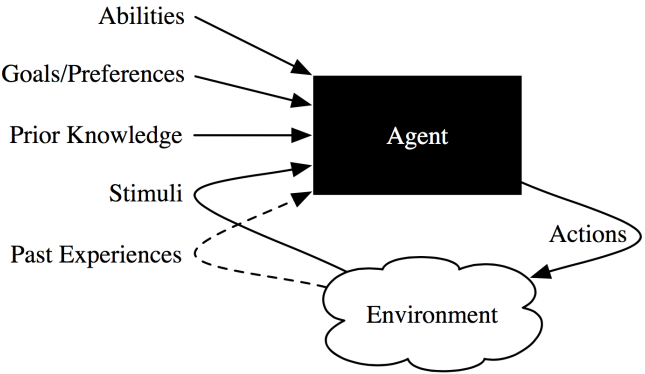
Lecture 2 - Agents and Abstraction

Jesse Hoey
School of Computer Science
University of Waterloo

May 4, 2022

Readings: Poole & Mackworth 1.3-1.10

Situated Agent



- Agent+Environment= world
- Inside black box: belief state

Knowledge Representation

- non-AI:
  - specify how to compute something
  - specify what the next step is
  - programmer figures out how to do the computation
- AI:
  - specify what needs to be computed
  - specify how the world works
  - agent figures out how to do the computation
- Knowledge: information 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: How accurate is the model

Representations: Symbol Systems

- A symbol is a meaningful physical pattern that can be manipulated.
- A symbol system creates, copies, modifies and destroys symbols.

physical symbol system hypothesis (Newell & Simon, 1976):

> *A physical symbol system has the necessary and sufficient means for general intelligent action.*

implies that : AI on a computer is possible in theory, but not necessarily feasible in practice
most connectionist approaches are still symbolic at their core

Searle's Chinese Room

- What would you do?

Searle's Chinese Room

- What would you do?
- Start to make mistakes
- Look for correlations in subsequent inputs
- Establish a secondary communication based on the symbols
- but what are these correlations?
- psychology studies: 96% of samples come from 12% of the world (Henrich)
- understanding what's outside the Chinese room is understanding different cultures

Knowledge Representation

A good representation should be

- Rich enough to express the problem
- Close to the problem: compact, natural and maintainable
- Amenable to efficient computation
- Amenable to elicitation from people, data and experiences

also (not in book):

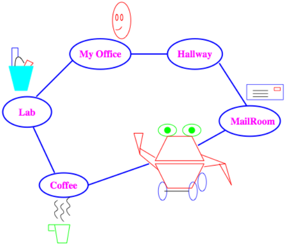
- explainable to humans

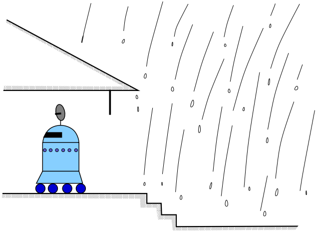
Four Example Application Domains (From Book)

- Autonomous delivery robot roams around an office environment and delivers coffee, parcels,. . .
- Diagnostic assistant helps a human troubleshoot problems and suggests repairs or treatments. E.g., electrical problems, medical diagnosis.
- Intelligent tutoring system teaches students in some subject area.
- Trading agent buys goods and services on your behalf.

Let's talk about the Autonomous Delivery Robot

Domain for Delivery Robot





robot must:

- deliver coffee & mail when needed
- avoid getting wet

Autonomous Delivery Robot

- Abilities: movement, speech, pickup and place objects, sense weather
- Observations: about its environment from cameras, sonar, sound, laser range finders, or keyboards.
- Prior knowledge: its capabilities, objects it may encounter, maps.
- Past experience: which actions are useful and when, what objects are there, how its actions affect its position.
- Goals: what it needs to deliver and when, tradeoffs between acting quickly and acting safely, effects of getting wet.

What does the Delivery Robot need to do?

- Determine where user is. Where coffee is. . .
- Find a path between locations.
- Plan how to carry out multiple tasks.
- Make default assumptions about where user is.
- Make tradeoffs under uncertainty: should it go near the stairs or outside?
- Learn from experience.
- Sense and act in the world, avoid obstacles, pickup and put down coffee, deliver mail

Dimensions of Complexity

- Research proceeds by making simplifying assumptions, and gradually reducing them.
- Each simplifying assumption gives a dimension of complexity
  - Can be multiple values in a dimension: values go from simple to complex
  - Simplifying assumptions can be relaxed in various combinations
- Much of the history of AI can be seen as starting from the simple and adding in complexity in some of these dimensions.

Dimensions of Complexity

- Flat $\rightarrow$ modular $\rightarrow$ hierarchical
- Explicit states $\rightarrow$ features $\rightarrow$ objects and relations
- Static $\rightarrow$ finite stage $\rightarrow$ indefinite stage $\rightarrow$ infinite stage
- Fully observable $\rightarrow$ partially observable
- Deterministic $\rightarrow$ stochastic dynamics
- Goals $\rightarrow$ complex preferences
- Single-agent $\rightarrow$ multiple agents
- Knowledge is given $\rightarrow$ knowledge is learned from experience
- Perfect rationality $\rightarrow$ bounded rationality

Succinctness and Expressiveness

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

- explicit states — a state is one way the world could be

Succinctness and Expressiveness

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

- explicit states — a state is one way the world could be
- features or propositions.
  - It's often more natural to describe states in terms of features.
  - 30 binary features can represent $2^{30} = 1,073,741,824$ states.

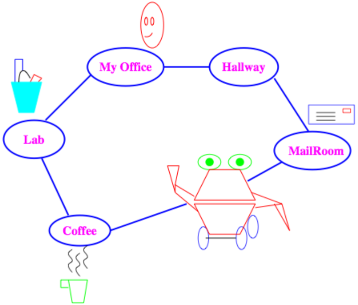
Succinctness and Expressiveness

Much of modern AI is about finding compact representations and exploiting that compactness for computational gains.

A agent can reason in terms of:

- explicit states — a state is one way the world could be
- features or propositions.
  - It's often more natural to describe states in terms of features.
  - 30 binary features can represent $2^{30} = 1,073,741,824$ states.
- individuals and relations
  - There is a feature for each relationship on each tuple of individuals.
  - Often we can reason without knowing the individuals or when there are infinitely many individuals.

Example: Delivery Robot



- Explicit: enumeration of all worlds: s1,s2,s2,...
- Features: robot location, user location, robot has coffee?, ...
- Relations: robot moves (clockwise + or counter-clockwise -)
  $\forall m \in \{+, -\}, l \in \{1, 2, 3...\}, move(m) : l' \leftarrow (l \ m \ 1)\%5$

Planning horizon

...how far the agent looks into the future when deciding what to do.

- Static: world does not change
- Finite stage: agent reasons about a fixed finite number of time steps
- Indefinite stage: agent is reasoning about finite, but not predetermined, number of time steps
- Infinite stage: the agent plans for going on forever (process oriented)

Uncertainty

What the agent can determine the state from the observations:

- Fully-observable : the agent knows the state of the world from the observations.
- Partially-observable : there can be many states that are possible given an observation.

Defining a Solution

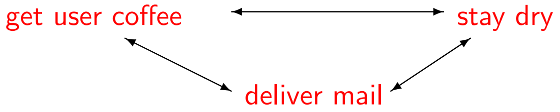
- Optimal solution (utility)
- Satisficing solution (good enough)
- Approximately optimal solution (how far off?)
- Probable solution (how likely not?)

Goals or complex preferences

- achievement goal is a goal to achieve. This can be a complex logical formula.
- maintenance goal is a goal to be maintained.
- complex preferences that may involve tradeoffs between various desiderata, perhaps at different times.
  Either ordinal or cardinal (e.g., utility)
- Examples: coffee delivery robot, medical doctor

Example: Complex Preferences

Delivery Robot



- Goals may conflict

  e.g. can't deliver mail and coffee at the same time
- Goals may be combinatorial

  e.g. user may not want coffee if he doesn't get mail
- Goals may change

  e.g. - when wet, robot can't deliver mail

  - user switches from coffee to kale juice

Single agent or multiple agents

- Single agent reasoning is where an agent assumes that any other agents are part of the environment. (delivery robot)
- Multiple agent reasoning is when an agent needs to reason strategically about the reasoning of other agents. (robot soccer, trading agents)

Agents can have their own goals: cooperative, competitive, or goals can be independent of each other

Next:

- Read Poole & Mackworth chapter 2.1-2.3
- Uninformed Search (Poole & Mackworth chapter 3)
- Informed Search (Poole & Mackworth chapter 4)