On the Broad Implications of Reinforcement Learning based AGI

SCOTT LIVINGSTON, JAMIE GARVEY, ITAMAR ELHANANY

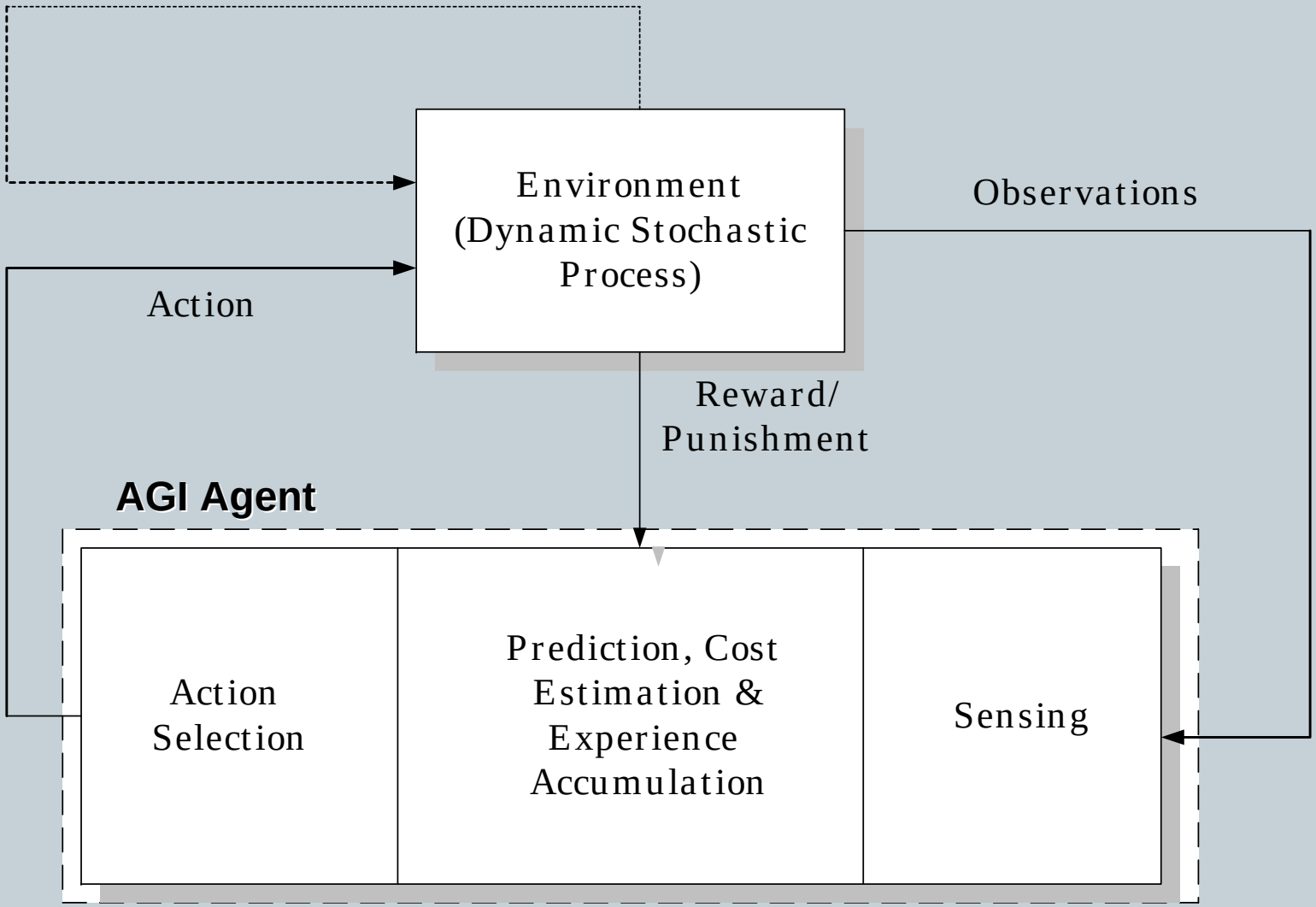
MACHINE INTELLIGENCE LAB
THE UNIVERSITY OF TENNESSEE

> mil.engr.utk.edu <

Outline

- Reinforcement Learning (RL) **is** a viable approach to facilitate AGI
- What is missing in current RL theory?
- Assuming RL-based AGI, what are some ethical and practical implications?

The Reinforcement Learning Framework



What's missing in current RL theory?

- ***Rich state representation*** – partial observability; dealing with high dimensional inputs and multi-scale spatio-temporal dependencies
- ***Reward function*** – multi-dimensional (vector-based) reward measure
- ***Action selection*** – high-dimensional, complex state-action associations
- ***Value function*** – shift away from sum of (infinite) discounted rewards; needs to truly reflect the future prospect spanning different time scales

Can RL-based AGI development be controlled?

- We can limit available resources
  - Computing power
  - Internal models
  - Architected reward definitions
  - RL development can be restricted
- RL-AGI's capabilities may be limited due to …
  - Cost limitations
  - Application goals (i.e. virtual pets)
  - Pragmatic concerns

Is limiting AGI intelligence morally permissible?

- ***Intelligence threshold*** below which one cannot recognize its own limitations and above which one can imagine its growth potential
- Limiting capacity below this threshold is deemed morally permissible
  - Since agent is unaware of its lost potential
- Once AGI surpasses this threshold, further limitations are immoral
  - Development of hierarchical goals, including desire for higher intelligence

Legal and Moral Status of AGI agents

- AGI considered an end in itself rather than a means/tool for human purposes
- AGI entrance into human moral communities
- Acceptance of AGI may lead to extra-human moral theories

Closing Thoughts …

- Several limitations persist for facilitating RL-based AGI
- Resource limitations may be used as growth control
- Such control may be considered immoral
  - Intelligence threshold
  - Agent developed hierarchical goals
- **Now** is the time to consider moral and legal implications of AGI