Inverse Cognition - Inferring What Kind of Mind Someone Has:

Motivation:

Current inverse planning models (e.g. Baker et al. 2009, Jara-Ettinger et al. 2016) assume that observers interpret others' actions by inverting a rational planner, inferring the goal that would make the observed behavior most efficient. However, real people differ in not only what they want but also how they think: some are more/less impulsive, strategic, or exploratory.

This project will extend inverse planning into inverse cognition modeling, inferring an agent's cognitive architecture (rationality, exploration rate, boundedness) in addition to its goals.

Research question:

How can a Bayesian observer infer both what an agent wants and what kind of mind it has, given a sequence of actions?

Can different "cognitive styles" (ex: exploration, impulsivity, bounded rationality) be recovered from observed behavior through probabilistic inference?

Hypothesis:

Observers keep a hierarchical prior over goals and cognitive styles. They infer an agent's type M alongside its goal G given observed actions $a_{1:T}$

$$ P(G,M∣a1:T)∝P(a1:T∣G,M) P(G) P(M) $$

Here M parameterizes the agent's decision policy (inverse-temperature β in a softmax model or exploration parameter ϵ in ϵ-greedy decision making).

Generative model:

Latent variables:

G: goal state in a gridworld (ex: reach target A, B, or C).

M: cognitive style (optimal, bounded, exploratory, impulsive).

Agent model:

Chooses actions according to policy $$ \pi(a \mid s, G, M) \propto \exp!\big(\beta_M , Q(s, a; G)\big) $$

Observer model:

Uses Bayes' rule to infer $P(G,M∣a1:T)$ given observed actions.

Simulation plan:

Generate synthetic behavior for agents with different M values ($\beta = 1, 3, 10, $ etc.) in a simple gridworld or navigation task.

Compute the posterior $P(G,Mla1:T)$ via importance sampling or MCMC.

Evaluate:

Identifiability:

Can the observer recover the correct MMM?

Human comparison:

Present short clips to human participants, ask them to label the actor as “impulsive,” “curious,” or “strategic,” and compare to model posteriors.

Expected contribution:

This will create a formal framework for inverse cognition modeling: inferring minds, not just goals.

Demonstrates how inverse planning extends to meta-inference about cognitive style, bridging rational analysis and theory of mind psychology.