Synthesizing Explainable Behavior for Human-AI Collaboration

Subbarao Kambhampati
Arizona State University
rao@asu.edu

ABSTRACT

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. This requires AI systems to exhibit behavior that is explainable to humans. Synthesizing such behavior requires AI systems to reason not only with their own models of the task at hand, but also about the mental models of the human collaborators. Using several case-studies from our ongoing research, I will discuss how such multi-model planning forms the basis for explainable behavior.

ACM Reference Format:
Subbarao Kambhampati. 2019. Synthesizing Explainable Behavior for Human-AI Collaboration. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 2 pages.

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 agent interacting with humans–be they embodied or virtual–will also need to take this implicit mental modeling into account. In order for the AI agent to show behavior that makes sense to the human, AI agents thus 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 humans in the loop, but includes the human's model of the AI agent's 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 $\widetilde{M}^H_r$, 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 it's goals/capabilities M^R_h, and needs an approximation of this, denoted $\widetilde{M}^R_h$. Synthesizing explainable behavior then becomes a challenge of supporting planning in the context of these multiple models. (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. [14].)

Proactive help: Left to itself, the AI agent will use M^R to synthesize its behavior. When the agent has access to $\widetilde{M}^H_r$, 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. [1]) or avoiding resource contention with them (c.f. [8]).

Explicability: When the agent has access to $\widetilde{M}^R_h$, 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 $\widetilde{M}^R_h$. This optmization can be done either in a model-based fashion, where the distances between π and π' are explicitly estimated (c.f. [10]) 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. [18]). 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 [3].

Explanation: In some cases, $\widetilde{M}^R_h$ 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^R_h 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^R_h, such that the robot behavior π is explicable with respect to $M^R_h + \mathcal{E}$, thus in essence making the behavior interpretable to human in light of the explanation. In [7] we show that computing such explanations can be cast as a *meta search* in the space of models spanning M^R and $\widetilde{M}^R_h$ (which is the AI agent's approximation of M^R_h). 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.

Balancing Explicability & Explanation: While the foregoing presented explicable behavior and presenting 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 $\widetilde{\pi}$ (where $\widetilde{\pi}$ is still not explicable with respect to M^R_h), and provide an explanation $\mathcal{E}'$ such that $\widetilde{\pi}$ is explicable with respect to $M^R_h + \mathcal{E}'$. In [4], 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 [12], we provide an approach to handle the incomplete model, viewing it as a union of complete models. More generally, the AI agent may

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

have to learn the model from the past traces of interaction with the human. In [16, 17], we discuss some efficient approaches for learning shallow models.

Multiple Humans & Abstraction:: The basic framework above can be generalized in multiple ways. In [15], we show how we can handle situations where the human and AI agent have models at different levels of abstraction. In [15] we consider explanations in the context of specific "foils" (e.g. "*why not this other type of behavior?*") presented by the humans. In [12], we consider how the AI agent can handle multiple humans–obviously with different models ($M^R_{h_i}$)– in the loop, and develop the notions of "conformant" *vs.* "conditional explanations."

Self-Explaining Behaviors: While the foregoing considered explanations on demand, it is also possible to directly synthesize *self explaining* behaviors. In [6], 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 [13], 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.

Validity & Evaluation: The explanations computed in our model reconciliation framework satisfy several desiderata–such as selectivity and contrastiveness that are seen as essential according to psychological theories. We have applied this framework in the context of human-robot interaction (e.g. [6]) and interaction between humans and virtual decision support systems (e.g. [11]). We have also conducted principled human-subject studies. In [5], we show that people indeed exchange the type of explanations we compute, and that the need for explanations diminishes when the behavior is explicable.

Manipulation & Ethical Considerations: 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 [9], we present 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 M^R_h closer to a model different from M^R. In [2], we explore the question of whether and when it is reasonable for AI agents to lie.

ACKNOWLEDGMENTS

The 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 and colleagues Nancy Cooke, Matthias Scheutz and David Smith. Thanks also to Behzad Kamgar-Parsi, Jeffery Morrison and Marc Steinberg for sustained support. 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.

REFERENCES

[1] T. Chakraborti, G. Briggs, K. Talamadupula, Y. Zhang, M. Scheutz, D. Smith, and S. Kambhampati. 2015. Planning for Serendipity. In *IROS*.
[2] T. Chakraborti and S. Kambhampati. 2019. (When) Can AI Bots Lie?. In *AIES*.
[3] T. Chakraborti, A. Kulkarni, S. Sreedharan, D. Smith, and S. Kambhampati. 2019. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. In *ICAPS*.
[4] T. Chakraborti, S. Sreedharan, and S. Kambhampati. 2018. Explicability Versus Explanations in Human-Aware Planning. In *AAMAS*.
[5] T. Chakraborti, S. Sreedharan, and S. Kambhampati. 2019. Plan Explanations as Model Reconciliation – An Empirical Study. In *HRI*.
[6] T. Chakraborti, S. Sreedharan, A. Kulkarni, and S. Kambhampati. 2018. Projection-Aware Task Planning and Execution for Human-in-the-Loop Operation of Robots in a Mixed-Reality Workspace. In *IROS*.
[7] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*.
[8] T. Chakraborti, Y. Zhang, D. Smith, and S. Kambhampati. 2016. Planning with resource conflicts in human-robot cohabitation. In *AAMAS*.
[9] A. Kulkarni, A. Srivastava, and S. Kambhampati. 2019. A Unified Framework for Planning in Adversarial and Cooperative Environments. In *AAAI*.
[10] A. Kulkarni, Y. Zha, T. Chakraborti, S. Vadlamudi, Y. Zhang, and S. Kambhampati. 2019. Explicable Planning as Minimizing Distance from Expected Behavior. In *AAMAS*.
[11] S. Sengupta, T. Chakraborti, S. Sreedharan, S. Vadlamudi, and S. Kambhampati. 2017. RADAR - A Proactive Decision Support System for Human-in-the-Loop Planning. *AAAI Fall Symposium on Human-Agent Groups* (2017).
[12] S. Sreedharan, T. Chakraborti, and S. Kambhampati. 2018. Handling model uncertainty and multiplicity in explanations via model reconciliation. In *ICAPS*.
[13] S. Sreedharan, T Chakraborti, C Muise, and S. Kambhampati. 2019. Planning with Explanatory Actions: A Joint Approach to Plan Explicability and Explanations in Human-Aware Planning . *ArXiv e-prints* abs/1903.07269 (2019).
[14] S. Sreedharan, A Olmo, A Mishra, and S. Kambhampati. 2019. Model-Free Model Reconciliation. *ArXiv e-prints* abs/1903.07198 (2019).
[15] S. Sreedharan, S. Srivastava, and S. Kambhampati. 2018. Hierarchical Expertise Level Modeling for User Specific Contrastive Explanations.. In *IJCAI*. 4829–4836.
[16] X. Tian, H. Zhuo, and S. Kambhampati. 2016. Discovering underlying plans based on distributed representations of actions. In *AAMAS*.
[17] Y. Zha, Y. Li, S. Gopalakrishnan, B. Li, and S. Kambhampati. 2018. Recognizing Plans by Learning Embeddings from Observed Action Distributions. In *AAMAS*.
[18] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. Zhuo, and S. Kambhampati. 2017. Plan Explicability and Predictability for Robot Task Planning. In *ICRA*.

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 and was a trustee of IJCAI. He was the program chair for IJCAI 2016 , ICAPS 2013, AAAI 2005 and AIPS 2000. He served on the 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.