Natural, effective collaboration between intelligent agents and humans on complex tasks is the next frontier in artificial intelligence (AI) and machine learning (ML) research [1]. Intelligent agents are often envisioned as an enabling technology for personalizing education, improving public health outcomes, advancing human learning, increasing scientific productivity etc. To address these challenges, intelligent agents must be designed to operate effectively within human systems. Generative AI (GenAI) methods have enabled human-machine interaction using natural modalities such as language, diagrams, images etc. Building upon this breakthrough, for effective human-AI collaboration, agents must be imparted with capabilities that enable them to reason about the structure and dynamics of world, diversity of tasks, and individual needs of humans.

I have over 15 years of experience in developing intelligent agents that can collaborate with humans in various roles. In my work, I apply a systems lens and build agent architectures with diverse AI & ML components. Additionally, I adopt insights about human decision making, behavior, and learning from social sciences to build agents that are both human-like and human-aware. This work is interdisciplinary and has been published at 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 government agencies (DARPA, ARPA-E, AFOSR, and NSF/NIH) as well as corporations (Xerox, Kaiser Permanente). Further, it supports a growing patent portfolio.

1 Research Context, Vision, and Experience

A Systems Lens Dominant view in AI & ML studies methods for specific capabilities (vision, language understanding, planning) in isolation with each other. A minority view studies AI with a systems lens [19] to understand how diverse intelligent capabilities can be brought into an integrated system for complex behavior. As AI & ML methods mature, the systems lens is becoming important and prominent. Human-agent collaboration necessarily requires a systems lens [20] - not only must an agent make productive decisions given its observations and action space, it must also exchange information with a human partner naturally. I take a four-pronged approach towards developing collaborative agents. First, I build agents that are *human-like*; I draw upon the insights in cognitive science about the nature of the human mind to build agent systems that have multiple intelligent capabilities including vision, goal-oriented reasoning, AI planning, natural language processing, control etc. Second, I develop methods for *human-aware* behavior; I adapt descriptive models of human decisions, behavior, and learning in social sciences - behavioral economics, psychology, education - into prescriptive models that can be used within an agent's decision making processes. Along the third prong, I embed agents in interfaces and embodiments to study the principles of natural collaborative human-AI interaction. For evaluation, I relinquish computation-centric metrics (e.g, accuracy, efficiency) and adopt human-centered metrics (e.g, flexibility, safety, acceptability) as well as experimental methods from social sciences, advancing the fourth and final prong.

Advances in Agent Architectures A persistent dogma of AI & ML technology is the design-and-deploy cycle that assumes that the structure and dynamics of the deployment environment are known at design time and are stationery post deployment. Classical AI systems are programmed by an AI designer based on their understanding of the deployment environment. Along similar lines, ML systems are trained on datasets presumed to reflect the generative processes in the deployment environment. If the assumptions, that deployment environment is known and is stationery, are violated, AI & ML technology has to be taken offline and reprogrammed or retrained. In a stark contrast, humans adapt and learn with volition whenever the need arises.

I build intelligent agents that can adapt to an evolving world and changing task requirements; autonomously and with human teaching. I served as principal investigator for Open-World Learning (OWL DARPA SAIL-ON) and Interactive Task Learning (ITL DARPA GAILA). The agent architectures I developed commit to the vision that complex intelligent behavior results from an interplay of diverse reasoning and learning methods. With John Laird, I received the AAAI 2018 Blue Sky award [21] for a framework of autonomous learning that integrates lower-level ML processes with higher-level learning strategies under the volitional control of an agent.

OWL studies how agents can autonomously adapt in evolving, non-stationery environments. We introduce the idea of a *novelty* [2] - a meaningful change in the operational characteristics of the environment (e.g., change in gravity or a new tool is made available) that occurs after an agent has been deployed. Model-free learning architectures, such as deep reinforcement learning, experience catastrophic failures when novelties are introduced. Our agent architecture [2, 3] builds upon an explicit representation of a *world model* (e.g., a planning domain) encoding the structure and dynamics of the environment. The world model is reasoned and adapted with model-based reasoning (e.g., AI planning) and related machine learning methods. Our approach can elegantly handle novelties without the need for full retraining or reprogramming. Inspired by human cognition, we pioneered a meta-cognitive reasoning process that maintains explicit expectations about the agent behavior in canonical, non-novel settings given its world model. Violations of those expectations indicates the presence of a novelty that the architecture characterizes in terms of changes to its world model. It then, accommodates through a novel model diagnosis and repair process. Our architecture is resilient, quick (learns $20x$ faster than deep reinforcement learning), and interpretable, encouraging human trust in learning agents.

ITL investigates teachable agents that can dynamically learn *task models* through natural human-AI interaction. At the University of Michigan, I led the development of Rosie, a world model-based agent that learns interactively. It was built upon a cognitive architecture [22] and was the first in the literature to demonstrate interactive learning of task-relevant knowledge (e.g., elements of a planning domain) in a single, integrated agent system. We introduced a new paradigm for comprehensive task model acquisition [18, 6] from situated, task-oriented dialog [17]. To support language understanding, we proposed a comprehension model [15] that grounds language 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 such iterative, incremental teaching, we developed an embodied agent architecture [14] that learns new task-relevant concepts (aka elements of a planning world model) using graph inference and generalization [23]. Most recently, we are investigating how large-language models can be used to understand task-related natural language in embodied agents [24].

Applications of Agent Technology I have developed intelligent agents for various applications 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 the design of intelligent systems. In this research, I adapt descriptive models of human decisions, behavior, and learning from social sciences to develop *prescriptive human models* that enable agents to reason about their human partners.

In my recent work [25], we investigate if generative AI systems (GenAI) can support humans in sensemaking - understanding their medical scans and reports. We found that in addition to being incorrect frequently, the responses produced were characteristically different from how a physician responds. While a physician focused on specifics of the case being discussed, GenAI produced general diagnostic knowledge about the disease. For better alignment, we are applying collaborative theory of discourse [20] to adapt GenAI's constitution on the fly.

Under NSF/NIH Smart and Connected Health, I developed interactive, coaching agents embodied in a mobile interface that help people develop healthy exercise and nutirion behaviors [26, 9]. The agents used a parameterized, prescriptive, adaptive model of humans' aerobic capability in conjunction with AI scheduling methods. We pioneered a novel four-pronged staged evaluation approach for collaborative agents [9, 13]. The evaluation approach 1) characterized alignment with human experts, 2) assess efficacy of the user interface [13], 3) benchmark the space of agent adaptation with simulated profiles, and 4) measured the agent's capability to promote behavior change through a deployment that lasted several weeks. The agents were built [8, 12, 11] upon insights from behavioral psychology (adaptive goal setting [27], self-efficacy [28]) and cognitive science (the Common Model of Cognition [22]). My work [26, 9] was the first and is one of the very few demonstrations of AI operating with humans in ecological settings for long-time horizons.

Under ARPA-E TransNet, I built an agent that influences people to adopt sustainable modes of transport [5, 10] to bring down a city's energy consumption. This work brings together interdisciplinary insights from human factors research, behavioral economics, AI & ML, and transportation systems to address a complex societal issue. We identified factors underlie people's transit-related choices through a set of semi-structured interviews and surveys [8]. Then, we developed a deep learning prescriptive model of traveler mode adoption drawing upon the rational choice theory [29]. This model biases plan selection in an AI planning framework [30] to generate energy-efficient plans for each individual traveler that acceptable given their personal travel context. Through choice experiments [5] and transit simulations, we demonstrated energy and time savings in Los Angeles.

2 Future Directions

At Microsoft Research, I aim to continue developing collaborative agents with a systems lens. And, leverage my expertise in diverse methods for model-based reasoning and machine learning to augment GenAI agentic architecture for flexible, trusted, and reliant autonomy. Specifically, I want to advance AI systems science along three critical thrusts: *advanced autonomy*: that incorporates models of the world, task, and humans within GenAI agentic architectures; *cooperative multi-dimensional inference*: that leverages statistical inference and structured reasoning together; and *unified architectures* comprised of deep learning and model-based reasoning & learning components that are capable of complex collaborative behavior in the real world.

Advanced Autonomy Current generation of GenAI agentic frameworks (AutoGen[31], LangGraph[32]) enable a flexible orchestration of various capabilities in service of a complex, multi-step task. While these frameworks are malleable, the current state of art implements level 0 autonomy or *autonomy of behavior* where an agentic system can perform a task on a user's behalf. However, the execution itself has steps specified by the AI designer or user. I want to build agentic systems with *autonomy of reasoning* that themselves can determine which steps to execute and when. To support contextual, flexible, and intentional reasoning, we need to develop and integrate predictive models within agentic frameworks. *World models* encode the structure and dynamics of the world, enabling agentic systems to condition their behavior on expectations of future states (level 1 autonomy). *Task models* encode parameters, soft & hard constraints, and goals, enabling an agentic system to reason about task execution reliably (level 2 autonomy). *Human models* encode the beliefs, desires, and intentions of human partners as well as drive expectations about their behavior, decision making, and learning; enabling agentic systems to individualize execution to a user's needs (level 3 autonomy). Levels 4+ are *autonomy of learning* where the agentic system can acquire these predictive models autonomously through experience or teaching.

Cooperative Multi-dimensional Inference Foundation models [33] have strengths that are complementary to classical AI methods (e.g., knowledge graphs, search, planning etc.). They are robust to noise, uncertainty, and the variation inherent in the real world. However, unlike classical AI, they implement implicit inference that is not easily understood, structured, or controlled, limiting their use in critical cases. For example, 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, motivating the use of foundation vision models. Emergence of methods such as retrieval augmented generation (RAG) demonstrates that inference within foundation models is not sufficient for complex tasks and must be augmented. I want to study the space of configurations of foundation models and reasoning approaches such that they can benefit from the strength of others. A variety of configurations are possible, each differing in how the inference load is distributed between the two kinds of systems. From reasoning systems augmenting foundation models' context with additional information, reasoning systems adapting foundation models' consitution based on task and conversation status status, reasoning systems validating foundation model responses (LLM-modulo [34]), to foundation models as user/world interfaces to reasoning systems, and foundation models as source of knowledge when reasoning systems are incomplete. Through a principled study, I will uncover the tradeoffs in using different configurations in terms of data needs, inference time, accuracy, assurability etc. Understanding of these tradeoffs will be useful in developing design guidance for agentic systems for real-world problems.

Unified Architectures Human intelligence comprises multiple intelligent capabilities: perception, planning, action & control, long/short-term memory, learning etc. in an integrated architecture [35]. The earlier generation of cognitive architectures [22] sought to build a similar infrastructure for machines using symbolic reasoning methods. While these architectures had contextual, flexible behavior, they were limited in their application to the real world with noise, uncertainty, and partial observability. The discovery of modern subsymbolic inference (transformers [36] and their applications as foundation models [33]) has opened up the possibility of unified cognitive architectures that balance subsymbolic and symbolic inference to operate flexibly and robustly in the real world. Going beyond the original goals of cognitive architectures research that focuses on the cognitive and rational bands [19], I want to develop architectures that are inherently collaborative, addressing the social band as well.

To situate unified 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 to design teaching agents that can support humans in learning novel tasks such as assembling a new artifact [37] using augmented reality embodiment or teaching humans new science and mathematics concepts using conversational and visual interfaces. One of the challenges facing the state of art in agents is that they act/generate language 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 a AI-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 agents to be reactive to human needs, I want to explore how models of human learning [38], capabilities [39], task-oriented dialog [40] etc. can be leveraged to modulate agent response in an unified architecture.

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] A. Newell. *Unified Theories of Cognition*. Harvard University Press, 1994.

[20] B. J. Grosz. "Collaborative systems (AAAI-94 presidential address)". In: *AI magazine* 17.2 (1996), pp. 67–67.

[21] 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.

[22] J. E. Laird. *The Soar Cognitive Architecture.* MIT press, 2012.
[23] K. D. Forbus, D. Gentner, and K. Law. "MAC/FAC: A Model of Similarity-Based Retrieval". In: *Cognitive Science* (1995).
[24] 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).
[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] 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).
[28] A. Bandura and S. Wessels. *Self-efficacy.* Cambridge University Press Cambridge, 1997.
[29] T. Domencich and D. McFadden. "Urban Travel Demand - A Behavioral Analysis". In: *Transport Research Laboratory* (1975).
[30] F. Dvorak, S. Mohan, V. Bellotti, and M. Klenk. "Collaborative optimization and planning for transportation energy reduction". In: *ICAPS Proceedings of the 6th Workshop on Distributed and Multi-Agent Planning (DMAP.* 2018.
[31] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang. "Autogen: Enabling next-gen llm applications via multi-agent conversation framework". In: *arXiv preprint arXiv:2308.08155* (2023).
[32] LangChain. *LangGraph.* https://www.langchain.com/langgraph. Accessed: 2024-11-08. 2024.
[33] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).
[34] S. Kambhampati, K. Valmeekam, L. Guan, M. Verma, K. Stechly, S. Bhambri, L. Saldyt, and A. Murthy. *LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks.* arXiv:2402.01817. June 2024. URL: http://arxiv.org/abs/2402.01817 (visited on 11/14/2024).
[35] 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).
[36] A. Vaswani. "Attention is all you need". In: *Advances in Neural Information Processing Systems* (2017).
[37] 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.
[38] K. VanLehn. "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems". In: *Educational psychologist* 46.4 (2011), pp. 197–221.
[39] B. Wilder, E. Horvitz, and E. Kamar. "Learning to complement humans". In: *arXiv preprint arXiv:2005.00582* (2020).
[40] B. J. Grosz and C. L. Sidner. "Attention, intentions, and the structure of discourse". In: *Computational linguistics* 12.3 (1986), pp. 175–204.