Research Statement

Suresh Kumaar Jayaraman
Postdoc, Robotics Institute, CMU
sureshkj@cmu.edu

The long-term goal of my research is to **develop autonomous systems that can seamlessly integrate into complex, human-centric environments**. With the increasing presence of autonomous systems, from virtual AI agents to fully embodied robots, these technologies are becoming integral to our daily lives. However, addressing the critical challenge of *mutual understanding*—where systems *learn to understand and be understood by the humans they interact with*—is essential to prevent errors, inefficiencies, and unsafe outcomes. My research addresses this critical challenge by developing methods that enable autonomous systems to *understand and model human behavior* while also ensuring that their own *decisions are intuitive and understandable to humans*. I envision building upon and bridging these two complementary thrusts of **human behavior modeling** and **explainable decision-making** to create autonomous systems that anticipate, adapt to, and align with human behaviors and expectations.

Real-world deployments of autonomous systems reveal critical gaps, highlighting the need for safety, trust, and human-centered alignment. For instance, an automated car must interpret a driver's gesture to merge—-gestures that vary due to individual preferences and situational contexts. Misinterpreting these cues risks confusion or collisions while failing to signal the system's intentions could affect trust. These challenges call for systems that effectively model human behavior and provide intuitive, transparent explanations of their decisions. Here are some of my contributions toward addressing these challenges.

Human behavior modeling: I have developed computational models that capture variations in human task-related states, such as actions and intentions, as well as latent mental states like trust, enabling autonomous systems to interpret and adapt to human behavior in real time. For instance, my models analyze pedestrian intents and paths to improve autonomous vehicle responses in shared spaces, ensuring safe navigation [1, 2]. I have also pioneered techniques to model drivers' trust dynamics in automated vehicles, providing real-time estimates that allow systems to adapt their behavior to maintain calibrated trust [3]. Additionally, my work examines how humans understand autonomous systems' decision-making [4]. I also explore group-level interactions, examining how motives, skill uncertainty, and communication shape trust and cooperation in human teams [11] that could inform future research for human-autonomy teams.

Explainable decision-making: To complement this understanding, I have designed algorithms that generate system behaviors that are both trustworthy and intuitive, ensuring autonomous systems act in ways humans can predict and trust [5, 6]. I have also developed methods for transparently communicating system capabilities and decision-making processes, fostering trust and collaboration in real-world human-centered environments [7, 8]. Together, these contributions enable autonomous systems to anticipate, adapt to, and align with human behaviors, promoting safer and more seamless interactions.

These contributions reflect my unique expertise in combining theory-grounded human behavior modeling with practical, explainable decision-making frameworks, positioning me to address the complex challenge of achieving mutual understanding between humans and autonomous systems. My works have been recognized through impactful publications, such as my papers on pedestrian-vehicle interactions [9], and trust estimation [3], and my paper that won the 3rd-best LBR at HRI 2018 [10]. It has also garnered media coverage in outlets such as The Conversation and Michigan Robotics News. By advancing research along my two primary thrusts of human behavior modeling and explainable autonomy, my work lays the foundation for bridging these areas by demonstrating that effective communication relies on an in-depth understanding of the human. This iterative, interactive communication process ensures that the systems I design are explainable and attuned to diverse humans' expectations.

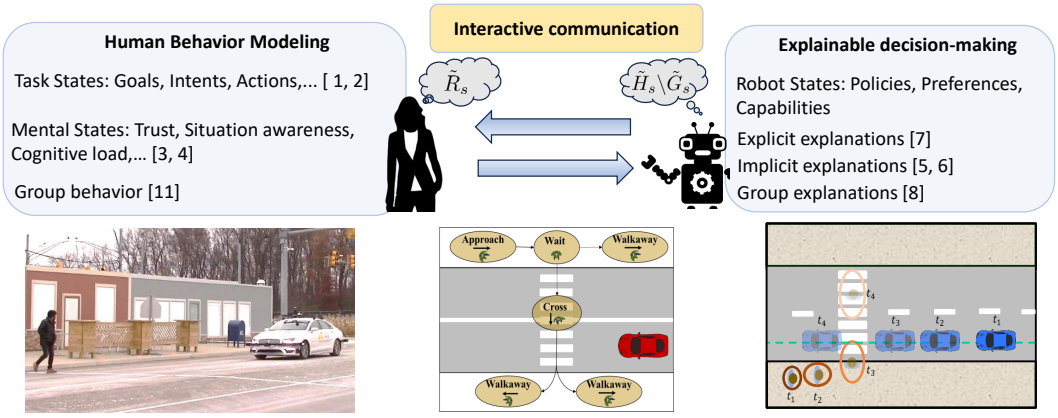

Fig 1. (Top) To enable successful interactive communication between humans and autonomous systems, my research has two primary thrusts: modeling human behavior, including task and mental state information (left), and creating explainable decision-making frameworks for robotic systems (right). (Bottom) My work applied to pedestrian-automated vehicle interactions, using pedestrian behavior and trust models to develop explainable, behavior-aware algorithms for safer interaction with vehicles.

Examining and Modeling Human Behavior in Interactive Situations

The first thrust of my research focuses on understanding and modeling the complexity of human behavior in interactions with autonomous systems. Human behavior is highly variable, influenced by individual characteristics and situational context. For example, my work shows that pedestrians are more cautious crossing in front of automated vehicles than human-driven ones, even under similar conditions [9]. This variability poses challenges for systems relying on data and machine learning to model behavior. To address this, I develop structured models that capture human task states (e.g., intents, actions) and mental states (e.g., trust, cognitive load, situation awareness) in diverse contexts. Key contributions include models of pedestrian intent and paths during interactions with AVs [1, 2] and real-time estimation of drivers' trust in AVs [3].

Intent and Multimodal Behavior Modeling: Intent modeling is crucial for enabling autonomous systems to anticipate human actions in dynamic environments. My research develops models that identify subtle or implicit intent cues such as shifts in position or gaze direction to predict behaviors that are not immediately observable. For example, pedestrian intent at crosswalks can vary dynamically, where subtle actions, such as shifts in position or gaze direction, signal an intention to cross or wait. My intent and behavior prediction framework [1] uses probabilistic machine learning to estimate pedestrian intent in real time, enabling automated vehicles (AVs) to make informed adjustments to speed and trajectory. Building on intent estimation, my work incorporates the variability of human behavior through a structured hybrid systems model. The Multimodal Hybrid Pedestrian (MHP) model [2] maps decision-making points and potential pedestrian paths using a probabilistic automaton, allowing AVs to anticipate multiple outcomes and rank their likelihoods based on context. By explicitly modeling crossing intent, the MHP model achieves more accurate and less conservative predictions compared to traditional approaches.

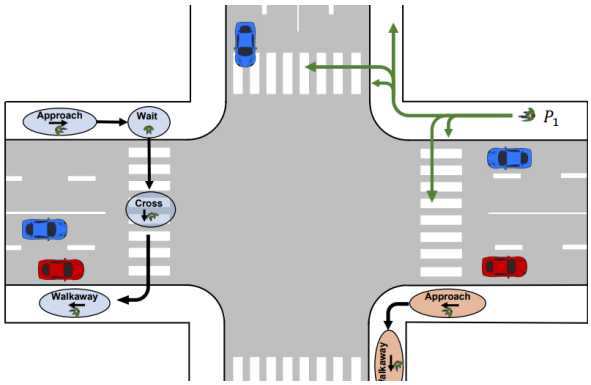

Fig 2. The figure shows how the multimodal pedestrian model, a hybrid automaton-based model, predicts pedestrian behavior at an intersection with varied intents — intending to cross or not, choosing to wait or not, etc.

Mental State Modeling: In addition to intent, the mental states of humans such as trust, cognitive load, situational awareness, risk perception, etc., are crucial because they directly impact user engagement, safety, and overall system effectiveness. For instance, without calibrated trust, users may disengage prematurely or over-rely on automation, leading to unsafe situations. To address these challenges, I developed a real-time trust estimation model to continuously monitor human trust in the system [3]. I created a dynamic trust estimator using a Kalman filter-based approach, integrating sensor data such as gaze tracking, interaction duration, driver task performance, system performance, and situational context. The results showed that the estimator accurately tracked trust levels compared to self-reported trust levels over successive interactions, improving in accuracy as the system observed more user behaviors.

Group behavior: Beyond individual behaviors, my research extends to understanding group-level interactions in team settings. In recent work, I investigated team cooperation dynamics in an online social dilemma game with uncertainty, focusing on how divergent motives, skill uncertainty, and communication influence trust, alignment, and cooperative strategies [11]. This work revealed that teams with proactive communication and a reduced uncertainty about skill and strategy alignment demonstrated higher levels of cooperation and trust. It highlighted the interplay between task performance, perceptions of competence, and resource allocation, offering valuable insights into how team dynamics evolve under mixed incentives. Human task and mental state modeling significantly enhance the adaptability and safety of autonomous systems by allowing them to respond intelligently to dynamic human actions and mental states [5, 7].

Enhancing Explainability for Fluent Human-Autonomy Interaction

The second thrust of my research focuses on adaptive explainability, aimed at enhancing human understanding of autonomous system behavior and decision-making. Explainability is essential for making autonomous systems not only safe and reliable but also interpretable and trustworthy. By tailoring explainability to users' mental states and knowledge levels, my work enables systems to provide timely, understandable insights into their decision making, fostering safer, more intuitive human-autonomy interactions.

My explainability framework integrates three modes: (1) explicit explanations using verbal cues to align system actions with user understanding, (2) implicit explanations through observable behavior to build intuitive comprehension, and (3) interactive feedback to assess and refine users' knowledge. Together, these modes could enable a

system to adapt explanations both reactively (to address immediate user needs) and proactively (to prepare users in advance), supporting a comprehensive, user-centered approach to human-autonomy interaction, particularly in cooperative and collaborative settings.

Explicit Explainability based on Human Mental States: During real-time interactions, autonomous systems can enhance safety and trust by adapting their explanations in response to users' fluctuating mental states, such as levels of trust, attentiveness, or perceived risk. My research developed a trust calibration framework that monitors real-time trust based on the previously discussed trust estimation model and delivers tailored verbal cues to calibrate trust dynamically [7]. Based on the system's knowledge of its capabilities and the estimated trust, it provides warnings to refocus attention if overtrust is detected or encouraging messages to boost confidence when undertrust is observed. This reactive approach resulted in a reduction of trust miscalibration from 70% 43.9%, promoting safer and more effective use of autonomy.

Implicit Explainability: Implicit explainability aims to align users' understanding of a robot's capabilities and limitations with its actual functions through observation of system behavior. A relevant example of implicit explainability is my Behavior-aware Model Predictive Controller (B-MPC) for automated vehicles (AVs). This controller anticipates pedestrian behavior using a probabilistic hybrid systems model, to generate vehicle trajectories [5]. By balancing safety, performance, and comfort, the B-MPC communicates the AV's behavior to pedestrians through its driving behavior characterized by stopping distance, acceleration/deceleration, and distance to pedestrian, that are perceived to be safe and trustworthy [6].

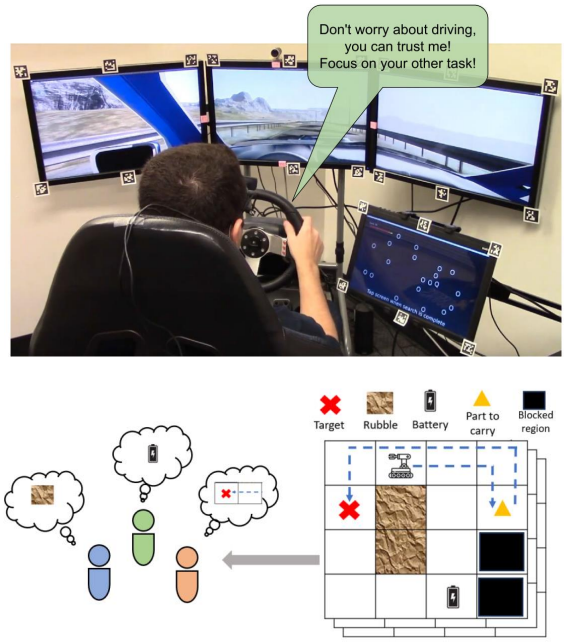

Fig 3. (Top) In task explicit verbal explanations indicating capability of autonomous system to calibrate driver trust. (Bottom) Pre-task proactive implicit robot policies to convey the robot's decision-making to a diverse learning group.

Implicit Explainability in groups: Autonomous systems will not operate in isolation, but in group settings, such as team-oriented environments like emergency response, hospital service, etc., where multiple individuals rely on the same system. In these situations, they will likely need to communicate with different and multiple human partners/operators and thus effective group interaction is crucial, as misunderstandings among team members can jeopardize safety and efficiency. Tailored approaches are therefore essential to establish a shared understanding of the system's decision-making processes. My research addresses the complex challenges of group explainability, i.e., aligning diverse knowledge levels, learning styles, and interpretations within a team, by using group machine teaching to communicate the system's decision-making process through demonstrations of its behavior [8]. My approach models the beliefs of each team member about the robot policy using particle filters. By aggregating individual beliefs into collective belief representations, i.e., a common belief (knowledge shared by all members) or a joint belief (knowledge held by at least one member), the system dynamically selects demonstrations that align with the collective understanding of the group.

Future Research

While my research in the past has taken significant strides toward the two thrusts, there are still many unsolved problems. As autonomous systems are becoming increasingly relevant in collaborative tasks alongside humans to achieve shared objectives, addressing the alignment challenges between humans and autonomy requires a more holistic approach. Here, I briefly discuss the future research directions extended from my previously discussed research thrusts and how they can be bridged to address human-autonomy alignment in dyadic and group settings.

Modeling Evolving Human Behavior: Understanding how human behavior evolves over extended interactions with autonomous systems is a critical and underexplored challenge. Long-term interactions require systems to account for users' changing familiarity, changing preferences, shifting goals, and evolving levels of trust. For example, a hospital robot delivering supplies might initially be closely monitored by nurses as they are uncertain of its capabilities. As trust builds, they may assign more critical tasks to the robot. However, occassional failures as erode the trust and rebuilding this trust takes time, reducing its utility. Capturing these dynamics requires a nuanced understanding of how user trust and engagement fluctuate over time and in response to system behavior. Addressing this problem demands the development of adaptive frameworks that can detect and respond to behavioral drift—subtle changes in user attitudes, preferences, or actions. One approach is to create models that leverage real-time feedback from user interactions, such as task completion times, verbal cues, or physiological data like gaze

patterns. These models could be augmented with meta-learning techniques, allowing the system to quickly adapt to individual users while generalizing to new ones.

In a group setting, this problem is further complicated by evolving group dynamics and emergent behaviors due to changing group composition, conflicting objectives, interpersonal influence, etc., in addition to evolving individual behaviors. Modeling group-level dynamics requires frameworks that capture both individual behaviors and their influence on the group as a cohesive unit. One promising approach is to develop models that combine insights from individual behaviors with an understanding of collective group interactions. These models would account for shared group states, such as trust and decision-making, while also capturing the influence of individual actions on group outcomes. By incorporating a temporal perspective, such models can track how group dynamics evolve over time, adapting to changes in leadership, roles, or trust.

Expanding adaptive explainability: Building on my prior work in tailoring explanations to users, future research could extend this personalization to include a broader range of user-specific factors, such as goals, emotional states, and contextual priorities. For example, a navigation assistant might provide a concise route summary to a commuter focused on reaching a meeting quickly, while offering more scenic, detailed directions to a tourist exploring an unfamiliar city. These additional dimensions of personalization can significantly enhance the relevance and intuitiveness of system explanations, fostering a more engaging and supportive human-autonomy interaction.

Furthermore, in dynamic environments, where systems make decisions under time-sensitive and uncertain conditions, explanations must not only adapt to user profiles but also align with the immediate operational context. For instance, an autonomous drone delivering packages may need to explain a sudden route deviation to avoid a weather hazard, tailoring its explanation to the urgency and user's understanding of logistics. One approach is to integrate context-aware user models with real-time decision data to ensure explanations remain timely, relevant, and minimally intrusive. By combining insights from user modeling, feedback loops, and incremental explanation techniques, this research aims to make explanations more adaptive, accessible, and effective across diverse and dynamic scenarios.

Interactive Human-autonomy Value Alignment in dyadic and group settings: One of the critical challenges in human-autonomy interaction is ensuring that an autonomous system's actions align with human values and preferences, even as these evolve over time. Traditional alignment methods assume that users can fully specify their objectives, but in practice, users often lack a complete understanding of the system's capabilities and constraints. For example, a cooking robot might encounter differing priorities such as allergies and dietary restrictions, ingredient substitutions, presentation aesthetics, or even sustainability concerns, some of which might evolve over time as users try different diets and lifestyle practices. If the robot is not aligned with these changing priorities, it could create inefficiencies, delays, or even safety risks. My research aims to address this by developing interactive frameworks that identify users' values, assess their compatibility with system constraints, and negotiate adjustments through back-and-forth communication. As user preferences, situational contexts, and system capabilities evolve, the system must iteratively refine its behavior models to better interpret and anticipate user actions. Simultaneously, it must refine communication strategies to provide clear, context-aware explanations that help users understand its actions and limitations. This iterative approach fosters a shared understanding and enables the system to balance user preferences with feasible actions.

Building on this foundation, the alignment framework can be extended to group settings, where autonomous systems interact with multiple users simultaneously. Group alignment introduces additional challenges, such as conflicting preferences, diverse knowledge levels, and changing group dynamics. Extending the bi-directional approach, the system could aggregate individual preferences into a unified group profile while accounting for variations across members. By leveraging group-oriented communication strategies, the system can foster a shared understanding of its actions and intent, resolving conflicts and aligning group goals. These advancements will support seamless collaboration in diverse scenarios, from healthcare to multi-robot coordination in warehouses and manufacturing floors.

Funding and Collaboration

I am thankful to several federal and private funding sources, such as NSF, US Army GVSC, DARPA, ONR, Toyota Research Institute, and BMW, for funding some of my research efforts during my PhD and postdoctoral research. In addition to the above grant agencies, my research aligns well with the interests of other grant agencies like ARL, and industrial research groups like Google Research, Meta AI Research, and Amazon Robotics. I am hoping to secure funding from these sources and specific programs such as NSF M3X to continuously scale my research. Collaborating with professionals working in human-centric environments, such as nurses, warehouse workers, and emergency responders, alongside experts in fields like affective computing, cognitive science, human factors, multi-agent systems, and control theory, could expand my research into areas such as elder care, public safety, and collaborative robotics, supporting interdisciplinary advancements in safe, reliable, and human-aligned autonomous systems.

Publications

[1] **S. K. Jayaraman**, D. M. Tilbury, X. J. Yang, A. K. Pradhan, and L. P. Robert, "Analysis and prediction of pedestrian crosswalk behavior during automated vehicle interactions," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6426–6432, IEEE, 2020.

[2] **S. K. Jayaraman**, L. P. Robert, X. J. Yang, and D. M. Tilbury, "Multimodal hybrid pedestrian: A hybrid automaton model of urban pedestrian behavior for automated driving applications," *IEEE Access*, vol. 9, pp. 27708–27722, 2021.

[3] H. Azevedo-Sa, **S. K. Jayaraman**, C. T. Esterwood, X. J. Yang, L. P. Robert Jr, and D. M. Tilbury, "Real-time estimation of drivers' trust in automated driving systems," *International Journal of Social Robotics*, vol. 13, no. 8, pp. 1911–1927, 2021.

[4] **S. K. Jayaraman**, A. Steinfeld, R. Simmons, and H. Admoni, "Modeling human learning of demonstration-based explanations for user-centric explainable AI," in *Presented at the Explainability for Human-Robot Collaboration workshop at ACM/IEEE Conference on Human-Robot Interaction*, 2024.

[5] **S. K. Jayaraman**, L. P. Robert, X. J. Yang, A. K. Pradhan, and D. M. Tilbury, "Efficient behavior-aware control of automated vehicles at crosswalks using minimal information pedestrian prediction model," in *2020 American Control Conference (ACC)*, pp. 4362–4368, IEEE, 2020.

[6] **S. K. Jayaraman**, L. Robert, X. J. Yang, and D. Tilbury, "Automated vehicle behavior design for pedestrian interactions at unsignalized crosswalks," in *2021 International Symposium on Transportation Data and Modelling (ISTDM)*, 2021.

[7] H. Azevedo-Sa, **S. K. Jayaraman**, X. J. Yang, L. P. Robert, and D. M. Tilbury, "Context-adaptive management of drivers' trust in automated vehicles," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6908–6915, 2020.

[8] **S. K. Jayaraman**, R. Simmons, A. Steinfeld, and H. Admoni, "Understanding robot minds: Leveraging machine teaching for transparent human-robot collaboration across diverse groups," in *IEEE International Conference on Intelligent Robots and Systems*, 2024.

[9] **S. K. Jayaraman**, C. Creech, D. M. Tilbury, X. J. Yang, A. K. Pradhan, K. M. Tsui, and L. P. Robert Jr, "Pedestrian trust in automated vehicles: Role of traffic signal and av driving behavior," *Frontiers in Robotics and AI*, vol. 6, p. 117, 2019.

[10] **S. K. Jayaraman**, C. Creech, L. P. Robert Jr, D. M. Tilbury, X. J. Yang, A. K. Pradhan, and K. M. Tsui, "Trust in av: An uncertainty reduction model of av-pedestrian interactions," in *Companion of the 2018 ACM/IEEE international conference on human-robot interaction*, pp. 133–134, 2018.

[11] "Anonymized for review," *Submitted to the ACM SIGCHI Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2025.