It has been at least a decade since the wave of deep learning first washed over the field of natural language processing (NLP). Since then, we have seen deep learning methods support exciting and rapid progress across a spectrum of NLP tasks, from classics like syntactic parsing and part-of-speech tagging to new (or newly revived) tasks such as question answering and entailment reasoning. In many of the most popular and most critical shared tasks, high-capacity deep learning systems are now at the state of the art. Many of these models exhibit a striking unity of architecture: they are variations on what is mostly a shared formula, combining neural network sequence models with continuous vector representations of words. Various elaborations of this architecture have fueled many of the successes which have led artificial intelligence luminaries to proclaim NLP as the research world's "next frontier" to be conquered by neural networks and deep learning.

There is no question that this class of models has helped push the field forward in substantial ways. But I am convinced that there is much more to language learning and language use which these models do not currently capture. Much of what we currently fail to capture are just those qualities which are necessary for deploying robust and safe systems for natural language understanding.

The next decades will see increasing deployment of systems for dialogue and information retrieval in environments from preschools to nursing homes. With real-world deployment in mind, it is critical to examine how these bleeding-edge algorithms perform beyond simple measures of precision and recall evaluated on cleaned test sets. In particular, after working with such models for several years in deep learning for NLP, I am motivated to investigate three recurring deficits:

Interpretability and evaluation. When models work, it's often not clear why they do. When models don't work, it's also often not clear why they don't. We cannot make guarantees about performance or draw theoretical conclusions from models whose behavior cannot be explained.

Generalization and transfer. Humans can quickly and robustly deal with novel words and novel concepts in the language they encounter. High-capacity deep learning models are often brittle in the face of new data, and fail to transfer between tasks or even between domains. We have no theory for reasoning about how these large-capacity models will or will not generalize to novel circumstances.

Grounding and common-sense reasoning. Many models exhibit a deficient understanding of the real world [5]. We ought to select tasks and model architectures which promote the capacity to reason abstractly and draw on sense data just as humans do.

I believe that the first solution to the above problems is to **design and promote new tasks** which require us to explicitly address these issues (cf. my recent argument in [2, 3]). In MIT's Department of Brain and Cognitive Sciences, I do just this by bridging between NLP and the field of computational cognitive science. Within cognitive science, one of the primary research goals is to construct new evaluation paradigms of just this sort. In particular, cognitive scientists support paradigms which allow us to compare the performance of our models to *theoretically informed* notions of intelligent behavior. We acknowledge that a significant step in building intelligent and robust AI systems is to first characterize what sort of intelligence and robustness it is which *humans* exhibit.

My current project is an example of bridging between these two fields in a way which addresses these high-level problems. I am constructing a virtual world paradigm which consists of a 3-D world of simple shapes, which can be dynamically generated in any spatial configuration. These controlled environments have recently become popular within the NLP dialogue and question answering communities [see e.g. 6, 4]. This paradigm allows us to test new approaches to rapid inference and generalization in NLP systems, while simultaneously exploring theoretical questions in cognitive science.

My first project within this virtual world paradigm is designed to address the issues of generalization and common sense reasoning from first principles. I am designing a model which captures a phenomenon known in language acquisition as **fast mapping**: the ability to make rapid inferences about the meanings of completely novel words [1]. Children are able to fast-map word meanings before they are two years old, and they exploit the skill as they rapidly acquire new words in their early years.

A system which implements fast mapping must make informed predictions about how meanings generalize from a very limited number of examples. Decades of research in language acquisition suggest that children exploit secondary sources of evidence — prior knowledge, information from the physical environment, social cues, and much more — to perform this inference. This project promises to offer a computational description of these fast-mapping behaviors, and at the same time offer a ready-to-use implementation of one-shot inference in grounded language learning contexts. The figure to the right illustrates an example of these ideas.

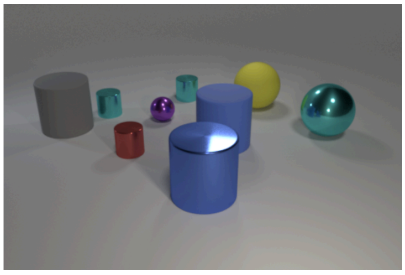

Figure 1: A word-learning scene from the CLEVR dataset [4]. Suppose someone asks you to *"touch the red blicket in front of the sphere."* What is the meaning of *blicket*? Syntactic cues suggest it is a noun. From reasoning about the intentions of the speaker, we can guess that there is indeed a *blicket* in the scene. Finally, by looking at the scene, we can guess that a *blicket* is a cylinder.

I believe this bottom-up approach to modeling language acquisition is the most promising way to make surefire, non-incremental progress in systems which can successfully and robustly learn from and interact with real humans. After several years as a researcher in natural language processing and machine learning, I've taken on this focus in cognitive science in order to better understand exactly *what it is* that we as language researchers are attempting to model in the first place.

It is slowly becoming clear what happens when artificial intelligence researchers attempt to engineer a solution before fully understanding *what it is* they are actually attempting to solve. Take the fundamentally ill-posed problems of "image recognition" and "question answering."[1] Both of these fields have likewise discovered puzzling instances of *adversarial examples:* cases in which trivial modifications to an input cause a system to produce an incorrect answer with high confidence. Such examples reveal that, contrary to superficial appearances, these systems actually fail to acquire a correct model of the world.

But what is a "correct" model of the world in the first place? **The answer to this question — and the long-horizon path to success in artificial intelligence — requires us to understand *what it is* humans are doing in the first place when they see, speak, and listen.**

The long-term solutions to massive problems like model interpretability and effective model transfer won't come from engineering hacks screwed onto existing systems. They will come first from a *proper understanding* of the tasks themselves we are attempting to solve. The goal of my research is to help specify those tasks — first by taking a close look at how humans actually acquire language, and then by building models to capture this behavior on a low level.

1. S. Carey and E. Bartlett. Acquiring a single new word. 1978. 2. J. Gauthier. On "solving language". 2016. 3. J. Gauthier and I. Mordatch. A paradigm for situated and goal-driven language learning. *NIPS Machine Intelligence Workshop*, 2016. 4. J. Johnson and et al. CLEVR. 2016. 5. L. Lucy and J. Gauthier. Are distributional representations ready for the real world? 2017. 6. S. Wang, Percy Liang, and Christopher D Manning. Learning language games through interaction. 2016.

[1] Space limits do not allow me to explain just how ill-posed these ideas are. You can read more about these ideas in [2].