Natural, effective collaboration between intelligent systems and humans on complex tasks is the next frontier in artificial intelligence (AI) and machine learning (ML) research [1]. This advance is critical for deployment of AI & ML solutions in various domains including improving public health outcomes, supporting sustainable lifestyles, advancing human learning, increasing scientific productivity, etc. Generative AI systems have enabled human-machine interaction using natural modalities such as language, diagrams, images etc. Building upon this breakthrough, for effective human-AI collaboration, intelligent machines must be imparted with capabilities that enable them understand the needs of their human partners and provide timely, useful support. With this aim, I design intelligent agents that collaborate with humans in various roles. My work is interdisciplinary and has been published in venues for AI [2, 3, 4, 5, 6], HCI/HRI [7, 8, 9], AI & society [10, 11, 12, 13], and human cognition & cognitive systems [14, 15, 16, 17, 18]. It has been supported by DARPA, ARPA-E, AFOSR, and NSF/NIH.

1 Research Context, Vision, and Experience

Strides made in AI and ML in the last decade have been fueled by computational and algorithmic breakthroughs. While these advances have been exciting as an AI researcher, the general public, civil society organizations, and government entities have sounded alarms about the negative impact AI technology can have on our communities. It is crucial for an AI scientist and for AI organizations to pursue research agendas where we study how AI and ML technology can support humans in our endeavors and enhance human lives. Such an agenda motivates leveraging insights from social sciences - particularly about how humans think, learn, and collaborate - to guide the design of intelligent computational systems. As an AI scientist, I want to advance a *human-aware* approach to AI [19] that centers human experiences in the design of intelligent systems.

I am interested in collaborative agents that can model, reason, and learn about their human collaborators in addition to modeling, reasoning, and learning about the world or tasks. I take a three-pronged approach towards this goal. I study real world problems where reasoning about human partners is crucial for success. Along the first prong, I build agent architectures for end-to-end complex behavior achieved through orchestration of multiple intelligent capabilities including computer vision, goal reasoning, AI planning, natural language processing, control etc. Along the second, I embed agents in interfaces (mobile applications, conversational interfaces) and embodiments to study the principles of natural, collaborative human-AI interaction. For system evaluation, I relinquish the computation-centric metrics (e.g, accuracy, efficiency) and adopt human-centered metrics (flexibility, safety, acceptability) and experimental methods from social sciences, advancing the third prong.

Intelligent Agents with Complex Behavior My ongoing research studies the design of intelligent agents that can adapt to an evolving world and changing task requirements autonomously, and with human teaching. Under this theme, I led research as a principal investigator on Open-World Learning (OWL) under DARPA SAIL-ON and on Interactive Task Learning (ITL) under DARPA GAILA as well as during my graduate training. The agent architectures I developed for these efforts commit to the idea complex intelligent behavior results from an interplay of diverse reasoning and learning methods.

OWL studies how agents can operate in an evolving environment that violates the closed-world assumption common in AI/ML agent design. We introduce the idea of a *novelty* [2] - a meaningful change in the operational characteristics of the environment (e.g., change in gravity) that occurs after an agent has been deployed. Model-free agent architectures, such as deep reinforcement learning, experience catastrophic failures when novelties are introduced and need to be retrained from scratch. Our agent architecture [2, 3] builds upon model-based reasoning methods (including planning) and can elegantly handle novelties without the need for retraining or reprograming. Inspired by insights from human cognition, we pioneered a meta-cognitive reasoning process that maintains explicit expectations about the agent behavior in canonical, non-novel settings. Violations of those expectations indicates the presence of a novelty that the architecture accommodates through a novel model diagnosis and repair process. We demonstrated that our architecture is resilient, quick (learns $20x$ faster than reinforcement learning), and interpretable, encouraging human trust in learning agents.

ITL studies *teachable agents* - agents that can dynamically extend their task & domain knowledge through natural human-AI interaction. Learning new domain concepts and task knowledge online, post deployment, is critical to the adoption of embodied agents and robots. At the University of Michigan, I led the development of Rosie, a model-based reasoning agent that learns interactively. It was built with Soar [20] and was the first in the literature to demonstrate interactive learning of a variety of concepts and tasks in a single, integrated agent

architecture. We introduced a new paradigm for learning domain concepts [18] and task knowledge [6] from situated, task-oriented dialog [17]. To support language understanding, we proposed a comprehension model [15] that grounds language semantics by using non-linguistic contexts (cognitive, attentional, and task-oriented). Adopting a human-centered perspective, we studied how human teachers naturally teach [7]. We found that teaching is an intentional process in which teachers introduce new concepts, define them and provide examples, evaluate the competency of the learner, and expand what was taught previously. To exploit iterative, incremental teaching, we developed an embodied agent architecture [14] that learns new task-relevant concepts using analogical reasoning and generalization [21]. Most recently, we are investigating how large-language models can support language understanding in embodied planning agents [22]. Our research has led to an emerging, inter-disciplinary, scientific inquiry on Interactive Task Learning (ITL [23]).

With John Laird, I was awarded the AAAI 2018 Blue Sky award [24] for proposing a framework of learning in generally intelligent agents that integrates lower-level automatic learning processes with higher-level learning strategies under the volitional control of the agent.

Human-Agent Collaboration I have developed intelligent agents for various real-world application domains where the agents adopt an assistive role to a human, supporting their sensemaking, learning, and decision-making. This research brings together methods from human factors research (such as need finding studies, cognitive task analysis, quantitative/qualitative human participant studies) with design of intelligent systems. In our most recent work [25], we investigated if conversational systems built with generative multi-modal language models can support humans as they perform a sensemaking task - understanding medical scans and reports. We found that in addition to being incorrect frequently, the responses generated by conversational systems were characteristically different from how a physician responds. While a physician focused on specifics of the case being discussed, generative systems generated general diagnostic knowledge about the disease. For better alignment, we are developing prompt engineering and reinforcement learning with AI feedback (RLAIF) methods that use an external model-based reasoning system to provide policy guidance to the generative models.

Previously, under the NSF/NIH Smart and Connected Health program, we developed long-living, interactive, coaching agents that help people pursue exercise and nutritional goals and develop healthy behaviors. Building upon theories from behavioral psychology, we designed an interactive coaching agent embodied in a mobile application [26, 9]. The agent used a parameterized, prescriptive, adaptive model of growth in aerobic capability 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-living intelligent interactive agents [9, 13]. 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, we [8] leveraged the Common Model of Cognition [20] as an integrative framework for explaining several behavior change theories from psychology and used it to provide design recommendations for interactive coaching systems. This line of research has been published at venues for medical informatics [13, 11, 12] in addition to being highlighted as key technical advancement on the roadmap to robust interactive intelligence [27] at an NSF workshop. It is one of the first demonstrations of long-term interactive, adaptive behavior that was evaluated with human participants in ecological settings.

Under the ARPA-E TransNet program, we explored how people can be influenced to adopt sustainable modes of transport [5, 10]. We identified what factors underlie people's transit-related choices through a set of semi-structured interviews and survey studies [8]. We used deep learning methods to build a prescriptive model of traveler mode adoption, drawing insights rational choice theory from economics [28]. We used this model to bias plan selection in an AI multi-modal planning framework to generate personalized, energy-efficient plans for each individual traveler. Through transportation modeling simulations, we demonstrated that our approach could achieve small but significant energy and time savings in Los Angeles. This line of research demonstrates how insights from human factors, behavioral economics, AI & ML, and transportation systems can be brought together to address a complex, societal issue.

2 Future Directions

I want to advance the human-aware AI systems agenda and build intelligent agents that can collaborate with humans in a variety of roles. There are two advances that are critical to achieving this goal: first, is developing intelligent agent architectures for collaboration, and second, developing computational models of human decision making, behavior, and learning that can be integrated with AI & ML systems. My background in intelligent agent architectures in addition to experience in human-factors research has built a strong foundation for me to successfully develop this agenda upon.

Intelligent Architectures for Human-AI Collaboration Over several decades, AI & ML research has produced a variety of computational methods that capture some aspect of intelligence. For instance, the recent breakthroughs in large-language models (LLMs) enable natural interactions in machines. These diverse computational methods have complimentary strengths. While LLMs can handle natural variation in human expression, they are unable to reason methodically about action and causation like a planning system. Planning systems, on the other hand, cannot deal with noise and partial-observability. They need to be augmented with foundational vision models for robust operation in physical worlds. I envision these computational methods as building blocks for complex, intelligent behavior and want to study how they can be interfaced together in various configurations. My research will lead to design principles behind multi-representational, hybrid, AI architectures that seamlessly integrate statistical and symbolic inference methods. I am particularly interested in exploring how LLMs can be brought within larger reasoning and learning frameworks. I want to study this integration from three different perspectives. First, LLMs as a mechanism mediating humans and formal reasoning/learning frameworks; second, as a source of knowledge when formal frameworks don't have sufficient knowledge to advance inference; and finally, as statistical reasoners whose inferences are validated through further formal analysis.

To ground agent architecture research, I want to study 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 intelligent agents that can learn novel domain concepts and task knowledge through natural interaction, online and during performance. Additionally, I want design teaching agents that can support humans in learning novel tasks such as assembling a new artifact [29] using augmented reality embodiment or teaching humans new science and mathematics concepts using conversational and visual interfaces.

Modeling Humans in Collaborative Intelligent 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 human-AI collaboration needs explicit, causal models human behavior and learning 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. A teachable agents should be able to exploit the full range of information in varying human teaching strategies [7]. Similarly, for agents to participate in human-machine teams, they must model their human partner's goals and skillset.

In my research, I want to advance hybrid modeling methods for human behavior, problem-solving, learning, and teaming that incorporate both symbolic and statistical mechanisms. I will use theoretical understanding of human behavior from social sciences (e.g., goal setting theory [30], Common Model of Cognition [31], knowledge tracing [32]) to provide symbolic structural scaffolds over which quantitative information can be overlaid using modern machine learning methods. While the structural scaffolds ensure explicability and diagnosability of the models, the quantitative information reflects stochasticity from individual variability and non-modeled factors.

I will incorporate human models in collaborative agents to enable them to reason about the needs and preferences of their human partners to provide proactive support. One of the challenges facing current generation of LLMs is that they generate language/information with limited understanding of the human interacting with them. Human needs vary significantly based on their prior knowledge and the task they are performing. For instance, a novice learning how to program needs significant help in not only understanding programming constructs but also in evaluating if an LLM-generated solution is correct or not. On the other hand, an expert programmer can

evaluate generated solutions and consequently, is looking to quickly access the space of plausible solutions to pick an appropriate one. To enable LLMs to be reactive to human needs, I want to explore how models of human skill acquisition can be used to bias generation in LLMs. Previous research on interactive tutoring systems [33] provides insights on how a student's problem solving behavior can be used to estimate their knowledge gaps and adapt lessons accordingly. I want to bring leanings from this research to the modern era.

The recent successes of AI and ML are now accompanied with an ever increasing expectation of using those methods to support human goals including public health, learning and education, sustainable living etc. Studying AI & ML algorithms in isolation will not lead us to solutions for these challenging problems. An effective intelligent solution requires an interdisciplinary approach that brings together insights from human-factors research, social sciences, and AI & ML. My research will advance our understanding of the human-AI ecosystems towards developing effective collaborative intelligent systems.

References

[1] E. Schmidt, R. O. Work, Y. Bajraktari, S. Catz, E. J. Horvitz, S. Chien, A. Jassy, M. L. Clyburn, G. Louie, C. Darby, et al. "National Security Commission on artificial intelligence". In: (2021).

[2] S. Mohan, W. Piotrowski, R. Stern, S. Grover, S. Kim, J. Le, and J. De Kleer. "A Domain-Independent Agent Architecture for Adaptive Operation in Evolving Open Worlds". In: *Artificial Intelligence Journal (preprint)* (2023).

[3] W. Piotrowski, R. Stern, Y. Sher, J. Le, M. Klenk, J. deKleer, and S. Mohan. "Learning to operate in open worlds by adapting planning models". In: *International Conference on Autonomous Agents and Multi-Agent Systems*. 2023.

[4] W. Piotrowski, Y. Sher, S. Grover, R. Stern, and S. Mohan. "Heuristic search for physics-based problems: angry birds in PDDL+". In: *Proceedings of the International Conference on Automated Planning and Scheduling*. Vol. 33. 1. 2023, pp. 518–526.

[5] 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.

[6] 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.

[7] 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.

[8] 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* (2021).

[9] 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).

[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] 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).

[12] 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).

[13] 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 Conference of the IEEE Engineering in Medicine and Biology Society*. 2016.

[14] 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.

[15] 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).

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

[17] 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.

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

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

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

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

[22] S. Grover and S. Mohan. "A Demonstration of Natural Language Understanding in Embodied Planning Agents". In: *Proceedings of the International Conference on Automated Planning and Scheduling* (2024).

[23] 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.

[24] 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.

[25] S. Rajagopal, S. Hazarika, S. Kim, Y.-m. Chiou, J. H. Sohn, H. Subramonyam, and S. Mohan. "Exploring How Generative Visual Question Answering Systems can Support Patients' Understanding of their Medical Reports". In: *arXiv preprint arXiv:2402.00234 (under review at ACM FAccT)* (2024).

[26] 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.

[27] J. Oakley. "Intelligent Cognitive Assistants (ICA) NSF Workshop Summary and Research Needs". In: *Semiconductor Research Corporation* (2018).
[28] T. Domencich and D. McFadden. "Urban Travel Demand - A Behavioral Analysis". In: *Transport Research Laboratory* (1975).
[29] 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.
[30] 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).
[31] 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).
[32] A. T. Corbett and J. R. Anderson. "Knowledge tracing: Modeling the acquisition of procedural knowledge". In: *User modeling and user-adapted interaction* 4 (1994), pp. 253–278.
[33] K. VanLehn. "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems". In: *Educational psychologist* 46.4 (2011), pp. 197–221.