Prelude: The Vision AI systems are increasingly used to predict, explain, and simulate human behavior. From recommender systems to digital agents, modern models often succeed at reproducing behavioral patterns—but they do so without a principled account of how people reason, learn, or revise their beliefs. This gap has led me to a central research question that now anchors my work: ***When modeling human behavior, what cognitive structures and inductive biases—such as causal structure, belief persistence, and hypothesis update rules—must be represented explicitly, rather than approximated statistically, in order for machines to reason, learn, and adapt in human-like ways?***

Rather than treating behavior as the final object of prediction, my research approaches behavior as the surface expression of underlying cognitive processes: how people form explanations, update beliefs under intervention, and generalize across contexts. I am particularly interested in heterogeneity at the cognitive level—how differences in reasoning priors and hypothesis spaces lead different individuals to arrive at distinct judgments for principled reasons rather than noise. Understanding and modeling this diversity is essential both for cognitive science, where it reveals the structure of human reasoning, and for machine learning, where it exposes inductive biases missing from current models.

From Behavioral Prediction to Cognitive Representation My intellectual trajectory is shaped by training in design and urban systems, where I learned that observed behavior is inseparable from the cognitive and contextual structures that generate it. Translating this perspective into AI research, I became increasingly dissatisfied with dominant "demographics in, behavior out" paradigms in behavioral modeling. These approaches capture correlations, but they obscure the reasoning processes that make human behavior flexible, interpretable, and revisable under change.

This concern motivated my work at the MIT JTL Urban Mobility Lab with Prof. Jinhua Zhao, where I contributed to the NeurIPS 2025 position paper *Simulating Society Requires Simulating Thought*. We argued that modeling populations without modeling reasoning leads to brittle explanations and limited generalization. Building on ideas from cognitive science and causal modeling, I helped develop GenMinds, a framework that represents individual reasoning as structured causal graphs rather than latent vectors. These structures function as explicit hypotheses about how beliefs, values, and contextual factors generate decisions.

To test whether such hypotheses are meaningfully preserved under change, we introduced RECAP, a benchmark that evaluates reasoning fidelity through targeted interventions. Rather than measuring surface-level behavioral accuracy, RECAP asks whether a model maintains coherent reasoning hypotheses when conditions are perturbed. This work reframed my research agenda: if individual reasoning is a cognitive object, it must be represented directly, tested under perturbation, and studied as a learning problem—not inferred post hoc from behavioral aggregates. Underlying this approach is a cognitive assumption that human reasoning operates over structured, revisable causal hypotheses.

Anchoring Human Reasoning through the Think-Aloud Method To study how humans form beliefs and make decisions, I collaborated with Chance Li at the MIT City Science Center to design an adaptive interviewing chatbot inspired by the think-aloud method. Rather than treating behavior alone as data, the system treats people's articulated reasoning—their explanations, justifications, and intermediate hypotheses—as first-class observations.

Instead of static surveys, the chatbot dynamically selects questions to maximize information about a participant's underlying reasoning structure. My contributions focused on designing the interviewing logic and adaptive mechanisms that guide inquiry toward the most informative questions for each person. This work demonstrated that individual reasoning can be collected systematically at scale, while treating people as sources of explanation rather than variables. It suggests that humans maintain explicit internal hypotheses that can be probed and refined with minimal prompting.

Evaluating Individualized Belief Adaptation Collecting reasoning traces raised a deeper question: can modern models actually track an individual's beliefs as they evolve? To investigate this, I co-developed HugAgent, a benchmark designed to test individualized belief adaptation rather than population-level Theory of Mind accuracy.

HugAgent pairs interview-style reasoning traces with controlled interventions, asking whether models update beliefs consistently when new information is introduced. I contributed to task design,

experimental protocols, and analysis. Across models, we observed systematic failures: even strong LLMs often default to population-level priors rather than maintaining individual-specific belief trajectories over time. This pattern points to a failure of inductive bias rather than model capacity.

These results suggest that current models lack inductive biases for persistent, person-specific reasoning. From a cognitive perspective, this highlights a gap between human belief tracking and machine adaptation. From a machine learning perspective, it raises the question of how structured priors over reasoning hypotheses might be distilled into flexible neural models.

Scaling Cognitive Diversity through Natural Discourse While interview-based reasoning data is powerful, it is costly. To study reasoning diversity at scale, I turned to naturalistic discourse, where people spontaneously articulate judgments, justifications, and disagreements. I am currently leading SUITE, a benchmark derived from Reddit's Am I The Asshole forum, which evaluates how models interpret and adapt to heterogeneous moral reasoning.

Unlike controlled tasks, these narratives reveal how different individuals apply distinct causal frames—such as fairness, safety, or obligation—to similar situations. SUITE operationalizes this variation into a large-scale evaluation framework for probing individual differences in reasoning. Accepted as a Spotlight at the AAAI 2026 Theory of Mind Workshop, this project connects cognitive heterogeneity to challenges in generalization and representation learning.

Together, GenMinds, HugAgent, and SUITE form a coherent empirical program: measuring how reasoning varies across individuals, how it changes under intervention, and where current models fail to capture these dynamics.

Toward Prior-Enhanced and Human-Like Reasoning Models Across these projects, my work has focused on measurement: representing individual reasoning explicitly, eliciting it efficiently, and evaluating whether models preserve it under change. The next stage of my research therefore moves from evaluation to modeling.

Within Prof. Lake's lab, my research would focus on two tightly linked questions: how structured representations of individual reasoning can be distilled as explicit priors into neural models, and how those priors are revised during learning under intervention. Using my existing benchmarks as stress tests, I aim to evaluate whether prior-enhanced models better preserve individual-specific reasoning and exhibit more human-like generalization, rather than collapsing to population-level averages. Central to this effort is interpretability—understanding how hypotheses are encoded in the model, how gradient updates correspond to belief revision, and how individual priors interact with new evidence over learning.

By grounding model design in the structure of human reasoning, I seek to advance both sides of the cognitive–computational loop: using insights from human cognition to inform model design, and using model behavior as a testbed for evaluating cognitive hypotheses. Prof. Lake's lab provides the theoretical and technical framework needed to pursue this integration of cognitive theory, learning dynamics, and generalization.

References

Zhenze Mo*, Chance Jiajie Li*, Ao Qu, Yuhan Tang, Luis Alberto Alonso Pastor, Kent Larson, and Jinhua Zhao. "SUITE: Scaling Up Individualized Theory of Mind Evaluation in Large Language Models." *Advancing Artificial Intelligence through Theory of Mind Workshop @ AAAI 2026* (Spotlight).

Chance Jiajie Li*, **Zhenze Mo***, Yuhan Tang*, Ao Qu, Jiayi Wu, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Hang Jiang, Paul Pu Liang, Jinhua Zhao, Luis Alonso, and Kent Larson. "HugAgent: Benchmarking LLMs for Simulation of Individualized Human Reasoning." *PersonaLLM Workshop @ NeurIPS 2025* (Oral); *Language, Agent, and World Models (LAW) Workshop @ NeurIPS 2025* (Spotlight).

Chance Jiajie Li*, Jiayi Wu*, **Zhenze Mo**, Ao Qu, Yuhan Tang, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, Paul Pu Liang, Luis Alonso, and Kent Larson. "Simulating Society Requires Simulating Thought." *NeurIPS 2025.*

*Equal contribution.