My research goal is to understand the computational mechanisms that enable people to plan while building models of the world and to instantiate these mechanisms on robots. Planning and building models of the world are tightly coupled problems since agents must plan in spite of their uncertainty about how the world works. Indeed, agents must choose actions that improve their world models so that they may achieve their goals. I aim to provide an account of how this is possible by developing computational models of planning and world modeling, testing them against human behavior, and deploying them on robots.

My research principally draws upon the tools of probabilistic generative modeling, Bayesian inference, and model-based planning. Probabilistic generative models provide a flexible representation for modeling the world, Bayesian inference gives a framework for maintaining beliefs over these models, updating them as new evidence arrives, and model-based planning provides a mechanism for selecting actions to achieve a goal given the agent's beliefs. Together, these tools can provide a framework for modeling how people act while learning how the world works.

My current direction is motivated by three principles: first, that agents construct and plan over bespoke, resource-rational world models tailored to their goals; second, that these models should be probabilistic and generative, supporting both inference from evidence and forward simulation of hypothetical actions; and third, that the role of learning is to accelerate inference and planning algorithms while preserving the ability to reason online.

Model Synthesis Architecture

I propose that intelligent agents construct bespoke, resource-rational models tailored to the decisions they face. Rather than maintaining a global high-fidelity model of the world, agents should represent the world only to the extent necessary to achieve their goals. This perspective is closely related to the Model Synthesis Architecture proposed by Wong, Collins et al. (2025), which synthesizes small, structured probabilistic models in response to particular inference queries. I aim to extend this perspective towards agents that must not only answer inference queries, but also plan, act, and revise their models online on the basis of experience.

I am additionally interested in the setting where agents must model other people as part of their environment. I forward that other people should themselves be represented as agents with uncertain world models and planning processes, producing nested models of approximately rational agents as in the recursive reasoning paradigm (Chandra et al., 2025). Importantly, the resource rationality principle should apply to models of others too: the agent must determine how to model others tractably given computational constraints. The central research question is therefore how agents can construct world models tailored towards their goals without knowing how to achieve the task in advance.

Probabilistic Generative Models

I suggest that probabilistic generative models provide a natural representation for the world models synthesized by intelligent agents. Probabilistic generative models represent uncertainty over possible states and models of the world, can be used both to simulate possible futures and, when conditioned on observations, can be used to update beliefs about the world. It is important that these models can be used both for forward simulation and inference. Forward simulation is needed for planning since an agent must reason about the consequences of hypothetical actions to decide what to do, and inference is necessary for incorporating evidence about the world to update the agent's beliefs.

These capabilities become particularly important when another agent is represented within the world model. A model of another agent could assign a probability distribution over that agent's actions, allowing those actions to be generated during forward simulation and assigned likelihoods when observed. This allows an observer to condition on an agent's behavior to infer its latent beliefs. Importantly, this same property allows these models to serve as computational theories of cognition: we can use them to predict how people will act in a given situation and compare those predictions with observed human behavior. Sequential Inverse Plan Search (Zhi-Xuan et al., 2020), which models people as boundedly rational planners, provides one example of this approach.

Learning and Inference

What, then, should the role of learning be for planning and building models of the world? Undoubtedly, learning-based systems have made remarkable progress in recent years: foundation models have shown that large-scale training can produce broad and flexible capabilities across many tasks. But to provide a computational account of human intelligence, we must also explain how people learn so much from comparatively so little and generalize effectively to new situations. These principles of planning and inference provide the beginning of such an account but are often prohibitively slow.

I propose that learning should capture reusable structure that makes future inference and planning more efficient, rather than using learned systems to complete tasks end-to-end. Learning can amortize computations that recur across many problems, while inference and planning allow an agent to adapt these learned priors to the particular situations it faces (Du, 2026). I conjecture that systems which combine learning and inference resource rationally might explain how we can generalize broadly while still reasoning efficiently at inference time. One research goal is to explore this trade-off to better understand what should be learned or left for inference.

Conclusion

In the era of the increasing capacity of the foundation model, it is tempting to ask why one should not work on end-to-end model-free approaches to learning. Such approaches are certainly valuable, especially if one's goal is to optimize performance on a set of benchmarks. As scientists, however, I believe we must adopt the more ambitious goal of understanding, and not just creating, intelligent systems. By developing computational models of planning and inference that can both explain human behavior and be implemented on robots, I hope to take a step in this direction.