Agents And Abstraction

Arnav Gupta

October 10, 2024

Contents

1 Agents **1**

2 Knowledge Representation **2**

3 Dimensions of Complexity **2**
3.1 Planning Horizon 3
3.2 Representation 3
3.3 Computational Limits 3
3.4 Learning from Experience* 3
3.5 Uncertainty 3
3.5.1 Uncertain World Dynamics 4
3.6 Goals or Complex Preferences 4
3.7 Reasoning by Number of Agents 4

1 Agents

Agent: an entity that performs actions in its environment

The agent and environment make up the world. The black box that comprises the agent contains the belief state.

Agents can have properties such as:

- **abilities**: what actions can the agent perform
- **stimuli**: what and how can the agent sense in its environment
- **prior knowledge**: knowledge about agent capabilities and environment

- **past experience**: which and when actions are useful
- **goals**: what the agent must achieve

2 Knowledge Representation

Knowledge: info used to solve tasks

Representation: data structures used to encode knowledge

Knowledge Base (KB): representation of all knowledge

Model: relationship of KB to world

Level of Abstraction: model accuracy level

For AI agents:

1. specify what needs to be computed
2. specify how the world works
3. agent figures out how to do the computation

3 Dimensions of Complexity

There are 9 dimensions of complexity that make up the agent design space:

1. **Modularity**: flat → modular → hierarchical
2. **Planning Horizon**: non-planning → finite horizon → indefinite horizon → infinite horizon
3. **Representation**: explicit states → features → individuals and relations
4. **Computational Limits**: perfect rationality → bounded rationality
5. **Learning**: knowledge is given → knowledge is learned
6. **Uncertainty**: fully observable → partially observable
   (a) world dynamics: deterministic → stochastic
7. **Preference**: goals → complex preferences
8. **Reasoning by Number of Agents**: single agent → adversarial → multi-agent

9. **Interactivity**: offline → online

3.1 Planning Horizon

How far the agent looks into the future when deciding what to do:

- **static**: world does not change
- **finite horizon**: agent reasons about a fixed finite number of time steps
- **indefinite horizon**: agent is reasoning about finite, but not predetermined, number of time steps
- **infinite horizon**: agent plans to go on forever (process oriented)

3.2 Representation

AI is about finding compact representations and exploiting compactness for computational gains. An agent can reason in terms of:

- **explicit states**: state directly represents one way the world could be
- **features/propositions**: describe states in terms of features
- **individuals/relations**: a feature for each relationship on each tuple of individuals

3.3 Computational Limits

- **perfect rationality**: agent always chooses the optimal solution
- **bounded rationality**: agent chooses an action given its limited computational capacity
  - satisficing solution is good enough
  - approximately optimal solution is has some defined acceptable error

3.4 Learning from Experience*

Knowledge is either **given** or **learned** from data/past experience.

3.5 Uncertainty

What about the state the agent can determine from observations:

- **fully observable**: the agent knows the state of the world from observations
- **partially observable**: can be many possible states given an observation

3.5.1 Uncertain World Dynamics

If the agent knew initial states and actions, is the resulting state known?

Dynamics can be:

- **deterministic**: state resulting from carrying out an action in state is determined from the action and the state
- **stochastic**: uncertainty over states resulting from executing a given action in a given state

3.6 Goals or Complex Preferences

Achievement Goal: goal to achieve, can be complex logical formula

Maintenance Goal: goal to be maintained

Complex Preferences: may involve trade-offs between desiderata, can be ordinal or cardinal

3.7 Reasoning by Number of Agents

Single Agent Reasoning: an agent assumes that any other agents are part of the environment

Adversarial Reasoning: considers other agents to be opposition

Multi-agent Reasoning: an agent needs to reason strategically about the reasoning of other agents

Agents can be cooperative, competitive, or have independent goals.