Inferring Minds, Not Just Goals: Limits of Inverse Cognition from Behavior

Dominic Gearing
Yale University
dominic.gearing@yale.edu

Abstract

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 sought to extend inverse planning into inverse cognition modeling, inferring an agent's cognitive architecture in addition to its goals. This problem was simulated using agents navigating a gridworld with hidden goals and rationality parameters (β). An "observer" then performed Bayesian inference over both goal and mind-types. Amortized neural inference as an approximation and additional analysis. The results validated previous work on inverse planning models by achieving near-perfect goal inference (˜99% accuracy). Mind-type was inferred at rates above chance, but was systematically limited. Posterior uncertainty over β remained, even as trajectory length increased. These results suggest that inverse cognition is possible, but imply that it may be fundamentally coarser than inverse planning. We propose one potential explanation for this result: minds may be inferred as regions in latent space, rather than precise parameters.

1 Introduction

Previous research on inverse planning asserts that observers infer an agent's goal by assuming that they will choose actions that are approximately rational with respect to their goals. Instead of directly reading goals from behavior, observers invert a generative model of action selection. They observe actions and infer a goal that rationalizes those actions. The major assumption is that agents act efficiently, given a set of beliefs and constraints, and this is used to perform probabilistic inference over goals.

In inverse planning, the observer assumes the agent selects their actions based on a planning policy conditioned on a goal G:

$$\pi(a \mid s, G) \propto \exp(\beta Q(s, a; G))$$

Where

- $Q(s, a; G)$ measures the expected progress toward goal G
- β captures the optimality of an action
- higher-value actions are exponentially more likely

Then, given an observed action sequence $a_{1:T}$, the observer computes:

$$P(G \mid a_{1:T}) \propto P(a_{1:T} \mid G)\, P(G)$$

The inversion from action back to goals is the defining feature of inverse planning. However, the standard inverse planning framework holds planning fixed and treats goals as the only latent variable.

Baker et al. (2009) showed that adults and infants interpret actions by reasoning backwards from efficiency. Observers in this experiment only expected agents to take longer paths when there were obstacles present to justify the less efficient movement. Goal inferences changed when environmental constraints changed, even if the same actions were maintained. This suggests that observers interpret actions relative to what is possible in the environment. The same action can therefore imply different goals depending on environmental circumstances. This demonstrates that observers implicitly represent the agent's planning problem.

Jara-Ettinger et al. (2016) extended these findings to also model cost-reward tradeoffs on top of goals. They modeled what agents want, and additionally how much they wanted a given goal. The results suggest that people expect agents to exert more effort for higher-value goals, and that people can reliably distinguish between inefficient actions that are due to high reward versus irrationality.

Both of these experiments assume a specific planning algorithm shared across different agents, known rationality parameters, and assume that differences in behavior come from differences in goals. This ignores any sort of differences in how agents actually plan (ie their personalities or planning styles). One agent could be more impulsive, noisy, or prefer a specific ratio of exploration versus exploitation. Observers assume inefficiency come from goals, and not necessarily from differences in mind.

These inverse planning models have been incredibly successful in explaining how observers infer what an agent wants, but assume they all plan in the same way. However, real agents differ in both their goals and how they go about achieving them. Agents have different preferences for consistency and strategy in their actions. This raises a question: can observers infer not only the goal of the agent, but also the kind of mind that may have produced that behavior?

This paper introduces a probabilistic framework for inferring cognitive style from actions:

$$P(G, M \mid a_{1:T}) \propto P(a_{1:T} \mid G, M)P(G)P(M)$$

Where:

- G: agent's goal
- M: mind-type (operationalized as rationality parameter β)
- $a_{1:T}$: observed action sequence

In this framework, the observer inverts the agents policy, and use both the agent's possible goals and an inferred mind type to jointly explain an agent's behavior. This project contributes a concrete implementation of inverse cognition, provides the probabilistic toolkit to implement this through exact Bayesian inference over goals and mind-types, and empirically demonstrates some of the potential limits of this framework. We also compare exact inference and amortized inference frameworks for inferring mind-type, and explore the implications of this model.

2 Methods

This project utilized a discrete gridworld made up of a 5x5 grid with an impassable wall at the center position (3,3) and two potential goal locations. Each episode was made up of an agent who started at a fixed location (bottom-left corner (1,1)), with the episode ending when the goal was reached or when the maximum step threshold was exceeded. The goals were located in two potential locations:

- $G = A$: top-right corner (5,5)
- $G = B$: top-left corner (1,5)

The goal was latent to the observer but known by the agent. Actions consisted of four deterministic directions,

$$\{\text{UP}, \text{DOWN}, \text{LEFT}, \text{RIGHT}\}$$

and actions that moved the agent into a wall or outside of the grid world left the agent stationary.

It was important to have two goals that induced qualitatively different trajectories for the agent in this paradigm, in order to make goal inference easier. The wall introduced planning constraints, to make sure that behavior was interpreted relative to the environment. Due to the fact that the task was relatively simple and known to the observer, ambiguity in inferring β was likely not due to task complexity. Instead, it reflected an identifiability limit to the model in inferring cognitive style.

The generative agent model was made up of latent variables

- Goal $G \in \{A, B\}$
- Rationality parameter $\beta \in \{0.5, 1.0, 3.0, 8.0\}$

An action-value function:

$$Q(s, a; G) = -\text{Manhattan distance to the goal after an action},$$

And a Boltzmann softmax policy:

$$\pi(a \mid s, G, \beta) \propto \exp(\beta Q(s, a; G))$$

Low β is the inverse temperature parameter, which controls how strongly action probabilities depend on their relative values. The Q-value encodes how good each action is for reaching a specific goal, and the β scales these values before they are exponentiated. That in turn determines how precisely the policy favors the best potential action. In this paradigm, low β values are interpreted as noisy/impulsive cognitive styles, and high β values are interpreted as near deterministic or rational planners. This definition of cognitive styles led to overlapping likelihoods, which supports why the model was good at identifying the more extreme β values but struggled to differentiate mid-range βs. This systematic confusion of β values at mid ranges does not disappear with greater amounts of data.

Data for this project was synthetically generated from a known generative model. This allowed for evaluation of inverse inference under near-ideal conditions, because the observer's assumptions perfectly matched the agent's actual decision process. Because of this, we were able to assume that any residual uncertainty in inference might actually reflect fundamental limits to the model, instead of a more trivial model mismatch or observational noise. Each episode was generated by sampling a pair of latent variables that included

- The goal $G \in \{A, B\}$,
- And Mind-type (rationality parameter) $\beta \in \{0.5, 1.0, 3.0, 8.0\}$,

Which controlled the randomness of the agent's policy through a softmax decision rule.

The dataset was then constructed by counting all of the possible combinations of goals and β values. For each pair of goals and β's (G, β), a number of independent episodes were generated. This captured any within-condition behavioral variability that came from the randomness of action selection. This led to a full dataset $\{G\} \times \{\beta\}$.

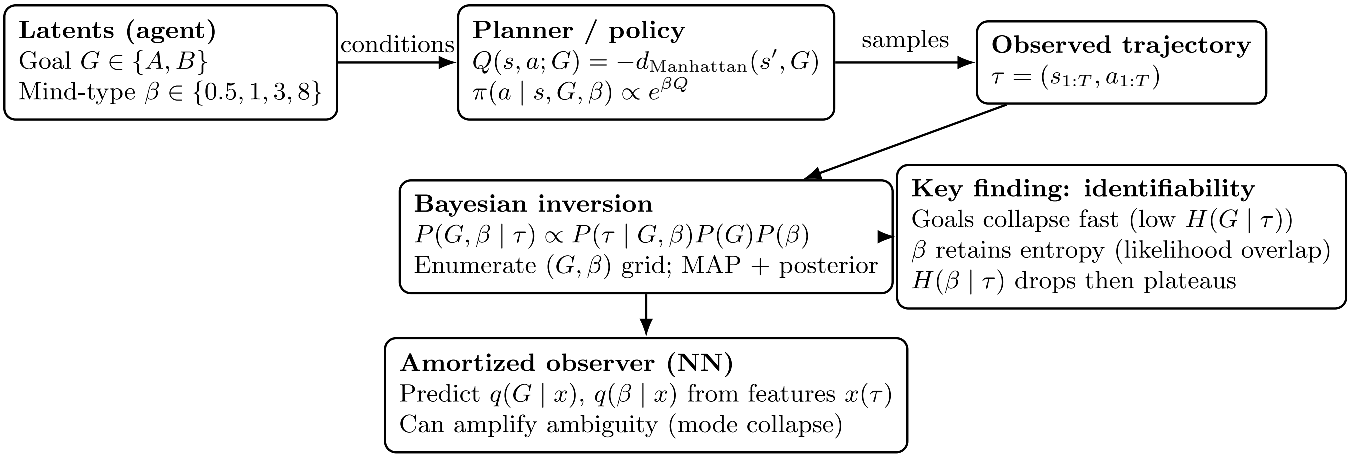

Figure 1: **Unifying schematic.** Forward: latent goal G and mind-type β generate behavior via softmax planning. Inverse: a Bayesian observer infers $P(G, \beta \mid \tau)$. Empirically, goal posteriors collapse quickly, but β remains partially ambiguous due to likelihood overlap, producing systematic mid-range confusions and saturation of information gain. An amortized observer approximates inference but can amplify this ambiguity.

To simulate an individual episode, the environment was initialized (a 5x5 gridworld with a single wall at the center of the grid, agent begins from buttom left corner (1,1)), and the agent was assigned a goal G. Then, the action-value function $Q(s, a; G)$ was computed for all of the states and actions. The agents policy was then defined as

$$\pi(a \mid s, G, \beta) = \frac{\exp(\beta Q(s, a; G))}{\sum_{a'} \exp(\beta Q(s, a'; G))}$$

At each timestep t, the agent samples an action from the policy $\pi(a \mid s_t, G, \beta)$, and the environment deterministically updated the state based on the chosen action. The episode was over (terminated) when the agent reaches its goal, or the maximum step limit was reached. Due to the randomness of action selection, repeated episodes with the same pair of goal and mind-type (G, β) could theoretically produce different trajectories, despite the underlying goal and mind-type being identical.

For each episode, State trajectory $s_{1:T}$ (the full sequence of visited gridworld states, Action sequence $a_{1:T}$ (the sequence of actions the agent takes), and latent ground truth (the true goal G and the true rationality parameter β) were recorded and stored explicitly, which allowed posterior inference results to be compared to the known ground truth.

Importantly, the observer uses the same model class as the agent does. The observer assumes the same state space, action space, and transition dynamics, and evaluates action likelihoods using the same Boltzmann softmax policy. The observer's hypothesis space includes the true values of G and β. Because of this, the inference problem is well specified, and failures to approximate β can't be due to bad assumptions, approximation error, or a lack of expressivity in the observer model. The leftover uncertainty instead is representative of the overlap in similar βs likelihood functions:

$P(a_{1:T} \mid G, \beta_1) \approx P(a_{1:T} \mid G, \beta_2)$ for multiple trajectories.

This is important to the claim that our results reveal structural limits on inverse cognition rather than simply being a result of the specific implementation.

The Bayesian Observer model used was:

Likelihood:

$$P(a_{1:T} \mid G, \beta) = \prod_{t=1}^{T} \pi(a_t \mid s_t, G, \beta)$$

Posterior:

$$P(G, \beta \mid a_{1:T}) \propto P(a_{1:T} \mid G, \beta)P(G)P(\beta)$$

Accuracy of the model was tested on the test set of a train-test split of the synthetically generated data using goal and β accuracy, as well as a joint goal, β accuracy measure. Confusion matrices revealed any structured errors. We also extended the framework past exact enumeration for inference to also include amortized inference, as a closer approximation to what humans actually do. For this, a neural network was trained to predict posterior marginals, taking trajectory features and goal conditioned regret statistics as inputs. The model output was $q(G \mid \tau)$ and $q(\beta \mid \tau)$, and the posterior was factorized.

3 Results

Our model achieved near perfect Goal accuracy (˜0.99), an average β accuracy of (˜0.62) and a joint accuracy of ˜(0.64). This reflects the previous research showing that Goal inference is almost fully solved by inverse planning frameworks. The substantially lower but nontrivial accuracy of β prediction reflects that inferring mind-type is

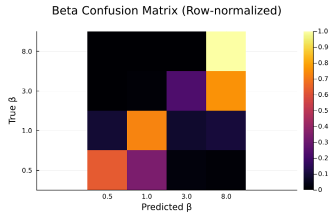

Figure 2: Confusion matrix of $\beta predictions$

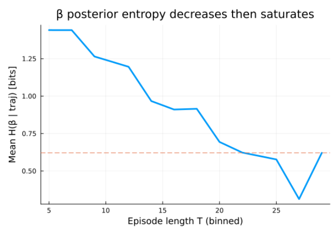

Figure 3: $\beta posteriorentropy$

a substantially harder task than goal inference. Further analysis of the confusion matrices confirm these results. Goal confusion matrices were almost perfectly diagonal, highlighting the high level of accuracy of Goal inference. The β confusion matrices, however, systematically confused the mid-range β values (1.0, 3.0), but easily identified the correct mind-type for extreme β values (0.5, 8.0). This shows that errors in mind-type prediction are not random or noise, but are instead structured and informative of the limitations of our model on inverse cognition modeling.

Entropy plots show that $H(\beta \mid \tau)$ decreases with T, and that there is large variance at short lengths T. After 20-25 steps, there is clear saturation where entropy no longer increases with increased information and T. This highlights that more data may help, but only up to a limit.

Additionally, mutual information analysis, defined by

$$I(\beta; \tau) = H(\beta) - H(\beta \mid \tau)$$

Showed that information increases sublinearly, and that it plateaus well below full identifiability. This suggests that behavior by itself contains limited information about the mind-type of the agent.

Finally, amortized inference maintained Goal accuracy (0.97), but saw a drop in β accuracy (0.37). This suggests

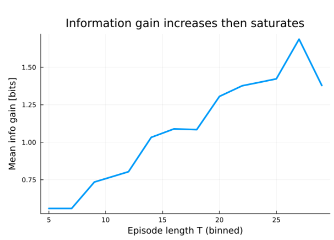

Figure 4: information gained

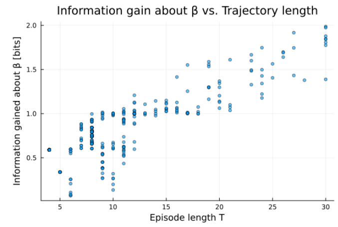

Figure 5: Information gained

that the neural network inherited some of the posterior ambiguity, and that amortization biases the model and leads to mode collapse. This low β accuracy reflects true uncertainty in the model, rather than a failure or design flaw in the model itself.

4 Discussion

Inverse cognition as an extension of inverse planning models is a promising method of inferring what kind of mind someone else has based on their actions and goals. It is possible, but may be fundamentally coarser than inverse planning. In this paradigm, minds are inferred as equivalence classes or regions of latent space, not as precise parameters. This shows the limits of the likelihood geometry for this sort of work. Even ideal bayesian observers are constrained by their goals and possible actions. This strengthens the theory that humans exhibit coarse trait attribution, and aligns with resource-rational accounts of cognition modeling. This also aligns with the intuitive experience of inferring broad categories like “impulsive” vs “deliberate”, as opposed to precise rationality parameters, suggesting that social cognition really does operate at a coarse-grained level, as compared to action planning or goal inference.

One major limitation of this project is that cognitive style is operationalized using a finite set of inverse-temperature values. This discretization enables exact Bayesian inference and clean interpretability, but it simplifies what is very likely a continuous and multidimensional space of cognitive variation in real agents. Human differences in impulsivity, consistency, and planning precision are probably not clustered around a small number of discrete β values. Discretization sharpens the distinctions between mind types. If the more fine grained distractions are already difficult to recover under a discrete hypothesis, they are probably even more difficult to identify in a continuous setting. Therefore, the observed residual posterior entropy over β probably underestimates the true ambiguity that is present in more realistic cognition models.

Another limitation is that the observer is assumed to know the entire environment, including all state transitions, obstacles, and goal locations. This design removes important uncertainty that is present in real-world social inference, where observers often have to reason about what others believe, how much they know, or if their world models are incorrect in any way. This makes the inference problem in this paradigm intentionally idealized. Eliminating environmental uncertainty ensured that limits on mind-type inference observed weren't attributed to not knowing the task structure, but instead represent ambiguity in how cognitive parameters map onto their associated behaviors, even under the best possible conditions.

Further, performance on our inverse cognition framework could be limited by the fact that all of the runs were conducted on the same navigation task in the same gridworld environment. While this was ok for the scope of this project, it would be interesting to see how observing a given agent under a range of different tasks and environments might influence an observer's ability to accurately infer the mind-type of the agent. Certain mind-type parameters may only be identifiable under specific tasks/environments, or may be more easily inferred given a range of conditions to observe the agent under. The conclusions drawn from this project are likely best interpreted as task-conditional. Identifying the mind-type of an agent is probably not an intrinsic property of just the agent, but also of the interaction between agent and task.

It would be interesting to model mind-type as a continuous latent variable or as a vector of cognitive traits in future work. This would allow researchers to investigate whether posterior uncertainty collapses along certain dimensions, or stays broad along others. These models would also help with the analysis of specific regions or manifolds of cognitive space that are indistinguishable from behavior alone. This reframes mind inference as a partitioning of latent space into equivalence classes induced by behavior instead of a recovery of parameters.

Finally, it would be more realistic and interesting to look at richer cognitive models that vary in their behavioral differences in more dimensions than just the randomness level of their actions. More realistic architectures could incorporate parameters like planning depth, lookahead horizon, differing explicit exploration mechanisms or algorithms, heuristics planning, or resource bounded approximations of value computation. This would allow researchers to identify traits that are easier to infer than others, and if there are traits that trade off in how they shape an agents behavior.

5 Conclusion

This project sought to extend inverse planning into inverse cognition modeling, inferring an agent's cognitive architecture in addition to its goals. Using a gridworld and inverse cognition inference, an observer was able to infer better than chance the mind-type of an agent navigating this world towards a specific goal, with specific limitations in the performance of this mind-type inference. While able to identify extreme mind-types well under specific conditions, the observer struggled with differentiating mid-level mind-types due to their overlapping likelihoods. This suggests a fundamental difference in mind-type inference compared to simple goal inference from inverse-planning, with mind-type inference being more coarse-grained and categorical. Future research should expand on the frameworks and ideas presented in this project by testing different mind-type parameters outside of randomness, making the environment and tasks more complex and realistic and increasing variation, and weakening assumptions around the knowledge of the agent/observer. Overall, the findings support the idea that inverse cognition is possible, but limited. This work could provide valuable information about what kinds of mental properties are behaviorally identifiable and which are not. This aligns with the broader goals of cognitive science of understanding cognition as more than just an idealized computation, but also as an inference problem that is constrained by the structure of the world, agent, and observer.

See additional figures and results on Github at: https://github.com/domgearing/Inverse$_Cognition$

6 References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. Cognition, 113(3), 329–349.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. Trends in Cognitive Sciences, 20(8), 589–604.