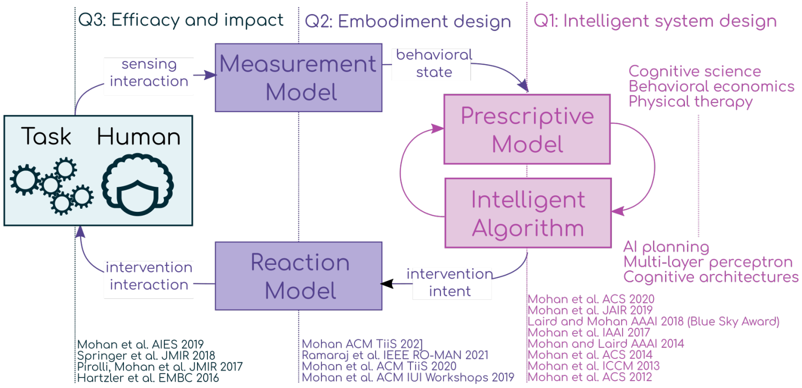
Natural, effective **collaboration between intelligent systems and humans** is the next frontier in artificial intelligence (AI) and machine learning (ML) research (NSCAI). Advances made on this agenda are crucial for technological solutions for a variety of **social good challenges** including improving health outcomes (NSF/NIH); supporting sustainable lifestyles (ARPA-E), and advancing human learning (NSF, DARPA). While a promising field, the science of human-AI collaboration is in its infancy due to the **dearth of human-centric methods** that enable intelligent systems to reason about their human partners. In my work, I explore an **interdisciplinary approach to designing human-AI collaborative systems**; leveraging insights from social sciences - cognitive science, psycholinguistics, economics - to structure sequential decision-making in complex intelligent agents. My research has been published at diverse venues including those for AI [1, 2, 3, 4], human cognition [5, 6], cognitive systems [7, 8, 9], AI and Society [10], human-computer interaction (HCI) [11, 12], human-robot interaction (HRI) [13], and medical informatics & engineering [14, 15, 16]. My work has been supported by both the government - ARPA-E, AFOSR, DARPA, NSF/NIH - and commercial - Xerox, Kaiser Permanente - entities.

1 Research Context, Vision, and Experience

Recent advances in AI & ML have mostly been driven through computational breakthroughs. As tremendous strides are made on developing efficient algorithms which can process voluminous data, it is valuable to study how humans and machines can collaborate to address complex societal challenges. To be effective collaborators, intelligent systems must be imparted with capabilities to model, reason, and learn about their human partner. However, the science of modeling other agents in AI systems is very limited in its scope [17]. Largely, human modeling research in AI originates in game domains where humans are modeled as opponents with fixed objective functions. These models are not sufficient for representing dynamic and evolving human behavior, decision making, and learning. There is a dearth of human model-based algorithmic methods that can be employed in developing intelligent solutions [18].



I seek to address this gap by bringing insights from social sciences into AI systems research, effectively putting reasoning about humans at the center of **end-to-end system development**. I take a three prong approach to answer several key questions about achieving a theoretic and systems-level integration of social sciences and AI & ML systems. Along the first prong, Q1, I explore methods for **human-centric decision making**. Some theories about human behavior are based on abstract constructs and are not computational (e.g, goal setting [19]). Several others that are computational (e.g; choice modeling [20]) are *descriptive*, i.e, they summarize observed human behavior. To be useful in AI systems, models must be *computational* and *prescriptive* i.e, they should characterize how human decisions and behavior evolve with situational and informational changes enabling evaluation of decisions. My research builds upon the abstract or descriptive theories to build prescriptive models that support reasoning in AI & ML systems. Along Q2, I embody decision making methods in interfaces (mobile applications, conversational user interfaces) to support natural **collaborative HCI & HRI**. Finally, I study a various real-world problems where reasoning about human partners is crucial for success. Identifying the right set of metrics and evaluation paradigms is critical to ensure relevant progress. I relinquish the computation-centric metrics (e.g, accuracy, efficiency) and **adopt human-centric metrics** (flexibility, safety, acceptability) and **experimental methods from social sciences**, advancing the final prong Q3.

Interactive Task Learning My doctoral research as well as the ongoing effort on DARPA GAILA study how to design intelligent agents so that they can dynamically extend their domain knowledge and skills through natural interactions with human collaborators. The capability of learning new domain concepts and task knowledge online, post deployment, is critical to the adoption of complex intelligent agents such as general-purpose robots. At the University of Michigan, I led the development of Rosie, a cognitive robotic framework for interactive learning. It is built with Soar [21] and was the first in the literature to demonstrate interactive learning of a variety of concepts in a **single, integrated agent architecture**. We introduced a new paradigm for learning domain concepts [9] and task knowledge [2] from mixed-initiative, task-oriented, dialog [6]. Our research has **led to an emerging, inter-disciplinary, scientific enquiry** on Interactive Task Learning (ITL [22]).

Natural, flexible interaction with human trainers is a core capability of any ITL system. To this end, we leveraged an abstract theory from psycholinguistics - the Indexical Hypothesis [23] - to develop a computational model of comprehension [8] for complex agent architectures. This model grounds language semantics by using non-linguistic contexts (cognitive, attentional, and task-oriented) to generate and disambiguate meaning representations. Currently, under DARPA GAILA, I am leading work that extends this approach [7] to learn language semantics through cognitive models of analogical reasoning and generalization [24]. A core commitment of our ITL effort is end-to-end behavior which motivates research on how different aspects of AI including perception and actuation, human-agent interaction, language processing, and learning interact and influence each other. With John Laird, I won the 2018 **Blue Sky award** [4] for proposing a framework to integrate low-level learning computations with high-level reasoning strategies.

Coaching Agents for Health Behavior Change Health behaviors - exercise and nutrition - account for an estimated 60% of the risks associated with chronic illnesses such as diabetes and cardiovascular disease. As the at-risk population, the challenge of developing and disseminating effective methods for improving health behaviors is becoming important. Intelligent adaptive systems present a unique opportunity in supporting healthy behaviors at scale. My research led to one of **first demonstrations of long-term interactive, adaptive behavior** that was evaluated with human participants in **ecological settings** [11, 12, 14, 15].

Under the NSF/NIH Smart and Connected Health program, we developed long-living, interactive, coaching agents that helped people pursue relevant exercise and nutritional goals and develop healthy behaviors. In Mohan et al. [1, 11], we proposed a computational and adaptive formulation of the abstract goal setting theory from behavioral psychology [19]. We designed an interactive coaching agent that was deployed through a mobile application and helped people with walking exercises. The agent used a parameterized, prescriptive model of growth in aerobic capability using principles of clinical practice in conjunction with AI heuristics-based scheduling methods. Through the mobile interface, the coaching agent assessed a human trainee's current exercise capability, assigned exercise goals, and revised them based on the trainee's performance. We proposed a novel evaluation paradigm for long-term intelligent interactive agents [11, 16]. We engaged with domain experts to determine if the agent's coaching strategy aligned with theirs along the dimensions of safety, acceptability, and likelihood of successful completion. We, then, studied if the coaching agent could promote safe and effective behavior change by deploying it for 6 weeks in a clinically relevant population of 21 people.

Extending these ideas further, Mohan [12] leverages the Common Model of Cognition [21] as an integrative framework for explaining several behavior change theories from psychology and uses it to provide design recommendations for interactive coaching systems. This line of research has been published at venues for medical informatics [16, 14, 15] in addition to being highlighted as a **key technical advancement on the roadmap to robust interactive intelligence** [25] at an NSF workshop.

Influencing Individual Behavior for Sustainable Transportation Transportation is one of the largest consumers of energy in the world - in the United States, it accounted for 29% of energy consumption in 2016. However, recent introduction of new transit services - bike/scooter sharing, carpools, car sharing etc. in addition to public transit - has created several opportunities for reducing energy consumption. ARPA-E TransNet aimed at developing solutions that incentivize people to adopt more sustainable modes of transit, greatly reducing the energy consumption in personal transportation. To this end, my research demonstrates how various theoretical insights and methods from **human-factors research, AI, economics, and transportation** can be brought together in a **single comprehensive system** for an effective human-AI collaborative solution.

In Mohan et al. [3, 10], we introduced the traveller influence problem for sustainable transportation planning and recommendation. We began with identifying what factors underlie people's transit-related choices through a set of semi-structured interviews and survey studies. Then, we leveraged descriptive models in economics on rational choice and preference [20] to develop a prescriptive model for traveler mode adoption. We demonstrated that ML approaches can be used to estimate the model parameters with a large, publicly available dataset. Next, we validated our proposed traveler mode adoption model through stated preference methods from behavioral economics. We demonstrated that mode adoption model can be used to bias plan selection an AI multi-modal planning framework to generate personalized, energy-efficient plans for each individual traveller. Finally, we engaged with transportation modeling research to simulate the impact of applying our approach in **Los Angeles demonstrating that it could achieve small but significant energy and time savings**.

2 Future Directions

I am interested in designing interactive AI systems that can collaborate with humans in a variety of roles. There are two fundamental scientific advances that are critical to achieving this goal: first, developing a set of **computational & prescriptive models of human decision making**, behavior, and learning that can be integrated with AI & ML algorithms; and second, developing **complex agent architectures** that support collaborative, task-oriented, long-term behavior. My background in designing complex AI systems in addition to experience in modeling human behavior has built a strong foundation for me to successfully develop this agenda upon.

Modeling Humans in Collaborative AI Systems Advances in AI have been enabled by the computational modeling of our physical world. It would have been impossible to develop computational algorithms that exploit this knowledge without languages (mathematical, qualitative, and quantitative) that describe how our physical world changes and evolves. Along similar lines, effective and generalizable human-agent collaborative solutions need explicit, causal **models of how human behavior and learning adapts** in evolving environments. Competent health behavior coaching agents must diagnose a trainee's behavior performance to identify their individual challenges and adapt their coaching strategy to suit each trainee's needs. Similarly, an effective ITL system should be able to exploit the full range of information in varying human teaching strategies. My research will contribute hybrid modeling methods for human behavior and will evaluate their efficacy by integrating them in AI systems. These methods will use the theoretical understanding of human behavior (such as the Common Model of Cognition [26] or the Belief-Desire-Intention framework [27]) to provide rule-based structural scaffolds upon which quantitative information can be overlaid. While the **structural scaffolds ensure explicability and diagnosability** of the models, the **quantitative information reflects the stochasticity** arising from individual variability as well as non-modeled factors. Models will be causal and prescriptive laying out the space of future outcomes and enable evaluation of those outcomes in AI systems. In addition to the usecases I discussed previously, these models can be integrated with agent-based modeling frameworks to greatly enhance counter-factual analyses of complex social systems (e.g, public policy analysis [28]).

Complex Agent Architectures for Collaboration My research will advance the **science of complex agent architectures**. Over several decades, AI & ML research has produced a variety of computational methods that capture some aspect of intelligence. Often they have complimentary strengths and weaknesses. I will investigate how different computational methods can be brought together in a single comprehensive hybrid AI system that has robust, complex, collaborative behavior. I will approach this question from **two perspectives: problem-driven and cognitive**. I am leading work on DARPA SAIL-ON to design intelligent systems that are robust to novelties introduced during deployment. Continuing this research thread, I will identify various computational subproblems that underlie complex behavior and develop novel hybrid AI systems. Through the second perspective, I will contribute to the **legacy of cognitive architectures** such as Soar [21] and Companions [29] that are the best examples of integration of multiple learning, memory, and reasoning algorithms.

I am motivated by **learning from social interactions** - one of the most fundamental forms of learning in the human society. Parents, teachers, experts enable effective and efficient learning in children, students, and novices. In these interactions, the facilitator trainer and the primary learner form a joint system, with the former helping the latter in achieving critical conditions of learning. I would like to study the human-agent collaborative learning dyad from a variety of perspectives. Continuing my ongoing research, I will develop **cognitive agents that can learn** novel domain concepts and task knowledge through natural interaction. Additionally, I would like to design teaching agents that can **support humans in learning** novel physical, procedural tasks such as repairing a machine or assembling a new artifact [30] using augmented reality embodiment. New HCI modalities have opened up exciting avenues and I will develop AI systems that use these modalities for novel human-AI collaborative behavior.

The recent successes of AI and ML are now accompanied with an ever increasing expectation of using those methods to **support human goals** in a variety of contexts. **Studying AI & ML algorithms in isolation**will not lead us to effective solutions for these challenging problems. Humans are key decision-makers in these ecosystems. An effective intelligent solution requires an **inter-disciplinary human-centric** approach that requires understanding how humans make decisions, what their goals and preferences are, and how to support their progress on their goals. My research will **advance our understanding of the human-AI ecosystems** towards developing effective collaborative intelligent systems.

References

[1] S. Mohan, A. Venkatakrishnan, M. Silva, and P. Pirolli. "On Designing a Social Coach to Promote Regular Aerobic Exercise". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[2] S. Mohan and J. Laird. "Learning Goal-Oriented Hierarchical Tasks from Situated Interactive Instruction". In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014.

[3] S. Mohan, H. Rakha, and M. Klenk. "Acceptable planning: Influencing Individual Behavior to Reduce Transportation Energy Expenditure of a City". In: *Journal of Artificial Intelligence Research* 66 (2019), pp. 555–587.

[4] J. Laird and S. Mohan. "Learning Fast and Slow: Levels of Learning in General Autonomous Intelligent Agents." In: *Proeedings of the AAAI Conference on Artificial Intelligence*. 2018.

[5] J. Laird and S. Mohan. "A Case Study of Knowledge Integration Across Multiple Memories in Soar". In: *Biologically Inspired Cognitive Architectures* (2014).

[6] S. Mohan, J. Kirk, and J. Laird. "A Computational Model for Situated Task Learning with Interactive Instruction". In: *Proceedings of the 2013 International Conference on Cognitive Modeling*. 2013.

[7] S. Mohan, M. Klenk, M. Shreve, K. Evans, and J. Maxwell. "Characterizing an Analogical Concept Memory for Architectures Implementing the Common Model of Cognition". In: *Annual Conference on Advances in Cognitive Systems*. 2020.

[8] S. Mohan, A. Mininger, and J. Laird. "Towards an Indexical Model of Situated Language Comprehension for Cognitive Agents in Physical Worlds". In: *Advances in Cognitive Systems* 3 (2016).

[9] S. Mohan, A. Mininger, J. Kirk, and J. Laird. "Acquiring Grounded Representations of Words with Situated Interactive Instruction". In: *Advances in Cognitive Systems* (2012).

[10] S. Mohan, F. Yan, V. Bellotti, H. Rakha, and M. Klenk. "On Influencing Individual Behavior for Reducing Transportation Energy Expenditure in a Large Population". In: *AAAI/ACM Conference on AI, Ethics, and Society*. 2019.

[11] S. Mohan, A. Venkatakrishnan, and A. Hartzler. "Designing an AI Health Coach and Studying its Utility in Promoting Regular Aerobic Exercise". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* (2020).

[12] S. Mohan. "Exploring the Role of Common Model of Cognition in Designing Adaptive Coaching Interactions for Health Behavior Change". In: *ACM Transactions on Interactive Intelligent Systems (preprint)* (2020).

[13] P. Ramaraj, C. Ortiz Jr., and S. Mohan. "Unpacking Human Teachers' Intentions For Natural Interactive Task Learning". In: *International Symposium on Robot and Human Interactive Communication (RO-MAN 2021)*. 2021.

[14] A. Springer, A. Venkatakrishnan, S. Mohan, L. Nelson, M. Silva, and P. Pirolli. "Leveraging Self-Affirmation to Improve Behavior Change: a Mobile Health App Experiment". In: *JMIR mHealth and uHealth* (2018).

[15] P. Pirolli, S. Mohan, A. Venkatakrishnan, L. Nelson, M. Silva, and A. Springer. "Implementation Intention and Reminder Effects on Behavior Change in a Mobile Health System: A Predictive Cognitive Model". In: *JMIR* (2017).

[16] A. Hartzler, A. Venkatakrishnan, S. Mohan, M. Silva, P. Lozano, J. D. Ralston, E. Ludman, D. Rosenberg, K. M. Newton, L. Nelson, and P. Pirolli. "Acceptability of a team-based mobile health (mHealth) application for lifestyle self-management in individuals with chronic illnesses". In: *Conference proceedings: 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2016.

[17] S. V. Albrecht and P. Stone. "Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems". In: *Artificial Intelligence* (2018).

[18] S. Kambhampati. "Challenges of Human-Aware AI Systems". In: *AAAI 2018 Presidential Address* (2019).

[19] M. K. Shilts, M. Horowitz, and M. S. Townsend. "Goal Setting as a Strategy for Dietary and Physical Activity Behavior Change: A Review of the Literature". In: *American Journal of Health Promotion* (2004).

[20] T. Domencich and D. McFadden. "Urban Travel Demand - A Behavioral Analysis". In: *Transport Research Laboratory* (1975).

[21] J. E. Laird. *The Soar Cognitive Architecture*. MIT press, 2012.

[22] J. E. Laird, K. Gluck, J. Anderson, K. D. Forbus, O. C. Jenkins, C. Lebiere, D. Salvucci, M. Scheutz, A. Thomaz, G. Trafton, R. Wray, S. Mohan, and J. Kirk. "Interactive Task Learning". In: *IEEE Intelligent Systems* 32.4 (2017), pp. 6–21.

[23] A. M. Glenberg and D. A. Robertson. "Indexical Understanding of Instructions". In: *Discourse Processes* (1999).

[24] K. D. Forbus, D. Gentner, and K. Law. "MAC/FAC: A Model of Similarity-Based Retrieval". In: *Cognitive Science* (1995).

[25] J. Oakley. "Intelligent Cognitive Assistants (ICA) NSF Workshop Summary and Research Needs". In: *Semiconductor Research Corporation* (2018).

[26] J. Laird, C. Lebiere, and P. S. Rosenbloom. "A Standard Model of the Mind: Toward a Common Computational Framework Across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics". In: *AI Magazine* (2017).

[27] M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge. "The Belief-Desire-Intention Model of Agency". In: *International Workshop on Agent Theories, Architectures, and Languages*. Springer.

[28] M. Tracy, M. Cerdá, and K. M. Keyes. "Agent-Based Modeling in Public Health: Current Applications and Future Directions". In: *Annual Review of Public Health* (2018).

[29] K. Forbus and T. Hinrichs. "Companion Cognitive Systems: A Step Toward Human-Level AI". In: *AI Magazine* (2006).

[30] S. Mohan, K. Ramea, B. Price, M. Shreve, H. Eldardiry, and L. Nelson. "Building Jarvis-A Learner-Aware Conversational Trainer". In: *In 2019 ACM IUI Workshops*. 2019.