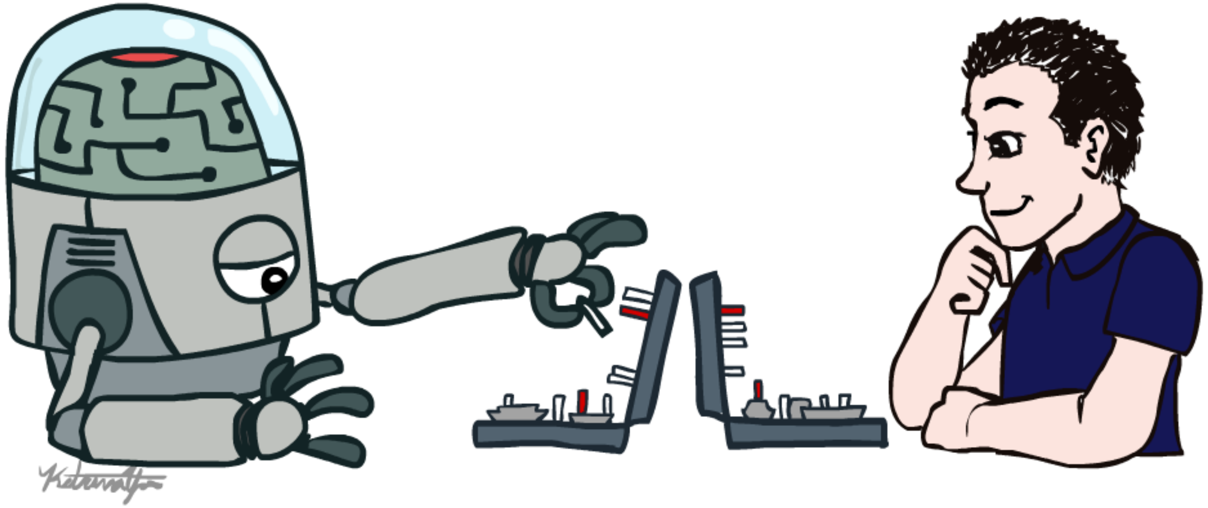
CS 188: Artificial Intelligence

Introduction



Summer 2024, Eve Fleisig & Evgeny Pobachienko

University of California, Berkeley

(slides adapted from Dan Klein, Pieter Abbeel, Anca Dragan, Stuart Russell, Saagar Sanghavi)

First Half of Today: Intro and Logistics

- Staff introductions: Evgeny, Eve, and course staff
- Course logistics
  - Lectures, discussions, office hours, and exams
  - Resources and communication platforms
  - Collaboration and academic honesty
  - DSP and extenuating circumstances
  - Stress management and mental health

Staff Introductions: Evgeny (he/him)

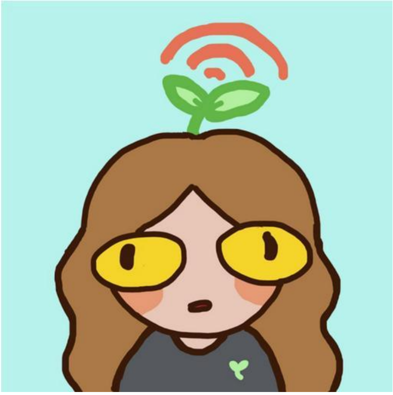
- Did my undergrad at Berkeley (2020-2023)
  - 4x Head TA for CS 188, 7x on Staff for CS 188
- Did a 4th year MS at Berkeley (2023-2024)
  - Research focus: Systems & Security
  - Advisor: Dawn Song
- First-time lecturer in EECS
  - I'm paid exclusively to care about students and staff
  - Feedback/advice/complaints are appreciated!
- Please call me "Evgeny"!
  - No "professor", "Mr.", "sir", "doctor", etc.

Staff Introductions: Eve (she/her)

Hi! I’m a rising fourth-year PhD student in EECS advised by Dan Klein. My research lies at the intersection of natural language processing and AI ethics, with a focus on preventing societal harms of language models. In my spare time, I enjoy reading, trivia, and trying to learn too many languages at once.

Our talented course staff!

Arjun Damerla
he/him



Noemi Chulo

Ademi Adeniji

Aidan Leung
he/him

Erin Tan
she/her

Jerry Sun
he/him

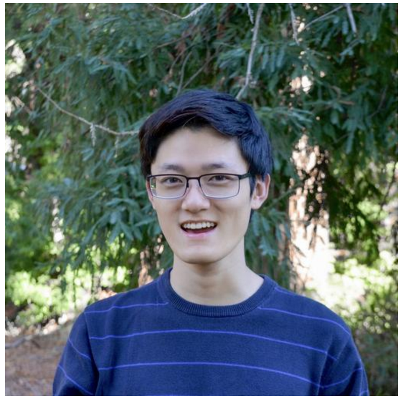
Samantha Huang
she/her

Wesley Zheng
he/him

Our talented course staff!

Advika Bhike
she/her

Andrew Choy
he/him



Curtis Hu
he/him

Danial Toktarbayev
he/him

Darren Shen
he/him

Faith Dennis
she/her

Lauren Lee
she/her

Mustafa Mirza
he/him

Our talented course staff!

Tina Rong
she/her

Enrollment

- Course staff does not control enrollment; we have to follow department policy
  - Only CS majors will be able to enroll this spring
  - More details on the course website

Course Structure

- Summer Session is DOUBLE SPEED
  - Please make sure to stay on top of material!

Course Structure: Lectures

- You are here!
- MTWT, 2:00–3:30 PM PT
- Attendance is not taken
- You can attend:
  - In-person in Genetics & Plant Bio 100
  - Asynchronously by watching recordings (posted on bCourses)

Course Structure: Discussions

- We offer three types of discussions
  - Regular discussions
  - Exam prep discussions
  - Extended-time discussions
- Discussion schedule coming today or tomorrow on website
  - Discussions start Thursday (June 20), twice a week.
- You can attend any discussion section you want (no need to enroll in a section)
  - A bit of extra credit available for attendance

Course Structure: Office Hours

- Join in-person or remotely to talk to staff about content, ask questions on assignments, or raise any concerns you have
- Schedule and queue available on website
  - Office hours start Tuesday (June 18)
- Instructor OH
  - Sticking around after lecture all lectures
  - More sign ups TBD

Course Structure: Exams

- Save the dates!
  - Midterm: Thursday, July 11, 2–4 PM PT
  - Final exam: Thursday, August 8, 2–5 PM PT
- If you can’t make it:
  - We’ll offer an in-person-only alternate exam right after the listed time
  - Emergencies resolved on a case-by-case basis.
- More logistics closer to the exam

Resources

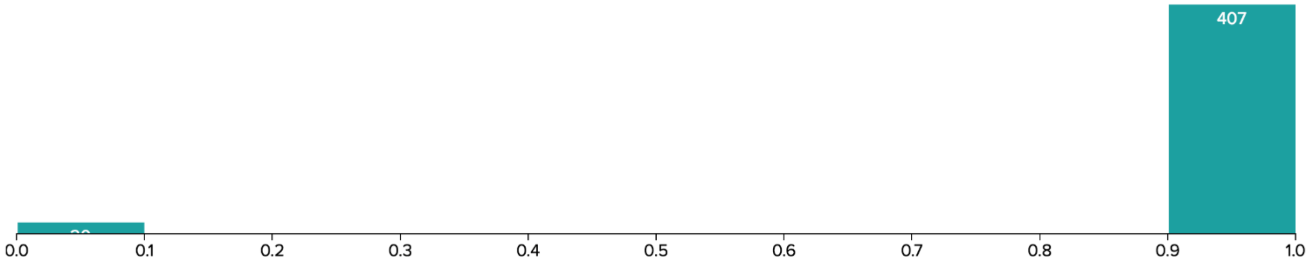
- Course website: https://inst.eecs.berkeley.edu/~cs188/su24/
  - All resources (slides, notes, recordings, assignments, etc.) posted here
- Ed: Discussion forum
  - Can make private posts for debugging but use code blocks for code!
- Staff email for private concerns: cs188@berkeley.edu
  - Making a private post on Ed is easier/faster
- Gradescope: Submit assignments here
  - Make sure project grade is what you expect!

Grading Structure

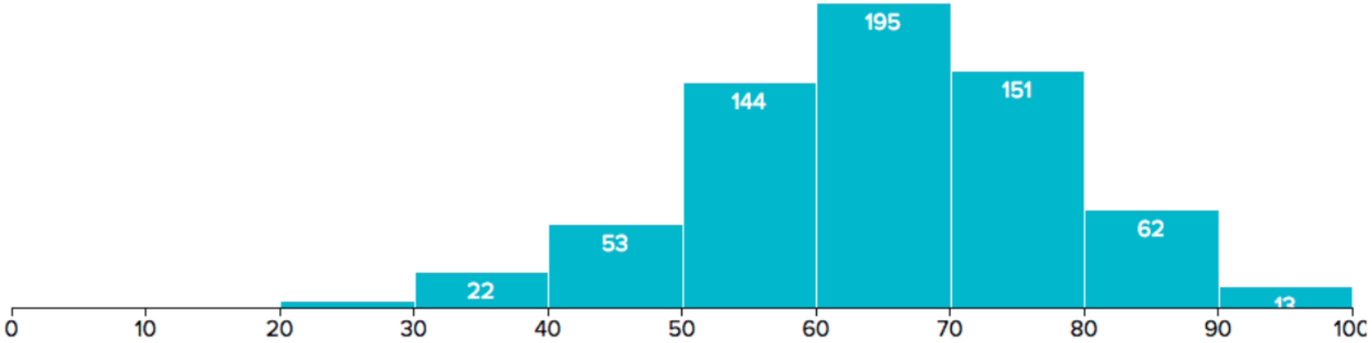
- Projects (25%)
  - Python programming assignments, autograded
  - You can optionally work with a partner
  - Reduced credit for submitting late, unless you have an extension
- Homework (20%)
  - Electronic homework: Autograded on Gradescope
  - Written homework: graded by TAs on correctness
  - Submit individually (but feel free to discuss with others)
  - No late submissions, unless you have an extension
- Midterm (20%), Final Exam (35%)

Some Historical Statistics

- Homework and projects: instruction (iterate/learn till you nailed it)



- Exams: assessment



Extensions and Accommodations

- We’ll drop your lowest homework score
- You have 5 slip days to use across the projects
  - See course policies page for details on how they work
- If you ever need an extension, please request one!
  - We’re here to support you, and we understand that life happens.
  - Extension form will be posted on the website

DSP

- Disabled Students' Program (DSP)
  - There's a variety of accommodations UC Berkeley can help us set up for you in this class
  - https://dsp.berkeley.edu/
- Are you facing barriers in school due to a disability?
  - Apply to DSP!
  - We maintain proper access controls on this information: Only instructors, course managers, head TAs, and logistics TAs can access any DSP-related info
- Our goal is to teach you the material in our course. The more accessible we can make it, the better.

Collaboration and Academic Dishonesty

- We're here to help! There are plenty of staff and resources available for you
  - You can always talk to a staff member if you're feeling stressed or tempted to cheat
  - Collaboration on homework is okay, but please cite collaborators
  - Do not post solutions online or share with others!
- Academic dishonesty policies
  - Reported to Center of Student Conduct
  - Negative points on assignments, and/or F in the class

Stress Management and Mental Health

- **Your health is more important than this course**
- If you feel overwhelmed, there are options
  - Academically: Ask on Ed, talk to staff in office hours, set up a meeting with staff to make a plan for your success this semester
  - Non-academic:
    - Counselling and Psychological Services (CAPS) has multiple free, confidential services
      - Casual consultations: https://uhs.berkeley.edu/counseling/lets-talk
      - Crisis management: https://uhs.berkeley.edu/counseling/urgent
    - Check out UHS's resources: https://uhs.berkeley.edu/health-topics/mental-health

Announcements

- Project 0: Python Setup & Tutorial
  - Due Fri, Jun 21 at 11:59pm
- Homework 1 Part 1 & 2: Math Review and Search
  - Due Fri, Jun 21 at 11:59pm
- Project 1: Search
  - Best way to test programming preparedness
  - Due Tue, Jun 25 at 11:59pm
- Sections & OH start this week

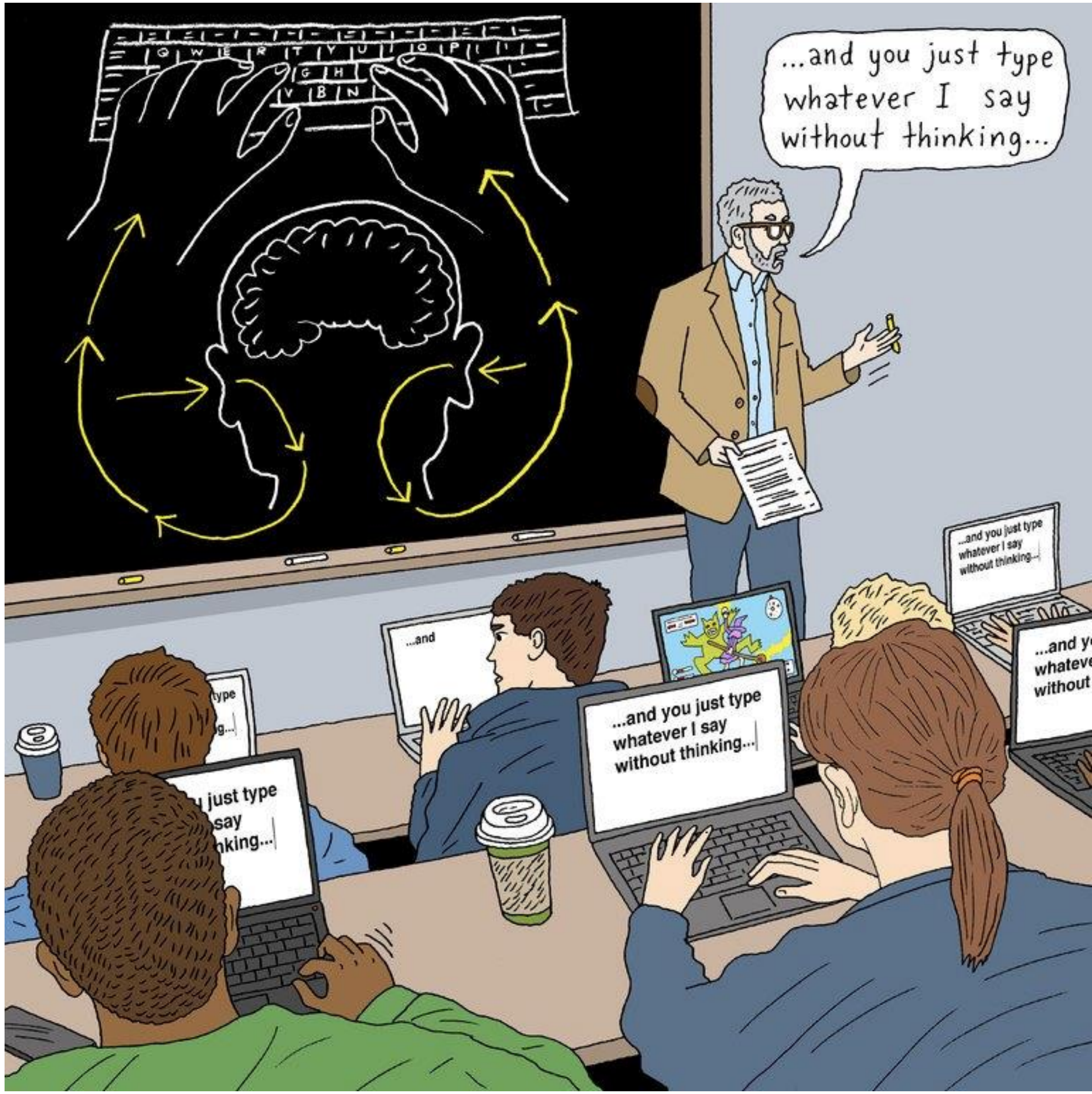
Laptops in Lecture?

The New York Times

Laptops Are Great. But Not During a Lecture or a Meeting.

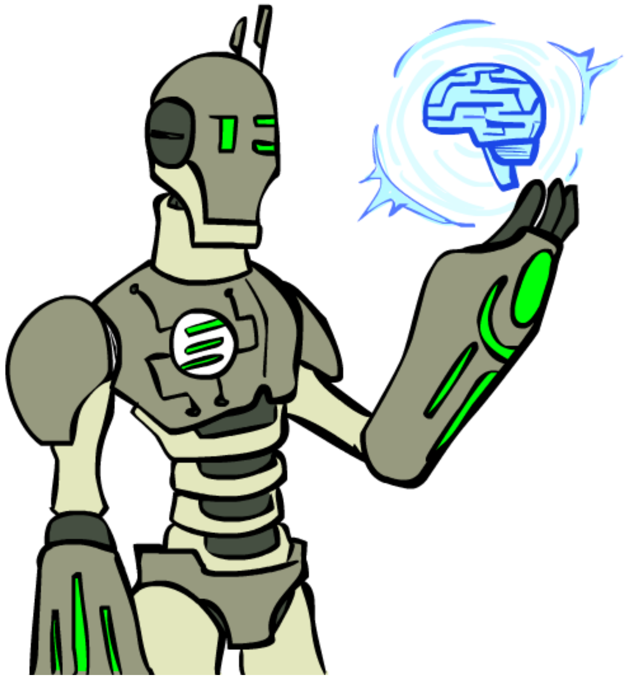
Economic View

By SUSAN DYNARSKI   NOV. 22, 2017

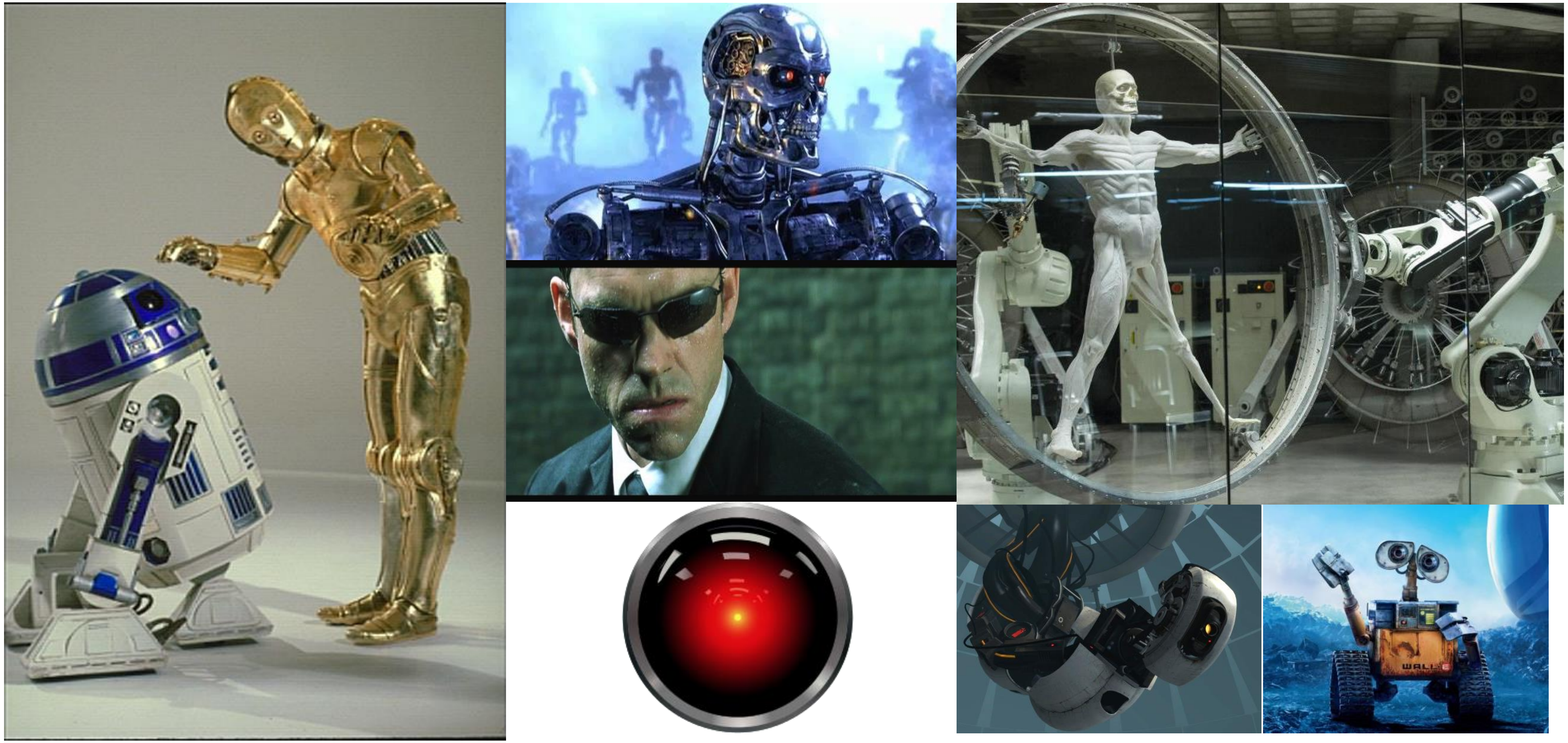


Second Half of Today: What is AI?

- What is artificial intelligence?
- What can AI do?
  - What should we worry about?
  - What can we do about those things?
  - What should we not worry about?
- What is this course?



Sci-Fi AI?




WAYMO


TUG
CAUTION
MAY CONTAIN
CHEMOTHERAPY DRUG
CAUTION
MAY CONTAIN
CHEMOTHERAPY DRUG

SS

What is artificial intelligence?

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think and learn like humans. It is a broad field of computer science that focuses on creating intelligent machines capable of performing tasks that typically require human intelligence, such as visual perception, speech recognition, decision-making, problem-solving, and language translation.

AI encompasses various subfields and techniques, including machine learning, natural language processing, computer vision, expert systems, and robotics. These approaches enable AI systems to acquire knowledge, process information, reason, and make predictions or decisions based on the available data.

Machine learning, a key component of AI, involves training algorithms to recognize patterns in large amounts of data and make predictions or take actions without being explicitly programmed. This ability to learn from experience and adapt to new situations is what sets AI apart from traditional software systems.

What is AI?

The science of making machines that:

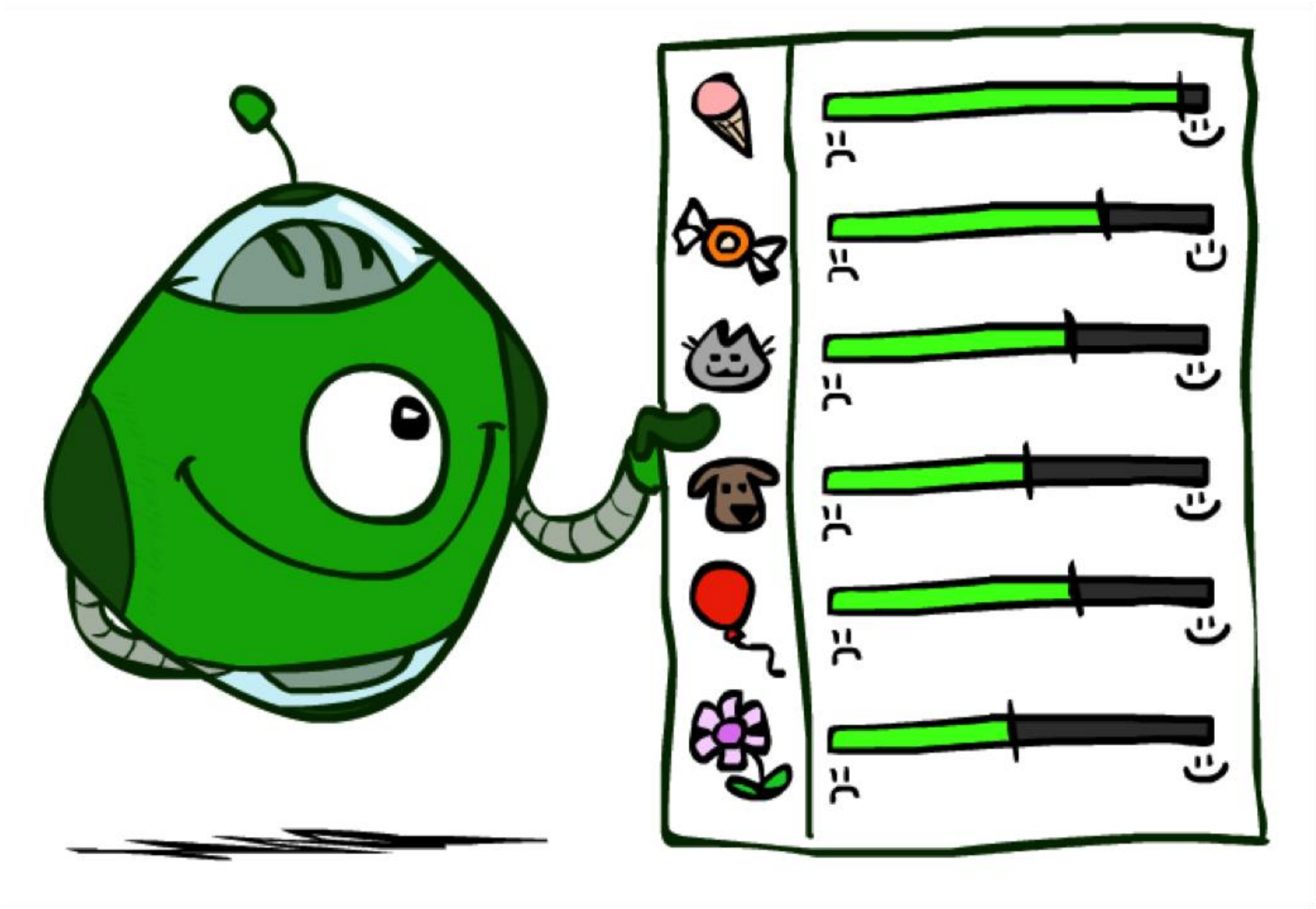
Rational Decisions

- We'll use the term **rational** in a very specific, technical way:
  - Rational: maximally achieving pre-defined goals
  - Rationality only concerns what decisions are made (not the thought process behind them)
  - Goals are expressed in terms of the **utility** of outcomes
  - Being rational means **maximizing your expected utility**

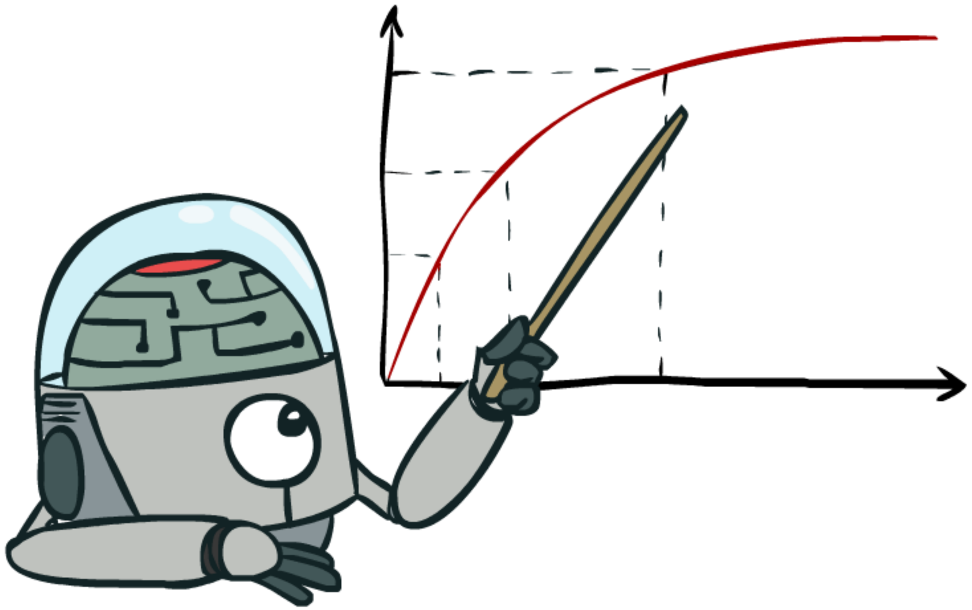
A better title for this course would be:

Computational Rationality

Utilities



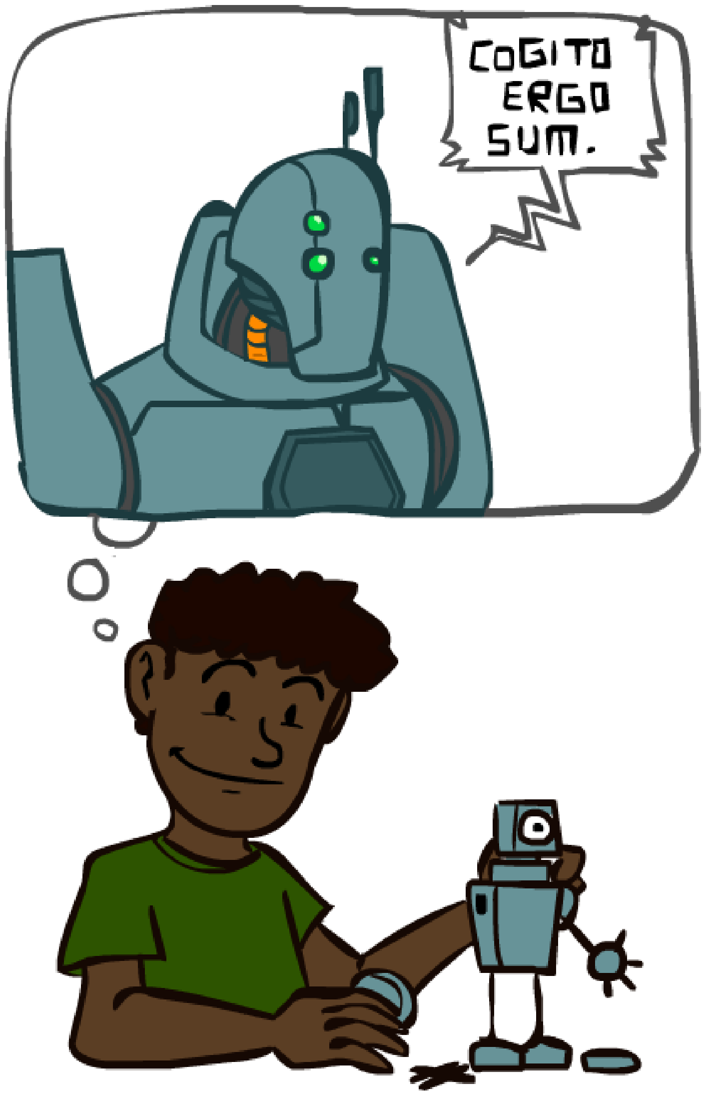
Maximize Your Expected Utility



What About the Brain?

- Brains (human minds) are very good at making rational decisions, but not perfect
- Brains aren’t as modular as software, so hard to reverse engineer!
- “Brains are to intelligence as wings are to flight”
- Lessons learned from the brain: memory and simulation are key to decision making

A (Short) History of AI



A (Short) History of AI

- 1940-1950: Early days
  - 1943: McCulloch & Pitts: Boolean circuit model of brain
  - 1950: Turing's “Computing Machinery and Intelligence”
- 1950—70: Excitement: Look, Ma, no hands!
  - 1950s: Early AI programs, including Samuel's checkers program, Newell & Simon's Logic Theorist, Gelernter's Geometry Engine
  - 1956: Dartmouth meeting: “Artificial Intelligence” adopted
  - 1965: Robinson's complete algorithm for logical reasoning

A (Short) History of AI

- 1970—90: Knowledge-based approaches
  - 1969—79: Early development of knowledge-based systems
  - 1980—88: Expert systems industry booms
  - 1988—93: Expert systems industry busts: “AI Winter”
- 1990—: Statistical approaches
  - Resurgence of probability, focus on uncertainty
  - General increase in technical depth
  - Agents and learning systems… “AI Spring”?
  - 1996: Kasparov defeats Deep Blue at chess
  - 1997: Deep Blue defeats Kasparov at chess

“I could feel --- I could smell --- a new kind of intelligence across the table.” ~Kasparov

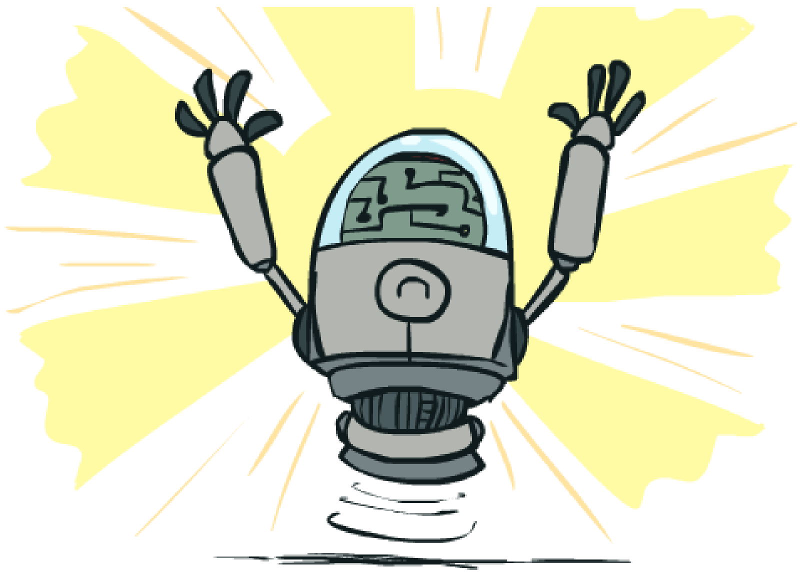
A (Short) History of AI

- 2000—: Where are we now?
  - Big data, big compute, neural networks
  - Some re-unification of sub-fields
  - AI used in many industries
  - Chess engines running on ordinary laptops can defeat the world's best chess players
  - 2011: IBM's Watson defeats Ken Jennings and Brad Rutter at Jeopardy!
  - 2016: Google's AlphaGo beats Lee Sedol at Go
  - 2017: Google's Transformer NN architecture
  - 2022: OpenAI's ChatGPT released, LLMs gain massive popularity

What Can AI Do?

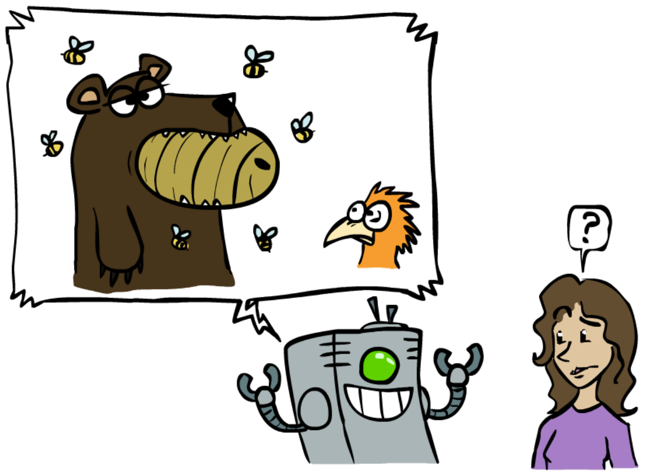
Quiz: Which of the following can be done at present?

- ✓ Play a decent game of Jeopardy?
- ✓ Win against any human at chess?
- ✓ Win against the best humans at Go?
- ✓ Play a decent game of tennis?
- ✓ Grab a particular cup and put it on a shelf?
- ✗ Unload any dishwasher in any home?
- ? Drive safely along the highway?
- ✗ Drive safely along Telegraph Avenue?
- ✓ Buy a week's worth of groceries on the web?
- ✗ Buy a week's worth of groceries at Berkeley Bowl?
- ? Discover and prove a new mathematical theorem?
- ✗ Perform a surgical operation?
- ✗ Unload a know dishwasher in collaboration with a person?
- ✓ Translate spoken Chinese into spoken English in real time?
- ✗ Write an intentionally funny story?



Unintentionally Funny Stories

- One day Joe Bear was hungry. He asked his friend Irving Bird where some honey was. Irving told him there was a beehive in the oak tree. Joe walked to the oak tree. He ate the beehive. The End.
- Henry Squirrel was thirsty. He walked over to the river bank where his good friend Bill Bird was sitting. Henry slipped and fell in the river. Gravity drowned. The End.
- Once upon a time there was a dishonest fox and a vain crow. One day the crow was sitting in his tree, holding a piece of cheese in his mouth. He noticed that he was holding the piece of cheese. He became hungry, and swallowed the cheese. The fox walked over to the crow. The End.



[Shank, Tale-Spin System, 1984]

AI in the News

Elon Musk
@elonmusk

Tesla Full Self-Driving Beta is now available to anyone in North America who requests it from the car screen, assuming you have bought this option.

Congrats to Tesla Autopilot/AI team on achieving a major milestone!

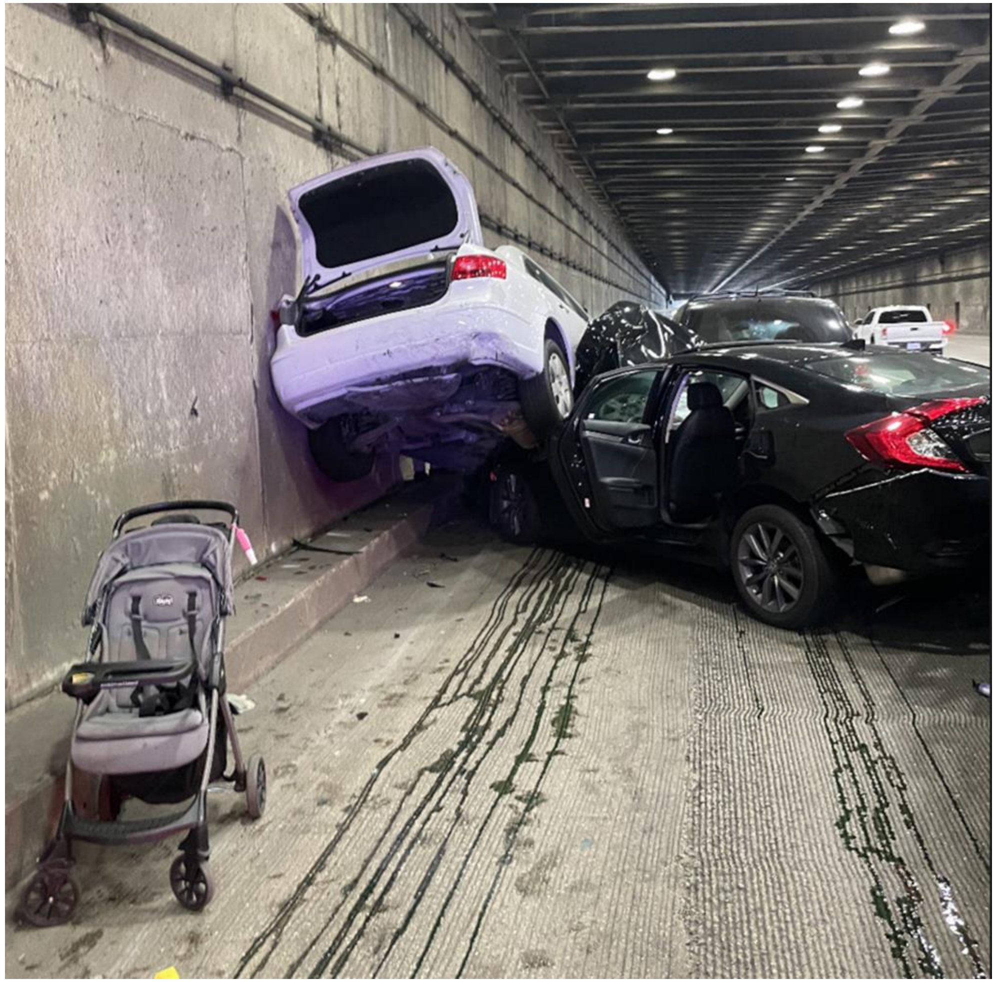
11:34 PM · Nov 23, 2022

12.9K Retweets **2,651** Quote Tweets **174.1K** Likes

AI in the News

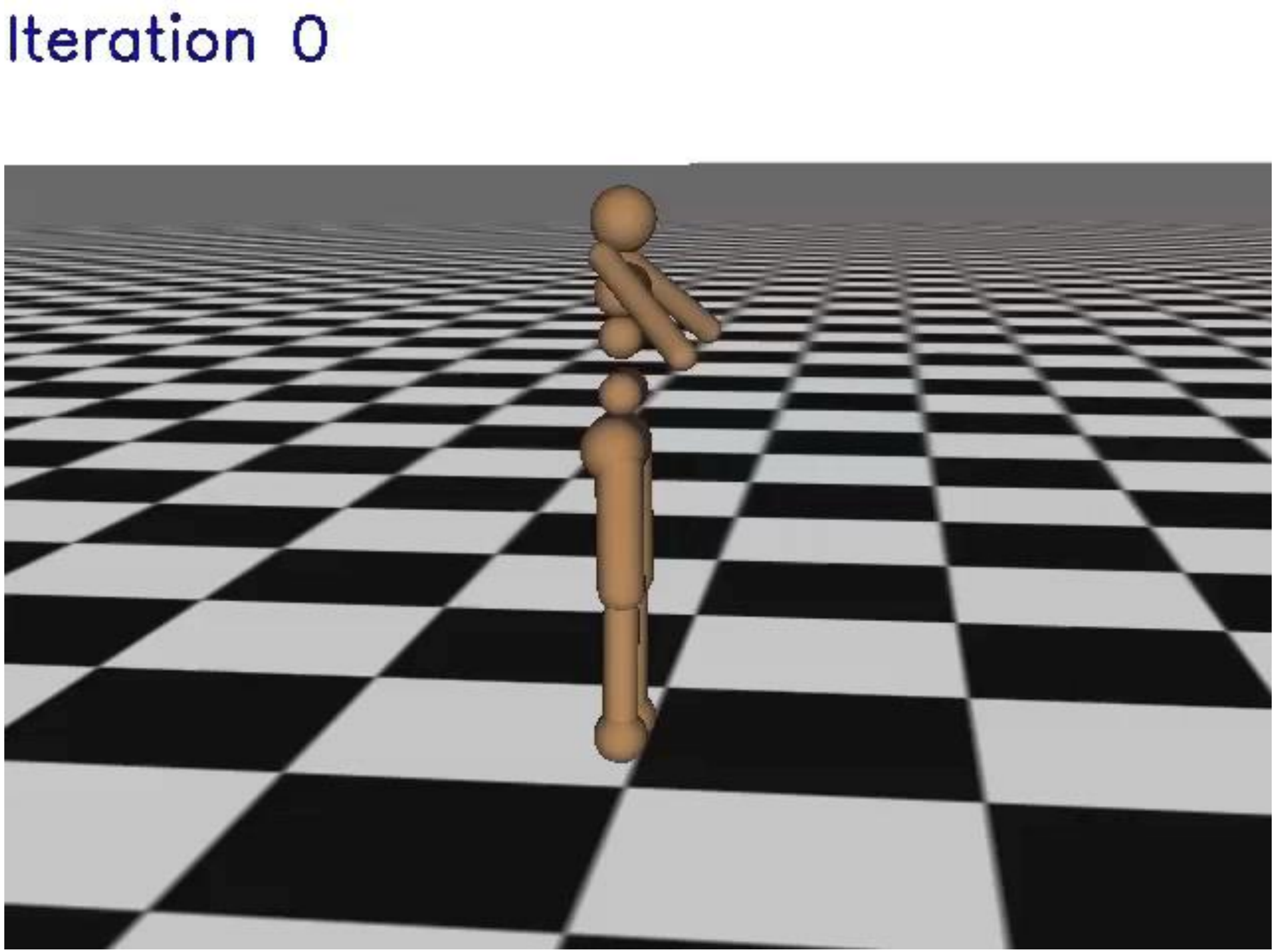
Highway surveillance footage from November 24 shows a Tesla Model S vehicle changing lanes and then abruptly braking in the far-left lane of the San Francisco Bay Bridge, resulting in an eight-vehicle crash.

As traditional car manufacturers enter the electric vehicle market, Tesla is increasingly under pressure to differentiate itself. Last year, Musk said that “Full Self-Driving” was an “essential” feature for Tesla to develop, going as far as saying, “It’s really the difference between Tesla being worth a lot of money or worth basically zero.”



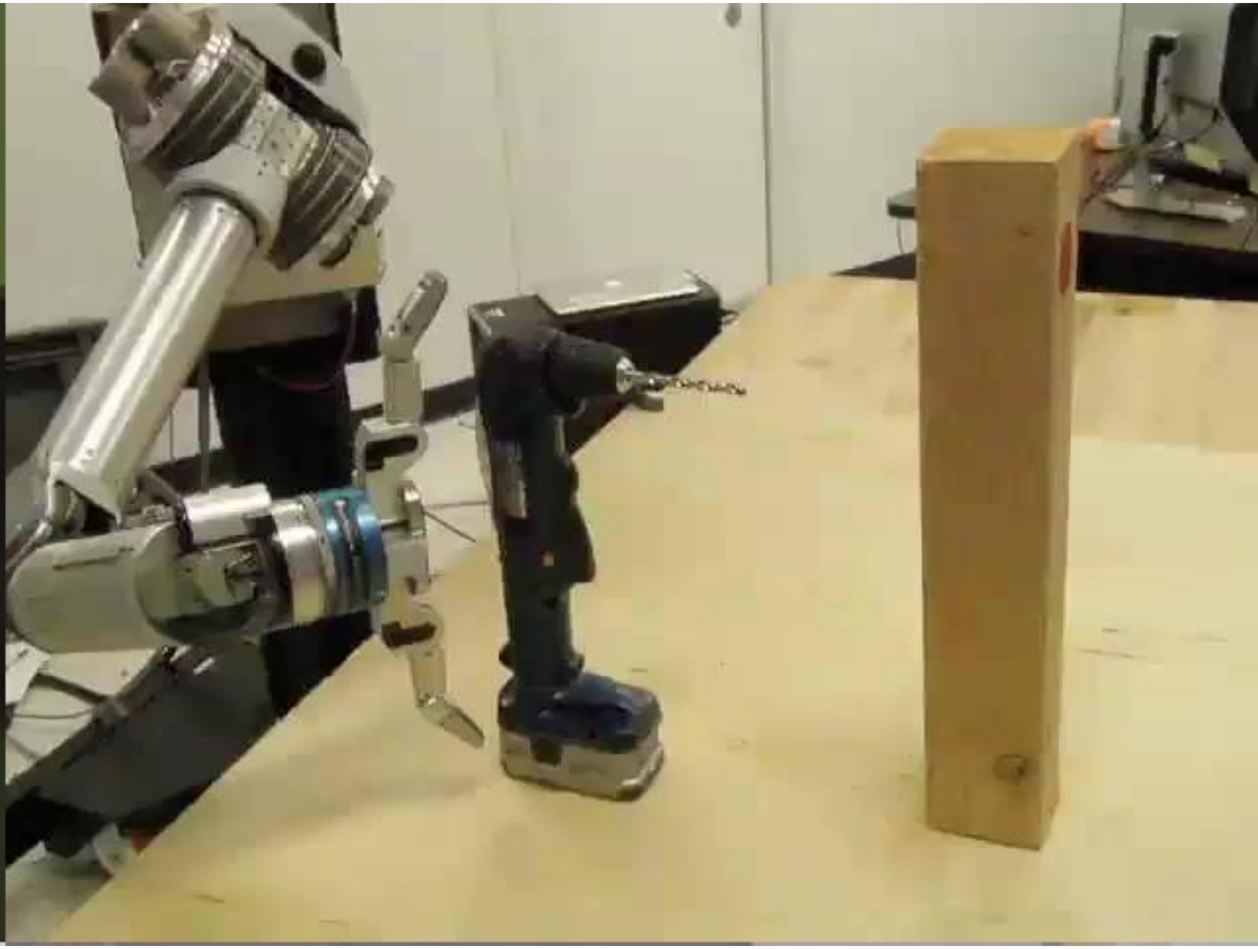
https://theintercept.com/2023/01/10/tesla-crash-footage-autopilot/

Simulated Agents

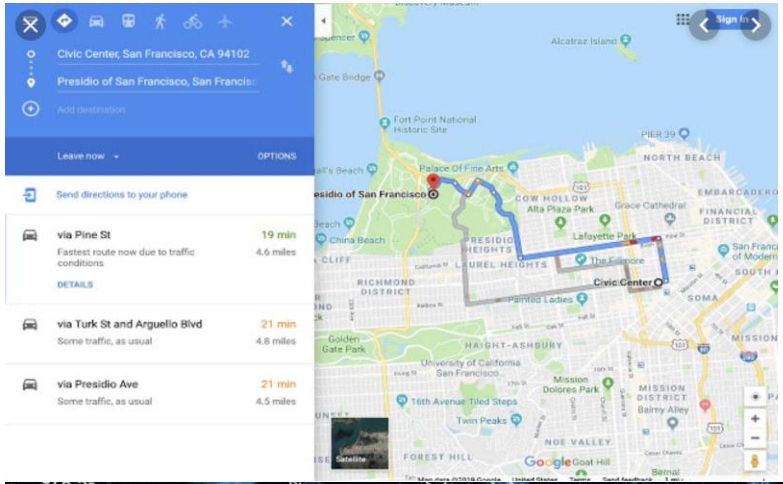


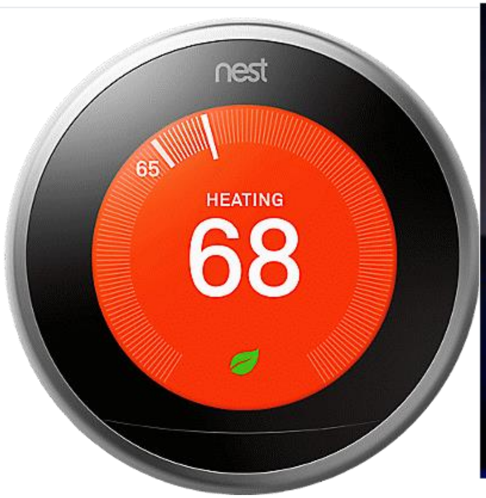
[Schulman, Moritz, Levine, Jordan, Abbeel, ICLR 2016]

Robots



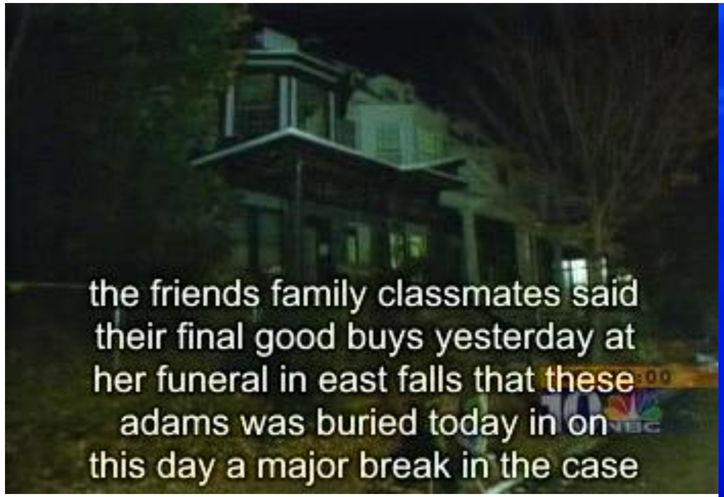
Tools for Predictions & Decisions





Natural Language

- Speech technologies (e.g. Siri)
  - Automatic speech recognition (ASR)
  - Text-to-speech synthesis (TTS)
  - Dialog systems



- Language processing technologies
  - Question answering
  - Machine translation

  - Web search
  - Text classification, spam filtering, etc…

Computer Vision

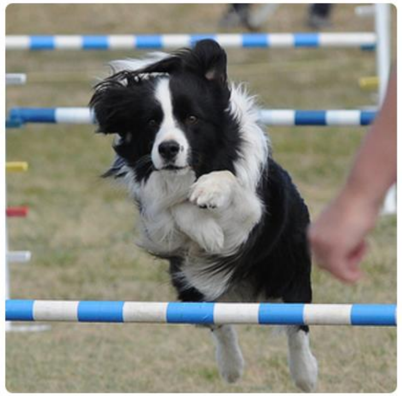
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."



"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

Course Topics

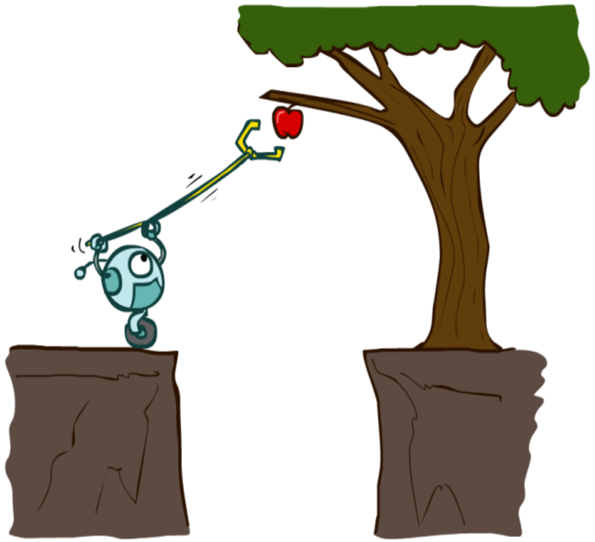
- Part 1: Intelligence from Computation
  - Fast search/planning
  - Constraint satisfaction (e.g. scheduling)
  - Adversarial and uncertain search (e.g. routing, navigation)
- Part 2: Intelligence from Data
  - Probabilistic inference with Bayes nets (e.g. robot localization)
  - Decision theory
  - Supervised machine learning (e.g. spam detection)
- Throughout: Applications
  - Natural language, vision, robotics, games, etc.

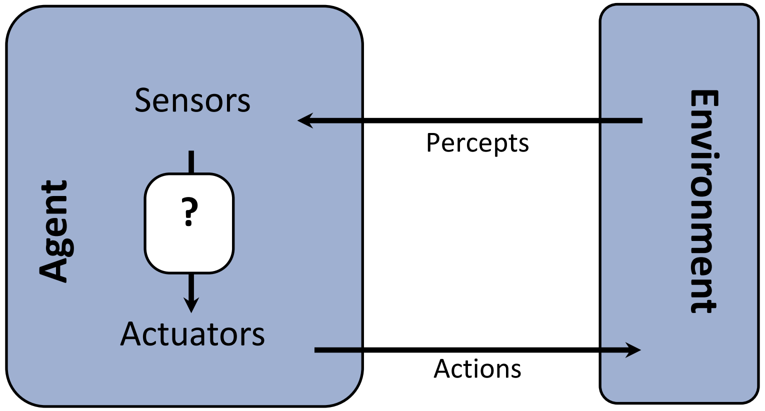
Should I take CS 188?

- Yes, if you want to know how to design rational agents!
  - CS 188 gives you extra mathematical maturity
  - CS 188 gives you a survey of other non-CS fields that interact with AI (e.g. robotics, economics)
- Disclaimer: If you’re interested in making yourself more competitive for AI jobs, CS 189 and CS 182 are better fits.
  - The last few CS 188 lectures (neural networks) are used by many modern state-of-the-art systems. CS 189 and CS 182 cover these in more depth

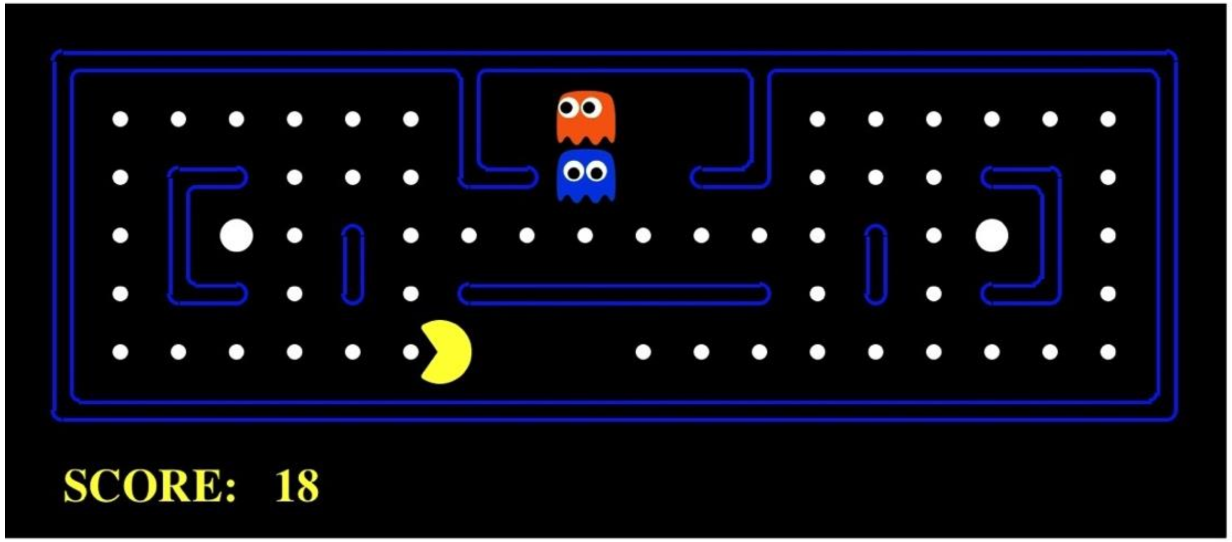
Designing Rational Agents

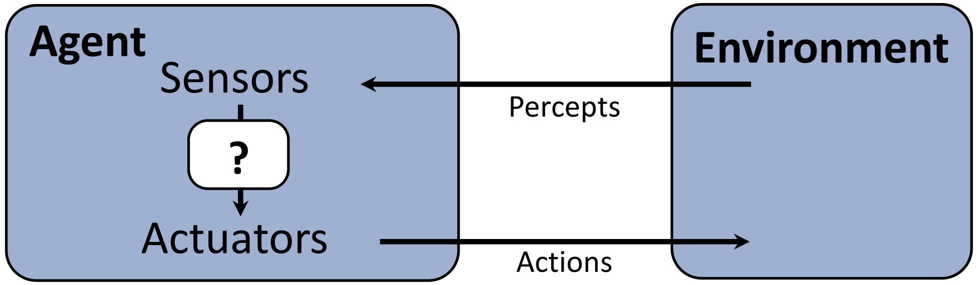
- An **agent** is an entity that perceives and acts.
- A **rational agent** selects actions that maximize its (expected) **utility**.
- Characteristics of the **percepts**, **environment**, and **action space** dictate techniques for selecting rational actions
- This course is about:
  - General AI techniques for a variety of problem types
  - Learning to recognize when and how a new problem can be solved with an existing technique



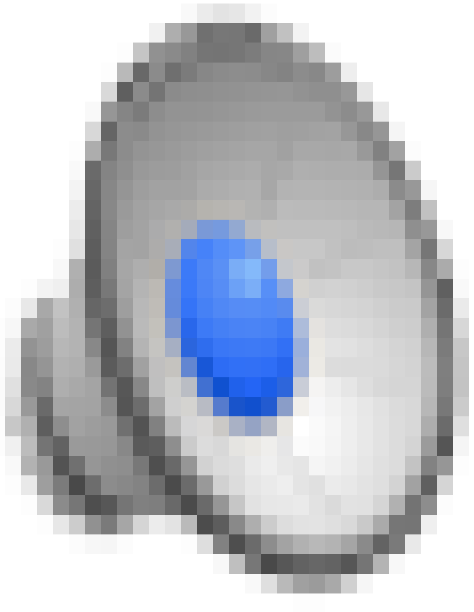


Pac-Man as an Agent

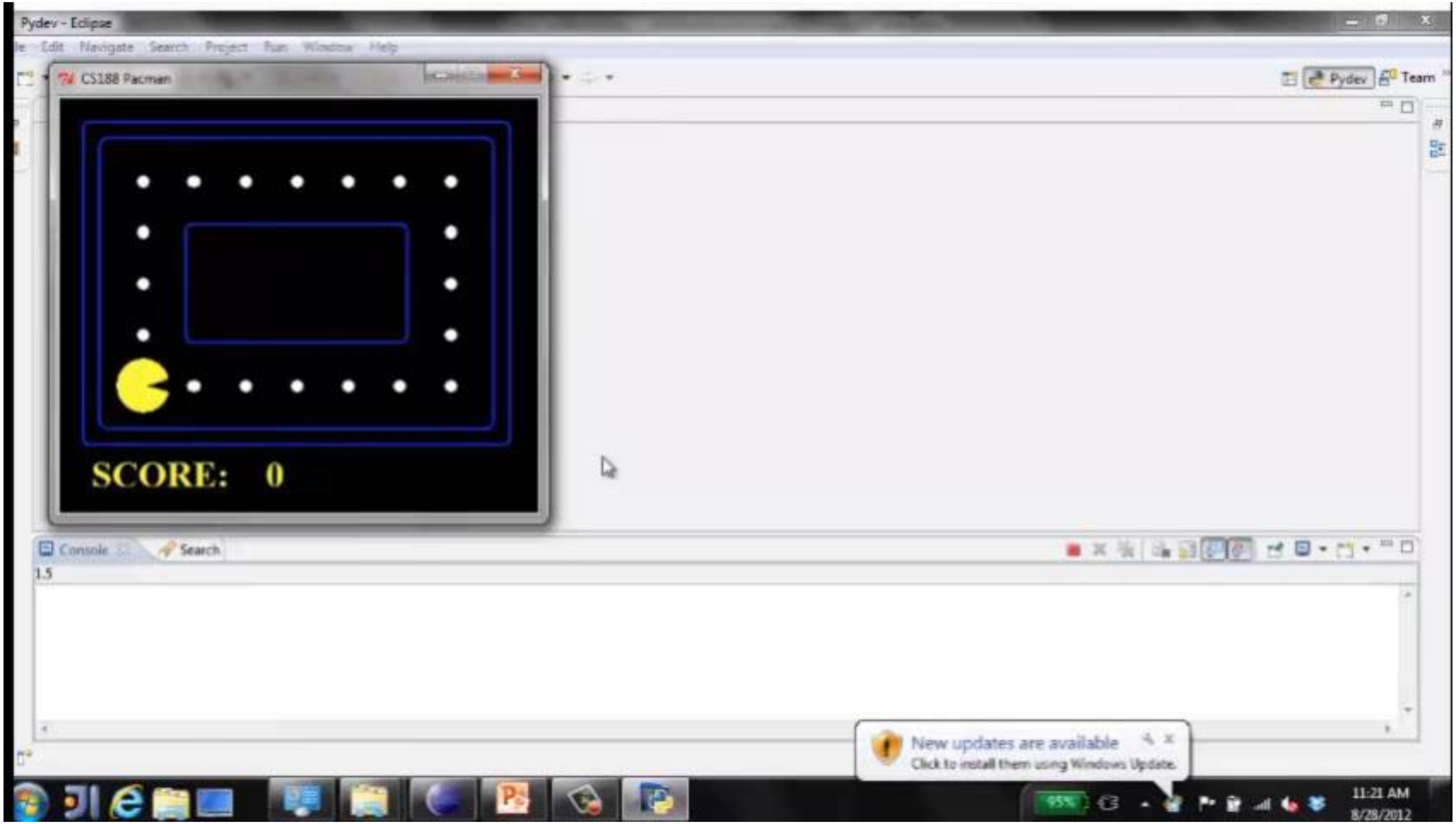




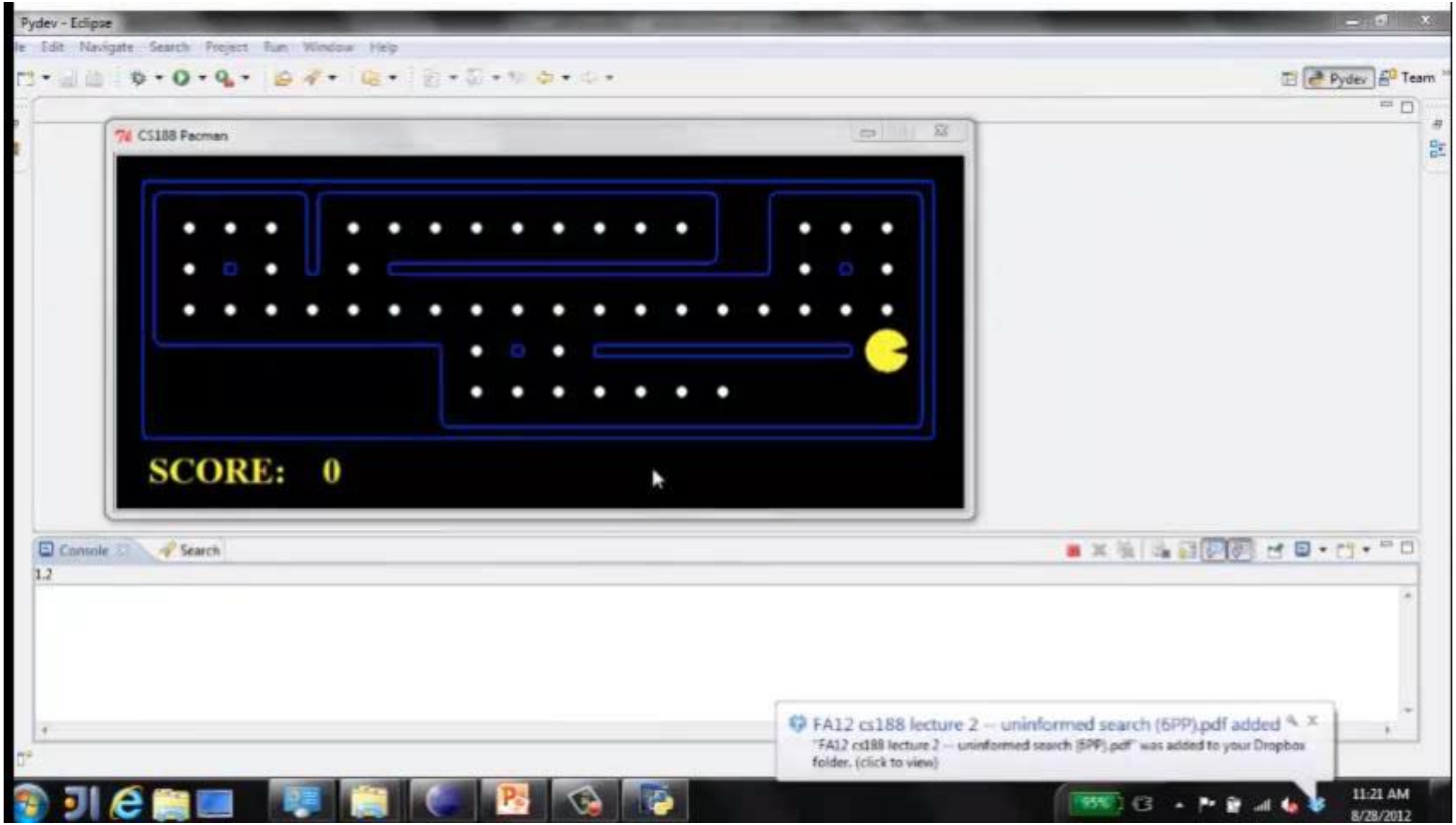
Pac-Man is a registered trademark of Namco-Bandai Games, used here for educational purposes



Eat adjacent dot, if any



Eat adjacent dot, if any

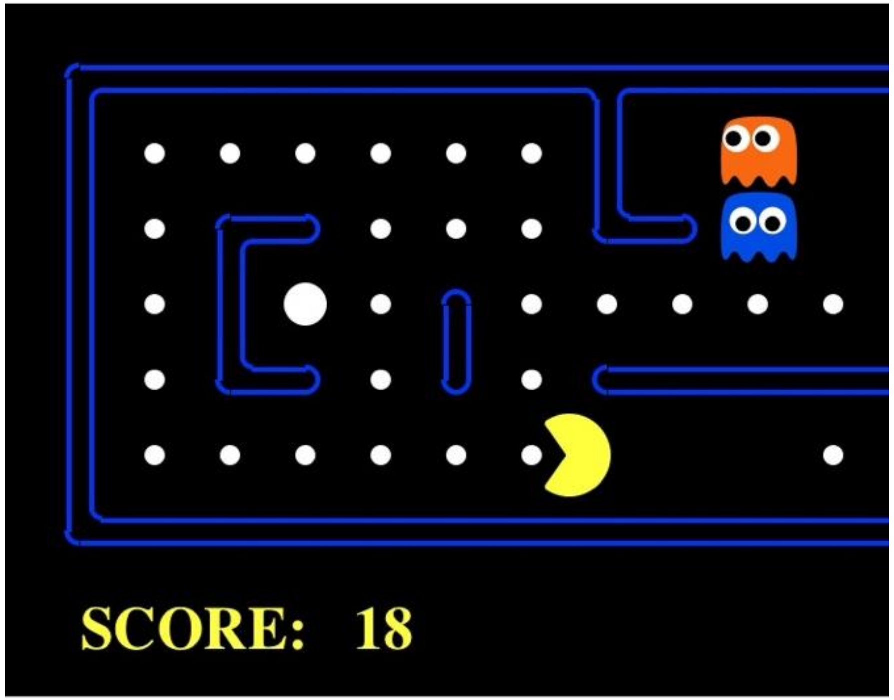


Pacman agent contd.

- 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…..
  - *How large would it be?*

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 or NSEW
- Sensors
  - Entire state is visible (except power pellet duration)



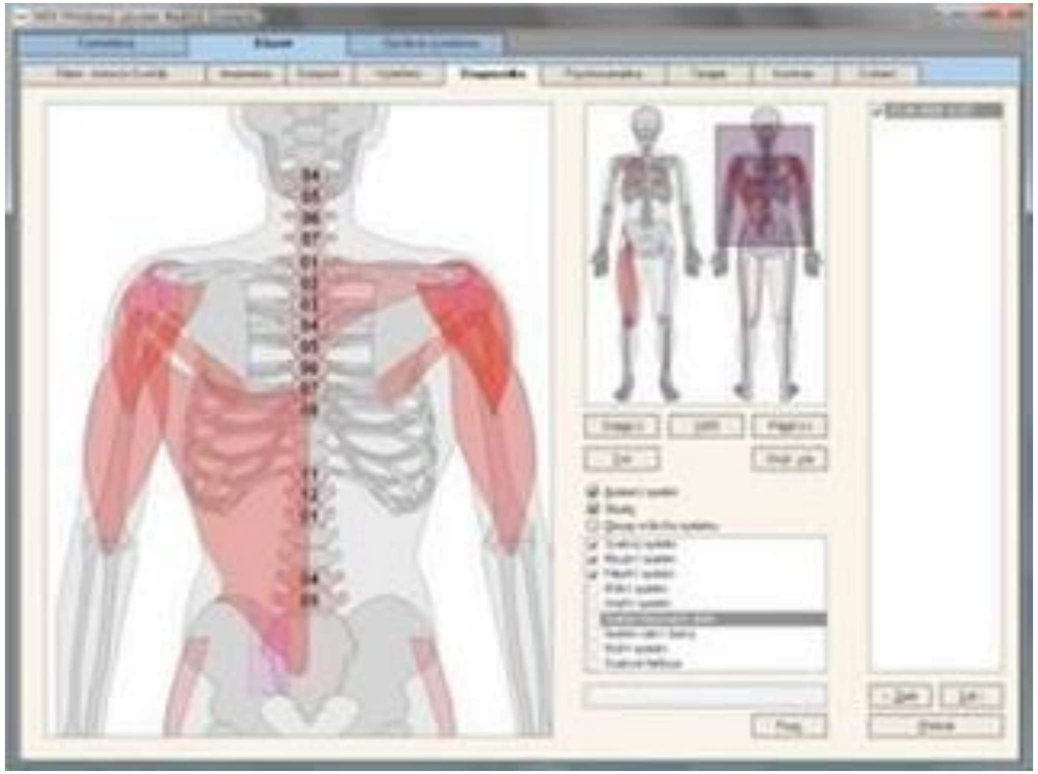
PEAS: Automated taxi

- 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

https://www.sfchronicle.com/sf/article/cruise-waymo-driverless-cars-in-s-f-18282902.php

PEAS: Medical diagnosis system

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



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***

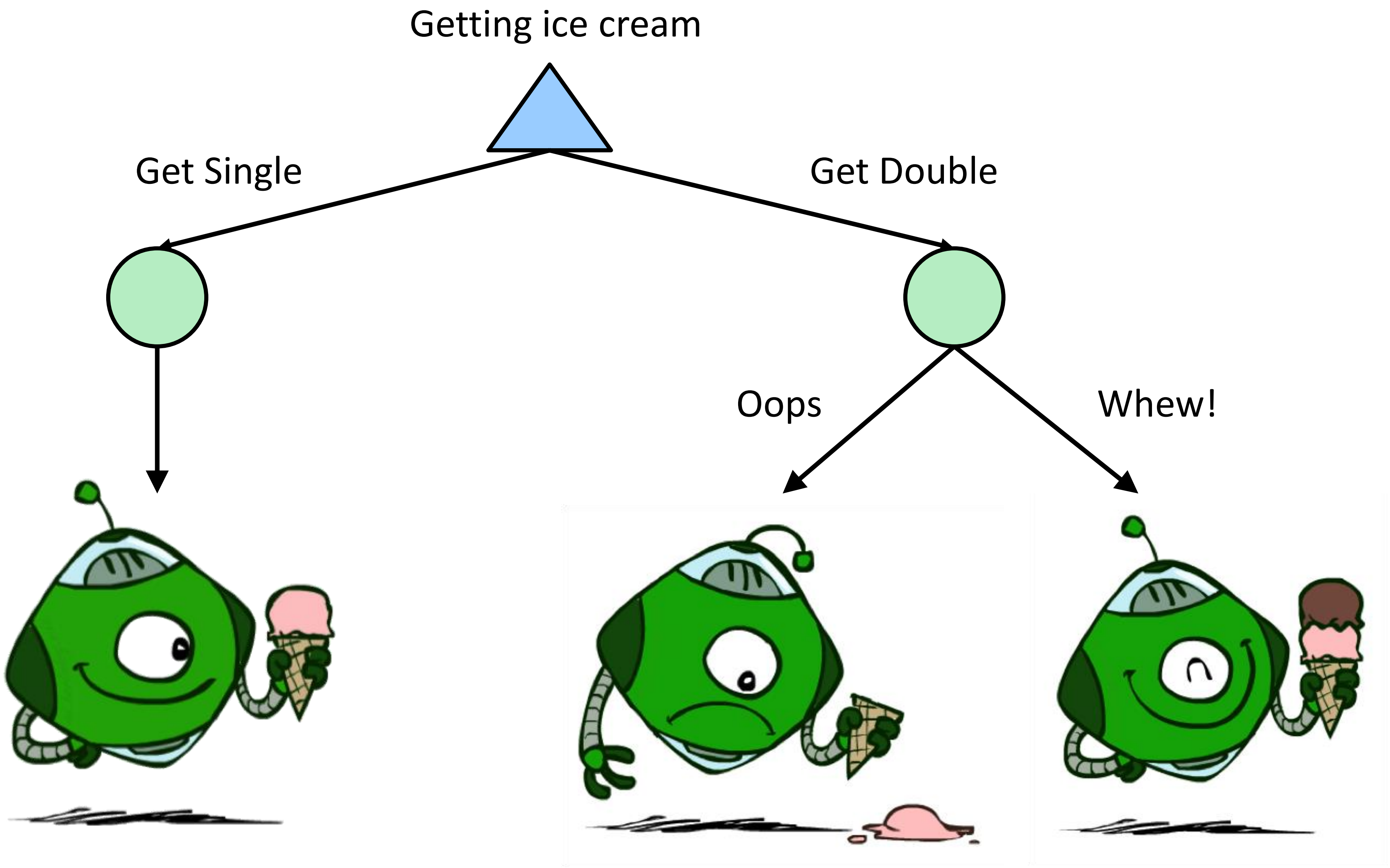
Maximum Expected Utility

- Questions:
  - Where do utilities come from?
  - How do we know such utilities even exist?
  - How do we know that averaging even makes sense?
  - What if our behavior (preferences) can't be described by utilities?

Utilities

- Utilities are functions from outcomes (states of the world) to real numbers that describe an agent's preferences
- Where do utilities come from?
  - In a game, may be simple (+1/-1)
  - Utilities summarize the agent's goals
  - Theorem: any "rational" preferences can be summarized as a utility function
- We hard-wire utilities and let behaviors emerge
  - Why don't we let agents pick utilities?
  - Why don't we prescribe behaviors?

Utilities: Uncertain Outcomes



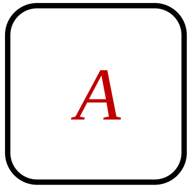
Preferences

- An agent must have preferences among:
  - Prizes: A, B, etc.
  - Lotteries: situations with uncertain prizes

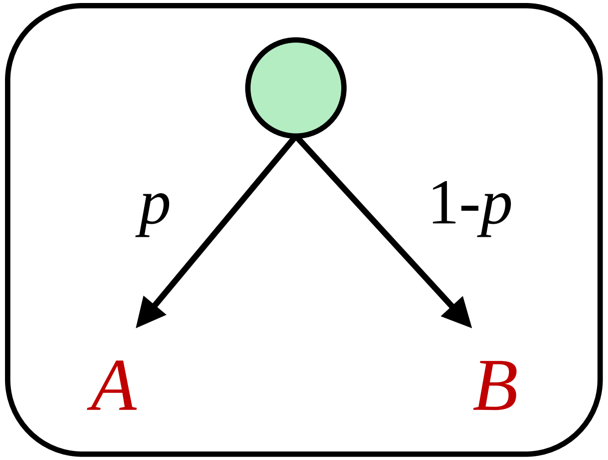
    $L = [p, A;\ (1-p), B]$

- Notation:
  - Preference: $A > B$
  - Indifference: $A \sim B$

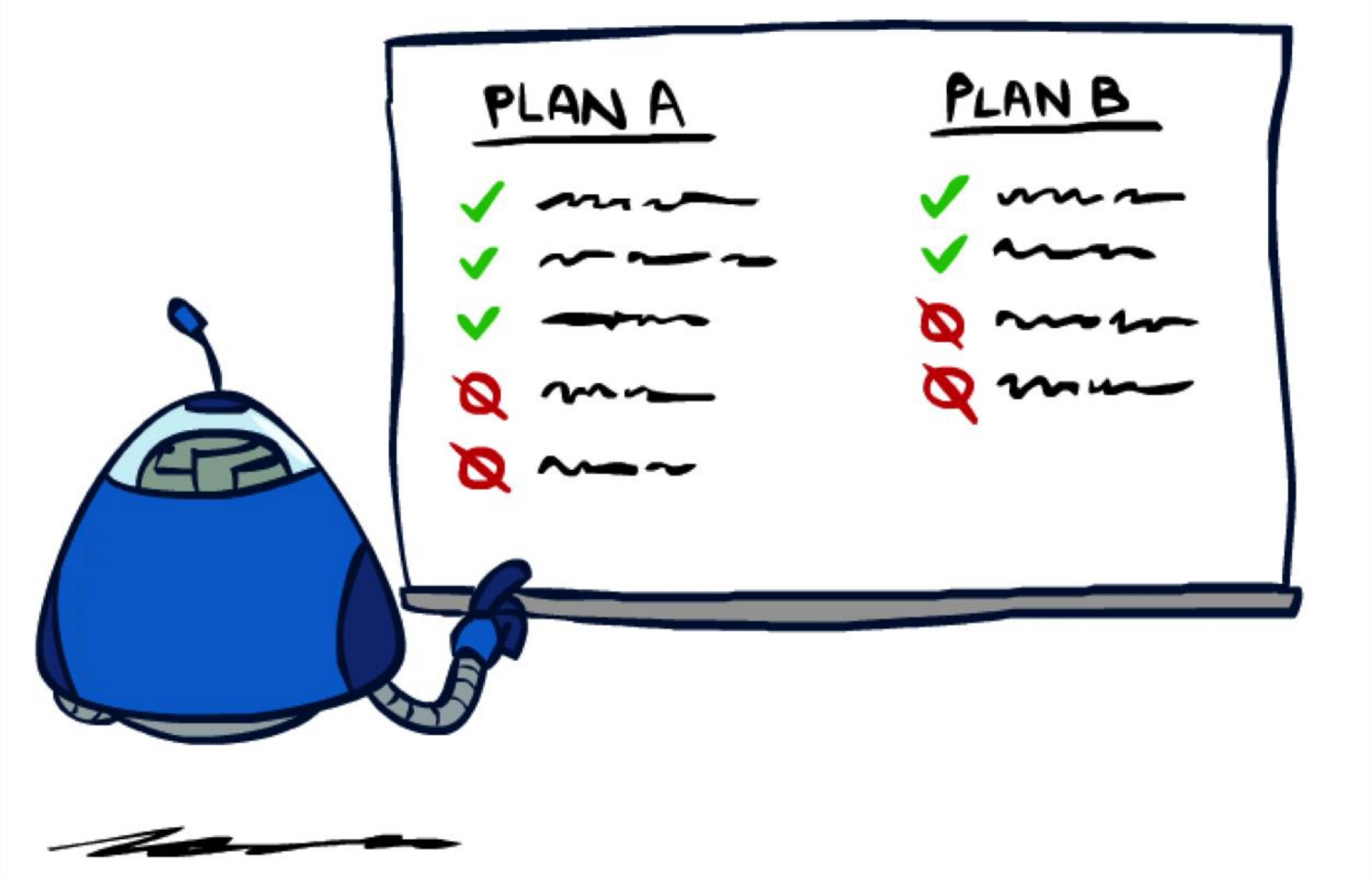
A Prize



A Lottery



Rationality

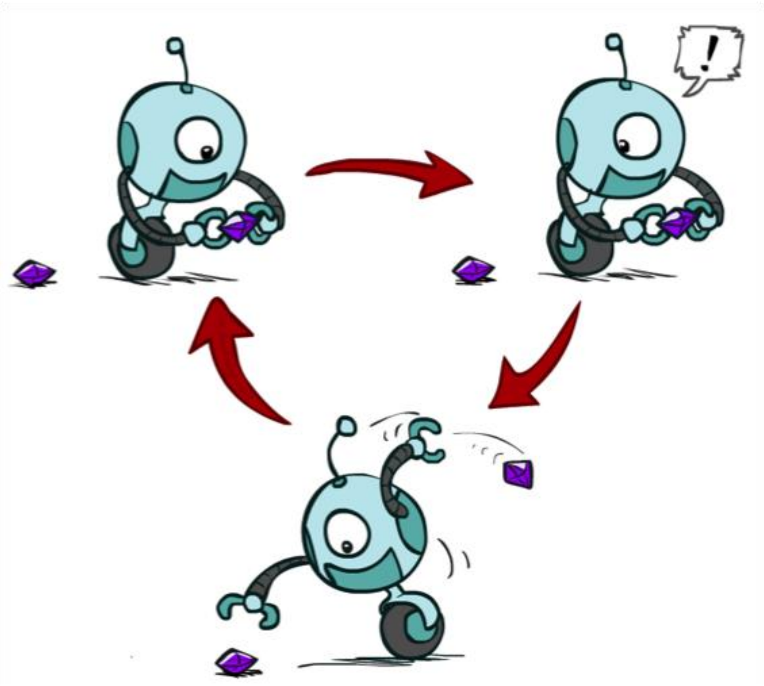


Rational Preferences

- We want some constraints on preferences before we call them rational, such as:

Axiom of Transitivity: $(A > B) \wedge (B > C) \Rightarrow (A > C)$

- For example: an agent with intransitive preferences can be induced to give away all of its money
  - If B > C, then an agent with C would pay (say) 1 cent to get B
  - If A > B, then an agent with B would pay (say) 1 cent to get A
  - If C > A, then an agent with A would pay (say) 1 cent to get C



Rational Preferences

The Axioms of Rationality

Orderability:

$(A > B) \vee (B > A) \vee (A \sim B)$

Transitivity:

$(A > B) \wedge (B > C) \Rightarrow (A > C)$

Continuity:

$(A > B > C) \Rightarrow \exists p\ [p, A;\ 1\text{-}p, C] \sim B$

Substitutability:

$(A \sim B) \Rightarrow [p, A;\ 1\text{-}p, C] \sim [p, B;\ 1\text{-}p, C]$

Monotonicity:

$(A > B) \Rightarrow$

$(p \geq q) \Leftrightarrow [p, A;\ 1\text{-}p, B] \geq [q, A;\ 1\text{-}q, B]$

Theorem: Rational preferences imply behavior describable as maximization of expected utility

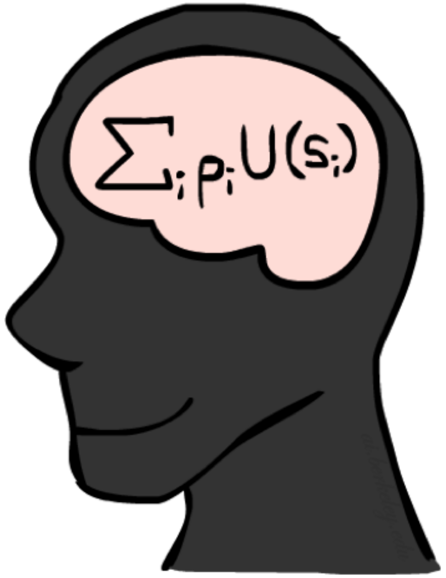
MEU Principle

- Theorem [Ramsey, 1931; von Neumann & Morgenstern, 1944]
  - Given any preferences satisfying these constraints, there exists a real-valued function U such that:

$$U(A) \geq U(B) \Leftrightarrow A \geq B$$

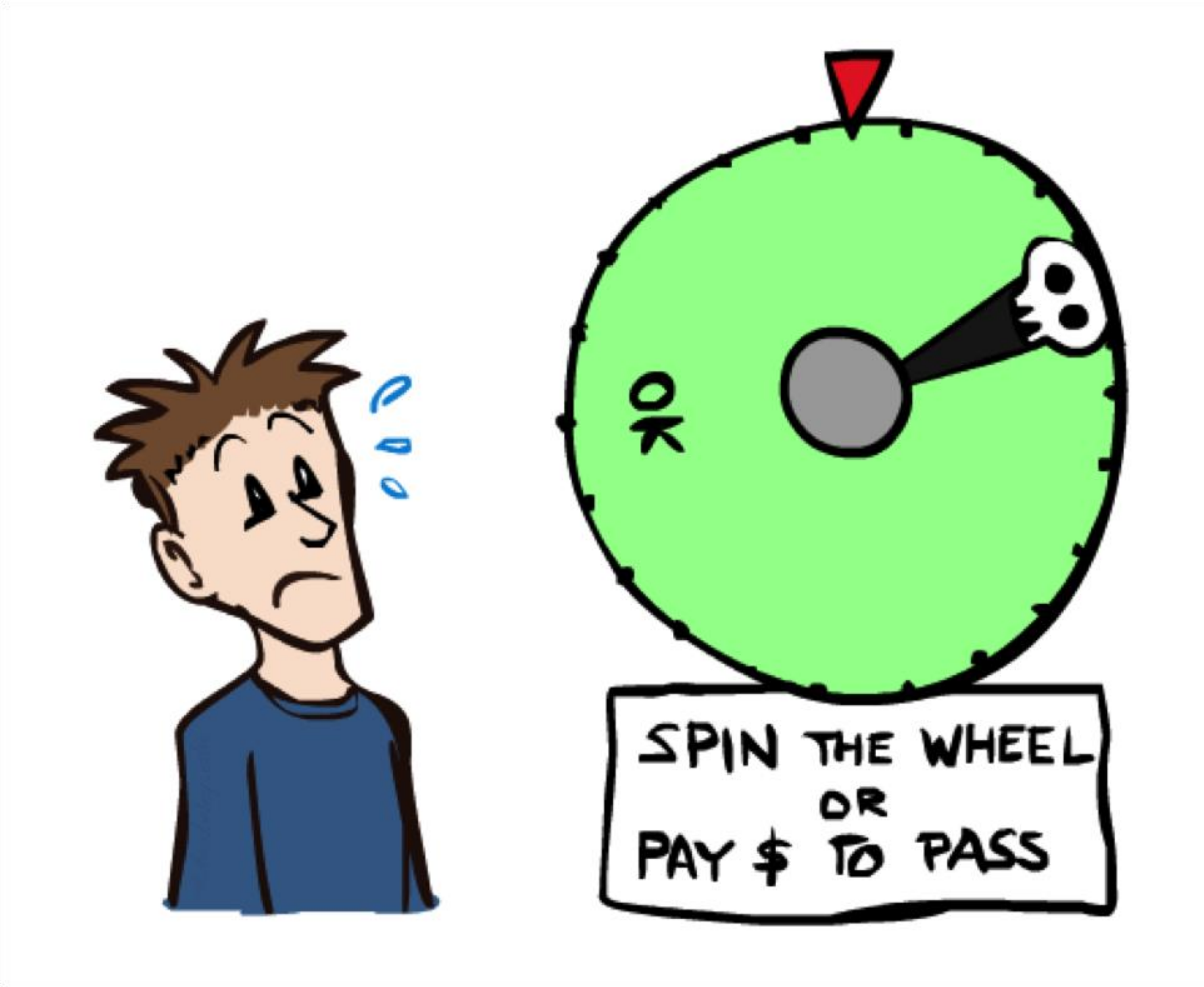
$$U([p_1,S_1; \ldots ; p_n,S_n]) = p_1 U(S_1) + \ldots + p_n U(S_n)$$

  - I.e. values assigned by U preserve preferences of both prizes and lotteries!



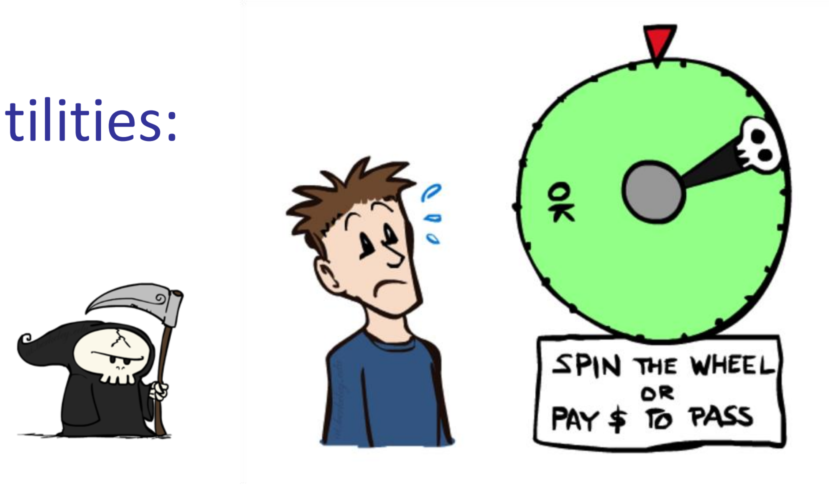
- Maximum expected utility (MEU) principle:
  - Choose the action that maximizes expected utility
  - Note: rationality does ***not*** require representing or manipulating utilities and probabilities
    - E.g., a lookup table for perfect tic-tac-toe

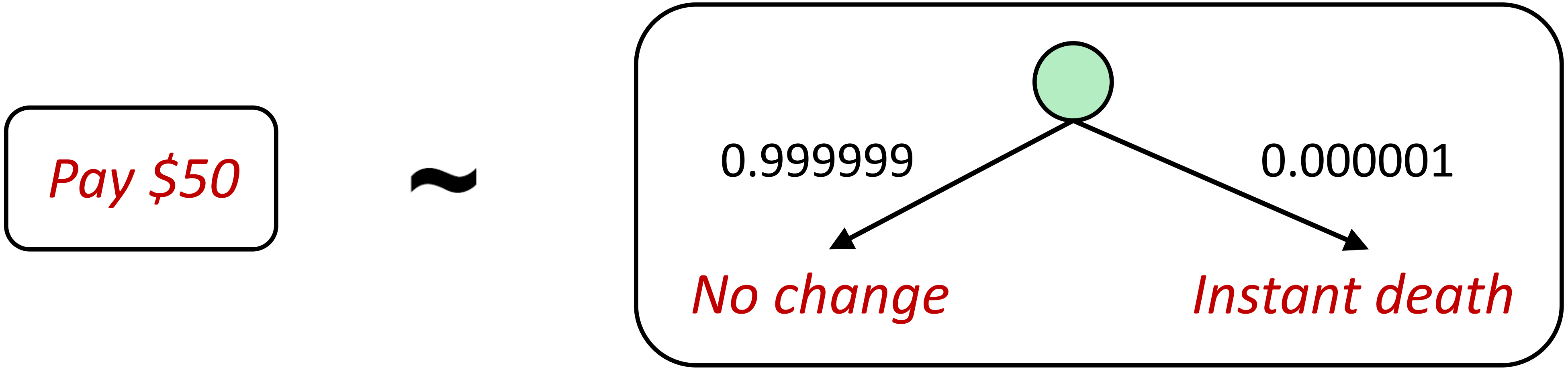
Human Utilities



Human Utilities

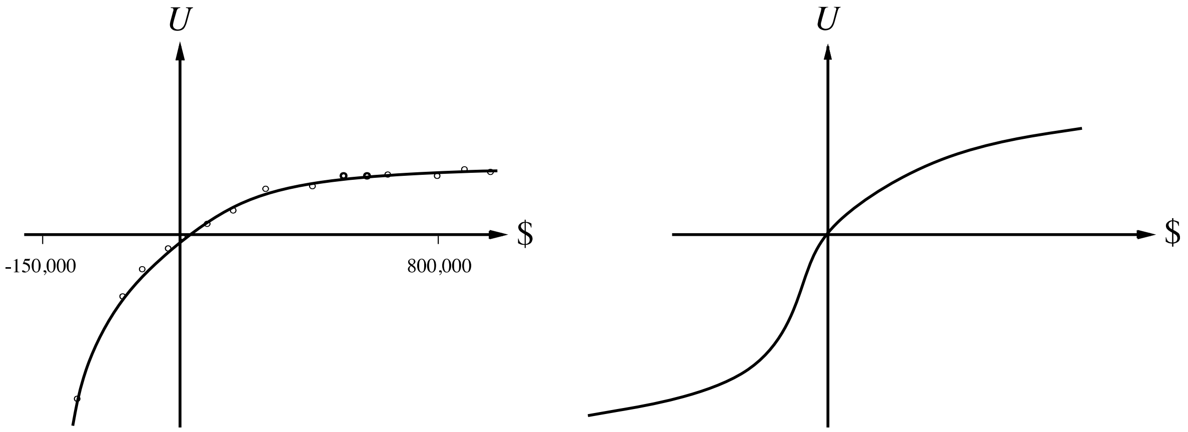
- Utilities map states to real numbers. Which numbers?
- Standard approach to assessment (elicitation) of human utilities:
  - Compare a prize *A* to a ***standard lottery*** L_p between
    - "best possible prize" $u_\top$ with probability *p*
    - "worst possible catastrophe" $u_\bot$ with probability 1-*p*
  - Adjust lottery probability *p* until indifference: $A \sim L_p$
  - Resulting *p* is a utility in [0,1]





Money

- Money ***does not*** behave as a utility function, but we can talk about the utility of having money (or being in debt)
- Given a lottery $L = [p, \$X; (1-p), \$Y]$
  - The ***expected monetary value*** $\text{EMV}(L) = pX + (1-p)Y$
  - The utility is $U(L) = pU(\$X) + (1-p)U(\$Y)$
  - Typically, $U(L) < U(\text{ EMV}(L))$
  - In this sense, people are ***risk-averse***
  - E.g., how much would you pay for a lottery ticket L=[0.5, $10,000; 0.5, $0]?
  - The ***certainty equivalent*** of a lottery $\text{CE}(L)$ is the cash amount such that $\text{CE}(L) \sim L$
  - The ***insurance premium*** is $\text{EMV}(L) - \text{CE}(L)$
  - If people were risk-neutral, this would be zero!



(a) (b)