CS 461
Introduction to AI

Basis and Concepts

What is AI?

- Does it involve thought? Behavior? To what standard?

- Acting *humanly*
  - Turing test: Can an AI system communicating via text interface mimic a human well enough to convince someone that it's human?
  - **General** knowledge, natural language processing, social *cues*

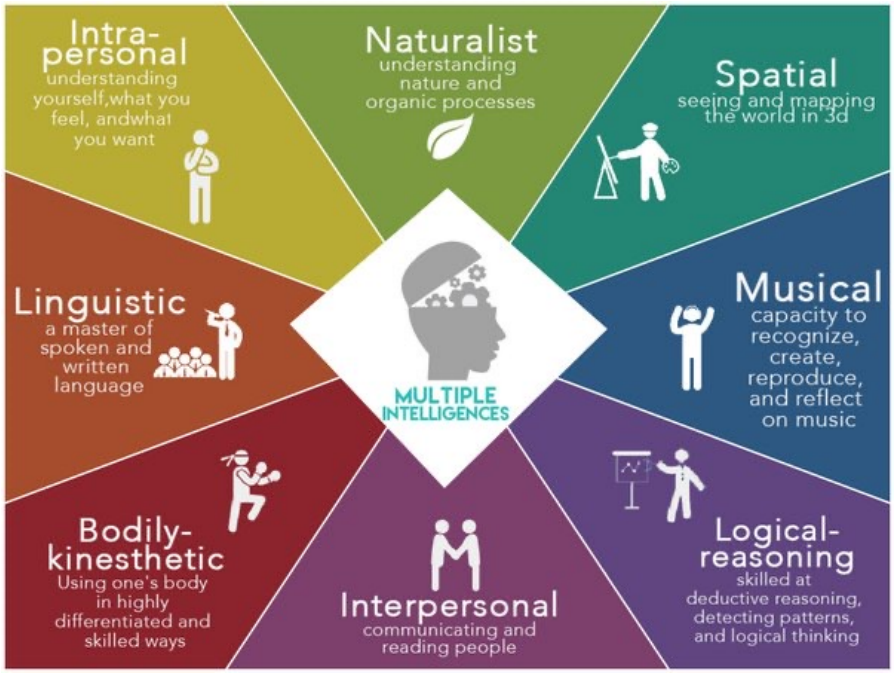
- Acting **Rationally**
  - Can a system **integrate** information about its **environment** and formulate a plan of action?
  - Sensing, reasoning in the presence of uncertain or incomplete information, identifying costs & tradeoffs

- Thinking *humanly*
  - Identifying "laws of thought" and human reasoning processes
  - Logic, sensing, learning via same **mechanisms** as humans use

- Thinking **rationally**
  - Logic, *perception*, drawing conclusions & new facts from known ones (learning)

What is AI?

- How do we determine if someone (or something) is intelligent?



Smart	VS	Intelligent
Definition		
Is a person who uses his intelligence practically and efficiently daily		Is something which a person is born with
Measurement		
Non-measurable		IQ Test (Intelligence Quotient)
Refers To		
Refers to intellect and the appearance		The intellect of a person
Nature		
Practical and has a good judgment		Not always practical

What is AI?

- Is it important that it use the same methods humans use?
  - What if humans can't really describe in depth how they reach a conclusion?
  - What if animal intelligence only emerges in groups?
    - Individual ants don't show signs of intelligence, but ant colonies demonstrate coordinated actions

- Working definition: *Artificial* Intelligence is the science of making machines **do** things that would require *intelligence* if done by humans

- Implications of more complex design?
  - "A year spent in Artificial Intelligence is enough to make one believe in God." –Alan Perlis, *Epigrams on Programming*

Identity

Leibniz's Law

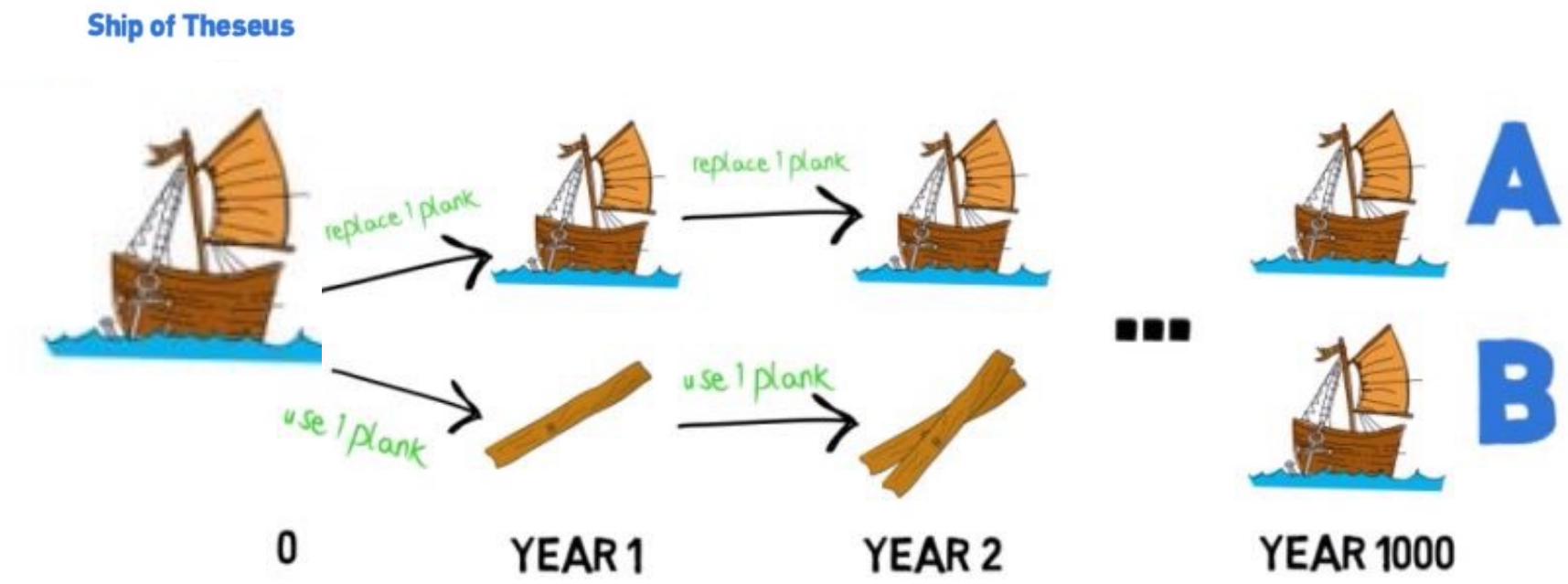
For any two objects X and Y: If X = Y, then X and Y share all the same properties.

As we've noted, this is logically equivalent to the following principle:

For any two objects X and Y: If X and Y do not share all the same properties, then X ≠ Y.

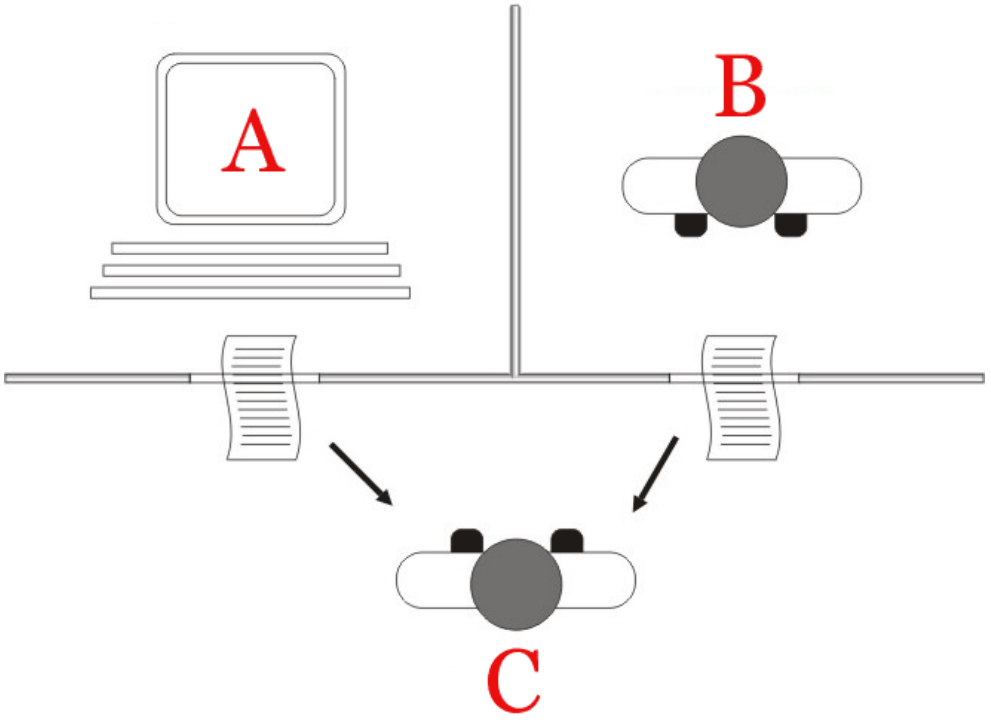
$$(a = b) \rightarrow \forall F\,(F(a) \leftrightarrow F(b))$$

Identity


Ship of Theseus
replace 1 plank
replace 1 plank
use 1 plank
use 1 plank
A
B
0
YEAR 1
YEAR 2
YEAR 1000

The Turing Test

- Turing proposed an imitation game.
- An interrogator interacts with 2 entities—1 human, 1 machine—via text interface
  - The human is assumed truthful; the machine can lie
- The **interrogator** must determine which is ***human*** and which is the ***machine***
  - So, for example, if asked for the square root of 23,921, the machine would not reply with the answer correct to 12 decimal places within microseconds, as humans don't do that; it might delay its answer, give an approximation, or make small errors
  - If asked to discuss today's weather (because the computer can't look out the window), the computer can perhaps look up the weather on the Web before responding
  - If asked about something emotional ("Tell me about your first romantic crush"), the machine would not have direct experience, but would know how humans have described these things; also, emotional depth isn't synonymous with intelligence



Objections to the Turing test

- The head in the sand objection: Intelligent machines would depose us, and thus are too horrible to think about
  - Not really a serious objection…
- Machines don't have **souls** and thus can't be "**really**" intelligent
  - We don't know what the limits are to what can and can't have souls; also, if the soul is by definition immaterial and immeasurable, how can we know for sure if it's there or not, i.e. how do we know that *we* have souls?
- A machine can't do something truly, surprisingly original
  - Machines have already surprised us with their capabilities.
  - This assumes that humans can immediately deduce all consequences of a given fact or action; this simply isn't true
- A machine will never, for example, make a human fall in love with it.
  - But we know how to make objects people come to love. Ever own a teddy bear?
  - Also, people have fallen in love with (and 'married') video game characters
- Or appreciate *beauty*…
  - That's your definition of intelligence? Aesthetic appreciation?

More basic objections to the Turing test

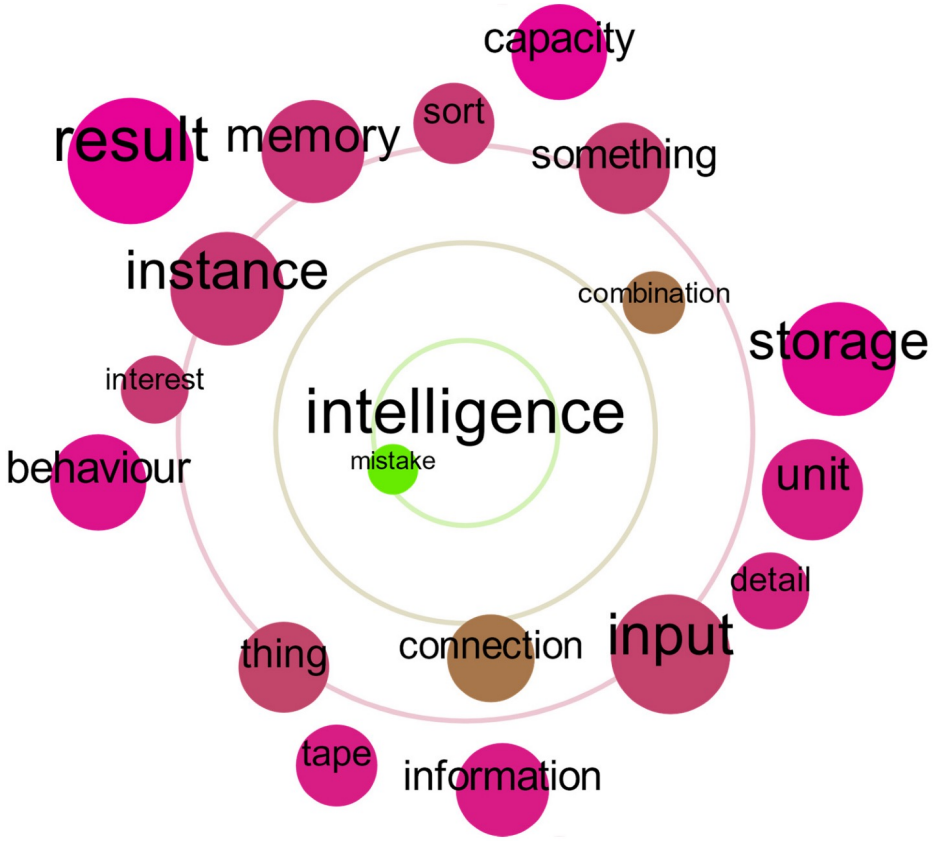
- Searle's Chinese Room: Users pass notes written in Chinese under a door. Inside, a human, who does not know Chinese, looks them up in a series of complex rulebooks and writes out some other symbols (which he also does not understand) and passes them back out under the door. People reading those responses on the other side believe he is answering in Chinese; yet neither the person nor the rulebooks can be said to understand Chinese.
  - Same setup, with 1000 users, each with their own rulebook, either what output to produce, or who to pass their note to. Again, where does the so-called "knowledge of Chinese" reside?
- These rely on the idea that we cannot deduce internal state purely from observing external behavior of a black box; but we know from some situations that we can
  - Consider a person who does know Chinese. The knowledge of Chinese can't be localized to a particular part of the brain or set of neurons.
    - There are parts of the brain that process language, but English and Chinese use the same parts of the brain

More basic objections to the Turing test

- **Block's** objection:
  Text is stored in binary.

Given a **large enough** database, it's possible to store a library of queries & plausible answers to mimic intelligence via table lookup.

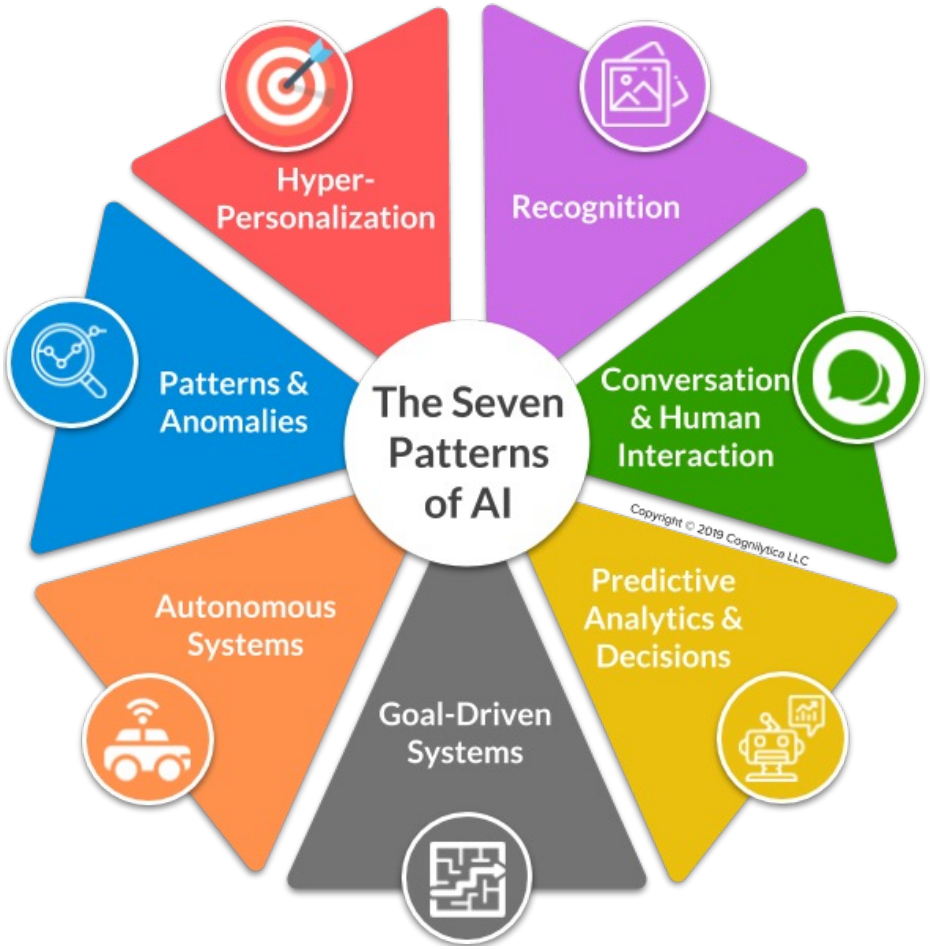
Do we ascribe intelligence to such a system?



Strong v. Weak AI

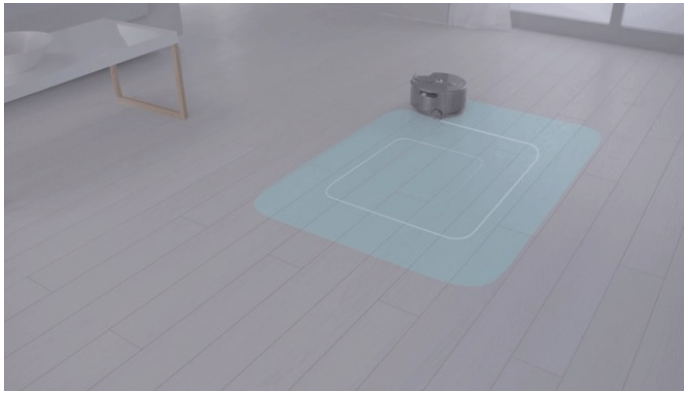
- One school of thought (dominated by MIT) argues that any system demonstrating intelligent behavior is AI, regardless of how it achieves its end. This is ***weak AI***.
- Another (dominated by Carnegie-Mellon) holds that a system should be based on the same methods of learning & cognition used by humans. This is ***strong AI***.
- Weak AI proponents argue that the purpose of AI is to solve difficult problems
  - if we obtain a solution, we are more concerned with the correctness & generalizability of that solution
  - If the process used to obtain it is different than a human would use, who cares?
- Strong AI proponents argue that the end goal is "true" intelligence, which would require self-awareness (consciousness)
- To date, there is **no** general-purpose strong AI in existence. All systems in existence today are weak AI, and often very limited
  - However, a system that can read MRI scans and diagnose cancers as well as an experienced radiologist, *and nothing else*, is still **very useful**

Common Implementations



Agents, Environments, and Perception

- **Agent**: Anything that can be viewed as perceiving its **environment** through **sensors** and acting on that environment through **actuators**.
- Percept: The agent's perceptual input at a given instant
- Percept sequence: Complete history of percepts
  - Choice of action can depend on percept or percept sequence, but not on anything it has not observed yet.
- Agent's behavior is described by an **agent function** that maps percepts or percept sequence to actions
  - Note that **agent function** does not base its actions on the environmental state, only on what has been perceived about that state
  - An agent function is implemented by an **agent program.**
  - Conceptually can be viewed as a table lookup or set of if-then rules; most implementations are more complex
- Vacuum cleaner world: 2 squares, A & B, each can be clean or dirty. Actions are GO LEFT, GO RIGHT, CLEAN
  - Can build up table: given this percept (or sequence), do this

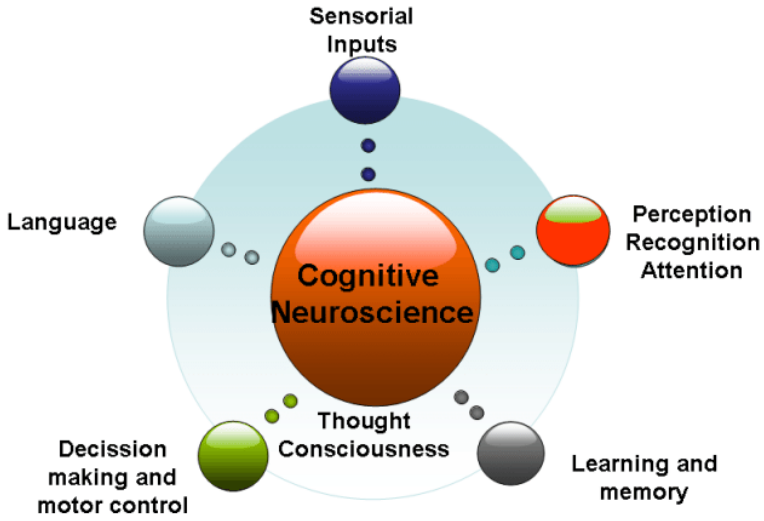


What makes an agent ‘rational’?

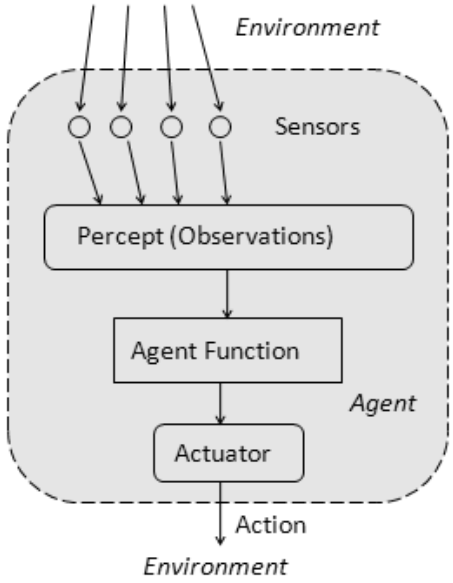
- One definition is that a rational agent does the right thing—with the proper “then” action for every “if” state
  - But what does “doing the right thing” mean?
- We consider consequences.
- The right action is the one that leads to the desired outcome based on the series of environmental states produced by interacting with the environment
  - Not agent states; otherwise our agent can delude itself by simply defining whatever it does (even if random) as the ‘correct’ thing to do
  - Compare the tendency among humans to tell themselves they don’t really want something that’s out of reach anyway (sour grapes)
- Agent function/program should be designed according to circumstances
  - For example, our vacuum cleaner should get credit for cleaning up dirt
  - But be careful! A vacuum could clean up some dirt, get credit, dump the dirt onto the floor, clean it up, get credit, dump it, clean it up, get credit…
  - Probably better to design it based on state of square (clean or dirty), perhaps with penalty for electricity use or noise, to discourage unnecessary work
- *In general, it’s better to define the goal state in terms of what is actually wanted, not on how the agent should behave.*
- Still issues to work out, based on how we reward cleanliness
  - Mediocre, kinda-sorta clean all the time?
  - Or immaculate as soon as it’s done, at the cost of long breaks (and therefore dirt buildup) between cleanings?

Foundations of AI

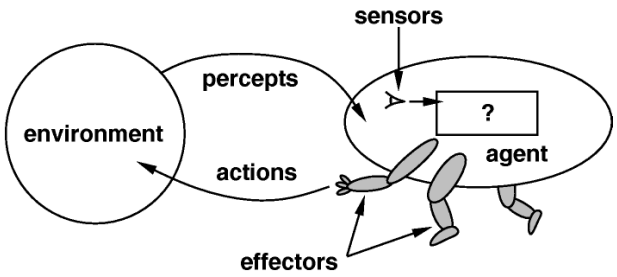
- Neuroscience
  - How do brains process information?
    - Brains are MUCH more energy efficient than any CPU
- Psychology
  - How do humans and animals think and act?
- Computer Engineering
  - How can we build an efficient computer?
- Control Theory & Cybernetics
  - How can artifacts operate under their own control?
    - Homeostasis, objective function
- Linguistics
  - How does language relate to thought?



Define “rational.”

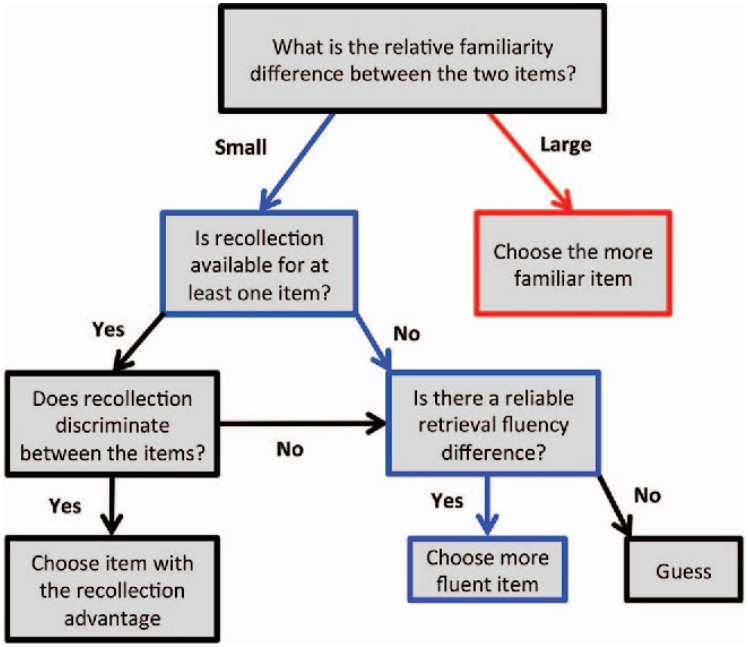


- Rationality depends on:
  - The performance **measure** that defines **success**
  - The agent’s **knowledge** of the **environment**
  - The **actions** the agent can perform
  - The agent’s **percept sequence**
- For each possible percept sequence, a rational agent **should** *select* an action that is expected to **maximize its performance measure**, given the evidence provided by the **percept sequence** and whatever built-in **knowledge** the agent has



Heuristics

- AI often relies on applying **heuristics** (which usually get a correct answer, at least approximately, and do it **quickly**) rather than algorithms (which always get a correct answer, exactly, **eventually**)
  - Solving a **simpler** but related problem
  - **Working backward** from a solution toward the starting state
  - Identifying similar solved problems & **testing** those solutions
  - Successive approximation (iteration)—if we can't find a solution, can we find a state that is in some way "closer"?
    - And if not, what do we do?



Problems suitable for AI

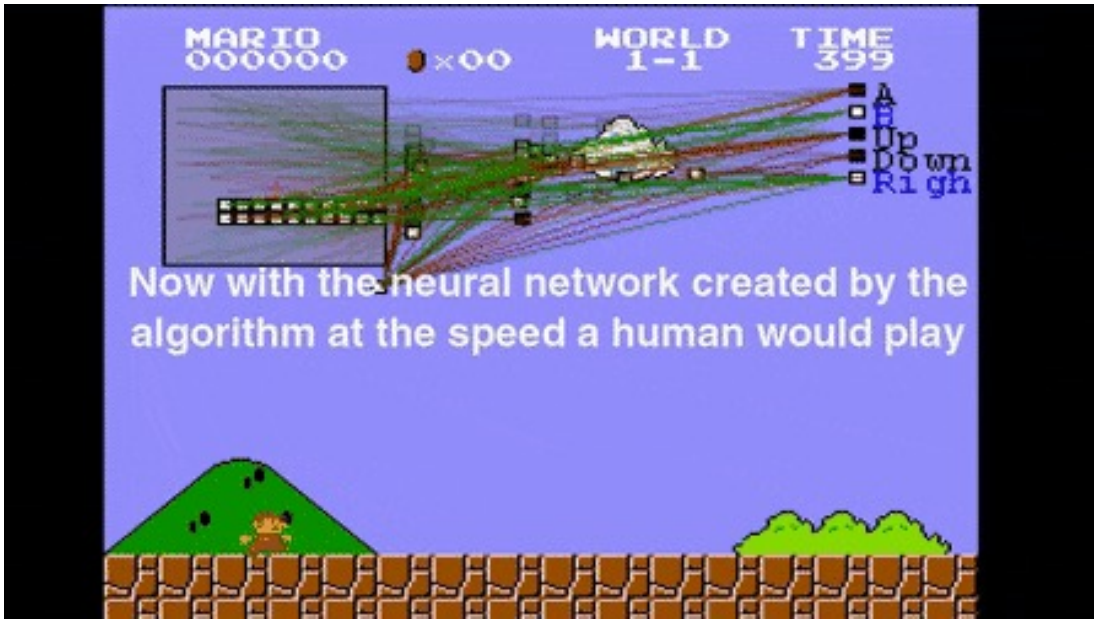
- Most AI problems are **large**
- Cannot be solved by straightforward algorithms
- Embody (encapsulate / integrate) a large amount of ***human expertise*** [representation]
- Examples:
  - Medical diagnosis (one of the first applications of ***expert systems***)
    - A lot of expertise, much of it in the form of if-then rules in a roughly hierarchical structure
    - Very complex, with complex interactions between rules & possible causes
  - Given a user's shopping history, what specials are likely to lure them into the store in the next week?
    - Given all of our shoppers' histories, what specials would bring the most total business in to the store?
    - Given this user's travel history, how much are they likely willing to pay for this airline ticket right now?
  - Given a user's spending history, what changes to their budget would we recommend? ("Do you really need to eat out quite so often?")
  - Given that chess has about 10^{42} "reasonable" games and about 10^{120} possible states, can we make a program that plays chess "well"?
    - How well?
    - With perfect play by both sides, who wins?
    - What is the best move in a particular position?
    - Note that this is still weak AI—a strong AI system could not only beat the very best humans, it could also explain the reasoning behind its moves and identify how to teach others to play better

Properties of Task Environments

- Fully v. Partially Observable
  - Can sensors detect all aspects of environment relevant to the choice of action?
    - May only be partially observable due to noise, distance, or inaccessibility (robot taxi can't tell what other drivers are thinking)

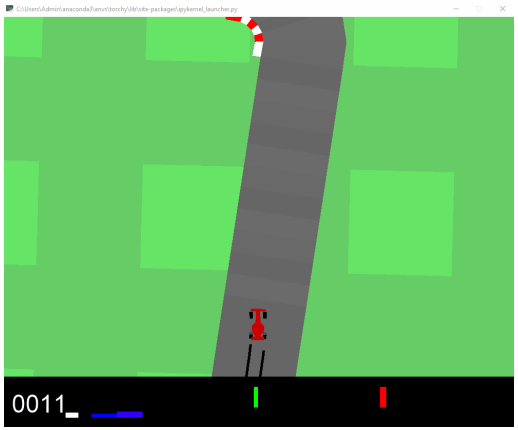
If there are no sensors at all the environment is **unobservable**

All is not lost, we can still behave rationally, and make certain deductions about our environment, without sensors



Properties of Task Environments

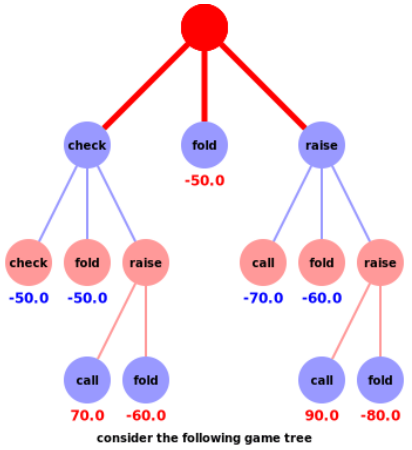
- Single Agent v. Multiagent
  - Is there more than one agent active in the environment?
    - Do we classify other drivers as agents, or semi-random features of the environment?
      - Is the other agent best described by assuming it's trying to maximize its own performance measure?
    - Are other agents **cooperative or competitive**?
      - Chess is competitive
      - Taxi-driving is partially both; vehicles cooperate to avoid collision, compete for limited parking spaces
    - Communication & cooperation emerge as rational strategies in cooperative environments; randomized behavior avoids predictability in competitive environments



Properties of Task Environments

- Deterministic v Stochastic
  - Is the environment's next (or immediate future) state completely determined by the current state (or state history) and the action executed by the agent?
    - Ignore the uncertainty from the actions of other agents; we deal with that elsewhere
  - Note that a fully deterministic but only partially observable environment might *seem* stochastic
    - Tires blow out; traffic can't be predicted exactly
  - Environments not fully observable or not deterministic are **uncertain**.
  - "Deterministic" implies we know something about the probabilities of possible outcomes. If we only know *possible* outcomes but not their probabilities, the situation is **nondeterministic.**
    - These agents usually have performance measures requiring success for *all possible* outcomes of its actions

Properties of Task Environments

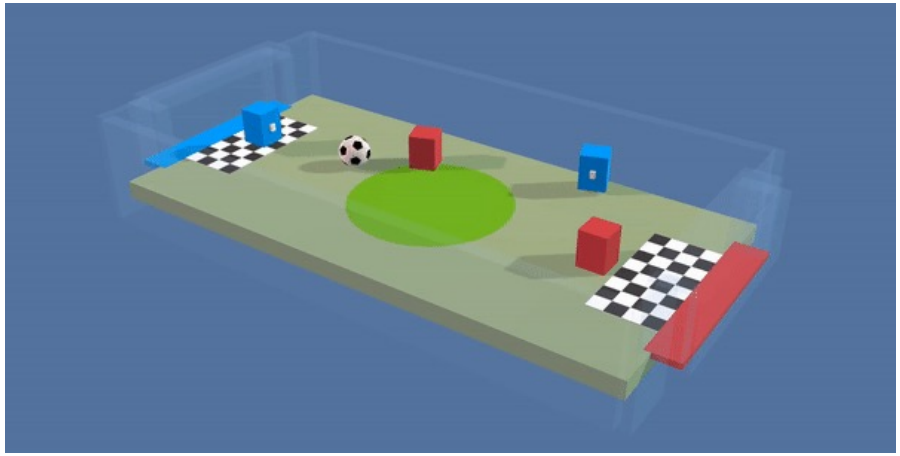


- Episodic v. Sequential
  - Does the agent need to remember the history of previous
  - states, or just attend to the current one?
  - Can this decision affect future ones?

a. Winning or losing this hand of poker doesn't affect the next hand

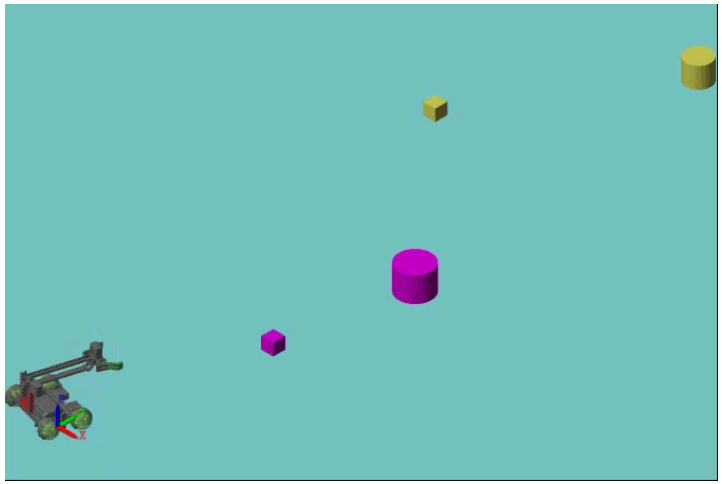
b. If the same position occurs 3 times in a game of chess, the game is a draw; thus our agent must remember the sequence of prior positions

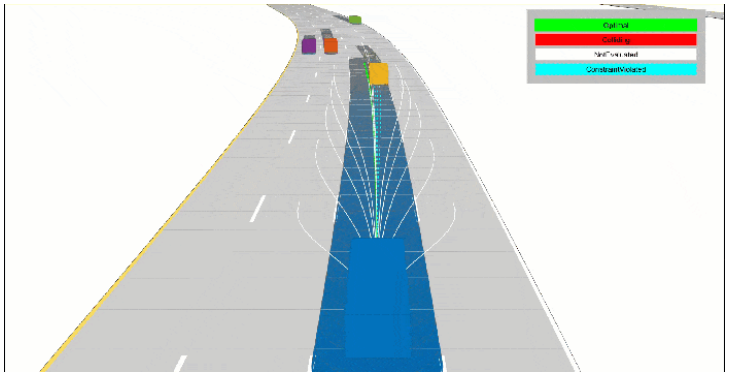
Episodic environments usually much easier to deal with



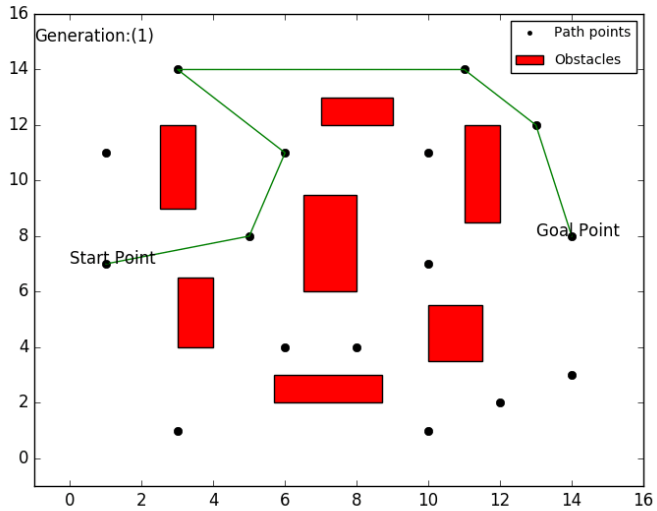
Properties of Task Environments

- Static v. Dynamic
  - Does the environment change while the agent is deliberating?
    - If yes, and an agent takes too long to decide, it counts as a decision to do nothing
  - If the environment doesn't change with time but the performance score does, the environment is **semidynamic**
  - Taxi driving is dynamic
  - Chess (played under tournament time-limit rules) is semidynamic
  - Sudoku is static
- Discrete v. Continuous
  - How are **state** and **time** handled? ***Discrete*** steps, or a **continuous** flow?
    - Some sensors discretize a continuous property; some input is technically discrete but treated as continuous (e.g. digital video)
- Known v. Unknown
  - The agent's (or designers') knowledge about the '**laws of physics**' related to a task.
    - In a known environment, the outcomes (or their probabilities) of all actions are known.
    - An unknown environment must be explored, or actions tried, to see what effect they have
  - A known environment can still be partially observable
    - A card-playing agent can't see other players' hands, but knows what game it's playing





Environments

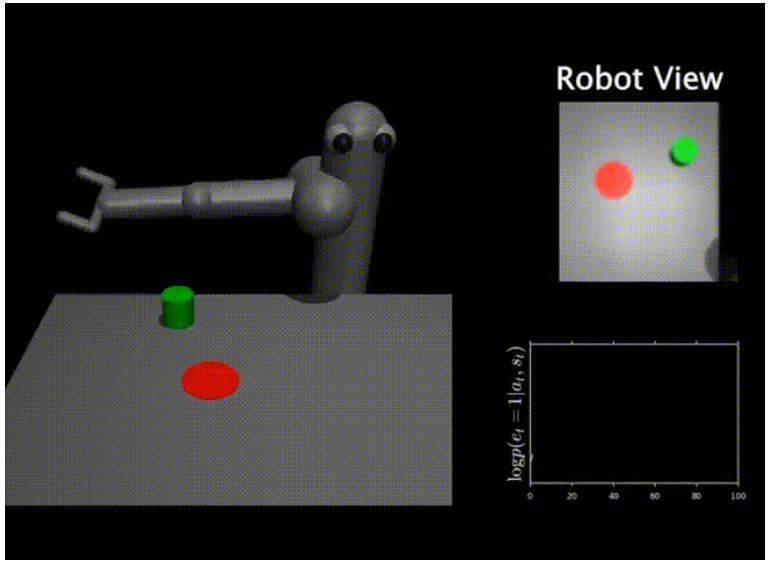


I. Most challenging cases:
   I. partially observable
   II. multiagent
   III. stochastic
   IV. sequential
   V. dynamic
   VI. Continuous
   VII. unknown
   - Driving a rental car in a new country with unfamiliar geography and traffic laws can be… interesting.

- Some environments can be classified multiple ways depending on how we define the boundaries of the problem
  - Is medical diagnosis single agent, or does the agent also have to deal with medical staff?
  - Is it single episode or sequential? (Can it propose a series of tests to refine the diagnosis? Review prior medical history from previous episodes?)

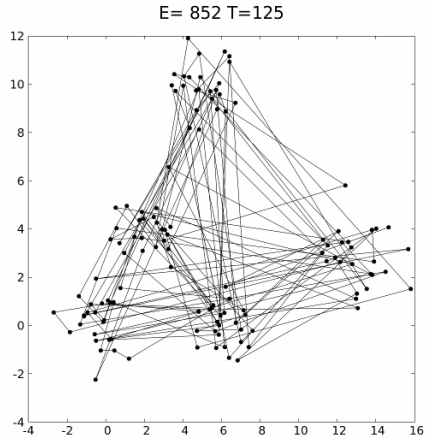
Foundations of AI

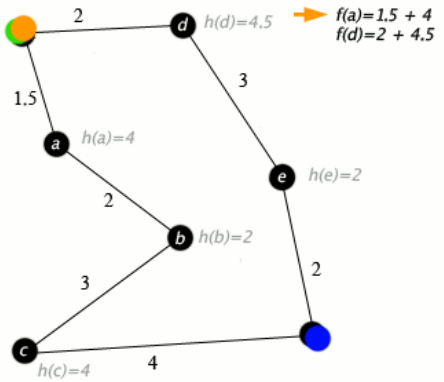
- Philosophy
  - Can formal rules be used to draw valid conclusions?
    - Rationalism
  - How does a mind arise from a physical brain?
    - Dualism (mind/body problem) v. Materialism (the body *is* the mind)
  - Where does knowledge come from?
  - How do we know what we know (epistemology)?
  - How does knowledge lead to action?
- Mathematics
  - What are the formal rules used to draw valid conclusions?
  - What can be computed?
    - Godel's Incompleteness Theorem
    - Halting problem
    - Tractability
  - How do we reason with uncertain information?
- Economics
  - How should we make decisions to maximize payoff?
    - Decision theory
  - How should we do this when others may not go along?
    - Game theory
  - How should we do this when the payoff may be far in the future?
    - Operations research, Markov processes



(Not covered in class) Looking Forward: Methods/techniques of AI

- Search algorithms
  - Unguided (blind)
    - Depth-first
    - Breadth-first
  - Guided (heuristic)
    - Hill-climbing
    - Beam search
    - Best first
    - Branch and bound
- Two-person games
  - Adversarial search
  - Iterated prisoner's dilemma
- Automated reasoning
  - Requires knowledge representation system, and inference engine
- Production rules & expert systems
- Cellular automata (complex behavior from simple rules, e.g. Conway's Game of Life)
- Neural Computation
- Genetic & Evolutionary computation
- Probabilistic & Fuzzy Reasoning





(Optional)Additional Readings

- Turing, *Computing Machinery and Intelligence*
  - http://phil415.pbworks.com/f/TuringComputing.pdf
    - Esp. Turing's treatment of various objections to AI, p. 443 & following
- Video - *Crash Course Philosophy: AI and Personhood*
  - https://www.youtube.com/watch?v=39EdqUbj92U&index=23&list=PL8dPuuaLjXtNgK6MZucdYldNkMybYIHKR