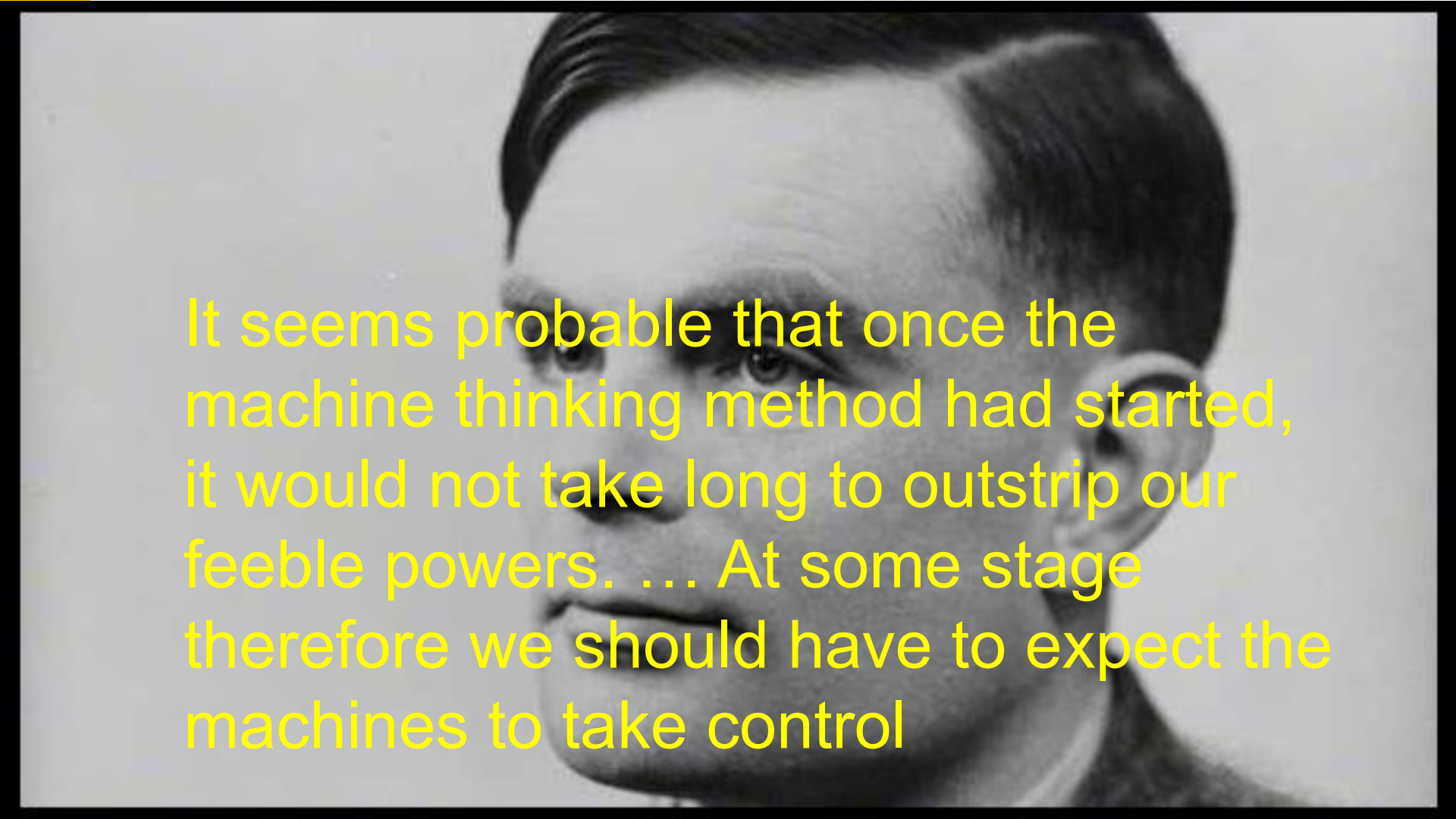
Announcements

- Upcoming due dates
  - P0 due today, 11:59 pm PST
  - HW0 due Monday, 10:59 pm PST

Future

- We are doing AI…
  - To create intelligent systems
    - The more intelligent, the better
  - To gain a better understanding of human intelligence
  - To magnify those benefits that flow from it
    - E.g., net present value of human-level AI ≥ $13,500T
    - Might help us avoid war and ecological catastrophes, achieve immortality and expand throughout the universe
- What if we succeed?


It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. … At some stage therefore we should have to expect the machines to take control

What's bad about better AI?

- AI that is incredibly good at achieving something other than what we really want
- AI, economics, statistics, operations research, control theory all assume utility to be ***fixed, known, and exogenously specified***
  - ~~Machines are intelligent to the extent that their actions can be expected to achieve their objectives~~
  - Machines are ***beneficial*** to the extent that ***their*** actions can be expected to achieve ***our*** objectives

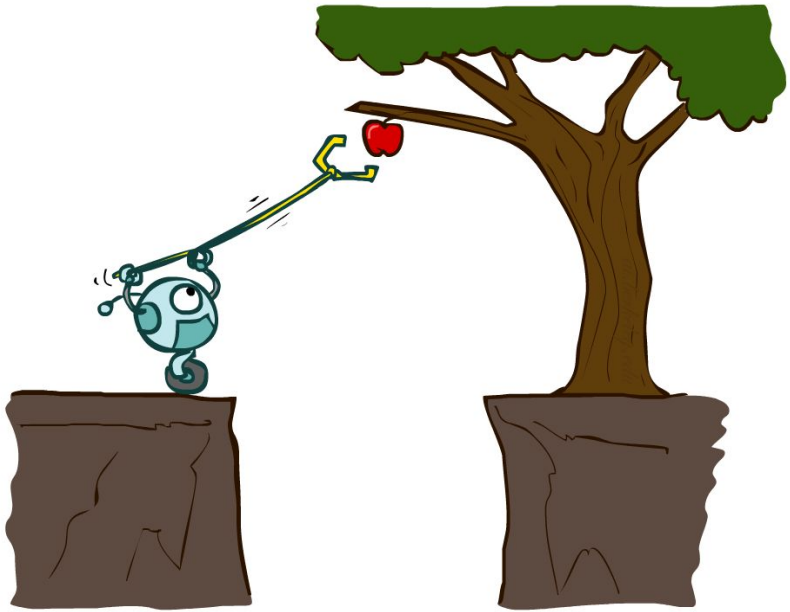
A new model for AI

1. The machine's only objective is to maximize the realization of human preferences
2. The robot is initially uncertain about what those preferences are
3. Human behavior provides evidence about human preferences

"The essential task of our age" [Nick Bostrom, Professor of Philosophy, Oxford]

CS 188: Artificial Intelligence

Agents and environments

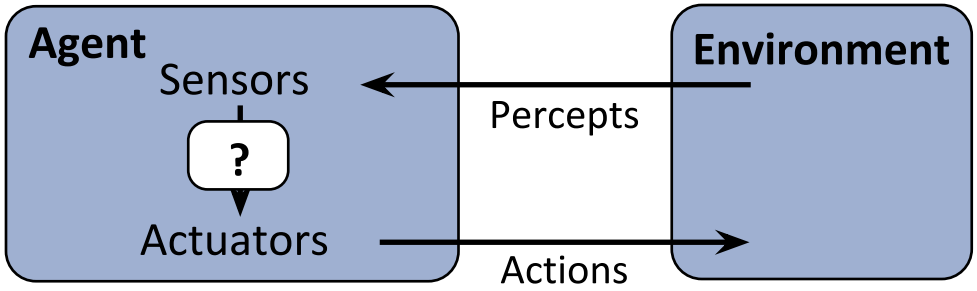


Instructors: Stuart Russell and Dawn Song

Outline

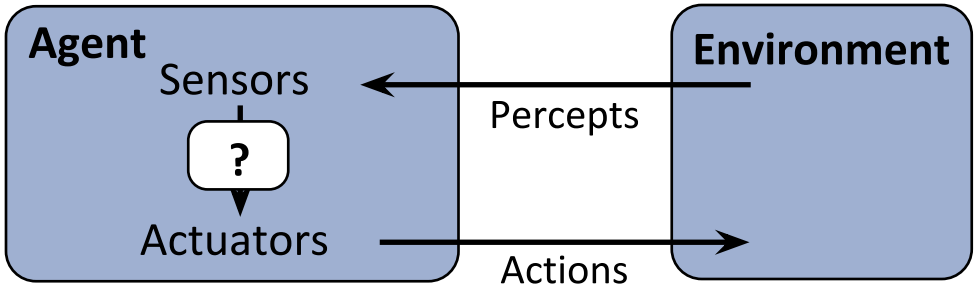
- Agents and environments
- Rationality
- PEAS (Performance measure, Environment, Actuators, Sensors)
- Environment types
- Agent types

Agents and environments



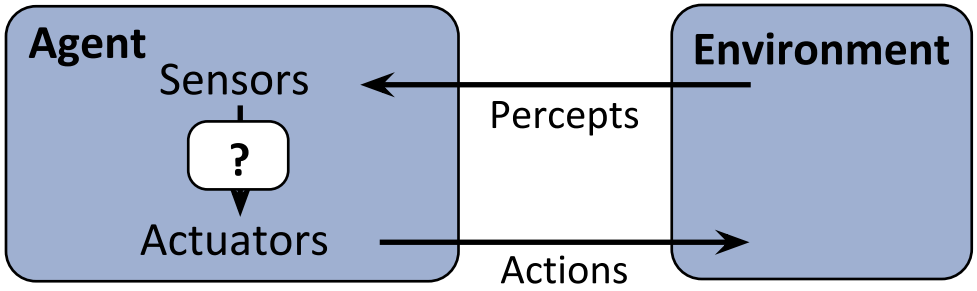
- An agent ***perceives*** its environment through ***sensors*** and ***acts*** upon it through ***actuators*** (or *effectors*, depending on whom you ask)

Agents and environments



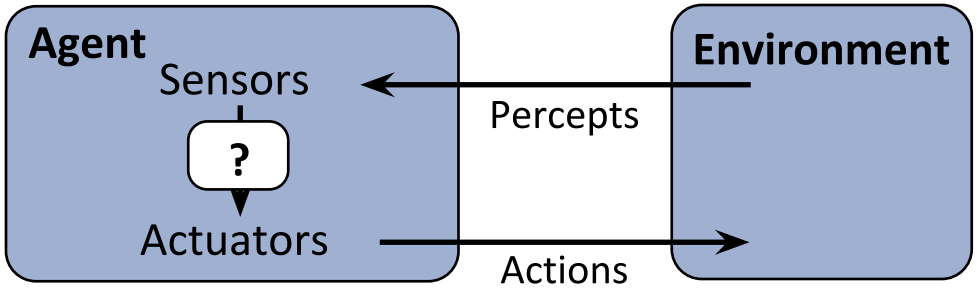
- Are humans agents?
- Yes!
  - Sensors = vision, audio, touch, smell, taste, proprioception
  - Actuators = muscles, secretions, changing brain state

Agents and environments



- Are pocket calculators agents?
- Yes!
  - Sensors = key state sensors
  - Actuators = digit display

Agents and environments

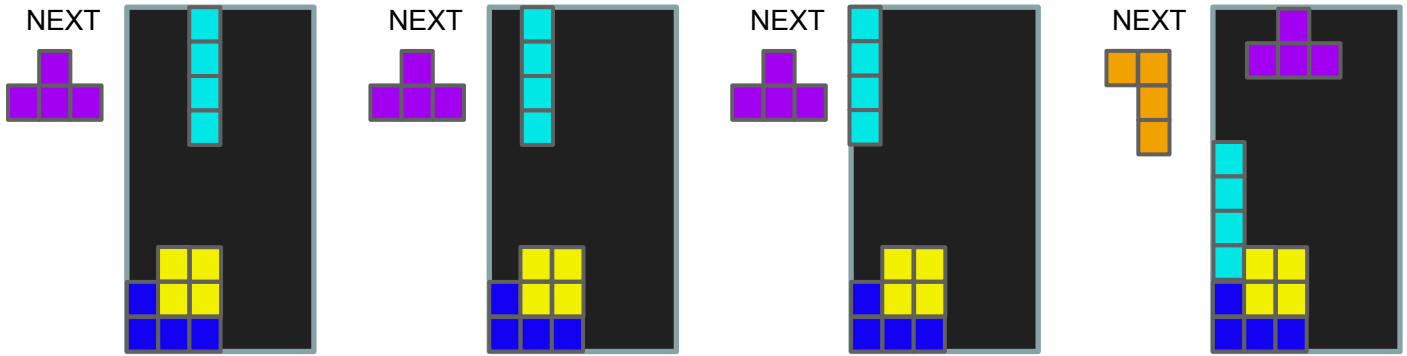


- AI is more interested in agents with large computational resources and environments that require nontrivial decision making

Agent functions

- The ***agent function*** maps from percept histories to actions:
  - $f : \mathcal{P}^* \rightarrow A$
  - I.e., the agent's actual response to any sequence of percepts

Percept

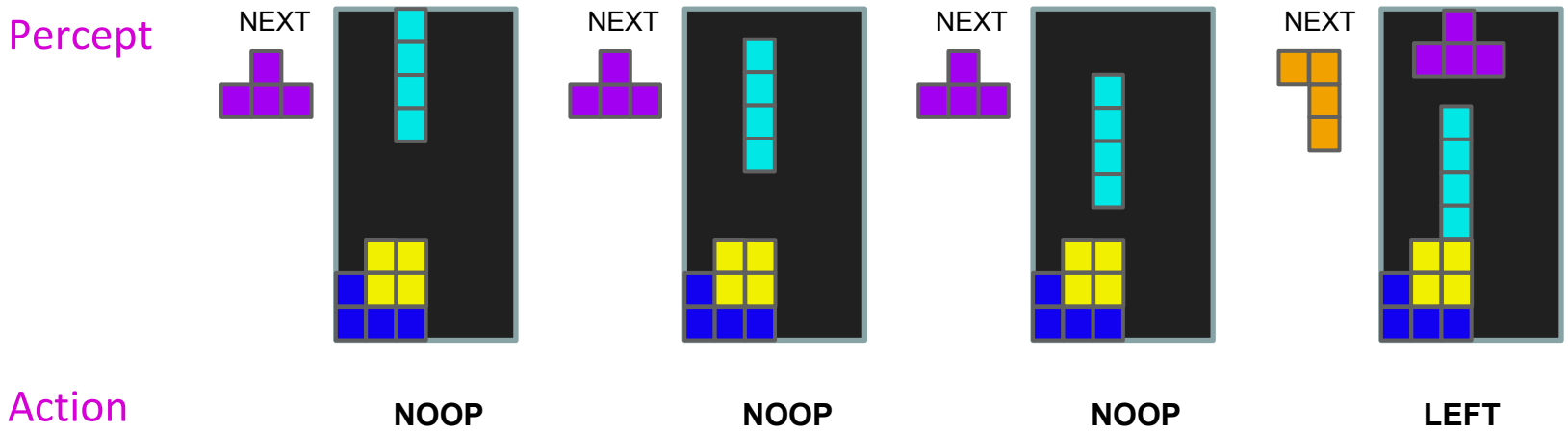


Action

LEFT LEFT DROP RIGHT

Agent programs

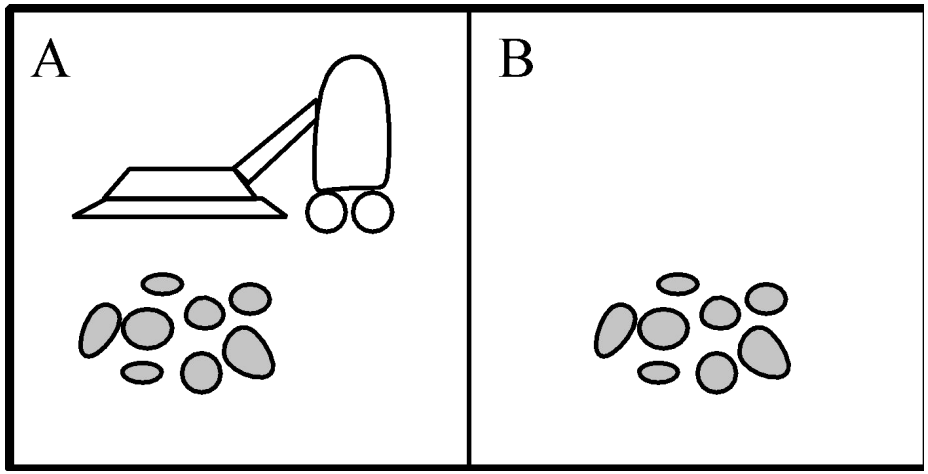
- The ***agent program*** l runs on some machine M to implement f :
  - $f = Agent(l,M)$
  - Real machines have limited speed and memory, introducing delay, so agent function f depends on M as well as l



Agent functions and agent programs

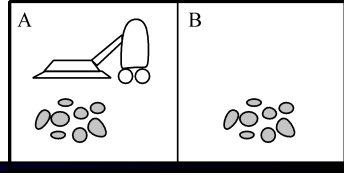
- Can every agent function be implemented by some agent program?
  - No! Consider agent for halting problems, NP-hard problems, chess with a slow PC

Example: Vacuum world



- Percepts: [location,status], e.g., [*A*,*Dirty*]
- Actions: *Left*, *Right*, *Suck*, *NoOp*

Vacuum cleaner agent



Agent function

Percept sequence	Action
[A,Clean]	Right
[A,Dirty]	Suck
[B,Clean]	Left
[B,Dirty]	Suck
[A,Clean],[B,Clean]	Left
[A,Clean],[B,Dirty]	Suck
etc	etc

Agent program

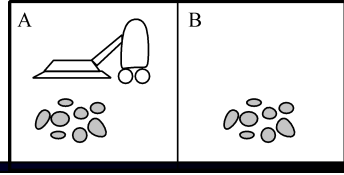
function Reflex-Vacuum-Agent([location,status])
    returns an action
if status = Dirty **then return** Suck
else if location = A **then return** Right
else if location = B **then return** Left

What is the ***right*** agent function?

Can it be implemented by a small agent program?

(Can we ask, “What is the right agent program?”)

Rationality



- Fixed ***performance measure*** evaluates the environment sequence
  - one point per square cleaned up?
    - NO! Rewards an agent who dumps dirt and cleans it up
  - one point per clean square per time step, for t = 1,…,T
- A ***rational agent*** chooses whichever action maximizes the ***expected*** value of the performance measure
  - given the percept sequence to date and prior knowledge of environment

Does Reflex-Vacuum-Agent implement a rational agent function?

Yes, if movement is free, or new dirt arrives frequently

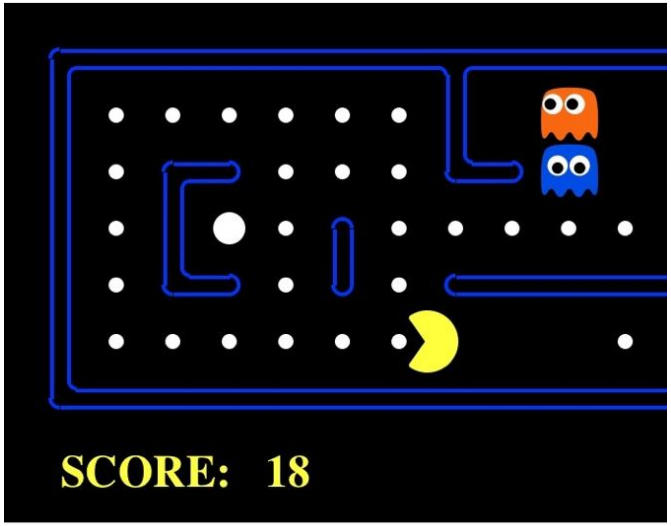
Rationality, contd.

- Are rational agents ***omniscient***?
  - No – they are limited by the available percepts
- Are rational agents ***clairvoyant***?
  - No – they may lack knowledge of the environment dynamics
- Do rational agents ***explore*** and ***learn***?
  - Yes – in unknown environments these are essential
- Do rational agents ***make mistakes***?
  - No – but their actions may be unsuccessful
- Are rational agents ***autonomous*** (i.e., transcend initial program)?
  - Yes – as they learn, their behavior depends more on their own experience

A human agent in Pacman

The task environment - PEAS

- Performance measure
  - -1 per step; + 10 food; +500 win; -500 die; +200 hit scared ghost
- Environment
  - Pacman dynamics (incl ghost behavior)
- Actuators
  - Left Right Up Down
- Sensors
  - Entire state is visible (except power pellet duration)



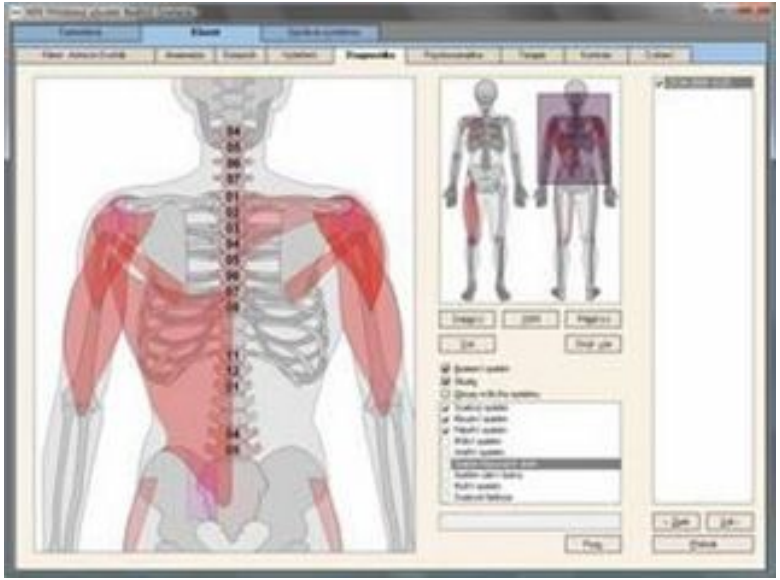
PEAS: Automated taxi

- Performance measure
  - Income, happy customer, vehicle costs, fines, insurance premiums
- Environment
  - US streets, other drivers, customers, weather, police…
- Actuators
  - Steering, brake, gas, display/speaker
- Sensors
  - Camera, radar, accelerometer, engine sensors, microphone, GPS

Image: http://nypost.com/2014/06/21/how-google-might-

PEAS: Medical diagnosis system

- Performance measure
  - Patient health, cost, reputation
- Environment
  - Patients, medical staff, insurers, courts
- Actuators
  - Screen display, email
- Sensors
  - Keyboard/mouse



Environment types

	Pacman	Backgammon	Diagnosis	Taxi
Fully or partially observable				
Single-agent or multiagent				
Deterministic or stochastic				
Static or dynamic				
Discrete or continuous				
Known physics?				
Known perf. measure?				

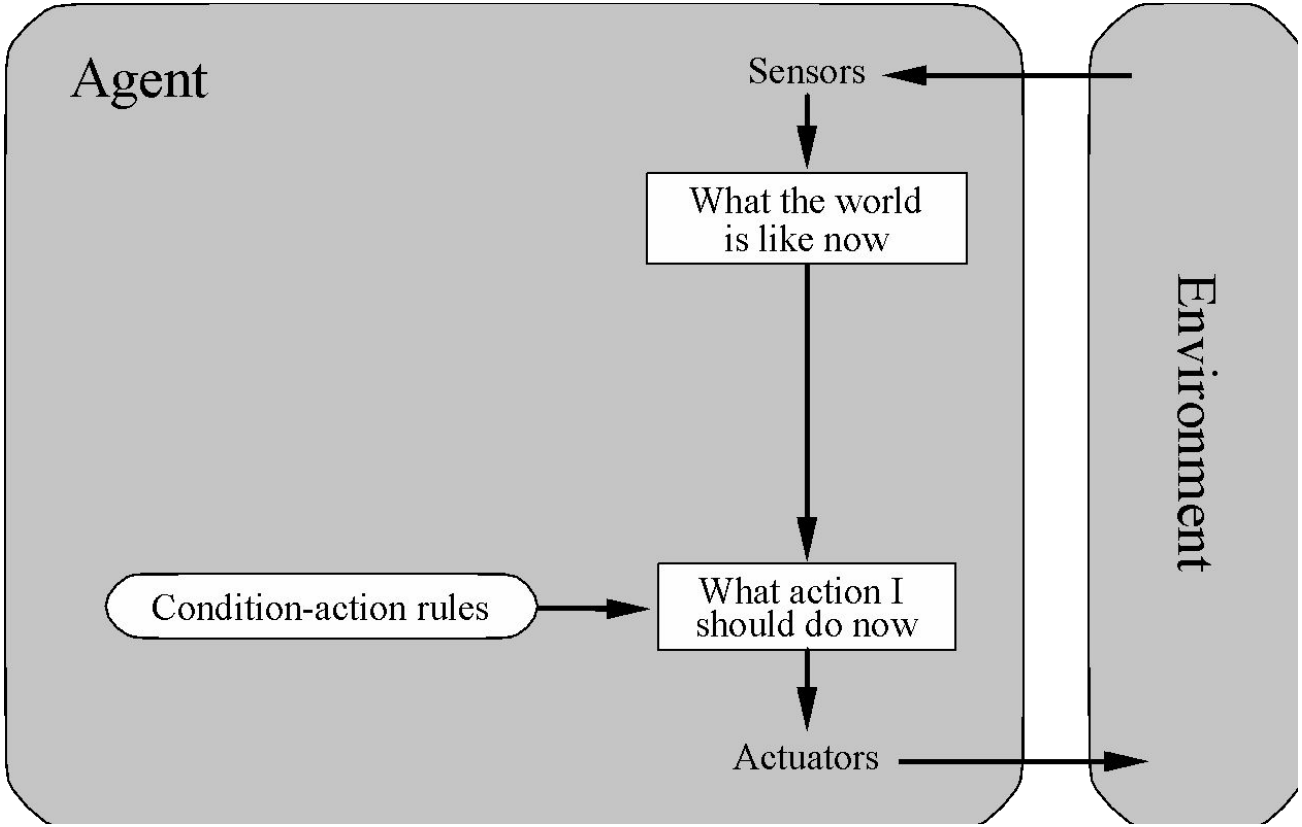
Agent design

- **The environment type largely determines the agent design**
  - ***Partially observable*** => agent requires ***memory*** (internal state)
  - ***Stochastic*** => agent may have to prepare for ***contingencies***
  - ***Multi-agent*** => agent may need to behave ***randomly***
  - ***Static*** => agent has time to compute a rational decision
  - ***Continuous time*** => continuously operating ***controller***
  - ***Unknown physics*** => need for ***exploration***
  - ***Unknown perf. measure*** => observe/interact with ***human principal***

Agent types

- In order of increasing generality and complexity
  - Simple reflex agents
  - Reflex agents with state
  - Goal-based agents
  - Utility-based agents

Simple reflex agents



Pacman agent in Python

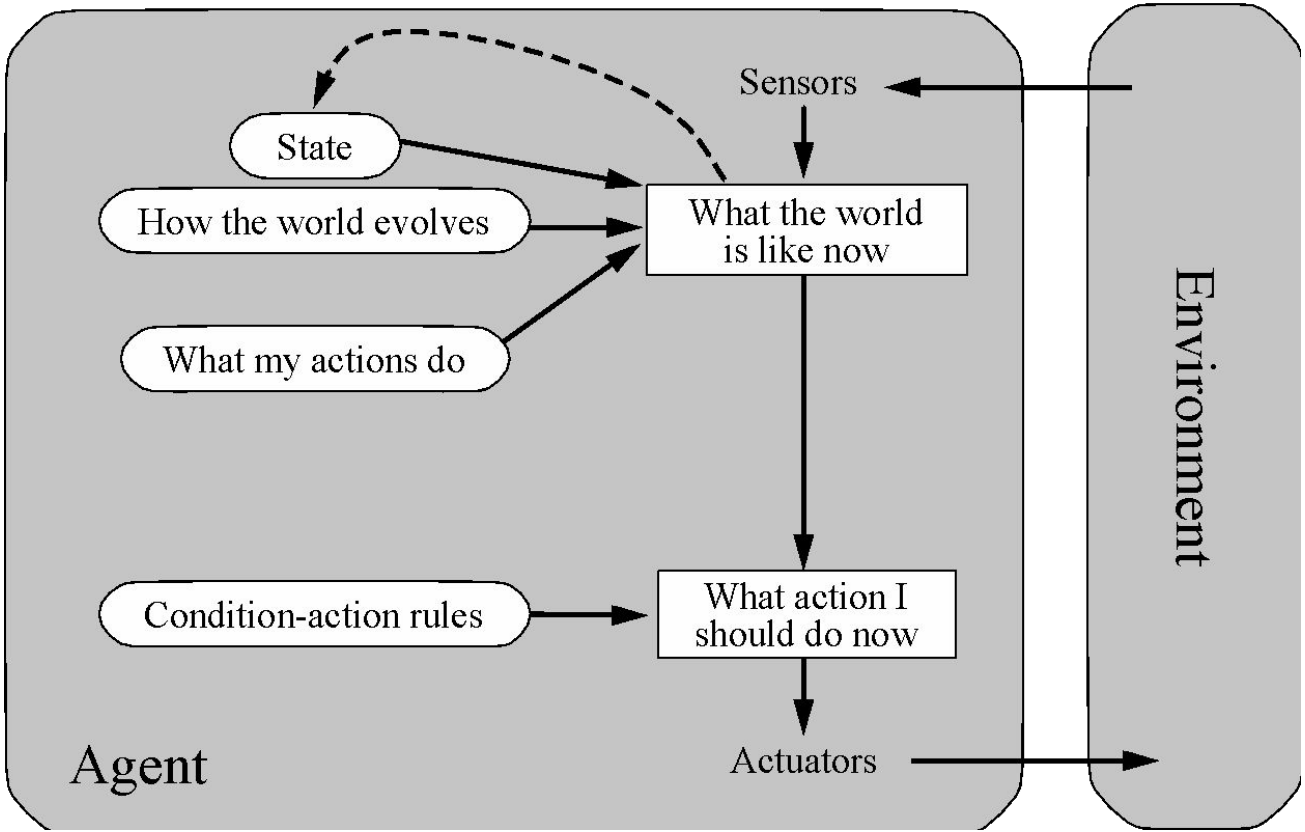
```
class GoWestAgent(Agent):

    def getAction(self, percept):
        if Directions.WEST in percept.getLegalPacmanActions():
            return Directions.WEST
        else:
            return Directions.STOP
```

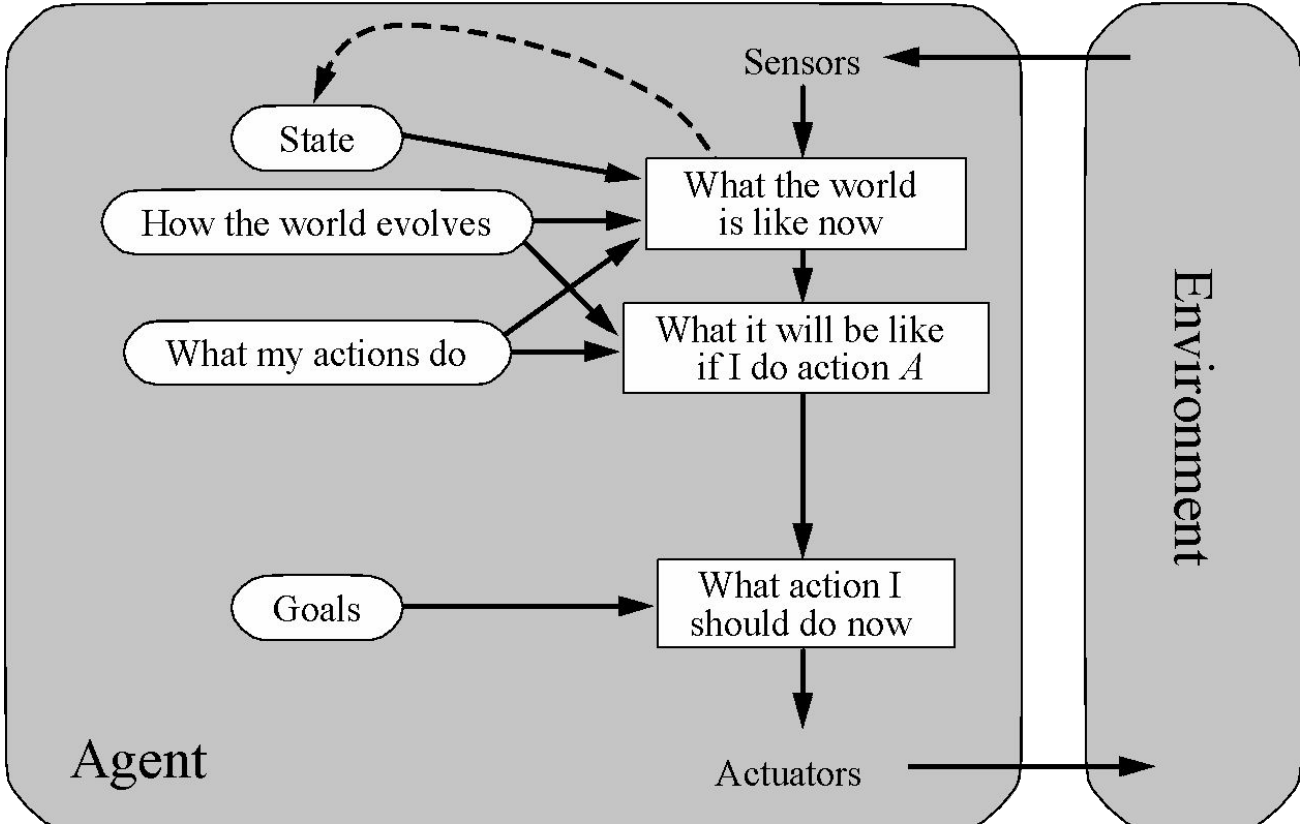
Pacman agent contd.

- Can we (in principle) extend this reflex agent to behave well in all standard Pacman environments?
  - No – Pacman is not quite fully observable (power pellet duration)
  - Otherwise, yes – we can (*in principle*) make a lookup table…..

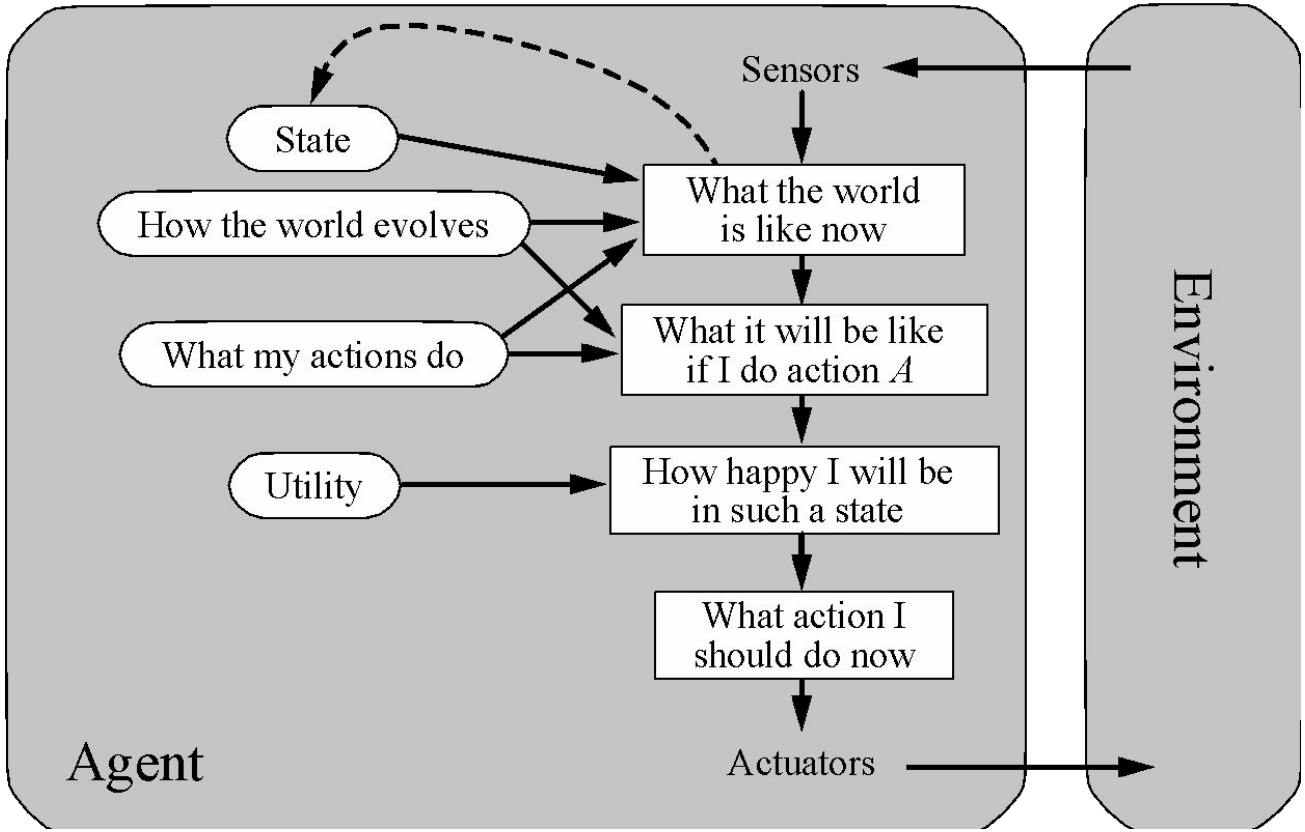
Reflex agents with state



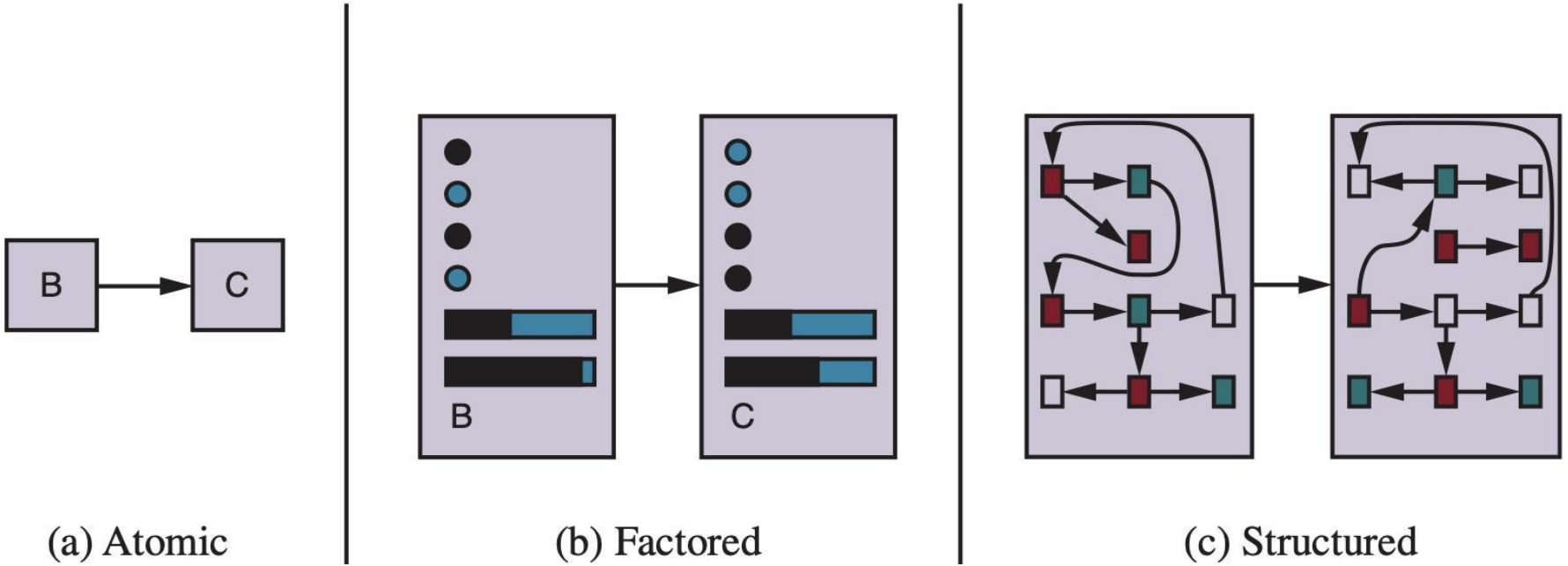
Goal-based agents



Utility-based agents



Spectrum of representations



Summary

- An ***agent*** interacts with an ***environment*** through ***sensors*** and ***actuators***
- The ***agent function***, implemented by an ***agent program*** running on a ***machine***, describes what the agent does in all circumstances
- Rational agents choose actions that maximize their expected utility
- PEAS descriptions define task environments; precise PEAS specifications are essential and strongly influence agent designs
- More difficult environments require more complex agent designs and more sophisticated representations