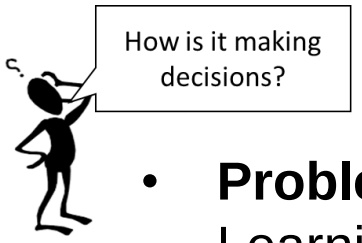
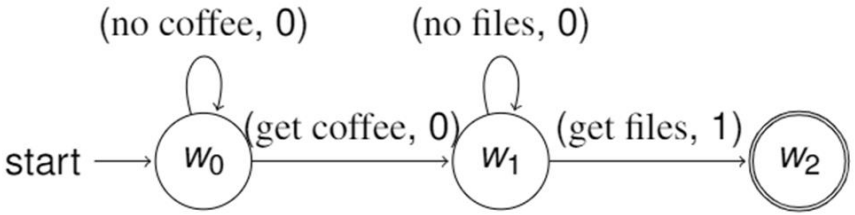
Active Automaton Inference for Reinforcement Learning using Queries and Counterexamples



- **Problem:** *Hard to explain* how Deep Reinforcement Learning (RL) agents make decisions.

- **Solution**: Having an agent distill knowledge in a finite reward automaton (FRA) makes the RL system more **interpretable** and data efficient.



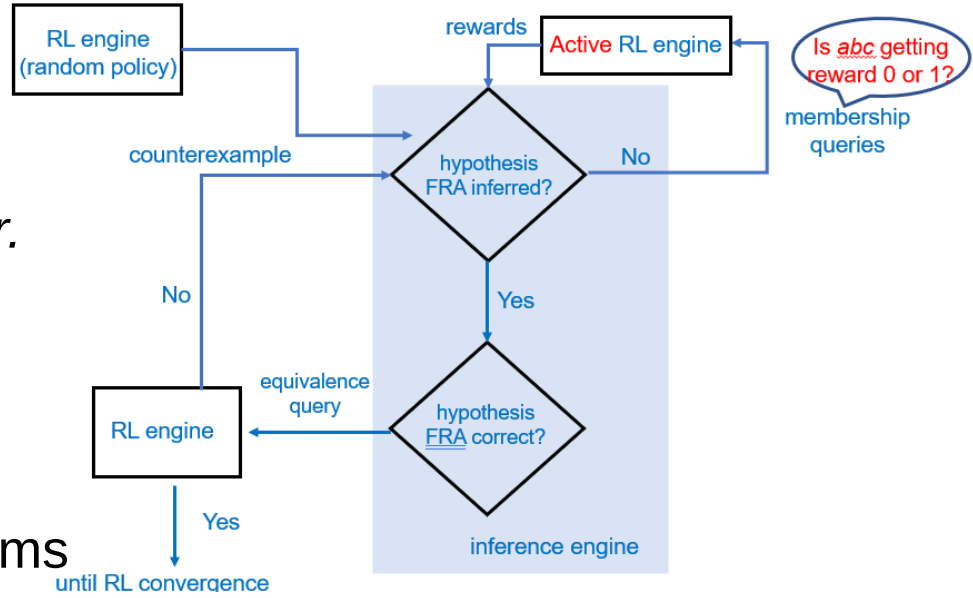
Example FRA

- **This abstract**:

  - Propose a new type of approach that allows an agent to actively infer non-Markovian reward functions *faster.*

  - Prove the expressivity of finite reward automation for non-Markovian rewards.

  - Compare new approach to state-of-the-art RL algorithms (LRM, JIRP, PPO2).



AFRAI-RL Block Diagram