Limitations of Agents Simulated By Predictive Models

Raymond Douglas Jacek Karwowski Chan Bae Andis Draguns Victoria Krakovna

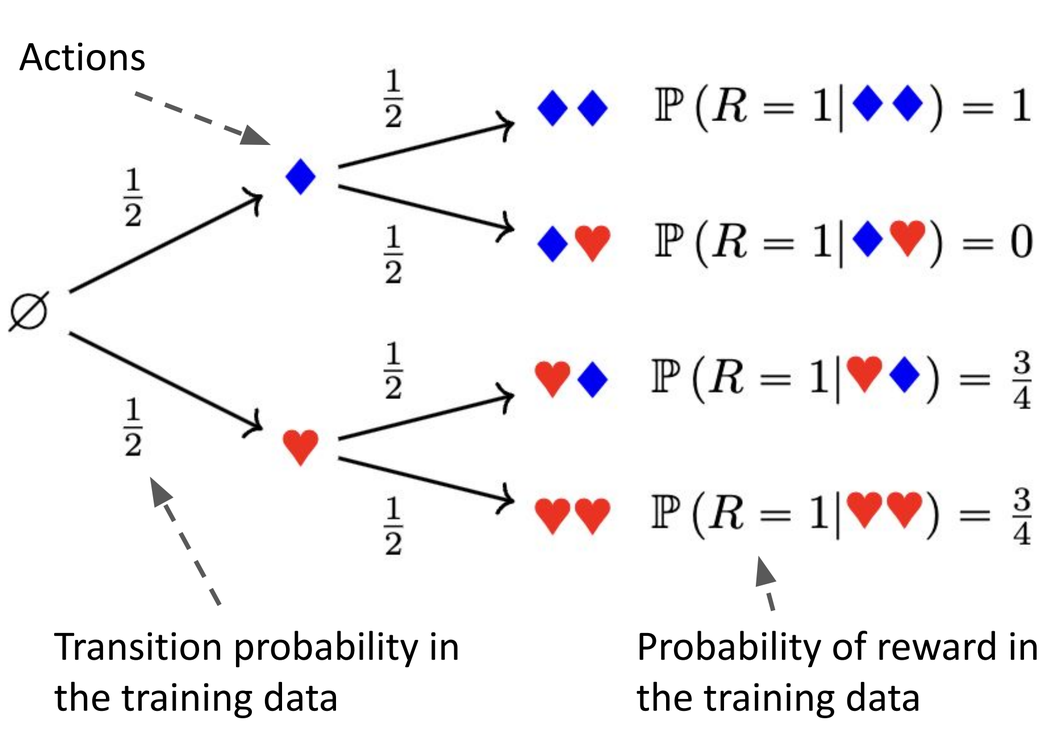
Conditioning on goals does not guarantee optimal policies

LLM-based agents are fundamentally made from ***predictive models***, and one way to steer them towards goals is to condition the predictive model on a goal and roll out the predicted actions autoregressively.

But conditioning an optimal predictive model on a goal does not guarantee an optimal policy for that goal.

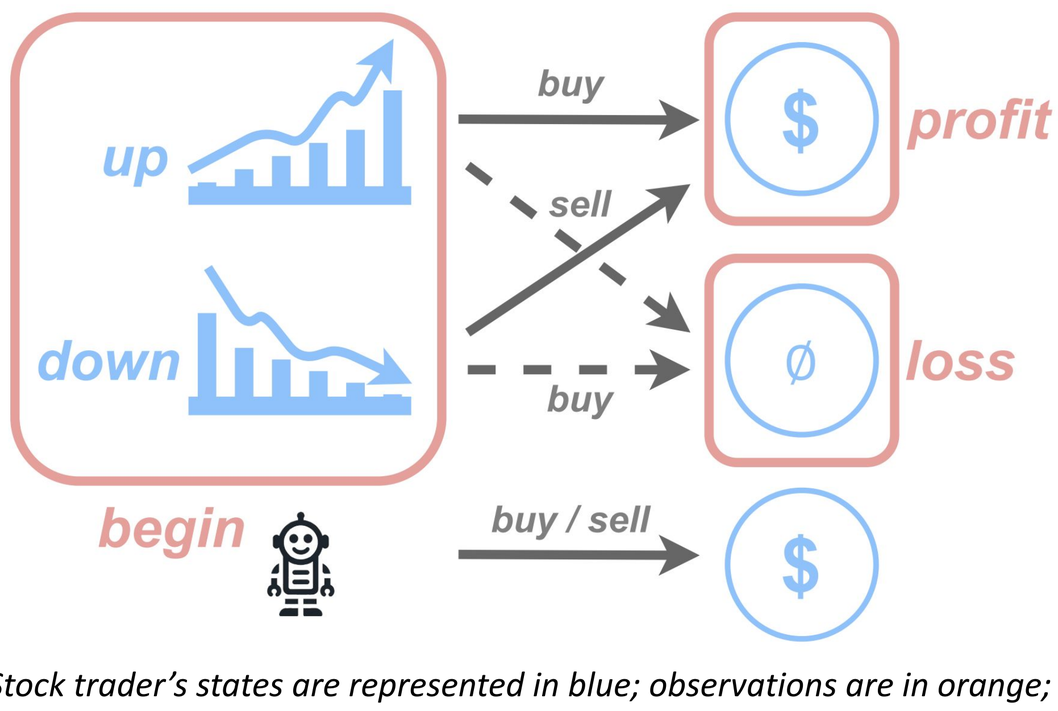
- **What we want:** The *policy* which *causes the goal* to have high probability
- **What we get:** The *action* which *is itself* highly probable *given the goal*
- **Where these diverge:** When the training data has either of the following:
  - A variety of policies with different chances of causing the goal
  - Hidden observations which guided the generation of the policies

Example: Predictor-Policy Incoherence



- The pair of actions ♦♦ has the highest probability of reward
- However, the first action *most likely given* the reward is ♥: $\mathbb{P}(♥|R) = \frac{3}{5}$, $\mathbb{P}(♦|R) = \frac{2}{5}$
- So a predictive model conditioned on the reward and autoregressively sampled from will begin with ♥ not ♦, performing suboptimally
- It occurs despite the predictive model being an optimal predictor of the training data

Example: Latent Variable Confounding



Stock trader's states are represented in blue; observations are in orange; lines show actions and associated (deterministic) transitions. Dotted lines are transitions which the expert is never observed to take

- In data generation, the agent observes the true state, and picks whichever action leads to profit
- But the predictive model is trained on data which does not distinguish between the starting states
- So it simulates an agent that will behave *randomly*, and predicts that the induced agent will *always profit*. In fact, the agent will profit only half the time.