Intelligent agents are expected transform the human society by enabling by enabling personalized education and health, automated enterprise workflows, tailored information access, advanced manufacturing etc. To materialize this vision, agents must be designed to operate effectively within human systems. Not surprisingly, natural and effective collaboration between intelligent agents and humans on complex tasks is the next frontier in artificial intelligence (AI) and machine learning (ML) research [1]. Generative AI (GenAI) methods have enabled human-machine interaction using natural modalities such as language, diagrams, images etc. Building upon this breakthrough, agents must be imparted with capabilities to reason about our world's structure and dynamics, diversity in tasks, and humans' individual needs.

I have 15 years experience in developing intelligent agents that collaborate with humans. I apply a systems lens and build agent architectures with diverse reasoning and learning 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. My interdisciplinary work 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 has resulted in a growing patent portfolio.

1 Research Context, Vision, and Experience

A Systems Lens Effective human-agent collaboration in the real world requires a systems lens [19]. An agent must understand the world, make productive decisions given its observations, and act to pursue its goals. In addition, it must also exchange information with human partners through natural modalities. I take a four-pronged approach towards developing collaborative agents. First, I draw upon the insights in cognitive science about the nature of the human mind to build *human-like* agent systems that have multiple intelligent capabilities - vision, learning, reasoning, planning, task execution, dialog etc. Second, I adapt descriptive models of human decisions, behavior, and learning in social sciences into prescriptive models to be used within an agent's decision making processes making it *human-aware*. Third, I embed agents in interfaces and embodiments to study the principles of natural collaborative human-AI interaction. Finally, I evaluate agent performance using social science methods and human-centered metrics (e.g, flexibility, acceptability). This approach goes beyond benchmark datasets and computation-centric metrics (e.g, accuracy).

Advances in Agent Architectures AI & ML technology is built on the design-and-deploy principle. This principle assumes that the deployment environment's structure and dynamics are known at design time and are stationary post deployment. An AI designer programs a classical AI system 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. When the assumptions that deployment environment is known and is stationary are violated, AI & ML technology is taken offline and reprogrammed. In contrast, humans adapt and learn with volition whenever the need arises.

I build intelligent agents that adapt to an evolving world and changing task requirements both 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 build commit to the vision that complex intelligent behavior results from an interplay of diverse reasoning and learning methods. I received the AAAI 2018 Blue Sky award [20] for a framework for autonomous learning that integrates lower-level ML processes with higher-level learning strategies under the agent's volitional control.

OWL studies how agents can autonomously adapt in evolving, non-stationary environments. We introduce the idea of a *novelty* [2] - a meaningful change in the environment's operational characteristics (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 facing novelties. Our agent architecture [2, 3] builds upon an explicit representation of a *world model* (e.g., a planning domain) encoding the environment's structure and dynamics. 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 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 that novelty 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 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 agent architecture [21] and implemented a new paradigm for task model acquisition [18, 6] based on situated task-oriented dialog [17, 15]. It was the first in the literature to demonstrate interactive learning of grounded, comprehensive task-relevant knowledge (elements of a planning domain) in a single integrated agent system. At PARC/SRI, I continued to build on this work. I developed a research agenda on embodied agents built with modern ML methods that learn from humans. We studied how humans naturally teach [7] and found that teaching is an intentional process in which teachers introduce new concepts, define them and provide examples, evaluate the learner's competency, and expand what was taught previously. To exploit such iterative, incremental teaching, we developed an embodied agent architecture [14] that learns new task models using graph inference and generalization [22]. Most recently, we exploit large-language models to understand task-related natural language in embodied agents [23].

Applications of Agent Technology I have built intelligent agents for various applications where the agents adopt an assistive role to a human, supporting their sense making, decision making, problem solving, and learning. This work brings together methods from human factors research (e.g., need finding studies, cognitive task analysis, quantitative/qualitative human studies) with AI systems engineering. Additionally, I adapt descriptive models of human decisions, behavior, & learning from social sciences - cognitive science, pyschology, behavioral economics - to develop *prescriptive human models* that enable agents to reason about their human partners.

In recent work [24], we investigate how generative AI systems (GenAI) can support humans in sensemaking - understanding their medical scans and reports. We found that in addition to being frequently incorrect, the responses produced were characteristically different from how a physician responds. While a physician discussed the specifics of the case focusing on information that help the patient make productive decisions, GenAI produced general diagnostic knowledge about the disease. For better alignment, we are applying collaborative theory of discourse [19] to adapt GenAI's constitution on the fly. Particularly, we are developing response-generation guidance for GenAI based on our analysis of the physicians answers.

Under NSF/NIH Smart and Connected Health, I developed interactive, coaching agents deployed on a mobile interface that help people develop healthy exercise and nutrition behaviors [25, 9]. The agents combined a parameterized, prescriptive, adaptive model of humans' aerobic capability with AI scheduling methods. Going beyond benchmark datasets typical in AI & ML research, we developed a novel staged approach for evaluating collaborative agents [9, 13]. The evaluation approach 1) characterized alignment with human experts, 2) assessed efficacy of the user interface [13], 3) benchmarked the agent adaptation space with simulated profiles, and 4) demonstrated a 20% increase in exercise volume over 6 weeks for 21 participants. We built the agents [8, 12, 11] upon insights from behavioral psychology (adaptive goal setting [26], self-efficacy [27]) and cognitive science (the Common Model of Cognition [21]). My work [25, 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 methods from human factors research, behavioral economics, AI & ML, and transportation systems to address a complex societal issue. We identified factors underlying people's transit-related choices through semi-structured interviews and surveys [8]. Then, we drew upon the rational choice theory [28] to develop a deep learning-based model of traveler mode adoption . This model biases plan selection in an AI planning framework [29] to generate energy-efficient plans for each individual traveler that are acceptable to them given their personal travel context. Through choice experiments [5] and transit simulations, we demonstrated 5% energy and 15% time savings in Los Angeles.

2 Future Directions

Leveraging my expertise in agent systems and architectures, I aim to extended GenAI agent frameworks such that they are flexible, reliable, and trusted. Specifically, I want to advance agent systems science along three critical thrusts. *Advanced autonomy* incorporates models of the world, task, and humans within GenAI agentic architectures. *Cooperative multi-dimensional inference* leverages both statistical inference and structured reasoning together to solve complex problems. And *unified architectures* support flexible, collaborative behavior in the real world relying on a principled integration of deep learning and model-based reasoning methods.

Advanced Autonomy Current generation of GenAI agentic frameworks (AutoGen[30], LangGraph[31]) enable a flexible orchestration of various capabilities in service of a complex, multi-step task. However, they only implement *autonomy of behavior* - while a user can delegate a task to an agentic system, the steps and order of execution are specified by an AI designer. I want to enable advanced autonomy in agentic systems. With *autonomy of reasoning*, the agentic system can itself determine which steps to execute and when. Utilizing my prior experience in world, task, and human models, I will extend agentic frameworks with predictive models that enable contextual, flexible, and intentional behavior. *World models* encode the world's structure and dynamics, enabling agentic systems to condition their behavior on expectations of future states. *Task models* encode parameters, soft & hard constraints, and goals, enabling an agentic system to reason about task execution reliably. *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. I will apply my research on OWL and ITL to impart *autonomy of learning* to agentic systems so that they can acquire and extend predictive models autonomously and through human teaching.

Cooperative Multi-dimensional Inference Foundation models' [32] strengths are complementary to classical AI methods (e.g., knowledge graphs, search, planning etc.). They are robust to noise, uncertainty, and variation 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. While LLMs can handle variation in human expression, they are unable to reason methodically about action and causation like a planning system [33]. Planning systems, on the other hand, cannot deal with noise and partial-observability and must rely on foundation vision models.

I will study the configuration space of foundation models and reasoning approaches with a problem-centered lens. Configurations differ in how onus of inference is distributed between the two systems and are appropriate for different usecases. For example, when a user wants to query for domain-specific information, inference can be driven by a foundation model with knowledge graph reasoning systems augmenting its context (LLM+KG [34]). When the user wants to evaluate various courses of action, a foundation model can be leveraged to generate them and a reasoning system to validate them, ensuring plausibility (LLM-modulo [35]). Where a user expects an agent to execute a task, a foundation model serves as an interface between the human and a task reasoner and executor (LLM+plan [23]). Through a structured exploration, I will uncover the tradeoffs in using different configurations in terms of data needs, inference time, accuracy, assurability etc. Further, I will relate the tradeoffs with problem characteristics, developing design guidance for agentic systems in the real world.

Unified Architectures Human intelligence comprises multiple intelligent capabilities: perception, planning, action & control, long/short-term memory, learning etc. in an integrated cognitive architecture [36]. Earlier cognitive architectures [21] sought to build a similar infrastructure for machines using symbolic reasoning methods. While these architectures had contextual, flexible behavior, real world with noise, uncertainty, and partial observability presented an operational challenge. The discovery of modern subsymbolic inference (transformers [37] and their applications as foundation models [32]) 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 [38], I want to develop architectures that are inherently collaborative, addressing the social band as well.

I want to study learning from social interaction as a motivating problem for unified architecture research. Learning with social constructs is the most fundamental form of human learning. Parents, teachers, experts en-

able 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. These learning interactions are characteristically different from ML. Humans trainers communicate structure, provide examples, evaluate the learner's competency bounds, provide feedback, adapt content, etc [7]. I will study the human-agent collaborative learning dyad from a variety of perspectives. Continuing my ongoing research, I will develop intelligent agents that learn novel domain concepts and task knowledge through natural interaction post deployment. In addition, I want to build agents that support humans in learning and upskilling, helping them become resilient our rapidly changing economy and the world. This vision includes helping humans learn new tasks such as assembling a new artifact [39] using augmented reality embodiment or teaching humans new science and mathematics concepts using conversational and visual interfaces. To enable agents to be reactive to human teachers and learners, I will incorporate models of human learning [40], capabilities [41], task-oriented dialog [42] etc. to modulate decision making and response in an unified architecture.

The recent successes of AI and ML are now accompanied with an ever increasing expectation of deploying them in real-world problems. I put forth a systems view of AI research and development, focusing on agent technology. I develop an inter-disciplinary approach to agent systems that builds upon prior state-of-art on world and task models and extends it to incorporate reasoning and learning about humans. My research will put forth a new generation of agents that are inherently collaborative and seamlessly integrate in human 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* (2024).

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

[20] 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.

[21] J. E. Laird. *The Soar Cognitive Architecture*. MIT press, 2012.
[22] K. D. Forbus, D. Gentner, and K. Law. "MAC/FAC: A Model of Similarity-Based Retrieval". In: *Cognitive Science* (1995).
[23] 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).
[24] S. Rajagopal, J. H. Sohn, H. Subramonyam, and S. Mohan. "Can Generative AI Systems Support Patients' & Caregivers' Informational Needs?" In: *arXiv preprint arXiv:2402.00234 (under review at ACM Intelligent User Interfaces)* (2024).
[25] 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.
[26] 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).
[27] A. Bandura and S. Wessels. *Self-efficacy*. Cambridge University Press Cambridge, 1997.
[28] T. Domencich and D. McFadden. "Urban Travel Demand - A Behavioral Analysis". In: *Transport Research Laboratory* (1975).
[29] 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.
[30] 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).
[31] LangChain. *LangGraph*. https://www.langchain.com/langgraph. Accessed: 2024-11-08. 2024.
[32] 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).
[33] M. Verma, S. Bhambri, and S. Kambhampati. "On the Brittle Foundations of ReAct Prompting for Agentic Large Language Models". In: *arXiv preprint arXiv:2405.13966* (2024).
[34] J. Cui, M. Ning, Z. Li, B. Chen, Y. Yan, H. Li, B. Ling, Y. Tian, and L. Yuan. "Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model". In: *arXiv preprint arXiv:2306.16092* (2024).
[35] 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. 2024. (Visited on 11/14/2024).
[36] 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).
[37] A. Vaswani. "Attention is all you need". In: *Advances in Neural Information Processing Systems* (2017).
[38] A. Newell. *Unified Theories of Cognition*. Harvard University Press, 1994.
[39] 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.
[40] K. VanLehn. "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems". In: *Educational psychologist* 46.4 (2011), pp. 197–221.
[41] B. Wilder, E. Horvitz, and E. Kamar. "Learning to complement humans". In: *arXiv preprint arXiv:2005.00582* (2020).
[42] B. J. Grosz and C. L. Sidner. "Attention, intentions, and the structure of discourse". In: *Computational linguistics* 12.3 (1986), pp. 175–204.