Research Statement

I am focusing on bridging machine learning and cognitive science. In cognitive science, building a computational model is a test of understanding; if people outperform all existing algorithms on certain types of problems, we have more to understand about how people solve them. In machine learning, these cognitive abilities are both significant open problems as well as opportunities to reverse engineer the human's solutions. I want to better understand humans and to build machines that learn in more.

A. Reasoning and Planning

I am interested in how networks could tackle cognitive ability (such as reasoning, planning, imagination) and how that can help a learner figure out high-level representations on both the perception and action sides. Thus, I aim to build systems that demonstrate a deep understanding of the world, integrating cognitive abilities for reading, learning, and reasoning.

1) Reading: Many questions require multiple pieces of information to be combined to arrive at an answer. I want to develop new multihop models capable of identifying and combining relevant facts to answer such questions. **2) Learning**: Language models (LMs) have dominated much of AI recently. But what kind(s) of reasoning are they capable of? And how can they be taught to do more? I want to develop an analytical datasets to probe LMs and help answer these questions. **3) Reasoning about Actions:** A key aspect of intelligence is being able to reason about the dynamics of the world. This requires modeling what state the world might be in, and how different actions might affect that state. Such capabilities are essential for understanding what happens during a procedure or process, for planning, and for reasoning about "what if..." scenarios.

B. Language Grounding

Language is grounded in experience. Unlike dictionaries which define words in terms of other words, humans understand many basic words in terms of associations with sensory-motor experiences. People must interact physically with their world to grasp the essence of words like "red", "heavy" and "above". Abstract words are acquired only in relation to more concretely grounded terms. Grounding is thus a fundamental aspect of spoken language, which enables humans to acquire and to use words and sentences in context.

In the future, I want to develop an interactive robot that learns and understands language via multisensory grounding and robotic embodiment. My goals are two-fold. **First**, I am interested in using computational models to gain insights into how humans process language. By building and testing models with realistic data, I am able to test theories that are difficult to assess using traditional methods based on observation and analysis. **Second**, I hope to build a new generation of spoken language interfaces with richer semantic representations leading to more intelligent machine behavior.

C. Decision-Making Applications

Decision-makers such as judges make crucial choices affecting human lives on a daily basis. Such high-stakes decisions have a significant and lasting impact on individuals as well as society. However, prior research has demonstrated that the judgments made by these

decision-makers are not only inconsistent and error-prone but also subject to a variety of biases - e.g., racial and gender biases. Can we help these decision-makers make better judgments?

I realize that there are certain fundamental challenges that hinder the applicability of existing AI techniques to improve high-stakes decision-making: **a)** The available data only captures the outcomes of the decisions made by human decision-makers and not the counterfactuals. **b)** The data is prone to selection biases and confounding effects. **c)** The successful adoption of AI in high-stakes decision-making relies heavily on how well decision-makers can understand and trust its functionality; however, most of the existing AI models are are not very interpretable.

Thus, I am developing machine learning tools and techniques which are not only accurate but also fair and interpretable so that human decision-makers can leverage them to make better decisions. More specifically, my research aims at the following fundamental questions pertaining to human and algorithmic decision-making: **(a)** How do we build interpretable models that can aid human decision-making? **(b)** How do we evaluate the effectiveness of algorithmic predictions and compare them with human decisions? **(c)** How do we detect and correct underlying biases in human decisions and algorithmic predictions?

In summary, I am interested in cognitive abilities, focusing on problems that are easier for people than they are for machines. The human mind is the best-known solution to a diverse array of difficult computational problems: learning new concepts, learning new tasks, understanding scenes, learning a language, asking questions, forming explanations, amongst many others. Machines also struggle to simulate other facets of human intelligence, including creativity, curiosity, self-assessment, and commonsense reasoning. And I know it is my great happiness to contribute towards it.