Research Statement

Over the past two decades, Artificial Intelligence (AI) technologies have been successfully applied to a wide range of complex problems and have achieved human level, or better, performance at tasks that were originally viewed as only within the purview of humans. Despite this progress, AI technologies are still prohibitively expensive to build. For example, it took more than a person century of AI expert development time to build the IBM Watson system that famously beat two Jeopardy! Champions (Laird et al., 2017) and building an AI-powered tutor to provide 1 hr. of classroom instruction can take as much as 200-300 hours of development time (Aleven et al., 2009). At their core, all AI systems are powered by knowledge (whether hand authored or learned). I would argue that one of the fundamental problems in the field is the *knowledge transfer problem*—mainly, how do we transfer knowledge into AI systems, so they can behave intelligently? As AI technologies become more widely used within human society, addressing this foundational problem has the potential to have a broad impact across many domains, such as education and medicine.

My research adopts a human-centered approach to the knowledge transfer problem and explores the development of teachable AI systems. These interactive learning systems integrate ideas from knowledge-based AI, ML, and Human-Computer Interaction (HCI) to enable end users, such as teachers or doctors, to build and personalize AI technologies through natural teaching interactions, similar to how they would teach another human. My research program blends use-inspired and foundational research to advance this concept along three thrusts (see Figure 1).

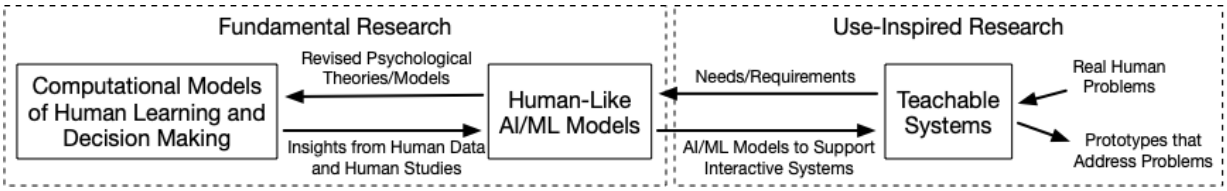

Figure 1. The three thrusts of my teachable AI research program and how they relate.

Thrust 1: Computational Models of Human Learning and Decision Making

My first research thrust focuses on building computational models of humans. This thrust aims to explore two questions: (1) how can we leverage human data to guide the development of human-like computational models? and (2) How can we leverage these human-like models to better understand human decision making and learning?

At the core of this research line is a theory of the computations that any agent (human or otherwise) needs to perform in order to learn from worked examples and feedback, which, in the spirit of Marr (1983), I refer to as a computational theory of *apprentice learning* (MacLellan & Koedinger, 2020). This theory applies approaches from expert systems, explanation-based learning, inductive logic programming, inverse reinforcement learning, and traditional reinforcement learning to support the induction and use of mixed symbolic and probabilistic structures. My preliminary theory posits three separate mechanisms that agents use to perform apprentice learning: how-learning, where-learning and when-learning. According to this theory, when an agent is faced with a problem to solve (e.g., what is 2+3?), a match is made between learned skills and the current problem state. If any skills match, the agent executes one with positive expected reward. If no skills apply (a typical initial response), then the agent requests a demonstration from the human instructor or learning environment (e.g., the instructor might enter a 5 in the answer field). The agent then performs *how-learning* to generate a sequence of mental operations (i.e., a procedure) that explains the provided demonstration, for example, the agent might explain the demonstration as the addition of the first and second numbers. Next, the agent uses

where-learning to induce a schema, or pattern, for extracting information from the environment necessary to execute the learned procedure and for recognizing when the procedure should be considered; for example, an agent would learn patterns for extracting the first and second numbers from the environment when there is an operator sign between them. Finally, the agent uses *when-learning* to estimate a reward function for the skill, which enables it to prioritize matching skills and determine which should be applied in any given situation; for example, the agent might learn that applying the add skill to two numbers produces a positive reward when the operator between them is a plus sign. The output of these steps (a procedure, a schema, and a reward function) constitute a new skill. On subsequent problems, the agent attempts to apply learned skills that it estimates will produce positive reward and receives correctness feedback. This feedback is then used in where- and when-learning to refine the skill's schema and reward function.

Based on this theory, I created the ***Apprentice Learner Architecture*** (MacLellan et al., 2016a), which affords the comparison of alternative computational models of apprentice learning. Within the architecture, a model presents as a set of algorithms to perform where-, when-, and how-learning. This architecture, and the underlying theory, defines a space of models that can be searched using different model evaluation criteria, and my research generates and tests models within this space, and, where necessary refines the theories underlying the models (see Figure 2).

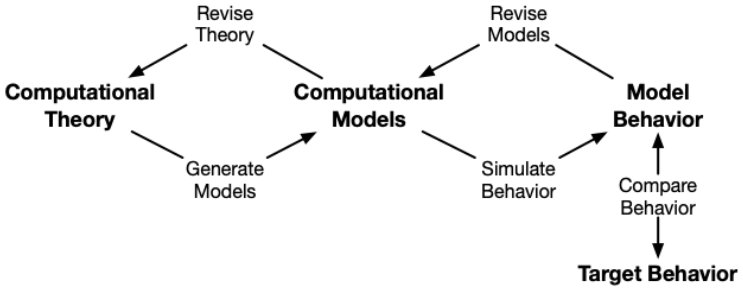

Figure 2. My computational theory refinement approach.

On my DARPA TAILOR project, I showed how these models could be individualized to better predict the learning trajectories of individual students and demonstrated how such a model could support instructional designers in counterfactually reasoning about how different individuals would respond to hypothetical cognitive training interventions (MacLellan, Stowers, and Brady, 2021). Similar to how bridge designers use parametric analysis to computationally simulate and test bridges prior to deploying them in the real world, I propose using computational models to simulate and test cognitive training interventions prior to running more costly human experiments. Purely statistical models of human learning (e.g., MacLellan et al., 2015) are very limited in their ability to generalize to interventions without existing human performance data. However, computational models of learning mechanistically model how a student's knowledge changes in response to an intervention and how their performance changes as a result. By leveraging cognitive learning theories within a unified computational model of learning (Newell, 1994), my work suggests that it is possible to make purely theory-driven predictions about human performance for alternative interventions, even when no existing human data are available (Maclellan et al., 2016a; Zhang & MacLellan, 2021).

Thrust 2: Human-Like AI/ML Models

Building on insights from my first research thrust, my second thrust aims to build human-like AI and ML models. Humans are incredible learners and problem solvers. While AI systems can exceed human performance for specific, narrowly defined tasks, they do not yet possess the kinds of flexible, general-purpose intelligence that humans are capable of. Even for tasks where computational systems outperform humans, learning for humans and ML systems is often qualitatively different (see table 1). In his recent book, *Becoming Human,* Tomasello compared human and ape development to identify

what makes us "uniquely human", and his work demonstrates the importance of considering the social nature of learning and knowledge transfer, such as humans' unique ability to learn from one another through situated imitation, explicit instruction, and collaboration. If we do not expand the current scope of mainstream ML research to investigate these important characteristics, then I argue that the best possible outcome for the field will be ape or animal-like—not human-like—AI/ML systems.

Table 1. Comparison of some key differences between humans and ML systems.

Humans	ML Systems
Able to learn completely new tasks with just a few examples (tens)	Require many examples to learn new tasks (tens of thousands)
Incrementally update their knowledge in light of new experiences	Must retrain on all past experiences, or face catastrophic forgetting
Can learn from multiple, mixed modalities, such as from demonstrations, feedback, verbal instruction	Typically, only support learning from a single modality
Can leverage their knowledge to explain their reasoning and behavior	Behavior is often opaque and unexplainable, especially for deep learning models

My research aims to better understand these key capabilities through the construction and evaluation of AI systems that emulate them. For example, I aim to develop new AI approaches that can learn incrementally from few examples and produce human relatable, explainable, and understandable outputs. My work explores the development of distinct AI and ML components (e.g., algorithms for incremental concept formation; MacLellan, et al., 2016b) as well as integrated cognitive systems that combine multiple components (e.g., my Apprentice Architecture; MacLellan and Koedinger, 2020). In a recent study, I compared my Apprentice Learner model to two state-of-the-art Reinforcement Learning (RL) models on two math learning tasks (MacLellan and Gupta, 2021). I found that while the RL models can learn these tasks, they require thousands to tens of thousands of examples to achieve mastery, whereas my Apprentice models only require tens to hundreds of examples. This is possible because Apprentice models support multiple modalities of learning; they learn incrementally from both demonstrations and feedback, whereas the RL models only learn from feedback. Additionally, they decompose the overall learning problem into multiple easier sub-problems (how-, where-, and when-learning). I currently have a paper under review with the Journal of AI Research that theoretically and empirically analyzes how this decomposition produces more efficient learning.

Thrust 3: Teachable Systems

My final thrust investigates the question, how do we design and build systems that people can teach and interact with like they would another human, while still taking advantage of key non-human features of AI/ML systems? This work is use-inspired and takes a human-centered approach to building teachable systems that address the knowledge transfer problem for real users and applications. For example, during my graduate studies at CMU, I explored the development of agents that teachers and other domain experts could use to author intelligent tutors via teaching rather than programming (MacLellan, et al., 2014; MacLellan et al., 2016a; MacLellan and Koedinger, 2021). While working as an industry research scientist, I developed teachable agent technologies for a broader range of military-related applications. For example, I worked closely with naval planners to create a teachable agent that could assist them in constructing operational plans for large numbers of unmanned assets. I also worked with expert fighter pilots to transfer their knowledge into an AI model that could fly an F16 during 1-vs-1 air combat (see DARPA AlphaDogFight competition: https://www.darpa.mil/news-events/2020-08-07). At Drexel, I am actively developing teachable agents to support adult learning (under my NSF-funded ALOE project) and to support medics in diagnosing battlefield injuries (under my DARPA POCUS project).

Human-AI Interaction is uniquely difficult to design for and requires a new set of design facilitators (Yang, Steinfeld, Rosé, and Zimmerman, 2020). While working across these different application areas, I have been extending foundational HCI theory and methodologies to support the design of interactive ML systems that people can naturally teach. A major outcome of these efforts is the development of my **Natural Training Interactions (NTI) framework** (MacLellan et al., 2018), which reviews the teachable agent literature and maps out the kinds of *knowledge* that users might transfer to an agent, as well as the *patterns*, *interaction types*, and *modalities* that they can use to transfer this knowledge. I hypothesize that similar to human learning (e.g., see the KLI framework; Koedinger, Corbett, and Perfetti, 2012), which teaching interactions are most natural and effective for end users will depend on the kind of knowledge a user intends to transfer (e.g., concept knowledge transfer will look different than skill transfer). To explore this hypothesis, I have developed a new prototyping methodology I refer to as dual-sided, limited perception Wizard-of-Oz (WoZ) experiments (Sheline & MacLellan, 2018; MacLellan et al., 2019), which adapts the standard WoZ approach to support prototyping of interactive learning systems (I use a naïve experimental participant rather than an experimental confederate to simulate the teachable system and they have to learn over the course of the experiment). Under my recent ARL STRONG project, I have also developed a new experimental platform for efficiently conducting online experiments using this paradigm, so I can rapidly prototype alternative teachable agent designs and further develop the theory needed to design effective teachable agents.

Future Work

My research program is a plan that I am actively putting into action. In just a single year at Drexel, I have been selected for six external awards totaling over $3.8M in funding. These research projects, which are being sponsored by multiple institutions including DARPA, ARL, and NSF, provide resources for me to advance all three of my research thrusts. My ARL STRONG project explores the development of task-general teachable agent capabilities inspired by Tomasello's work on collaborative learning (Thrust 2) and investigates how these capabilities can enable the creation of agents that can more effectively team with humans (Thrust 3). My work on the DARPA POCUS project investigates how we can reduce the amount of training data needed to build AI models that can diagnose injuries from ultrasound imagery (Thrust 2) as well as how teachable AI can be leveraged to effectively transfer knowledge from medical experts (Thrust 3). Finally, my NSF ALOE project explores the development of AI technologies to support adult learning—both in higher education and in the workforce. This project aims to improve our understanding of human learning (Thrust 1) and to develop teachable AI systems that can support teachers and students in building and customizing AI technologies (Thrust 3).

I am particularly excited about the possibility of joining Georgia Tech's School of Interactive Computing. My work spans diverse application areas (education, medicine, military) and disciplinary boundaries (AI, ML, HCI, and the Cognitive and Learning Sciences) and would be enhanced by an interdisciplinary institution, such as Georgia Tech, where I can find students and collaborators that also have a diverse range of backgrounds and skillsets. At Georgia Tech, I am particularly excited about the possibility of collaborating with faculty such as Ashok Goel, Mark Riedl, Sauvik Das, Munmun De Choudhury, Betsy DiSalvo, and Alex Endert, among many others. Finally, while I have been successful at Drexel University, Georgia Tech is a premiere research institution for AI and HCI, which would give me a platform to advocate more broadly for the cognitive systems and human-centered AI perspectives.

References (co-authors mentored by MacLellan are underlined)

- **MacLellan, C.J.**, Gupta, A. (2021). Learning Expert Models for Educationally Relevant Tasks using Reinforcement Learning. *Proceedings of the Fourteenth International Conference on Educational Data Mining*.
- Zhang, Q., **MacLellan, C.J.** (2021). Investigating Knowledge Tracing Models using Simulated Students. *Proceedings of the Fourteenth International Conference on Educational Data Mining*.
- **MacLellan, C.J.**, Stowers, K., Brady, L. (2020). Optimizing Human Performance using Individualized Computational Models of Learning. *Proceedings of the Eighth Annual Conference on Advances in Cognitive Systems*.
- **MacLellan, C.J.**, Koedinger, K.R. (2020). Domain General Tutor Authoring with Apprentice Learner Models. *International Journal of AI in Education*. doi: 10.1007/s40593-020-00214-2
- **MacLellan, C.J.**, Harpstead, E., Marinier III, R. P., Koedinger, K.R. (2018). A Framework for Natural Cognitive System Training Interactions. *Advances in Cognitive Systems*, 6, 177-192.
- Sheline, R. & **MacLellan., C.J.** (2018). Investigating Machine-Learning Interaction with Wizard-of-Oz Experiments. In *Proceedings of the NeurIPS 2018 Workshop on Learning by Instruction*.
- **MacLellan, C.J.**, Harpstead, E., Patel, R., Koedinger, K.R. (2016a). The Apprentice Learner Architecture: Closing the loop between learning theory and educational data. In *Proceedings of the 9th International Conference on Educational Data Mining*. Raleigh, NC: International Educational Data Mining Society.
- **MacLellan, C.J.**, Harpstead, E., Aleven, V., Koedinger K.R. (2016b). TRESTLE: A Model of Concept Formation in Structured Domains. *Advances in Cognitive Systems*, 4, 131-150.
- **MacLellan, C.J.**, Liu, R., Koedinger, K.R. (2015). Accounting for Slipping and Other False Negatives in Logistic Models of Student Learning. In O.C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining*. Madrid, Spain: International Educational Data Mining Society.
- **MacLellan, C.J.**, Koedinger, K.R., Matsuda, N. (2014). Authoring Tutors with SimStudent: An Evaluation of Efficiency and Model Quality. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (pp. 551-560). Switzerland: Springer International. doi: 10.1007/978-3-319-07221-0