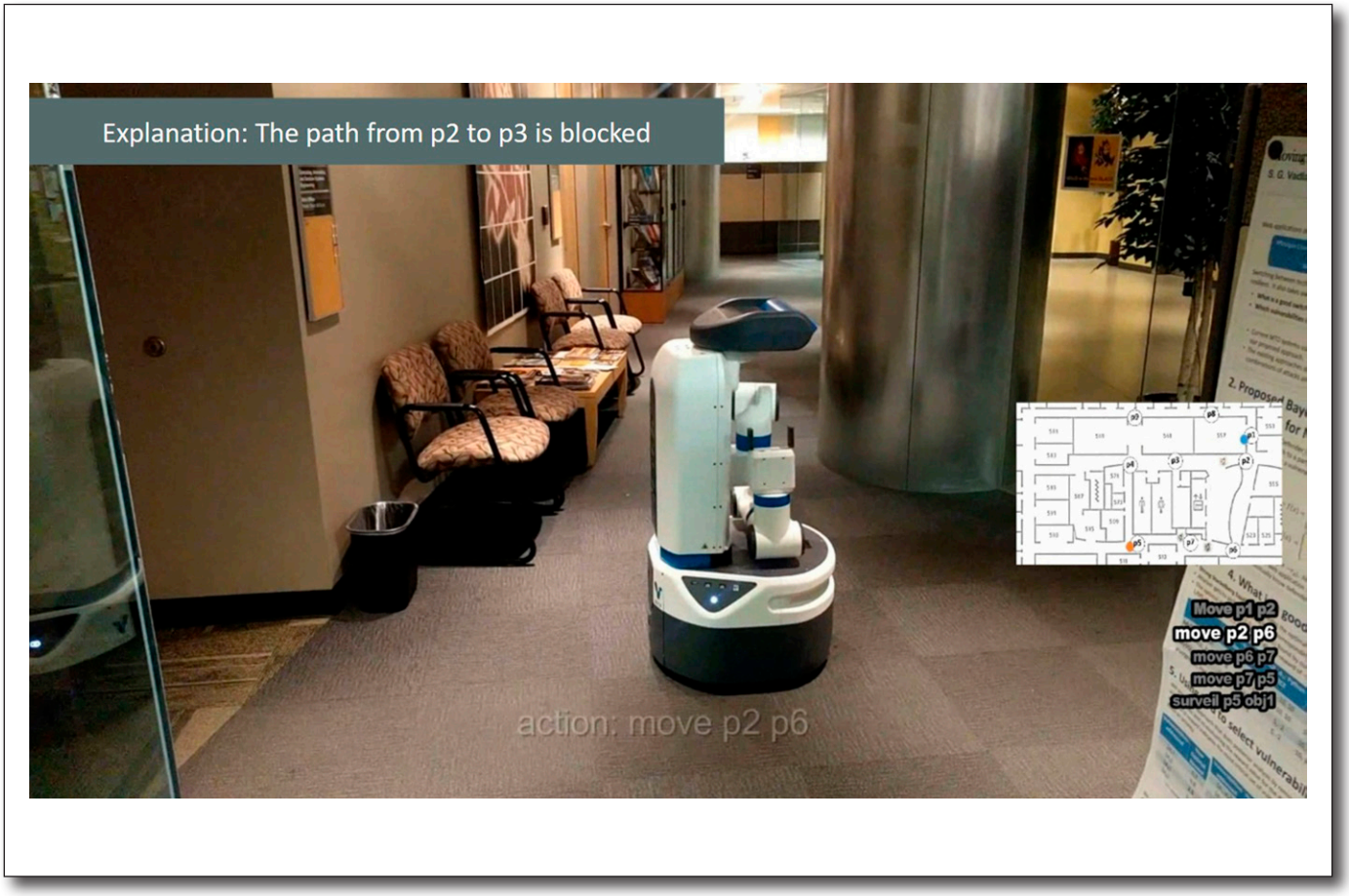

Figure 8. The Optimal Choice.

When explicable behavior is too costly or infeasible, the AI agent can take the path that is optimal to it (given that the original shortest path is blocked), and provide an explanation. The explanation involves communicating the model differences between M_h^R and M^R. For our case, this is just communicating that the shortest path is blocked (see the message at the top left).

Instead, the explanation should focus on minimal changes ε to M_h^R, such that the robot behavior π is explicable with respect to $M_h^R + \varepsilon$, thus in essence making the behavior interpretable to the human in light of the explanation. We show, in Chakraborti, Sreedharan, Zhang, and Kambhampati (2017), that computing such explanations can be cast as a meta search in the space of models spanning M^R and $\tilde{M}_h^R$ (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 — 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 the top left in figure 7).

Balancing Explicability and Explanation

While the foregoing presented showing an explicable behavior and giving an explanation as two different ways of exhibiting explainable behavior, it is possible to balance the tradeoffs 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),

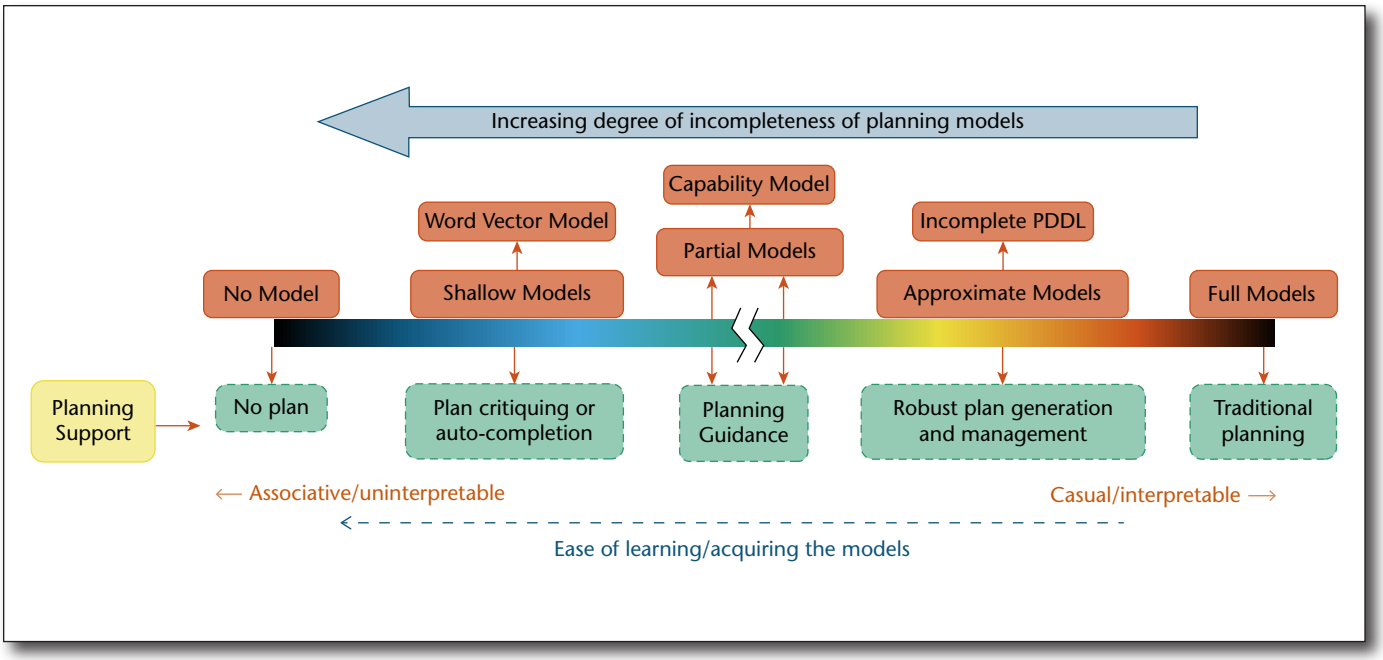

Figure 9. AI Agents Can Focus on Learning a Spectrum of Human Models.

Starting from fully causal specifications (for example, the Planning Domain Description Language) on one end to correlational or shallow models on the other.

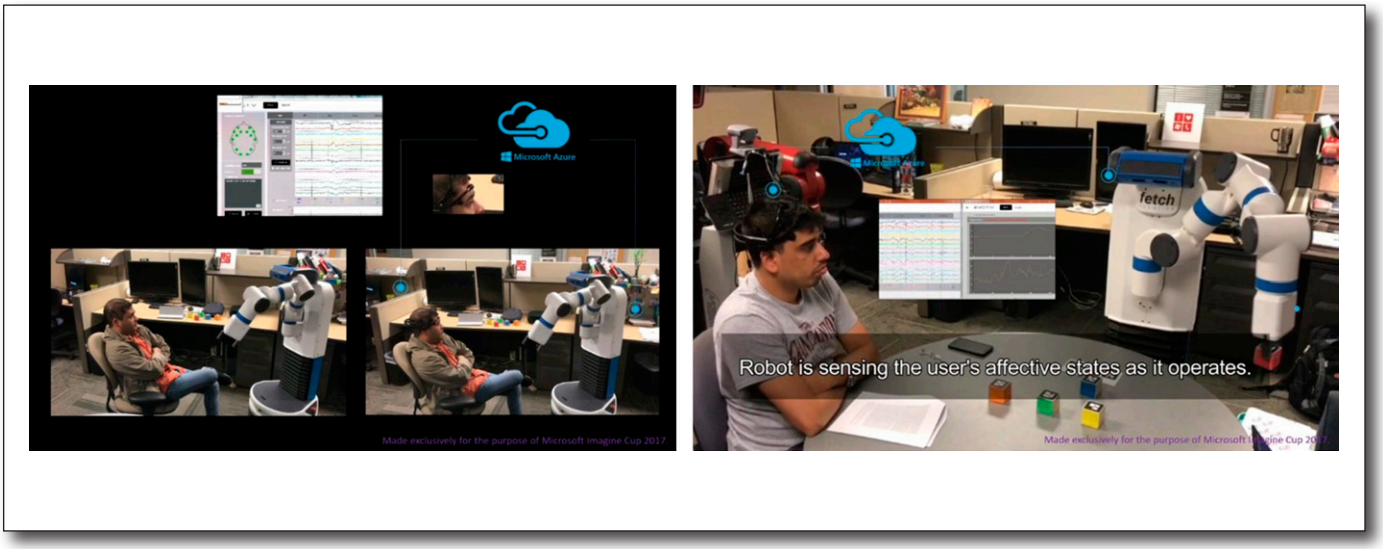

Figure 10. Facilitation Assessment of Human Affective States.

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.

and provide an explanation ε' such that $\tilde{\pi}$ is explicable with respect to $M_h^R + \varepsilon'$ (figure 8). In Chakraborti, Sreedharan, and Kambhampati (2018), we show how we can synthesize behaviors that have this tradeoff.

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 this as the default mental model. In other cases, the AI agent may have an incomplete model of the human; in Sreedharan, Chakraborti, and Kambhampati (2018), we provide an approach to handle the incomplete model, viewing it as a union of complete models. More generally, the AI agent may have

Figure 11. The AI Agent Can Project Its Own Intentions to the Human with the Help of Augmented Reality Technologies such as the Microsoft HoloLens.

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 (for example, the Planning Domain Description Language) on one end to correlational or shallow models on the other (see figure 9). In two articles (Tian, Zhuo, and Kambhampati 2016; Zha, Li, Gopalakrishnan, and Kambhampati 2018), we discuss some efficient approaches for learning shallow models.

Communicating with Humans

Much of our work focuses on the mechanics of synthesizing explainable behavior by assuming the availability of the human mental models. A closely related problem is sensing the affective 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 (Cantrell et al. 2012). 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 10 shows how off-the-shelf brain–computer interfaces supplement natural communication modalities in assessing human affective states. Figure 11 illustrates how the agent can project its intentions with the help of augmented reality technologies such as the Microsoft HoloLens (which projects the agent's intentions into human visual field). In Sreedharan, Chakraborti, Muise, and Kambhampati (2019), we look at the challenges involved in deciding when and what intentions to project.

Multiple Humans and Abstraction

The basic framework discussed previously can be generalized in multiple ways. In Sreedharan, Srivastava, and Kambhampati (2018), we show how we can handle situations where the human and AI agent have models at different levels of abstraction. In Sreedharan, Srivastava, and Kambhampati (2018), we consider explanations in the context of specific foils (for example, "Why not this other type of behavior?") presented by the humans. In Sreedharan, Chakraborti, and Kambhampati (2018), we consider how the

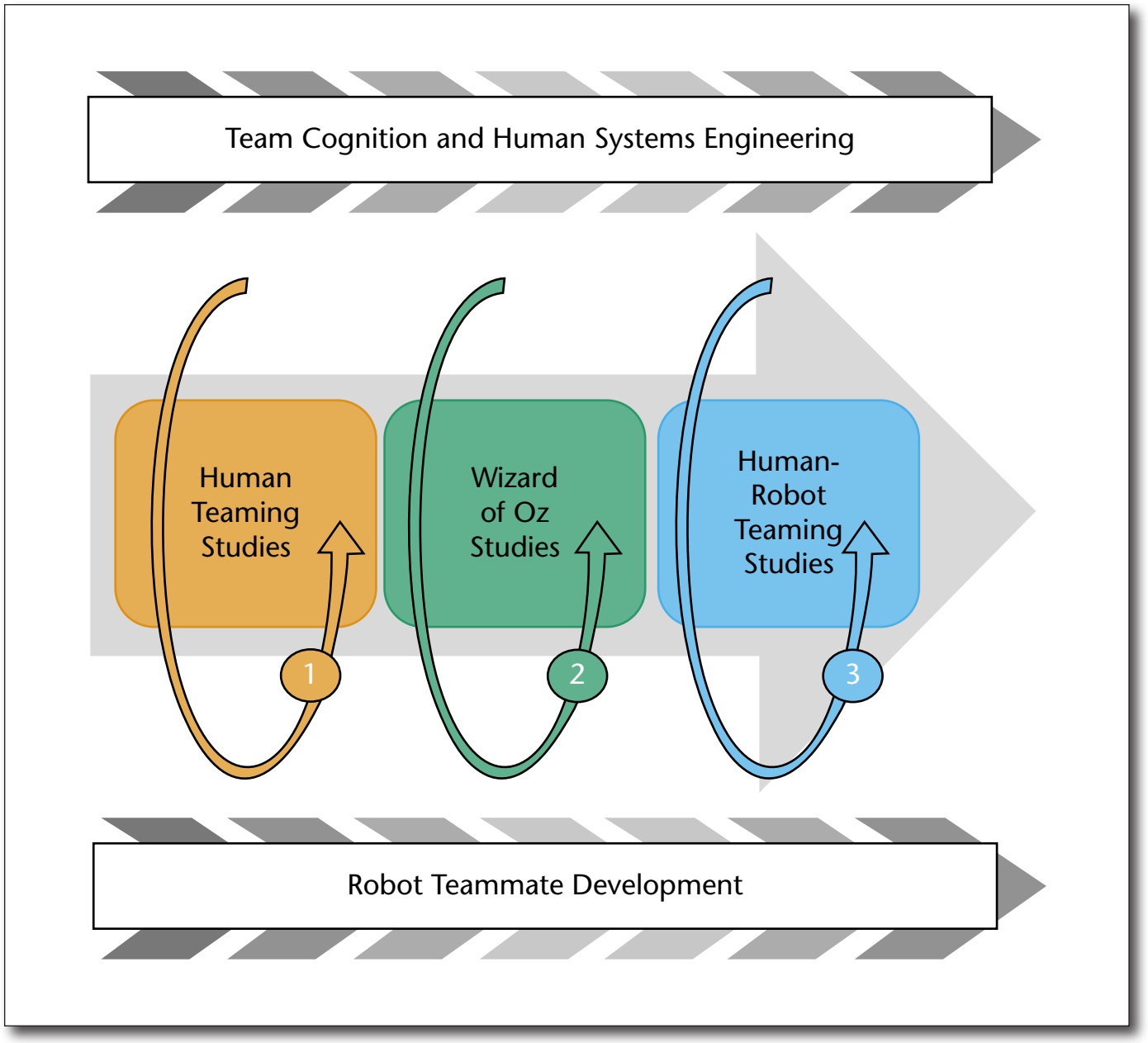

Figure 12. Evaluation Spirals for Human-Aware AI Systems.

AI agent can handle multiple humans — obviously with different models (M^R_{hi}) — in the loop, and develop the notions of conformant versus conditional explanations.

Self-Explaining Behaviors

While the foregoing considered explanations on demand, it is also possible to directly synthesize self-explaining behaviors. In Chakraborti et al. (2018), 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 Sreedharan, Chakraborti, Muise, and Kambhampati (2019), we show how we can synthesize self-explaining plans, where the plans contain epistemic actions, which aim to shift M^R_h, 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 human subjects. The temptation of a bunch of engineers to unilaterally decide what sort of support humans will prefer should be resisted. In our own work, we collaborate with researchers in

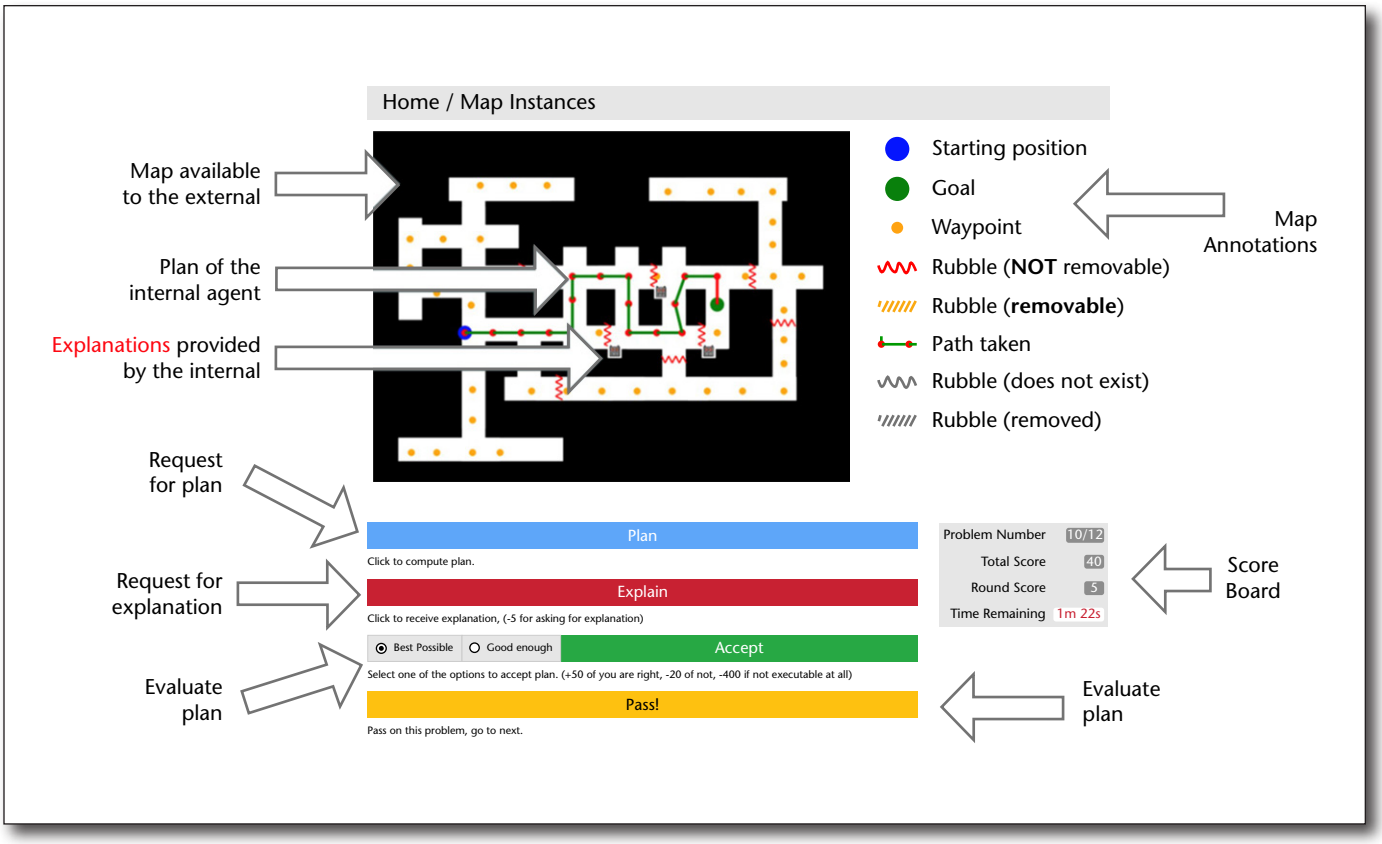

Figure 13. A Setup for Evaluating the Effectiveness of Explanations Produced by the AI Agent in a Simulated Search-and-Rescue Scenario.

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 Chakraborti, Sreedharan, and Kambhampati, 2019).

human-factors, and draw on their work in human–human teaming, as well as Wizard-of-Oz studies (Cooke, Gorman, Myers, and Duran 2012; McNeese, Demir, Cooke, and Myers 2017). We also evaluate the effectiveness of human-aware systems with systematic human-subject studies. Figure 12 displays the evaluation spirals. As illustrated in figure 13, we showed that people indeed exchange the type of explanations we compute, and that the need for explanations diminishes when the behavior is explicable (Chakraborti, Sreedharan, and Kambhampati 2019).

Explanations, Provenance, and Explainable AI

Explainable AI has become quite an active research topic in machine learning community 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 Explainable AI is concerned with pointing explanations — such as pointing the regions of an image that lead to it being classified as, say, an Alaskan Husky or a rare lung disease. Pointing explanations are, however, primitive. Imagine trying to explain or 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.[6] 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. A 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 psychologic theories of explanation (for example, Lombrozo 2006).

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 the Robot for the first time thought it could shoot hoops; yet the first people interacting with ELIZA[7] assumed it was a real doctor, and would pour their hearts out to it (prompting Joseph 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. We presented, in Kulkarni, Srivastava, and Kambhampati (2019), a framework for controlled observability planning, and show how it can be used to synthesize both explicable and obfuscatory 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 MR closer to a model different from M^R. For example, a personal assistant that has a good mental model of you can tell you white lies to make you eat healthy. In two articles (Chakraborti and Kambhampati 2019a, 2019b), 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:

> "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

This article is based on the AAAI 2018 Presidential Address I 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. 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 article 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 article for *AI Magazine* and provided helpful editorial comments.

This research is supported in part by the Office of Naval Research (grants no. N00014-16-1-2892, no. N00014-18-1-2442, and no. N00014-18-1-2840), the Air Force Office of Scientific Research (grant no. FA9550-18-1-0067), and the National Aeronautics and Space Administration (grant no. 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.

Notes

1. In a way, it thus follows in the footsteps of Barbara Grosz's AAAI Presidential Address (Grosz 1996), which talked about collaborative systems.

2. See fas.org/irp/agency/dod/jason/ai-dod.pdf

3. Available at https://www.whitehouse.gov/wp-content/uploads/2019/06/National-AI-Research-and-Development-Strategic-Plan-2019-Update-June

4. Available at www.whitehouse.gov/wp-content/uploads/2019/06/National-AI-Research-and-Development-Strategic-Plan-2019-Update-June-2019.pdf

5. A longer bibliography of work related to human-aware AI from other research groups can be found at rakaposhi.eas.asu.edu/cse591 as part of a graduate seminar at Arizona State University on the topic. A 4 hour AAAI 2020 tutorial on model-based synthesis of explainable behavior for human-AI interaction is available at https://rakaposhi.eas.asu.edu/haai-2020-tutorial.html

6. Pointing explanations are not even sufficient for image classification. Consider the case of adversarial examples, e.g. an adversarially doctored image of a school bus being classified as an ostrich. Imagine the futility of having an AI agent point to the region(s) of the school bus that explain why the agent thinks it is an ostrich!

7. A description of the ELIZA program can be found at en.wikipedia.org/wiki/ELIZA.

References

Allen, J. F. 1994. Mixed Initiative Planning: Position Paper. Presented at the ARPA/Rome Labs Planning Initiative Workshop.

Amershi, S.; Weld, D. S.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S. T.; Bennett, P. N.; Inkpen, K., et al. 2019. Guidelines for Human–AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI 2019)*. New York: Association for Computing Machinery.

Breazeal, C. 2003. Toward Sociable Robots. *Robotics and Autonomous Systems* 42(3-4): 167–75. doi.org/10.1016/S0921-8890(02)00373-1.

Breazeal, C. L. 2004. *Designing Sociable Robots*. Cambridge, MA: The MIT Press. doi.org/10.7551/mitpress/2376.001.0001.

Cantrell, R.; Talamadupula, K.; Schermerhorn, P. W.; Benton, J.; Kambhampati, S.; and Scheutz, M. 2012. Tell Me When and Why To Do It!: Run-Time Planner Model Updates via Natural Language Instruction. In *Proceedings of the International Conference on Human-Robot Interaction, HRI 2012*. 471–478. New York: Association for Computing Machinery. doi.org/10.1145/2157689.2157840.

Chakraborti, T.; Briggs, G.; Talamadupula, K.; Zhang, Y.; Scheutz, M.; Smith, D.; and Kambhampati, S. 2015. Planning for Serendipity. In *Proceedings of the 2015 IEEE/Robotics Society of Japan (RSJ) International Conference on Intelligent Robots and Systems, IROS 2015*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Chakraborti, T., and Kambhampati, S. 2019a. (How) Can AI Bots Lie? Presented at the ICAPS Workshop on Explainable Planning (XAIP), Berkeley, CA, July 11–15.

Chakraborti, T., and Kambhampati, S. 2019b. (When) Can AI Bots Lie? In *Proceedings of the 2019 Association for the Advancement of Artificial Intelligence (AAAI)/ACM Conference on AI, Ethics, and Society, AIES 2019*. New York: Association for Computing Machinery.

Chakraborti, T.; Kambhampati, S.; Scheutz, M.; and Zhang, Y. 2017. *AI Challenges in Human-Robot Cognitive Teaming*. arXiv:1707.04775. Ithaca, NY: Cornell University Library.

Chakraborti, T.; Kulkarni, A.; Sreedharan, S.; Smith, D.; and Kambhampati, S. 2019. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. In *Proceedings of the Twenty-Ninth International Conference on Automated Planning and Scheduling, ICAPS 2019*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence (AAAI).

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2018. Explicability Versus Explanations In Human-Aware Planning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2019. Plan Explanations As Model Reconciliation — An Empirical Study. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2019*. Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi.org/10.1109/HRI.2019.8673193.

Chakraborti, T.; Sreedharan, S.; Kulkarni, A.; and Kambhampati, S. 2018. Projection-Aware Task Planning and Execution for Human-In-The-Loop Operation of Robots in a Mixed-Reality Workspace. In *Proceedings of the IEEE/Robotics Society of Japan (RSJ) International Conference on Intelligent Robots and Systems, IROS 2018*. Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi.org/10.1109/IROS.2018.8593830.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI '17)*. International Joint Conference on Artificial Intelligence Organization. doi.org/10.24963/ijcai.2017/23.

Chakraborti, T.; Zhang, Y.; Smith, D.; and Kambhampati, S. 2016. Planning with Resource Conflicts in Human-Robot Cohabitation. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. New York: Association for Computing Machinery.

Cooke, N. J.; Gorman, J. C.; Myers, C. W.; and Duran, J. L. 2012. Interactive Team Cognition. *Cognitive Science* 37(2): 255–85. doi.org/10.1111/cogs.12009.

Grosz, B. J. 1996. AAAI-94 Presidential Address: Collaborative Systems. *AI Magazine* 17(2): 67–85.

Kulkarni, A.; Srivastava, A.; and Kambhampati, S. 2019. A Unified Framework for Planning in Adversarial and Cooperative Environments. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence. doi.org/10.1609/aaai.v33i01.33012479.

Kulkarni, A.; Zha, Y.; Chakraborti, T.; Vadlamudi, S.; Zhang, Y.; and Kambhampati, S. 2019. Explicable Planning as Minimizing Distance from Expected Behavior. In *Proceedings of the 2019 International Conference on Autonomous Agents & Multiagent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Lombrozo, T. 2006. The Structure and Function of Explanations. *Trends in Cognitive Sciences* 10(10): 464–70. doi.org/10.1016/j.tics.2006.08.004.

McNeese, N. J.; Demir, M.; Cooke, N. J.; and Myers, C. W. 2017. Teaming with a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors* 60(2): 262–73. doi.org/10.1177/0018720817743223.

Scassellati, B. 2002. Theory of Mind for a Humanoid Robot. *Autonomous Robots* 12(1): 13–24. doi.org/10.1023/A:1013298507114.

Sengupta, S.; Chakraborti, T.; Sreedharan, S.; Vadlamudi, S.; and Kambhampati, S. 2017. RADAR — A Proactive Decision Support System for Human-In-The-Loop Planning. *Journal of Artificial Intelligence* FS-17-01–FS-17-05: 269–76.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2018. Handling Model Uncertainty and Multiplicity in Explanations via Model Reconciliation. In *Proceedings of the Twenty-Eighth International Conference on Automated Planning and Scheduling (ICAPS 2018)*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence (AAAI).

Sreedharan, S.; Chakraborti, T.; Muise, C.; and Kambhampati, S. 2019. *Planning with Explanatory Actions: A Joint Approach to Plan Explicability and Explanations in Human-Aware Planning*. ArXiv preprint. arXiv:1903.07269. Ithaca, NY: Cornell University Library.

Sreedharan, S.; Olmo, A.; Mishra, A.; and Kambhampati, S. 2019. *Model-Free Model Reconciliation*. ArXiv preprint. arXiv:1903.07198. Ithaca, NY: Cornell University Library.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2018. Hierarchical Expertise Level Modeling for User Specific Contrastive Explanations. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI '18)*. International Joint Conference on Artificial Intelligence Organization. 4829–36. doi.org/10.24963/ijcai.2018/671.

Tian, X.; Zhuo, H.; and Kambhampati, S. 2016. Discovering Underlying Plans Based on Distributed Representations of Actions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. New York: Association for Computing Machinery.

VanLehn, K. 2006. The Behavior of Tutoring Systems. *International Journal of Artificial Intelligence in Education* 16(3): 227–65.

Wimmer, H., and Perner, J. 1983. Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition* 13(1): 103–28. doi.org/10.1016/0010-0277(83)90004-5.

Zha, Y.; Li, Y.; Gopalakrishnan, S.; Li, B.; and Kambhampati, S. 2018. Recognizing Plans by Learning Embeddings from Observed Action Distributions. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Zhang, Y.; Sreedharan, S.; Kulkarni, A.; Chakraborti, T.; Zhuo, H.; and Kambhampati, S. 2017. Plan Explicability and Predictability for Robot Task Planning. In *Proceedings of the 2017 IEEE International Conference on Robotics and Automation, ICRA 2017*. Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi.org/10.1109/ICRA.2017.7989155.

Subbarao Kambhampati (Rao) is a professor of computer science at Arizona State University. He received his BTech in electrical engineering (electronics) from the Indian Institute of Technology, Madras (1983), and his MS (1985) and PhD (1989) in computer science from the 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, Association for Computing Machinery (ACM), and the American Association for the Advancement of Science (AAAS), and was a National Science Foundation Young Investigator. He received multiple teaching awards, including a university "Last Lecture" recognition. Kambhampati is the past president of the AAAI, serving from 2016 to 2018. He served as a trustee of the International Joint Conference on Artificial Intelligence from 2013 to 2018. He was the program chair for the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 2016), the Twenty-Third International Conference on Automated Planning and Scheduling (ICAPS 2013), the Twentieth National Conference on Artificial Intelligence (AAAI 2005), and the Fifth International Conference on Artificial Intelligence Planning Systems (Artificial Intelligence Planning and Scheduling 2000). He served on the founding board of directors of the Partnership on Artificial Intelligence to Benefit People and Society (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.