Received: 1 October 2023 | Accepted: 13 November 2023

DOI: 10.1002/aaai.12142

HIGHLIGHT

Human-aware AI —A foundational framework for human–AI interaction

Sarath Sreedharan

Colorado State University, Fort Collins, Colorado, USA

Correspondence
Sarath Sreedharan, Colorado State University, Fort Collins, CO, USA.
Email: ssreedh3@colostate.edu

Abstract

We are living through a revolutionary moment in AI history. Users from diverse walks of life are adopting and using AI systems for their everyday use cases at a pace that has never been seen before. However, with this proliferation, there is also a growing recognition that many of the central open problems within AI are connected to how the user interacts with these systems. To name two prominent examples, consider the problems of explainability and value alignment. Each problem has received considerable attention within the wider AI community, and much promising progress has been made in addressing each of these individual problems. However, each of these problems tends to be studied in isolation, using very different theoretical frameworks, while a closer look at each easily reveals striking similarities between the two problems. In this article, I wish to discuss the framework of human-aware AI (HAAI) that aims to provide a unified formal framework to understand and evaluate human–AI interaction. We will see how this framework can be used to both understand explainability and value alignment and how the framework also lays out potential novel avenues to address these problems.

INTRODUCTION

In the past few years, we have seen a tremendous advancement in what AI systems can do. Recent progress in AI has enabled us to reach milestones that even the most optimistic futurists believed were beyond our reach for the time being. Many in the field, including several luminaries, now think that an AI with human-like intelligence is nearer than previously anticipated. However, as the capabilities of these systems have kept improving and more users are adopting these systems in their everyday lives, we, as AI researchers, are confronted by a set of new challenges, many of which are related to how people would work with these systems. Among the many problems that have received attention in recent years, the two this article is particularly interested in exploring are the problem of *generating explanations* (Gunning and Aha 2019) and the problem of *value alignment* (Hadfield-Menell et al. 2016). For explanation generation, the article is particularly interested in equipping an AI agent with the mechanisms necessary to explain why a specific decision it made was the right one for a given task (Langley 2019). As for value-alignment, we will follow the general outline sketched by Hadfield-Menell et al. (2016) and look at the problem of ensuring that the outcomes of an agent's behavior align with the true intent of its users. Developing robust

solutions to these problems is central to developing AI systems that users can trust and employ in safety-critical settings. In fact, for value-alignment, the problem is widely discussed in the context of the existential risk posed by AI systems. However, the utility and necessity of addressing these problems are not just limited to civilization-ending doomsday devices but are central to every human–AI interaction. To side-step any sensationalistic discussions and to further reinforce the fundamental nature of these problems, we will focus primarily on how these problems manifest in the context of a quotidian interaction setting. In particular, let us take a case of a user asking a robot to make a cup of tea. Now, the user might ask the robot to explain itself; if it does something unanticipated, say instead of pouring hot water into the cup, it pours cold water over the tea bag. On the other hand, we would say that there is a need for value alignment if the agent uses low-quality tea bags to make tea instead of using the ones the user likes.

Effectively solving each problem requires the agent to model what the user expects from the agent and how that differs from what the agent believes to be the best course of action. Returning to the tea example, for the agent to explain its choice of cold water, it needs to realize that the user is unaware of the fact that the robot is, in fact, incapable of safely handling a kettle. On the other hand, it can microwave a cup of water with a tea bag. Thus, an effective explanation must inform the user of this fact and how they can still make tea starting with cold water. Moving over to value alignment, the human expected the agent to pick the tea bag they liked because it was on a shelf next to the agent, while the low-quality tea bags were in the next room. As such, the user did not anticipate the need to be specific about what tea bags to use. However, unknown to the user, the robot could not directly reach the shelves and would have needed additional assistance to get the tea bags. In each of the above examples, we see the central role played by the user's belief about the agent and, by extension, their expectations from the agent. However, the need to support such modeling is important, not just to address these two specific challenges but to allow effective human–AI interaction and collaboration in general.

Toward this end, this article lays out some of the recent developments that have been made with respect to the development of the human–aware AI (HAAI) interaction framework (Sreedharan, Kulkarni, and Kambhampati 2022). HAAI is a psychologically feasible, multi-agent interaction framework that can be used to understand human–AI interaction. HAAI builds on many psychological concepts, such as the theory of mind (Apperly and Butterfill 2009), and makes the modeling of user beliefs and expectations as being central to modeling all user–AI interactions.

Modeling of other agents, their epistemological states and beliefs, are not novel within multi-agent decision-making literature. In fact, these form the basis of some of the most fundamental works within this area. What is novel, however, is that HAAI uses the insights from such normative work to propose a framework that can provide a unified account for problems such as explanation generation and value alignment. In doing so, it reveals novel ways of generating explanations and aligning the agent's objective with what the user expected.

In the rest of the article, we will start the discussion by laying out the basic HAAI setting. Next, we will discuss how we can use this basic setting to understand the problem of explanation generation and value alignment. Through these discussions, we will see the central role played by human expectations in shaping human–AI interaction dynamics. In particular, within *HAAI*, the explanation corresponds to the agent helping the humans adjust their behavioral expectations about the agent to match the agent's behavior. One way to achieve this might involve the agent updating the human's belief about the agent model (cf. Sreedharan, Chakraborti, and Kambhampati 2021). On the other hand, in value alignment, the agent tries to recognize the human's original expectations and match them. In opposition to the explanation case, this might involve the agent trying to learn human beliefs and preferences (for example, via model or preference elicitation (Goldsmith and Junker 2008)) or replicating the human inference process. We will also discuss all the work that has already been completed within the aegis of the framework and all the open problems and questions the framework surfaces.

BASIC HUMAN-AWARE AI SETTING

The basic *HAAI* setting is an interaction setting with just two agents (i.e., the human and the robot). We will look at one where the robot [1] takes the role of the actor, coming up with plans and executing them, and the human takes the role of a supervisor. The human provides instructions or objectives to the robot that they expect it to carry out. Figure 1 provides a graphical overview of this setting.

The first point of note is that both the human and the robot maintain models of the task at hand. The robot model here corresponds to a traditional AI model, one that encodes the state, objectives, and dynamics (which, in this case, encodes how the robot can influence the task state). The human model, on the other hand, is more interesting in that it represents the human belief in the task. It may encode the human beliefs about the robot's capabilities and the task state. Now, both the human and the robot would make use of their innate reasoning capabilities to derive

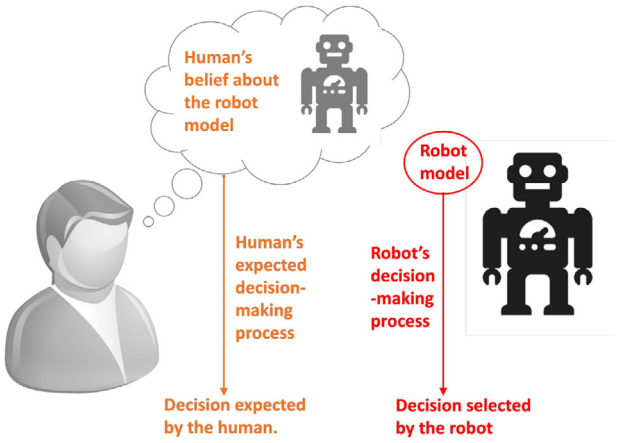

FIGURE 1 A diagrammatic representation of the overall framework of human-aware AI. The image highlights the various models that come into play in this setting and how the individual decision-making process helps shape the human–AI interaction.

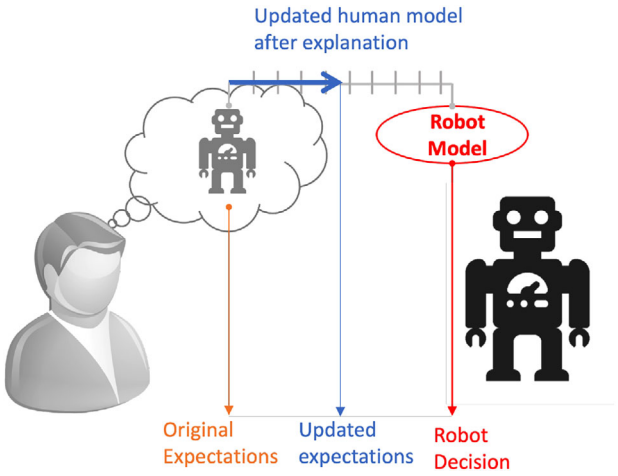

FIGURE 2 The image highlights the basic process of explanations as modeled within the *HAAI* framework, that is, a process of updating the expectations of the user so they better align with the agent's proposed plan.

what they believe the best course of action to address the task would be.

One of the primary assertions the *HAAI* framework makes is the fact that many problems in human–AI interaction, including inexplicability and value misalignment, arise from the mismatch in the human expectations about robot behavior and the decision the robot arrives at. In particular, the human would demand an explanation if the robot exhibits a behavior that is quite different from what they expected. Similarly, the human would have formalized the objective they presented to the robot based on what they thought the robot was capable of, and their belief about how it would carry out their specified objective. If the human expectations are wrong, it could result in behavior whose outcomes may not match the human's true intent, thus resulting in value misalignment. In this article, we will try to understand this overall expectation mismatch by focusing on three salient dimensions of asymmetry between the human and the robot. Specifically, asymmetry in knowledge: This dimension relates to the difference in the human's belief about the task and the contents of the agent model. The differences that could manifest as part of this dimension could range from the human's misunderstanding of the robot's capabilities to disagreement about the current state of the world. Asymmetry in inferential capabilities: This dimension corresponds to differences in human and robot capability to come up with effective decisions given their task models. There is a lot of evidence to support the fact that humans may be best understood as bounded-rational agents. On the other hand, depending on the use case, the robot could range from being an optimal decision-maker to an unsound or highly suboptimal one. Asymmetry in vocabulary: The final dimension relates to the differences in how the human and the robot may represent the task. While the human usually tries to make sense of the world in terms of high-level concepts, or in many physical tasks, in terms of objects and the relationship between objects, many modern AI systems may employ learned representation schemes that are inscrutable to users. Even if they have access to the same vocabulary, their grounding could differ significantly. Through the remaining sections of the article, we will see how these dimensions relate to the problem of explanation generation and value alignment. Finally, we will see how the framework of *HAAI* allows us to see how these two pieces connect to the overall problem of human–AI interaction.

EXPLANATION GENERATION

The first aspect we will investigate within this framework is that of explanations. As discussed previously, the central problem that arises within these settings is the mismatch between the behavior expected by the user and the one that the robot picks. Now, when presented with such unexpected behavior, the user may ask for an explanation, and thus, the role of explanation becomes to help the user understand why the course of action selected by the agent is the right one (at least from the agent's perspective). Figure 2 presents a visualization of this process of updating user expectations.

This style of explanation aligns with what has been referred to as preference explanation within the XAI literature (Langley 2019), in so far that the agent is trying to establish why the decision selected by it may be preferred over other alternatives (particularly the one the user had in mind). However, it is worth noting that other types

of explanations, such as process explanations, can be captured in this context. However, the article will focus on the former as most everyday explanation interactions fall under the category of preference explanations.

Now, going back to our three dimensions of asymmetry, we see that the first two (asymmetry in knowledge and asymmetry in inferential capabilities), correspond to potential sources for confusion in the first place, and the third dimension (asymmetry in vocabulary) could be a barrier for providing useful explanations. As such, effective explanation generation involves addressing all three dimensions.

Some of the earliest works to utilize the HAAI framework were explanation generation works that focused on addressing knowledge asymmetry. These works introduced and built on the model-reconciliation technique (Sreedharan, Chakraborti, and Kambhampati 2021), which generates explanations for users when their confusion about the system behavior primarily arose from a mismatch between the user's understanding of the agent model and the true model. The challenges to generating such explanations include finding concise and focused explanations that cover only parts of the model relevant to the current decision (Chakraborti et al. 2017) and unknown to the user. The system would also need to do so while not having an accurate or complete user model (Sreedharan, Kambhampati, and Chakraborti 2018). In our running example, the explanation could involve the robot specifying its limitations. For example, it could inform the user that it cannot perform the dipping action or use a kettle. This would leave microwaving a cup of water with a tea bag as the only course of action.

There have also been a number of works that explicitly evoke a HAAI framework that focuses on the second dimension, namely, addressing the asymmetry in inferential capabilities. Many of these works (cf. Sreedharan, Srivastava, and Kambhampati 2021, 2018) employ model minimization strategies like abstraction or problem decomposition and provide the user with the ability to ask more directed explanatory queries (Sreedharan et al. 2019). As for the running example, the agent utilizes a form of model abstraction by providing explanations that refer to temporally abstract skills such as dipping. In contrast to providing an explanation that refers to joint angles or manipulator positions, even though the robot may have originally reasoned about its actions at that level of detail.

Finally, coming to the third dimension, the explanations need to be provided to users in terms that are intuitive to them. Many of the state-of-the-art AI decision-making systems rely on internal representation schemes that are inherently inscrutable and unintuitive to lay users. However, in many cases, it may be possible to provide post hoc explanations about decisions in terms that are intuitive to end users. In our running example, even if the system generated its decisions using a neural network, it can choose to provide explanations framed in terms of objects the user understands (say, tea-bag, kettle, etc.) or the overall goal of making tea. Such methods have been employed both within the context of sequential decision-making problems (cf. Sreedharan et al. 2020) and single-shot decision-making problems (Kim et al. 2018).

HUMAN-AWARE VALUE ALIGNMENT

Now, we move on to the next problem tackled in this article, namely value alignment. In this section, we will see how value alignment is effectively the inverse of the problem presented in the previous section. As shown in Figure 3, the process starts with a user who wants the robot to perform a certain task. The user formulates an objective specification, whose successful completion the user believes will result in the robot successfully completing the task. Now, considering the three dimensions of asymmetry, we can easily see how the objective specification may not be the true encoding of the user's intent. For one, given the first two dimensions, the user's expectations about the agent's behavior and, by extension, the final outcomes could be quite different from the one obtained by the robot directly optimizing for the specified objective. As such, the robot's motivation should be to generate instead behavior whose outcomes align with the ones that the human originally expected when provided the specification.

While more works have started looking at how one could generate such aligned behaviors, the problem remains an open one. The primary challenge here is to use the specified objective and knowledge about the user's belief and reasoning capabilities to infer their true intent. Then, generating behavior that aligns best with the true intent and potentially informing the user about its inability to meet it (as appropriate). Works like Mechergui and Sreedharan (2023) have started looking at addressing this problem in the presence of knowledge asymmetry. In particular, the method described in this work tries to query the user about the outcomes (i.e., the final goal state) until it is certain that it can or cannot generate a state that satisfies human expectations. In particular, the work tries to revisit and reinterpret existing works in value alignment using the lens of HAAI. In our running example, this could correspond to the system leveraging its estimate of the user's belief to try to predict the outcomes the user would have expected. For example, the higher-quality tea bags are on the shelf right next to the robot, and as such, the user, who is unaware of the robot's incapability to reach those tea bags, would expect those to be used. The system could

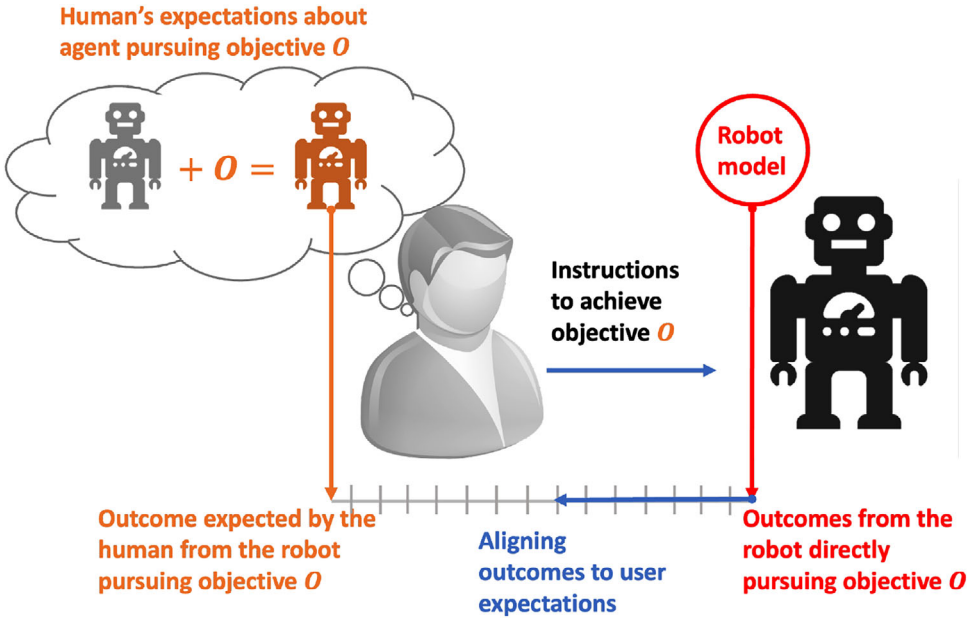

FIGURE 3 A graphical representation of the value alignment process as visualized within the *HAAI* framework.

use its estimate to predict such a potential expectation and query the user about it.

That said, considerable work remains to be done in handling value alignment. To start with, there is a clear lack of methods that are designed to handle issues that arise from asymmetry along the latter two dimensions. As with knowledge asymmetry, inferential asymmetry could lead to the user forming incorrect expectations about behavior the system may generate in response to a given objective. The asymmetry in vocabulary between the user and the system could hinder the effective communication of the human objective in multiple ways. For example, a lack of shared vocabulary or difference in the grounding of vocabulary items could make it hard for the user to communicate the objective they had in mind correctly. For example, if the system cannot differentiate between higher quality and lower quality tea bags, then a user would not be able to specify their true intent even if they were trying to express it explicitly.

As a final coda to this section, it is worth noting the similarity and difference between value alignment (as defined in this section) and explicable behavior generation (Kulkarni et al. 2019; Chakraborti et al. 2019). Explicable behavior generation focuses on generating behavior that aligns with user expectations. However, this problem is generally studied in settings where the human goal is assumed to be correctly and completely specified[2]. As such, the focus is to align the agent's behavior to what the user expects. On the other hand, the primary focus of value alignment work is to align the outcomes of the agent's behavior with what the user expected. This is particularly important because the problem setting considered is one where the objective provided to the robot is not necessarily the complete specification of the true user intent. One way this could be achieved might be by aligning the system's behavior with the user's expectations. However, in cases where the user's understanding of the task/robot model is incorrect, following the expected behavior need not result in the outcomes aligning with the user's expectations.

USING EXPECTATION-MISMATCH TO UNDERSTAND OVERALL HUMAN–AI INTERACTION

In the previous sections, we have looked at two problems widely recognized as being important for successful human–AI interaction and discussed how the framework of *HAAI* allows us to view them through the lens of expectation mismatch. In addition to looking at these problems in isolation, it is possible to see them as being related to individual steps in the overall human–AI interaction process. Figure 4 shows how these two problems connect to each other. In particular, the figure shows how the user develops a specific objective specification based on their current beliefs. In response, the robot chooses a behavior that best aligns with the user's underlying intent and proposes it to the user along with explanations or other relevant information (for example, information about the outcome). This could result in the user updating their beliefs and expectations, which in turn could cause the user to revisit their objective specifications. This interaction loop sketches a novel paradigm to understand value alignment and human–AI interaction in general.

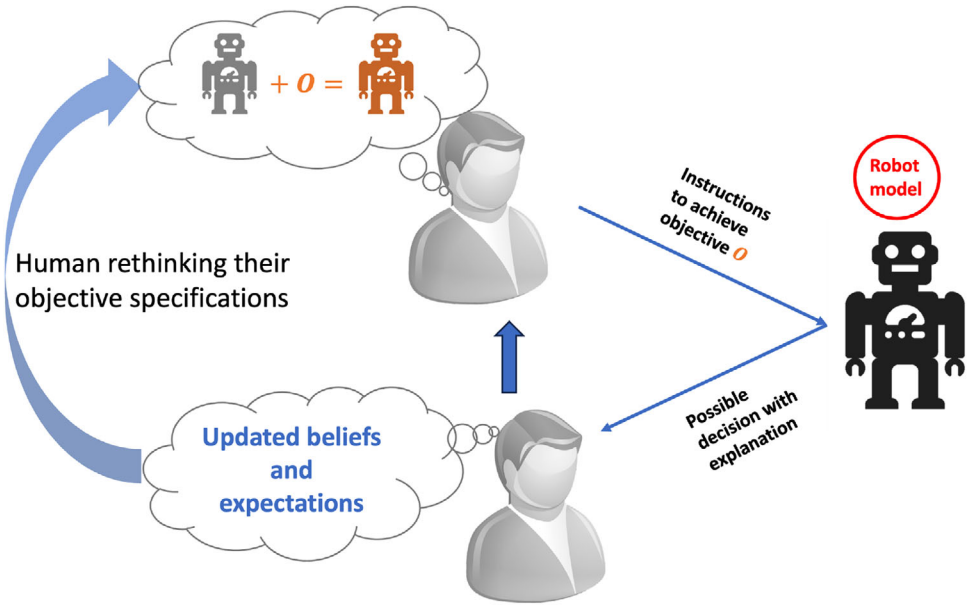

FIGURE 4 The image shows how the two problems discussed in the previous sections fit together in the overall context of human–AI interaction.

Under this paradigm, the robot tries to match the user's expectations. If the robot identifies potential opportunities that the user may have overlooked (and hence not expected), it would first try to make the user of said opportunity and only pursue the opportunity if the user updates their expectations and potentially the specified objective. While such comprehensive interaction loops are naturally expressed within *HAAI*, developing techniques to support such interactions remains an open challenge. One we hope will soon be tackled by the community.

CONCLUSION

In summary, the article provides an overview of the *HAAI* interaction framework proposed to model human–AI interactions. The proposed model places human expectations and potential asymmetry between expectation and agent choice at the center of human–AI interaction. In particular, the article discusses how two important problems within human–AI interaction, namely explainability and value alignment, could be understood in terms of human expectations and expectation mismatch. While many existing works within HAAI have mostly focused on capturing and providing tools to address the above-discussed problems, there has also been recent interest in capturing other phenomena related to human–AI interaction. Most prominently, recent works have tried to map trust in human-aware terms (cf. Zahedi, Sreedharan, and Kambhampati 2023). Also, while the basic psychological validity of the framework has been established through multiple user studies that have been performed over the years (Chakraborti et al. 2019; Grover et al. 2020), there is always scope to further build and expand the framework as we uncover more insights into the human mental processes driving these interactions. The author hopes that *HAAI* framework not only provides the community with a formal pedagogical framework to understand human–AI interaction but also allows for developing novel solutions to human–AI interaction problems. Specifically, by providing a model of how human expectations form and influence human–AI interaction, the framework provides new mechanisms and insights into how we can influence and improve the overall human–AI interaction.

CONFLICT OF INTEREST STATEMENT

The author declares that there is no conflict.

ORCID

Sarath Sreedharan https://orcid.org/0000-0002-2299-0178

ENDNOTES

[1]We loosely refer to our agent as a robot. No component of our framework requires our system to be physically embodied.

[2]In other words, the specified objective correctly encodes the true intent of the user.

REFERENCES

Apperly, I. A., and S. A. Butterfill. 2009. "Do Humans have Two Systems to Track Beliefs and Belief-Like States?" *Psychological Review* 116: 953–70.

Chakraborti, T., A. Kulkarni, S. Sreedharan, D. E. Smith, and S. Kambhampati. 2019. "Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior." In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, 86–96.

Chakraborti, T., S. Sreedharan, S. Grover, and S. Kambhampati. 2019. "Plan Explanations As Model Reconciliation—An Empirical Study." In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 258–66. IEEE.

Chakraborti, T., S. Sreedharan, Y. Zhang, and S. Kambhampati. 2017. "Plan Explanations As Model Reconciliation: Moving Beyond Explanation As Soliloquy." In *IJCAI 2017*, 156–63. ijcai.org.

Goldsmith, J., and U. Junker. 2008. "Preference Handling for Artificial Intelligence." *AI Magazine* 29: 9–12.

Grover, S., S. Sengupta, T. Chakraborti, A. P. Mishra, and S. Kambhampati. 2020. "Radar: Automated Task Planning for Proactive Decision Support." *Human–Computer Interaction* 35: 387–412.

Gunning, D., and D. W. Aha. 2019. "Darpa's Explainable Artificial Intelligence (XAI) Program." *AI Magazine* 40: 44–58.

Hadfield-Menell, D., S. Russell, P. Abbeel, and A. D. Dragan. 2016. "Cooperative Inverse Reinforcement Learning." In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, ed. D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett, 3909–17.

Kim, B., M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. 2018. "Interpretability Beyond Feature Attribution: Quantitative Testing With Concept Activation Vectors (TCAV)." In *International Conference on Machine Learning*, 2668–77. PMLR.

Kulkarni, A., Y. Zha, T. Chakraborti, S. G. Vadlamudi, Y. Zhang, and S. Kambhampati. 2019. "Explicable Planning As Minimizing Distance From Expected Behavior." In *AAMAS Conference Proceedings*, 2075–77.

Langley, P. 2019. "Varieties of Explainable Agency." In *XAIP Workshop*, 113–17. XAIP.

Mechergui, M., and S. Sreedharan. 2023. "Goal Alignment: Reanalyzing Value Alignment Problems Using Human-Aware AI." *In AAMAS Conference Proceedings, 2331-2333*.

Sreedharan, S., T. Chakraborti, and S. Kambhampati. 2021. "Foundations of Explanations as Model Reconciliation." *Artificial Intelligence* 301: 103558.

Sreedharan, S., S. Kambhampati, and T. Chakraborti. 2018. "Handling Model Uncertainty and Multiplicity in Explanations Via Model Reconciliation." In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 28, 518–26.

Sreedharan, S., A. Kulkarni, and S. Kambhampati. 2022. *Explainable Human-AI Interaction: A Planning Perspective*. Cham: Springer Nature.

Sreedharan, S., U. Soni, M. Verma, S. Srivastava, and S. Kambhampati. 2020. "Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems With Black Box Simulators." *CoRR*. abs/2002.01080. https://arxiv.org/abs/2002.01080. arXiv:2002.01080.

Sreedharan, S., S. Srivastava, and S. Kambhampati. 2018. "Hierarchical Expertise Level Modeling for User Specific Contrastive Explanations." In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 4829–36. ijcai.org.

Sreedharan, S., S. Srivastava, and S. Kambhampati. 2021. "Using State Abstractions to Compute Personalized Contrastive Explanations for AI Agent Behavior." *Artificial Intelligence* 301: 103 570.

Sreedharan, S., S. Srivastava, D. E. Smith, and S. Kambhampati. 2019. "Why Can't You Do That Hal? Explaining Unsolvability of Planning Tasks." In *IJCAI 2019*, 1422–30. ijcai.org.

Zahedi, Z., S. Sreedharan, and S. Kambhampati. 2023. "A Mental Model Based Theory of Trust." *In XAI Workshop at IJCAI*.

How to cite this article: Sreedharan, S. 2023. "*Human-aware AI* —A foundational framework for human–AI interaction." *AI Magazine* 44: 460–466. https://doi.org/10.1002/aaai.12142

AUTHOR BIOGRAPHY

Sarath Sreedharan is an Assistant Professor at Colorado State University. His core research interests include designing human-aware decision-making systems that can generate behaviors that align with human expectations. He completed his Ph.D. at Arizona State University, where his doctoral dissertation received the 2022 Dean's Dissertation Award for Ira A. Fulton Schools of Engineering and was also awarded ICAPS-2023 honorable mention for best dissertation. His research has been published in various premier research conferences, including AAAI, ICAPS, IJCAI, AAMAS, IROS, HRI, ICRA, ICML, and ICLR, and journals like AIJ. He has presented tutorials on his research at various forums and is the lead author of a Morgan Claypool monograph on explainable human–AI interactions. He was selected as a DARPA Riser Scholar for 2022 and as a Highlighted New Faculty by AAAI. His research has won multiple awards, including the Best System's Demo and Exhibit Award at ICAPS-20 and the Best Paper Award at Bridging Planning & RL workshop at ICAPS 2022. He was also recognized as a AAAI-20 Outstanding Program Committee Member, Highlighted Reviewer at ICLR 22, IJCAI 2022 and 2023 Distinguished Program Committee Member, and a Top Reviewer at NeurIPS 22.