**What the paper is doing:**

This paper explores how reinforcement learning (RL) operates differently in social versus non-social contexts, specifically examining the role of trait inference—the process of attributing stable personality characteristics (e.g., generosity)—in guiding decision-making.

Questions:

1. Could trait inference be modeled more explicitly in the RL architecture? For example, can one extend the state representation to include inferred traits as latent variables, and then update them with RL approaches?
2. Could this framework be applied to understanding social biases or stereotypes? Since trait inference is influenced by perceived social categories, is there a risk that reinforcement learning in social contexts could reinforce prejudices based on inferred (but not actual) traits?