Nov 2024

Sarath Sreedharan

Assistant Professor at Colorado State University

Contact: Sarath.Sreedharan@colostate.edu
Home: http://sarathsreedharan.com/
Google Scholar: https://bit.ly/2OdYCtn

Research Interests

- Human-aware planning for human-agent teams and decision support - Explicable/Interpretable behavior generation, modeling human decision making.
- Explanation Generation for Sequential Decision-Making Problems.
- Planning and Learning - Combining planning and learning methods for more effective long term sequential decision-making.
- Model-Lite Planning - Representation and learning of incomplete models for plan generation, recognition and recommendation under uncertainty.

Education

2016–2022	Ph.D. , Computer Science, Arizona State University
2014–2016	M.S., Computer Science, Arizona State University
2007–2011	B.Tech, Computer Science, CUSAT (Model Engg. College)

Professional Experience

8/2022 – Current	Assistant Professor	Colorado State University
5/2020 – 7/2020	AI Research Intern	IBM-Research AI
5/2019 – 7/2019	AI Research Intern	Invitae
8/2014 – 12/2016	TA CSE 471/CSE 301	Arizona State University
5/2015 – 8/2015	SRE Intern	LinkedIn
5/2011 – 6/2014	Snr. Software Engg./ Software Engg.	Zynga

Awards and Recognition

- Selected as 2024 EAISI visiting professor for Eindhoven University of Technology
- Best paper award IEEE TPS 2024
- Honorable Mention for the ICAPS 2023 Best Dissertation Award
- Selected as a Highlighted new Faculty at AAAI-2023
- 2022 Dean's Dissertation Award winner – Ira A. Fulton Schools of Engineering (ASU)
- Distinguished PC Members IJCAI 2022 and 2023 (top 3%)
- DARPA Riser Scholar 2022
- ICAPS 20 Best System Demonstration Award.
- AAAI-20 Outstanding Program Committee Member award (one of 12 recognized)
- Highlighted Reviewer at ICLR 22
- Recognized as a Top Reviewer at NeurIPS 22.
- Received CSE Outstanding Masters Student award from CIDSE, ASU.
- National Finalist US Imagine cup 2017 (organized by Microsoft).
- Received Champion Award, Emerging Star Award (Zynga).

Invited Talks and Tutorials

- "Human-Aware AI – A Foundational Framework for Human-AI Interaction" - HAXP ICAPS-24 and Agents-Vic Autumn Symposium 2024
- "Expectations and Mental Models: A Holistic Approach to Understanding Explainability, Value Alignment, and Trust" - RAD-AI AAMAS-24
- Invited talk at AAAI-23 highlighted young faculty track.
- ATAL AICTE Faculty Development Program on Explainable AI organized by IIIT Kota
- Invited speaker to SERB KARYASHALA Workshop on Understanding Machines: Explainable AI
- CNRS 2021 Summer School on Explainability
- AAAI 2020 tutorial on Synthesizing Explainable and Deceptive Behavior

Research Grants (Reverse Chronological Order)

- An AI Tutoring System for Pollinator Conservation Community Science Training. ($849,890). NSF Research on Emerging Technologies for Teaching and Learning (RETTL). PI. Team: PI: Sarath Sreedharan. Co-PIs: Nikhil Krishnaswamy, Nathaniel Blanchard, Jill Zarestky.

- Learning from Richer Feedback Through the Integration of Prior Beliefs.(€15,000). Humane-AI Network Microproject. Co-PI. Team: PI: Silvia Tulli. Co-PIs: Sarath Sreedharan, Mohamed Chetouani.
- TRACE: Transparency, Reflection, and Accountability in Conversational Exchanges. ($978,331) Co-PI. Team: Nikhil Krishnaswamy, Nathaniel Blanchard, Bruce Draper, James Pustejovsky.

Publications

As per google scholar, my papers have received a total of **3240** citations; I have an h-index of **28** and an i10-index of **47**.

Books

[1] **Explainable Human-AI Interaction: A Planning Perspective**. *S. Sreedharan*, A. Kulkarni and S. Kambhampati. Morgan and Claypool Publishers. (184 Pages). DOI: 10.2200/S01152ED1V01Y202111AIM050 ISBN: 9781636392899.

Journal Publications

[2] **Planning with Mental Models – Balancing Explanations and Explicability.** *S. Sreedharan*, T. Chakraborti, C. Muise, and S. Kambhampati . AI Journal.

[3] **Human-Aware AI – A Foundational Framework for Human-AI Interaction.** *S. Sreedharan.* AI Magazine - Volume44, Issue4.

[4] **Foundations of Explanations as Model Reconciliation.** *S. Sreedharan**, T. Chakraborti*, and S. Kambhampati. Artificial Intelligence Journal.

[5] **Using State Abstractions to Compute Personalized Contrastive Explanations for AI Agent Behavior** *S. Sreedharan*, S.Srivastava, and S. Kambhampati. Artificial Intelligence Journal.

[6] **Robust planning with incomplete domain models.** T. Nguyen, S. Kambhampati, *S. Sreedharan.* Artificial Intelligence Journal.

Conference Publications

[7] **Expectation Alignment: Handling Reward Misspecification in the Presence of Expectation Mismatch** M. Mechergui, *S. Sreedharan.* NeurIPS 2024.

[8] **Resiliency Graphs: Modelling the Interplay between Cyber Attacks and System Failures through AI Planning** S. Bashir, R. Podder, S. Sreedharan, I. Ray and I. Ray. IEEE TPS 2024 (Won the Best Paper Award)

[9] **HELP! Providing Proactive Support in the Presence of Knowledge Asymmetry.** T. Caglar, *S. Sreedharan.* AAMAS 2024.

[10] **Towards More Likely Models for AI Planning.** T. Caglar, S. Belhaj, M. Katz, T. Chakraborti, *S. Sreedharan.* . AAAI 2024.

[11] **A Human-Aware Method for Addressing Goal Alignment.** M. Mechergui, *S. Sreedharan.* AAAI 2024.

[12] **A Wireframe-based Approach for Classifying and Acquiring Proficiency in the American Sign Language.** D. Pallickara, *S. Sreedharan.* AAAI 2024. (Student Abstract Track).

[13] **Optimistic Exploration in Reinforcement Learning Using Symbolic Model Estimates** *S. Sreedharan*, M. Katz. NeurIPS 2023.

[14] **On the Planning Abilities of Large Language Models – A Critical Investigation.** K. Valmeekam, M. Marquez, *S. Sreedharan*, S. Kambhampati. NeurIPS 2023. (Accepted as Spotlight)

[15] **Leveraging Pre-trained Large Language Models to Construct and Utilize World Models for Model-based Task Planning.** L Guan*, Karthik Valmeekam*, *S. Sreedharan*, S. Kambhampati. NeurIPS 2023.

[16] **PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change.** K. Valmeekam, M. Marquez, A. Olmo, *S. Sreedharan*, S. Kambhampati. NeurIPS 2023. (Benchmark Track).

[17] **Human-Aware AI – A Foundational Framework for Human-AI Interaction** *S. Sreedharan.* AAAI 2023 (New Faculty Track)

[18] **Generalizing Action Justification and Causal Links to Policies.** *S. Sreedharan*, C. Muise, S. Kambhampati. ICAPS 2023.

[19] **Planning for Attacker Entrapment in Adversarial Settings.** B. Cates, A. Kulkarni, *S. Sreedharan.* ICAPS 2023

[20] **Goal Alignment: Re-analyzing Value Alignment Problems Using Human-Aware AI.** M. Mechergui, *S. Sreedharan.* AAMAS 2023 (Extended Abstract).

[21] **Trust-Aware Planning: Modeling Trust Evolution in Iterated Human-Robot Interaction.** Z. Zahedi, M. Verma, *S. Sreedharan*, S. Kambhampati. HRI 2023.

[22] **Leveraging Approximate Symbolic Models for Reinforcement Learning via Skill Diversity** L. Guan*, *S. Sreedharan**, and S. Kambhampati. ICML 2022.

[23] **On the Computational Complexity of Model Reconcilation** *S. Sreedharan*, P. Bercher, and S. Kambhampati. IJCAI 2022.

[24] **Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Inscrutable Representations** *S. Sreedharan*, U. Soni, M. Verma, S.Srivastava, and S. Kambhampati. ICLR 2022.

[25] **RADAR-X: An Interactive Mixed Initiative Planning Interface Pairing Contrastive Explanations and Revised Plan Suggestions** K. Valmeekam, *S. Sreedharan*, s. Sengupta, s. Kambhampati. ICAPS 2022.

[26] **Symbols as a Lingua Franca for Bridging Human-AI Chasm for Explainable and Advisable AI Systems** S. Kambhampati, *S. Sreedharan*, M. Verma, Y. Zha, L. Guan and AAAI 2022 (Blue Sky Track).

[27] **Not all users are the same: Providing personalized explanations for sequential decision making problems** U. Soni, *S. Sreedharan*, and S. Kambhampati. IROS 2021.

[28] **A Unifying Bayesian Formulation of Measures of Interpretability in Human-AI Interaction** *S. Sreedharan*, A. Kulkarni, D. Smith, and S. Kambhampati. IJCAI-Survey 2021.

[29] **Designing Environments Conducive to Interpretable Robot Behavior** . A. Kulkarni*, *S. Sreedharan**, S. Keren, T. Chakraborti, D. Smith, S. Kambhampati. IROS, 2020.

[30] **The Emerging Landscape of Explainable AI Planning and Decision Making**. T. Chakraborti*, *S. Sreedharan**, S. Kambhampati. IJCAI, 2020.

[31] **TLdR: Policy Summarization for Factored SSP Problems Using Temporal Abstractions.** S.Sreedharan, S.Srivastava, and S. Kambhampati. ICAPS 2020.

[32] **D3WA+: A Case Study of XAIP in a Model Acquisition Task .** *S. Sreedharan**, T. Chakraborti*, C. Muise, Y. Khazaeni, and S. Kambhampati. ICAPS 2020.

[33] **Expectation-Aware Planning: A Unifying Framework for Synthesizing and Executing Self-Explaining Plans for Human-Aware Planning** *S. Sreedharan*, T. Chakraborti, C. Muise, and S. Kambhampati. AAAI 2020.

[34] **Model-Free Model Reconciliation.** *S. Sreedharan*, A. Olmo, A. Mishra and S. Kambhampati. IJCAI 2019.

[35] **Why Can't You Do That HAL? Explaining Unsolvability of Planning Tasks .** S.Sreedharan, S.Srivastava, D. Smith, and S. Kambhampati. IJCAI 2019.

[36] **Balancing Explicability and Explanations for Human-Aware Planning**. T. Chakraborti*, *S. Sreedharan**, S. Kambhampati. IJCAI 2019.

[37] **CAP: A Decision Support System for Crew Scheduling using Automated Planning**. A. Mishra, S. Sengupta, *S. Sreedharan*, T. Chakraborti and S. Kambhampati. NDM 2019.

[38] **Plan Explanations as Model Reconciliation − An Empirical Study**. T. Chakraborti, *S. Sreedharan*, S. Grover and S. Kambhampati. HRI 2019.

[39] **Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior**. T. Chakraborti, A. Kulkarni, *S. Sreedharan*, D. Smith, S. Kambhampati. ICAPS 2019

[40] **Projection-Aware Task Planning and Execution for Human-in-the-Loop Operation of Robots in a Mixed-Reality Workspace**. T. Chakraborti, *S. Sreedharan*, A. Kulkarni, and S. Kambhampati. IROS 2018.

[41] **Hierarchical Expertise Level Modeling for User Specific Contrastive Explanations**. *S. Sreedharan*, S. Srivastava, and S. Kambhampati. IJCAI 2018.

[42] **Balancing Explicability and Explanations: Emergent Behaviors in Human-Aware Planning.**. T. Chakraborti*, *S. Sreedharan**, and S. Kambhampati. AAMAS 2018 (Extended Abstract).

[43] **Handling Model Uncertainty and Multiplicity in Explanations via Model Reconciliation**. *S. Sreedharan**, T. Chakraborti*, S. Kambhampati. ICAPS 2018.

[44] **Plan explanations as model reconciliation: Moving beyond explanation as soliloquy**. T. Chakraborti*, *S. Sreedharan**, S. Kambhampati, Y. Zhang. IJCAI 2017.

[45] **Plan Explicability and Predictability for Robot Task Planning**. Y. Zhang, *S. Sreedharan*, A. Kulkarni, T. Chakraborti, HH. Zhuo, S. Kambhampati. ICRA 2017.

[46] **Compliant Conditions for Polynomial Time Approximation of Operator Counts**. T. Chakraborti, *S. Sreedharan*, S. Sengupta, T. K. Satish, and S. Kambhampati. Symposium on Combinatorial Search (SOCS) 2016.

[47] **A Formal Analysis of Required Cooperation in Multi-agent Planning** . Y. Zhang, *S. Sreedharan* and S. Kambhampati. ICAPS 2016.

[48] **Capability Models and their application in Multi-agent planning.** . Y. Zhang, *S. Sreedharan* and S. Kambhampati. AAMAS 2015.

Workshops, Demos & Miscellaneous

[49] **Traversing the Linguistic Divide: Aligning Semantically Equivalent Fluents Through Model Refinement** K. Sykes, M. Fine-Morris, *S. Sreedharan*, M. Roberts. KEPS ICAPS 2024.

[50] **Reducing Human-Robot Goal State Divergence with Environment Design** K. Sykes, S. Keren, *S. Sreedharan.* HAXP ICAPS 2024.

[51] **Human-Modeling in Sequential Decision-Making: An Analysis through the Lens of Human-Aware A.** S. Tulli, S. Vasileiou, *S. Sreedharan.* HAXP 2024.

[52] **A Mental Model Based Theory of Trust.** Z. Zahedi, *S. Sreedharan*, S. Kambhampati. XAI IJCAI 2023.

[53] **A Mental-Model Centric Landscape of Human-AI Symbiosis.** Z. Zahedi, *S. Sreedharan*, S. Kambhampati. R2HCAI AAAI 2023.

[54] **Revisiting Value Alignment Through the Lens of Human-Aware AI.** *S. Sreedharan*, and S. Kambhampati. Virtual Workshop on Human-Centered AI, NeurIPS 2022.

[55] **Towards customizable reinforcement learning agents: Enabling preference specification through online vocabulary expansion.** U. Soni, *S. Sreedharan*, M. Verma, L. Guan, M. Marquez, and S. Kambhampati. Workshop on Human in the Loop Learning, NeurIPS 2022.

[56] **Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change)** K. Valmeekam*, A. Olmo*, *S. Sreedharan*, and S. Kambhampati

[57] **Why Did You Do That? Generalizing Causal Link Explanations to Fully Observable Non-Deterministic Planning Problems** *S. Sreedharan*, C. Muise, and S. Kambhampati ICAPS Workshop on Explainable AI Planning (XAIP) 2022

[58] **Leveraging PDDL to Make Inscrutable Agents Interpretable: A Case for Post Hoc Symbolic Explanations for Sequential-Decision Making Problems**. *S. Sreedharan*, and S. Kambhampati. ICAPS Workshop on Explainable Planning (XAIP), 2021.

[59] **Trust-Aware Planning: Modeling Trust Evolution in Longitudinal Human-Robot Interaction** . Z. Zahedi, M. Verma, *S. Sreedharan*, and S. Kambhampati. ICAPS Workshop on Explainable Planning (XAIP), 2021.

[60] **GPT3-to-plan: Extracting plans from text using GPT-3** . A. Olmo, *S. Sreedharan*, and S. Kambhampati. ICAPS Workshop on Planning for Financial Services (FinPlan), 2021.

[61] **A Bayesian Account of Measures of Interpretability in Human-AI Interaction**. *S. Sreedharan*, A. Kulkarni, T. Chakraborti, D. Smith, and S. Kambhampati. ICAPS Workshop on Explainable Planning (XAIP), 2020.

[62] **RADAR-X: An Interactive Interface Pairing Contrastive Explanations with Revised Plan Suggestions** . K. Valmeekam ,*S. Sreedharan*, S. Sengupta, and S. Kambhampati. ICAPS Workshop on Explainable Planning (XAIP), 2020.

[63] **Explainable Composition of Aggregated Assistants**. *S. Sreedharan*, T. Chakraborti, Y. Rizk, and Y. Khazaeni. ICAPS Workshop on Explainable Planning (XAIP), 2020.

[64] **A General Framework for Synthesizing and Executing Self-Explaining Plans for Human-AI Interaction**. *S. Sreedharan*, T. Chakraborti, C. Muise, and S. Kambhampati. ICAPS Workshop on Explainable Planning (XAIP), 2019.

[65] **Design for interpretability**. A. Kulkarni*, *S. Sreedharan**, S. Keren, T. Chakraborti, D. Smith, S. Kambhampati. ICAPS Workshop on Explainable Planning (XAIP), 2019.

[66] **Hierarchical Expertise-Level Modeling for User Specific Robot-Behavior Explanations.**. *S. Sreedharan*, M. Madhusoodanan, S. Srivastava, and S. Kambhampati. XAIP ICAPS 2018.

[67] **Balancing Explicability and Explanation in Human-Aware Planning.**. *S. Sreedharan**, T. Chakraborti*, and S. Kambhampati. AAAI Fall Symposium 2017.

[68] **Explanations as Model Reconciliation - A Mutli-Agent Perspective.**. *S. Sreedharan**, T. Chakraborti*, and S. Kambhampati. AAAI Fall Symposium 2017.

[69] **Alternative Modes of Interaction in Proximal Human-in-the-Loop Operation of Robots.**. T. Chakraborti, *S. Sreedharan*, A. Kulkarni, and S. Kambhampati. ICAPS 2017 UISP Workshop; and ICAPS 2017 System Demo and Exhibits.

[70] **Plan Explainability and Predictability for Robot Task Planning**. Y. Zhang; *S. Sreedharan*; A. Kulkarni;T. Chakraborti; H. H. Zhuo; and S. Kambhampati. In RSS 2016 Workshop on Planning for Human-Robot Interaction

Projects

- AERobots Framework - A framework for supporting coordination between humans and robots through structured communication mediums like AR/EEG. [website]
- Simoorg - Fault injection system to test highly scalable distributed systems. [github]