Definitions

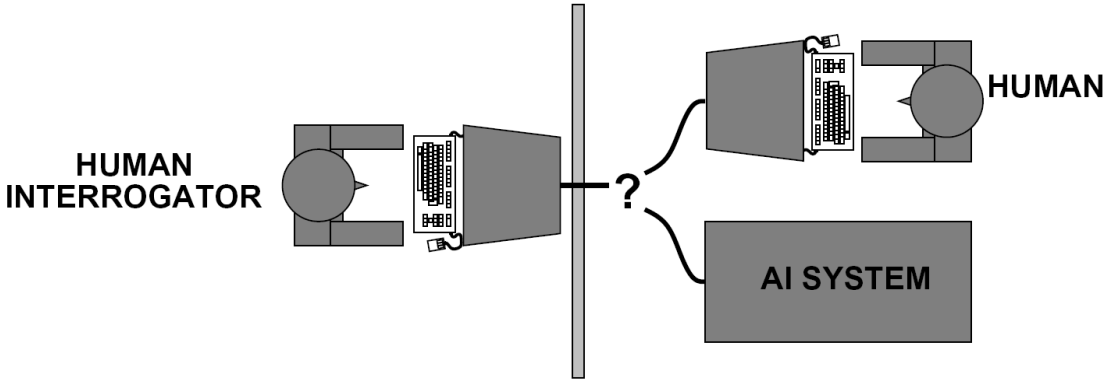
The science of making machines that:

Think like humans	Think rationally
Act like humans	Act rationally

This slide deck courtesy of Dan Klein at UC Berkeley

Acting Like Humans?

- Turing (1950) "Computing machinery and intelligence"
  - "Can machines think?" → "Can machines behave intelligently?"
  - Operational test for intelligent behavior: the *Imitation Game*



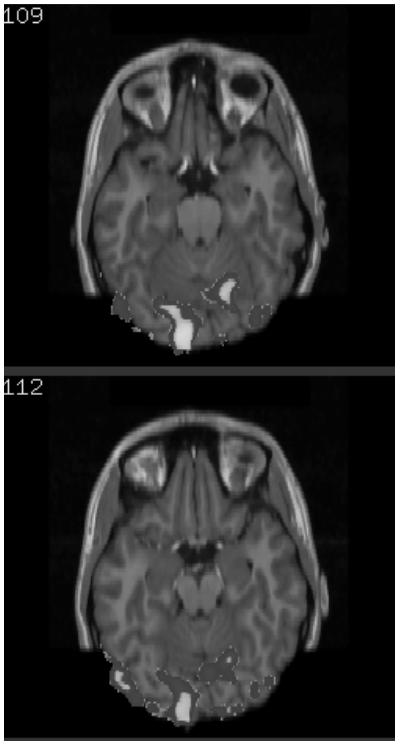
  - Predicted by 2000, a 30% chance of fooling a lay person for 5 minutes
  - Anticipated all major arguments against AI in following 50 years
  - Suggested major components of AI: knowledge, reasoning, language understanding, learning
- Problem: Turing test is not reproducible or amenable to mathematical analysis

Thinking Like Humans?

- The cognitive science approach:
  - 1960s ``cognitive revolution'': information-processing psychology replaced prevailing orthodoxy of behaviorism

- Scientific theories of internal activities of the brain
  - What level of abstraction? “Knowledge'' or “circuits”?
  - **Cognitive science:** Predicting and testing behavior of human subjects (top-down)
  - **Cognitive neuroscience:** Direct identification from neurological data (bottom-up)
  - Both approaches now distinct from AI
  - Both share with AI the following characteristic:

    The available theories do not explain (or engender) anything resembling human-level general intelligence

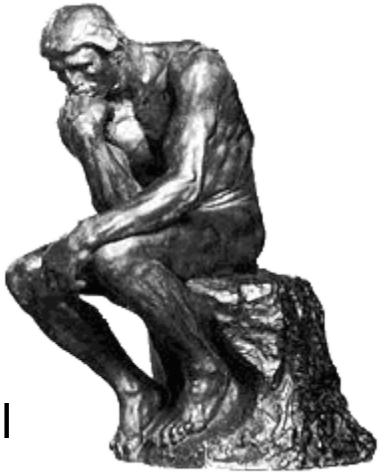


- Hence, all three fields share one principal direction!

Images from Oxford fMRI center

Thinking Rationally?

- The “Laws of Thought” approach
  - What does it mean to “think rationally”?
  - Normative / prescriptive rather than descriptive



- Logicist tradition:
  - Logic: notation and rules of derivation for thoughts
  - Aristotle: what are correct arguments/thought processes?
  - Direct line through mathematics, philosophy, to modern AI

- Problems:
  - Not all intelligent behavior is mediated by logical deliberation
  - What is the purpose of thinking? What thoughts should I (bother to) have?
  - Logical systems tend to do the wrong thing in the presence of **uncertainty**

Rational Decisions

We'll use the term **rational** in a particular way:

- Rational: maximally achieving pre-defined goals
- Rational only concerns what decisions are made
  (not the thought process behind them or their outcomes)
- Goals are expressed in terms of the **utility** of outcomes
- Being rational means **maximizing your expected utility**

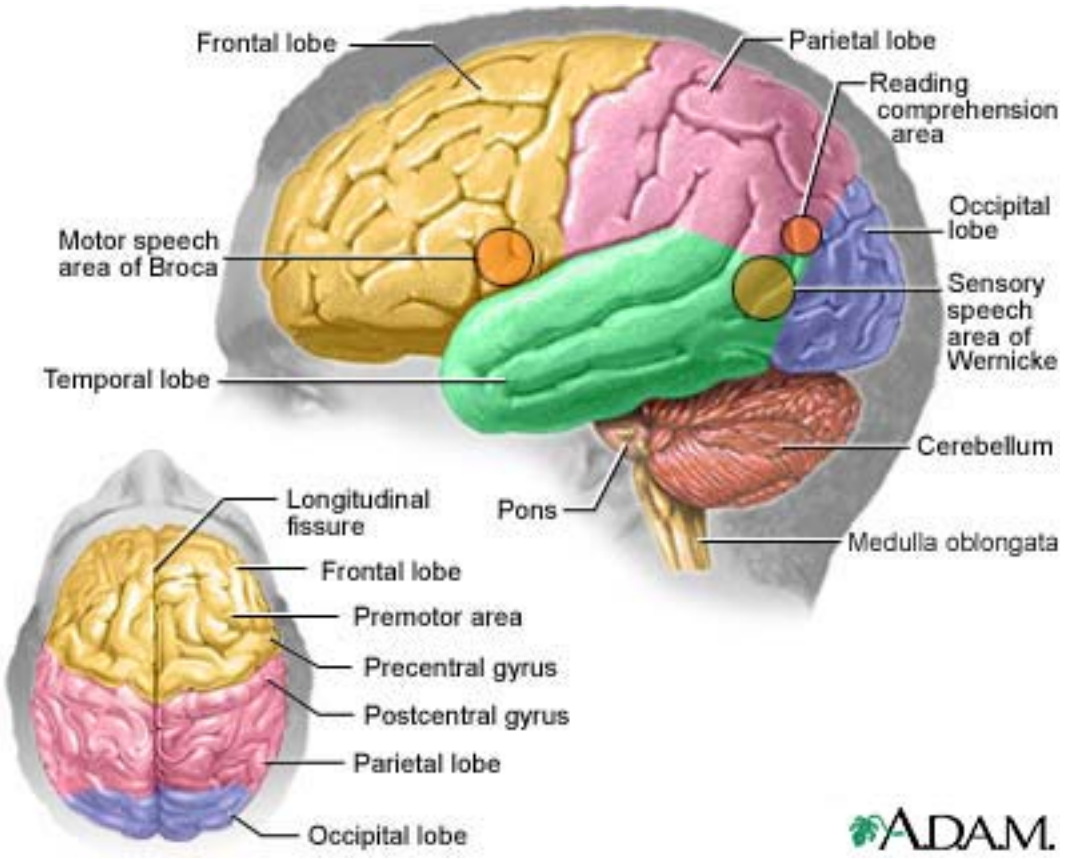
Another possible title for this course would be:

Computational Rationality

Maximize Your Expected Utility

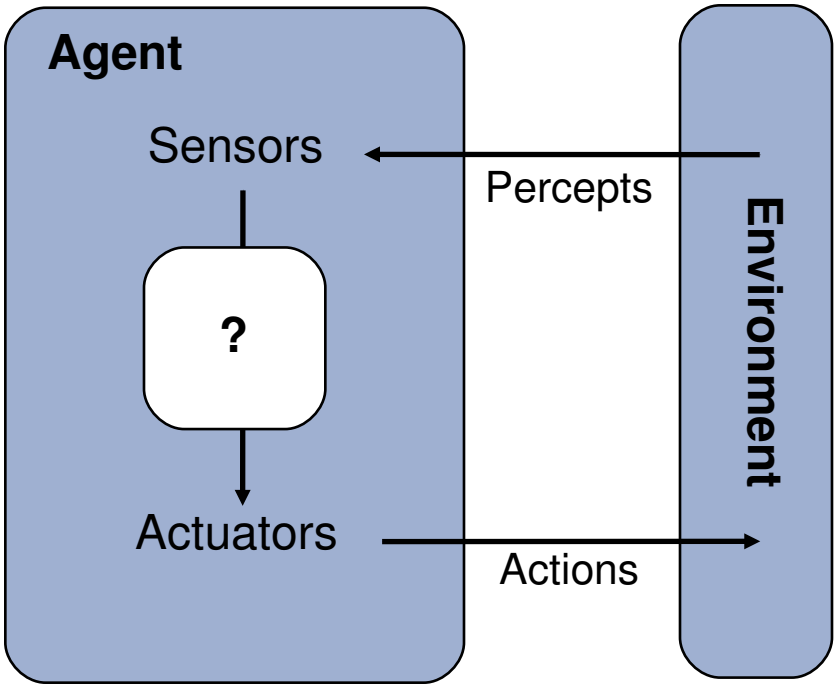
What About the Brain?

- Brains (human minds) are very good at making rational decisions (but not perfect)
- "Brains are to intelligence as wings are to flight"
- Brains aren't as modular as software
- Lessons learned: **prediction** and **simulation** are key to decision making



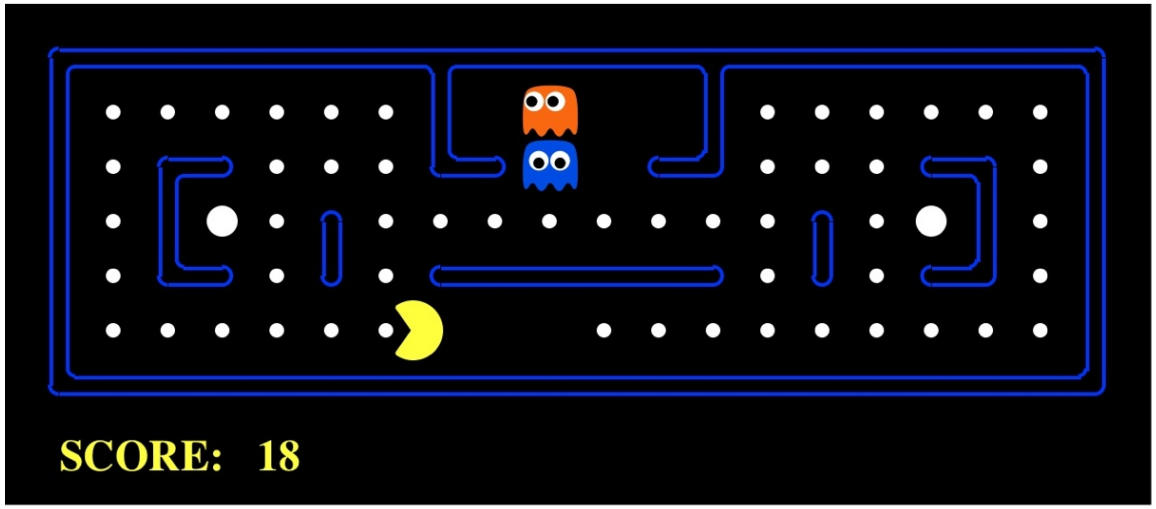
Rational Agents

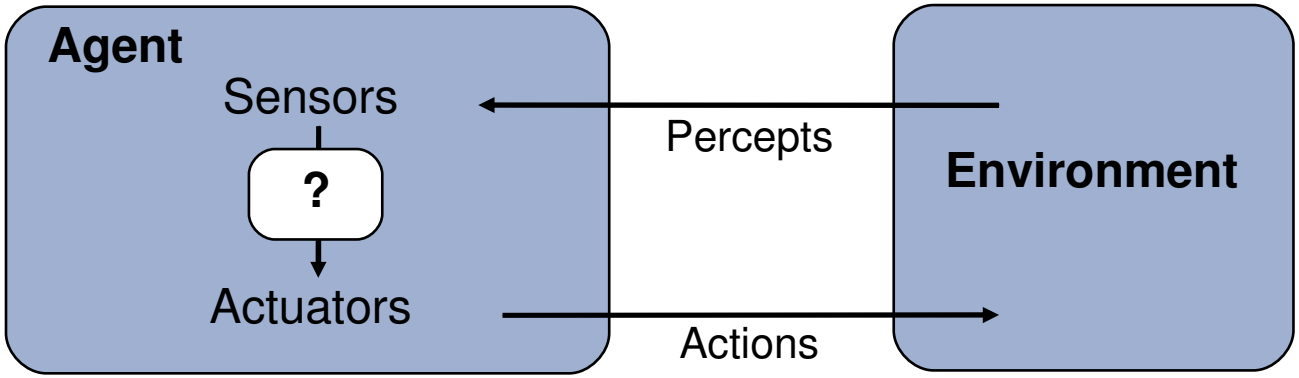
- An **agent** is an entity that *perceives* and *acts*.
- A **rational agent** selects actions that maximize its **utility function**.
- Characteristics of the **percepts, environment,** and **action space** dictate techniques for selecting rational actions.



- This course is about:
  - General AI techniques for a variety of problem types
  - Learning to recognize when and how a new problem can be solved with an existing technique

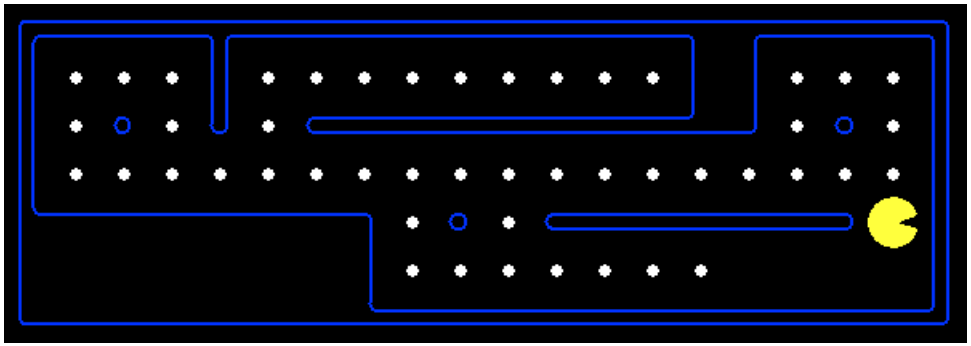
Pacman as an Agent

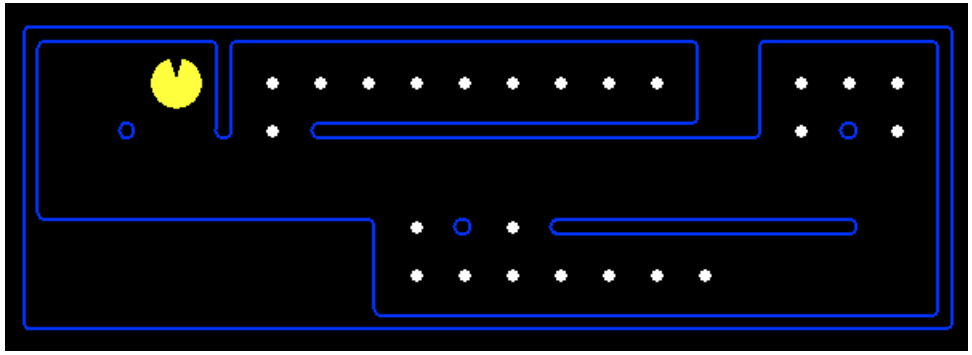




Reflex Agents

- Reflex agents:
  - Choose action based on current percept (and maybe memory)
  - May have memory or a model of the world's current state
  - Do not consider the future consequences of their actions
  - Act on how the world IS





- Can a reflex agent be rational?

Goal Based Agents

- Goal-based agents:
  - Plan ahead
  - Ask “what if”
  - Decisions based on (hypothesized) consequences of actions
  - Must have a model of how the world evolves in response to actions
  - Act on how the world WOULD BE