Trustworthy AI

Towards trustworthy decision making leveraging Formal Methods and RL
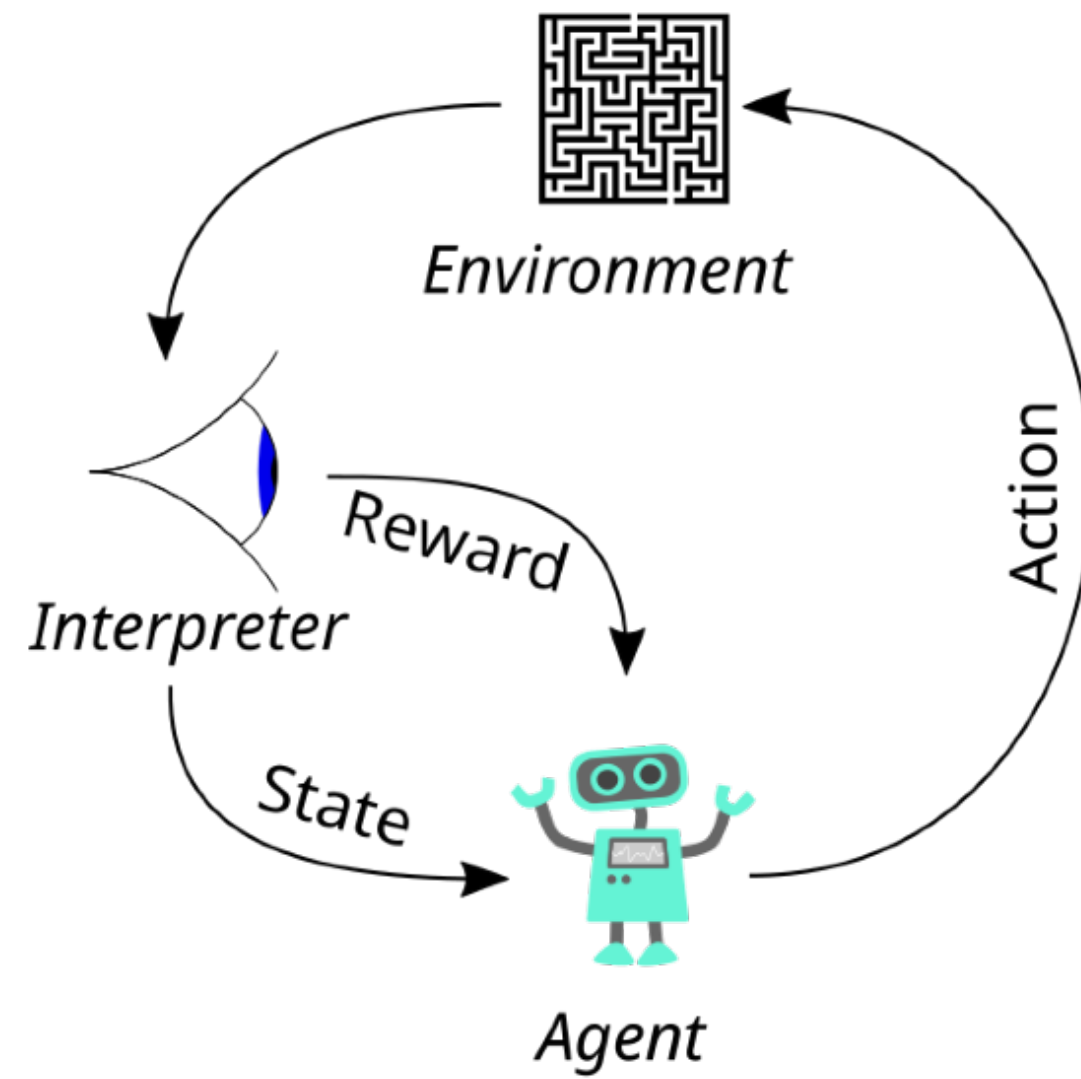
Specification Inference + RL

Verification
`Ensure controller satisfies the goal'

Adaptive Reactive Synthesis + RL

Reward Machines

Reinforcement Learning

Environment

Action

Reward

Interpreter

State

Agent

The feedback is in form of rewards!

Hard for sequential or long-horizon tasks!