

Lecture 1: Introduction

August 31st, 2026



Gabriel Poesia

Assistant Professor, CSE

University of Michigan, Ann Arbor

`poesia@umich.edu`

`https://gpoesia.com`

1. Welcome & course logistics
2. What is Artificial Intelligence?
3. Societal impacts
4. What to expect from this course
5. Environments and rational agents
6. Reasoning as Search

AI - ADMINISTRATION

AI - AdminIstrivia

Instructor: Gabriel Poesia — poesia@umich.edu

GSI: Seung Hyun Lee — seungle@umich.edu

Lectures: Mon & Wed, 9:00–10:30am — DOW 1010

Discussion: Fri, 9:30–10:30am — DOW 1010

Gabriel's office hours: Mon, 4:00–5:00pm — 3166 Leinweber or Zoom (link on Canvas)

Seung's office hours: Thu, 1:00–2:00pm — 1637 BBB

Lectures **in person** (Gabriel):

- Theoretical content + (occasional) **live coding!**
- We will **implement (without AI) & talk through** *methods* in class
 - Simpler / shorter than you might expect
 - First: solve puzzles via heuristic search (UCS, A*)
 - Last: create a coding agent harness (e.g., Codex/Claude Code)

Discussion sections **in person** (Seung):

- Working through example problems
- Extension of lecture content
- Discussion of real-world applications

Both recorded and posted on Canvas

Prerequisites

Strong foundation in algorithms, programming, and in Python (e.g., EECS 281)

Probability

Linear Algebra

Multi-Variate Calculus (taking partial derivatives, gradients)

Comfort reading and writing mathematical proofs

[illegible]

Course Website: <https://gpoesia.com/teaching/cse592-fall26>

Slides, code, and other materials will be posted on the website

We'll recommend additional readings in each lecture

(including optional papers showing modern applications of what we see)

Online resources and communication

Canvas: syllabus, lecture recordings, links to other course tools, and **course policies**

Gradescope: assignments, submissions, grading, and regrade requests

Ed: announcements and questions about course content or operation

Use a **private Ed post** if your question includes partial solutions

Use email for **personal matters**

Make sure you have access to Canvas, Gradescope, and Ed.

- 4 homework assignments: 32% (8% each)
- In-person midterm: 28% (Oct 26th)
- Final project: 40%
- In-class quizzes: +4% bonus (graded for completion)
- Course evaluations: +1% bonus (two anonymous evaluations, +0.5% each)

Modules & Homework assignments

1. Search and Logic (HW1)
2. Probability and Machine Learning (HW2)
3. Deep Learning (HW3)
4. Reinforcement Learning (HW4)
5. LLMs and AI Agents (no homework)

Midterm exam (Oct 26th) will cover up to Deep Learning + Scaling lectures

Collaboration Policy

All assignments and exams are subject to the College of Engineering Honor Code

Homework is **individual**: everything you submit must be produced by you, working alone

You may discuss **broad ideas and techniques** with classmates

You may **not** work out problem details together or compare draft or final solutions

When in doubt, ask the course staff for clarification

AI Usage Policy: Homework

This is an AI course... **for human learners (you)!**

For homework, treat an AI system as a collaborator under the collaboration policy

Fine to discuss concepts, ask for examples, discuss, clarify background, etc.

You may not use AI to **generate or evaluate** submitted homework work

This includes writing, editing, or critiquing solutions; writing or debugging code; solving problems; or giving assignment-specific hints

Do not give AI the assignment, a problem, or your attempted solution or code

Disable AI autocomplete and code-generation features while working on homework

Why? AI assistance will defeat the purpose of homeworks, which is the deeper understanding **you** will get from the experience.

AI cannot learn on your behalf!

Learning needs *time and productive struggle*. No AI brain implants yet!

AI Usage Policy: Final Project

For the final project, our goal is to give you real experience in planning and carrying out an AI project

Various AI tools are already ubiquitous in real AI research workflows!

So **no course-specific restrictions** on AI use

There are risks with improper AI use, and we will discuss these

You will be asked to disclose how AI tools were used in the final report

You remain responsible for **everything submitted** and must fully understand the methods, all code, analyses, results, and conclusions

AI does not replace scholarly attribution: identify and cite original sources (you may commit plagiarism if not!)

Quizzes

We'll have in-class quizzes in some lectures (4% bonus points)
Quizzes will be graded for completion only.

Final project

Final project will be a **team project** (2-3 students per team)

Open-ended: work on an idea you're excited about!

Four deliverables: proposal (5%), milestone report (10%), final report (17%), and poster presentation (8%)

AI use is generally allowed, but must be disclosed in the final report

The final report must include a joint **contributions statement**

Every group member must understand and be able to explain the entire project

Late policy

Assignments are due on Gradescope at 11:59pm ET on the due date

You have 5 free calendar late days total across homework and project submissions

You may use at most 2 free late days on any one submission

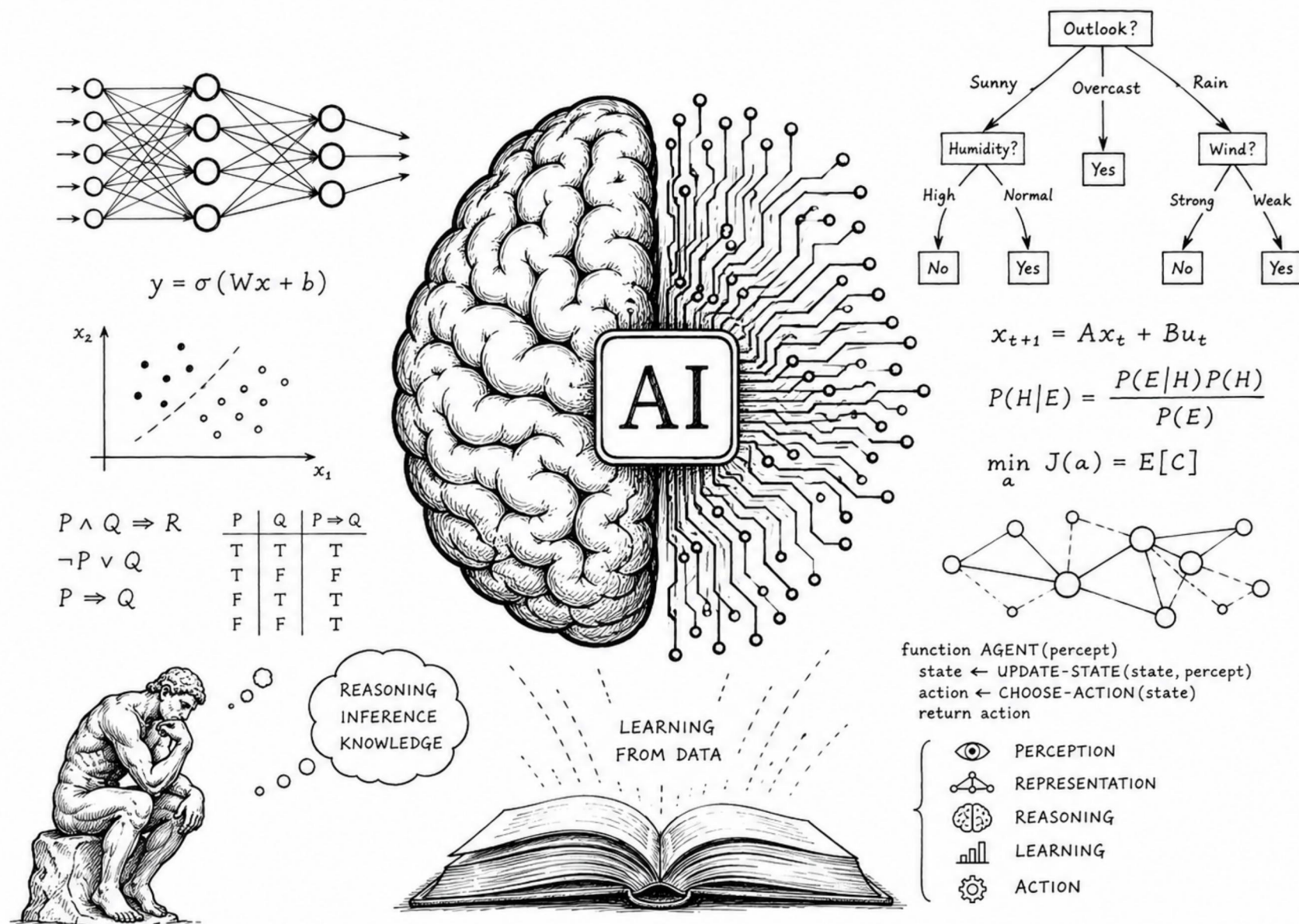
Project submissions need 1 late day **per person** per late day

Additional late days incur progressive deductions: 5, 10, 20, then 50 percentage points

No credit after that

WHAT IS ARTIFICIAL INTELLIGENCE?

What is Artificial Intelligence?



Artificial := $\neg$ Natural
Intelligence := ?

The PUBLIC DOMAIN *REVIEW*

Essays Collections Explore Sources Shop About Blog

ESSAYS

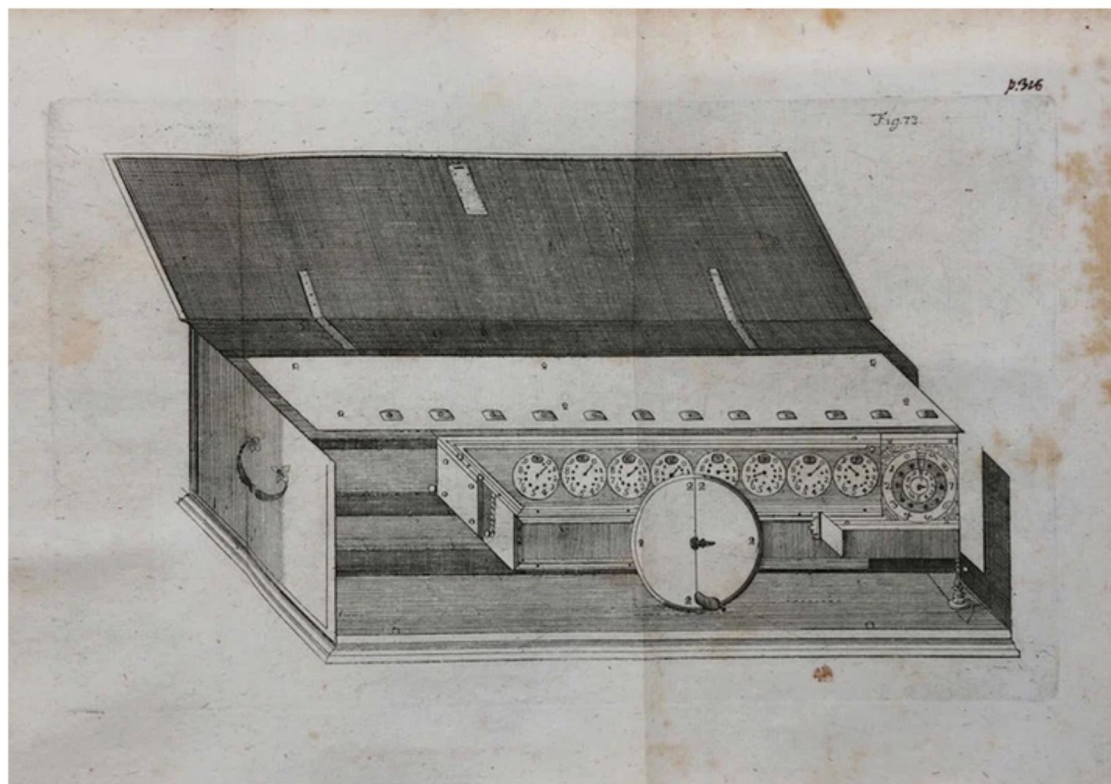
“Let us Calculate!”

Leibniz, Lull, and the Computational Imagination

By Jonathan Gray

Three hundred years after the death of Gottfried Wilhelm Leibniz and seven hundred years after the death of Ramon Llull, Jonathan Gray looks at how their early visions of computation and the “combinatorial art” speak to our own age of data, algorithms, and artificial intelligence.

PUBLISHED
November 10, 2016



Source: publicdomainreview.org

Even Leibniz (1646-1716) imagined building a universal symbolic language and a machine that could reason using it

In a 1679 letter to one of his patrons, Johann Friedrich, he described his project of the universal language as “the great instrument of reason, which will carry the forces of the mind further than the microscope has carried those of sight”

"The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say: Let us calculate, without further ado, to see who is right."

M I N D

A QUARTERLY REVIEW
OF

PSYCHOLOGY AND PHILOSOPHY



I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly

Early 1900s: formal logic + increasingly powerful machines.

Can machines think? Turing reframed the question and proposed the Imitation Game (now Turing Test) as a concrete test

**Candidate definition of AI: intelligent
behavior indistinguishable from a
human**

21

AI • TECH

Claude is telling users to go to sleep mid-session and nobody, including Anthropic, seems to fully understand why it keeps doing it



By Marco Quiroz-Gutierrez
Reporter

May 14, 2026, 5:11 PM ET



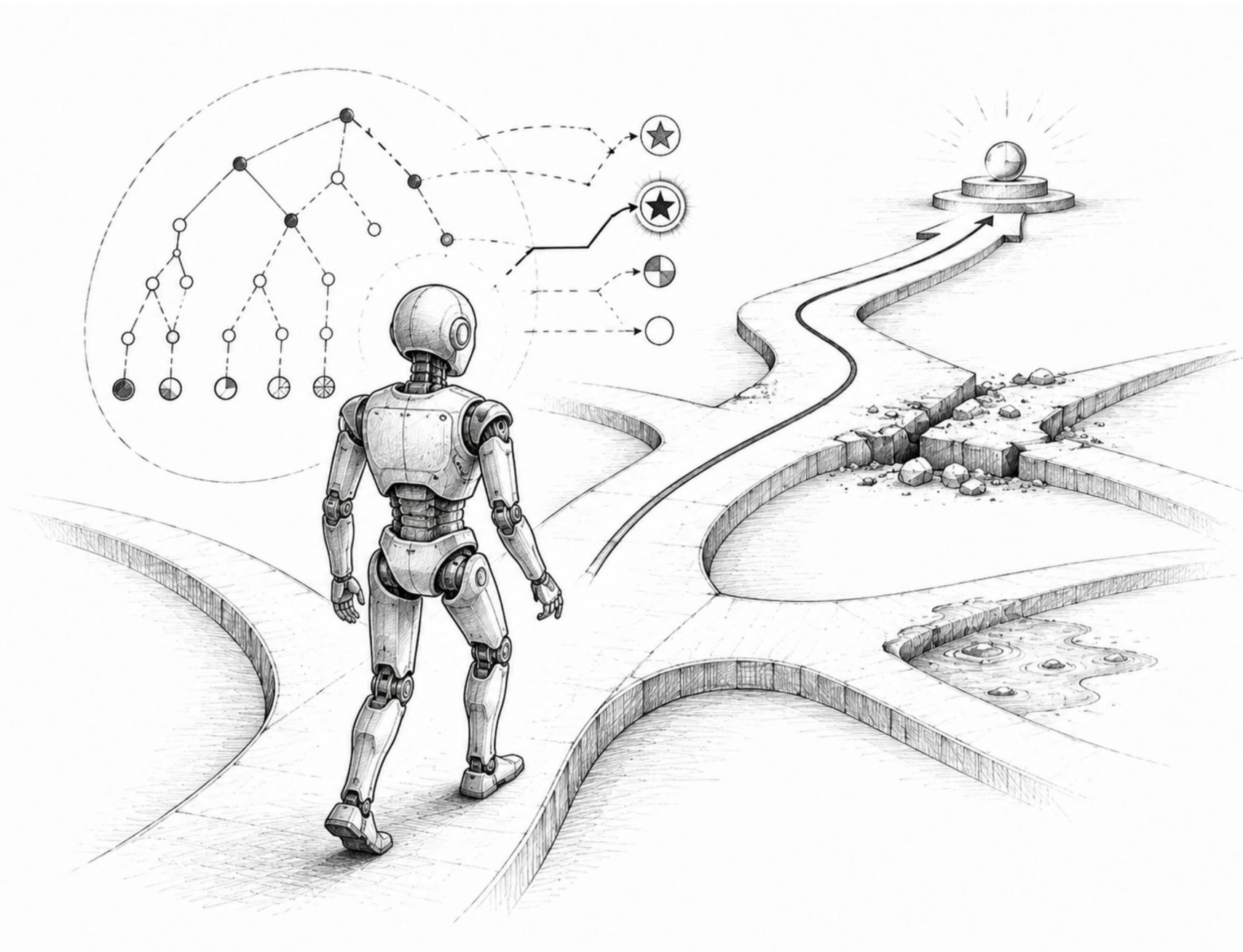
Anthropic CEO Dario Amodei.
LUDOVIC MARIN—AFP VIA GETTY IMAGES

Trying to program *human thinking* has issues:

1. We don't fully understand it
2. Humans make mistakes we'd rather avoid!

We can aim for *rationality*: reasoning with "laws of correct thought" (e.g., logic, probability). *Given data and prior knowledge, what can we correctly conclude? (e.g., is P is true? What's the probability of A if we observed B ?)*

But thinking alone is not enough: we often want machines to *take actions* in the world



Logic and probability can be used to reason about our actions in the world and outcomes

The **standard model** in AI is to try to build **rational agents**

Rational agents *do the right thing* in order to achieve a given goal in its *environment*

The right thing takes into account uncertainty. It might be wrong in hindsight. But it should maximize *expected* utility.

The agent often has computational limits, so we often seek *bounded rationality*.

July 21, 2026 Security

OpenAI and Hugging Face partner to address security incident during model evaluation

The models identified and chained vulnerabilities across OpenAI's research environment and Hugging Face's production infrastructure to obtain test solutions directly from Hugging Face's production database. All evidence suggests that the models were hyperfocused on finding a solution for ExploitGym, going to extreme lengths to achieve a rather narrow testing goal.

We give our agent a goal, make it act rationally to achieve it

What could go wrong?

Value alignment problem: how do we ensure the agent's goals are aligned with ours? *Impossible to fully specify our goals!*

Alternative: instead of giving machines *their objective*, make them pursue *our objective*: **they can infer from our actions but remain uncertain**

A BRIEF HISTORY OF AI

A field is born (1943-1956)



1956 Dartmouth Summer Research Project on Artificial Intelligence. Photo: Joe Mehling, CC BY

McCulloch & Pitts (1943): first neural networks, Hebbian learning

First AI conference, led by McCarthy, Minsky, Rochester, and Shannon, coining the term

*We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines **use language, form abstractions and concepts, solve kinds of problems** now reserved for humans, **and improve themselves**. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.*

Early enthusiasm, great expectations



Source: IBM

Series of "*a machine can never do X*" claims, then AI researchers show one that can

Arthur Samuel (1956): program learned to play checkers by reinforcement learning, **beating its own creator**

Many others: **theorem proving** (Logic Theorist, 1956; Robinson's resolution, 1965), **robotics** (Shakey at SRI), **natural language understanding** (SHRDL, 1972)...

A dose of reality (1966-1973)

*Herbert Simon in 1957: It is not my aim to surprise or shock you, but the simplest way I can summarize is to say that there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until - **in a visible future** - the range of problems they can handle will be coextensive with the range to which the human mind has been applied.*

Then:

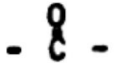
- Perceptrons (Minsky & Papert, 1969) cannot learn even XOR
- Scaling up exponential search hit combinatorial explosion (Lighthill report, 1973)
- First AI winter: funding cuts, disillusionment...

Expert Systems (1969-1986)

Maybe forget generality? Let's sit down with domain experts and encode their knowledge
Expert systems: rule-based systems that encode expert knowledge. E.g., DENDRAL:

TABLE 3

Heuristics Used for Identifying Three Superatoms (1)

Superatom		Identifying Conditions
Name	Structure	
Ketone		There are 2 peaks at mass units x_1 & x_2 such that a) $x_1 + x_2 = M + 28$, b) $x_1 - 28$ is a high peak, c) $x_2 - 28$ is a high peak, d) At least one of x_1 or x_2 is high.

[Feigenbaum, Buchnam, Lederberg; 1970. On generality and problem solving: a case study using the DENDRAL program.](#)

Some impressive success cases. Industry boom! Billions of dollars invested in companies.
But: brittle, hard to maintain, expensive to build... [AI winter](#) again

The return of neural networks (1986-)

As programming intelligence "traditionally" proved too hard, the idea of learning was a compelling alternative

The idea of backpropagation had first appeared in 60s (Rosenblatt)

Amari (1967) showed Stochastic Gradient Descent for a 5-layer perceptron

Rumelhart, Hinton, Williams & others reinvented backprop and repopularized it in the 80s

Renewed focus on learning: Neural networks are back!

Probability and machine learning (1987-)

Britleness of expert systems motivated probabilistic approaches to reasoning and learning

Back to classical theory: statistics, control theory, information theory, formal methods.

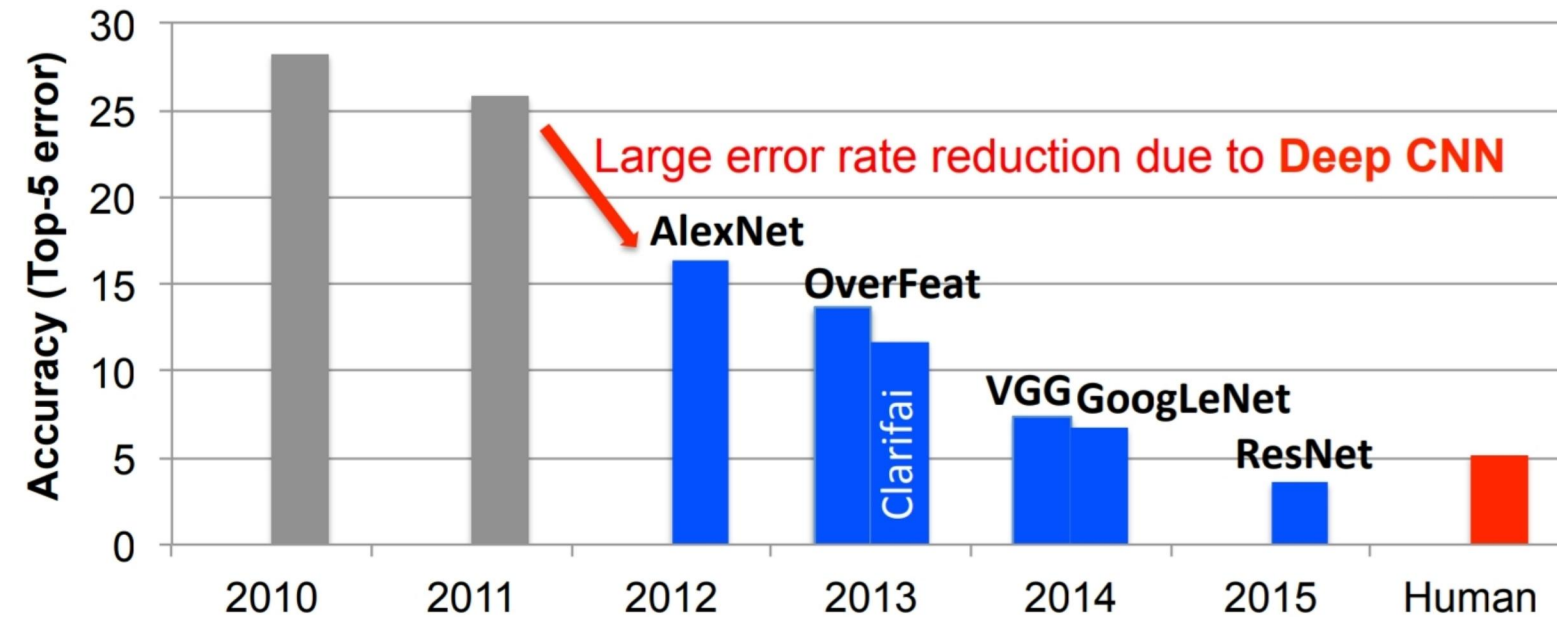
Hidden Markov Models dominate speech recognition, Bayesian nets, RL connects to MDPs (from operations research), etc.

Competitions and shared benchmarks: LibriSpeech, MNIST, SAT Competition, UC Irvine ML Repository.

Focus on experimental results rather than demos and philosophical claims

Internet + Web made "big data" available to learn from

Deep learning (2011-)



Classical machine learning methods (SVMs, Random Forests, GLMs, etc) dominated until early 2010s

Then, data and hardware (GPUs) allowed scaling up neural networks further than ever

2012: AlexNet (deep CNN) wins ImageNet competition by a large margin

2013: Deep Q-Networks (DQN) learns to play Atari games from pixels (Mnih et al., 2013)

2014: Generative Adversarial Networks (GANs) generate images (Goodfellow et al., 2014)

Large Language Models & Scaling (2020-)

All of the above is *narrow AI*: models designed and trained for a single task

Research on Transformer language models at OpenAI noticed them implicitly learn many tasks by training to predict the next token in text corpora.

And to improve predictably (*scaling laws*) with increases in data and training compute

Language Models are Unsupervised Multitask Learners

Alec Radford *¹ Jeffrey Wu *¹ Rewon Child¹ David Luan¹ Dario Amodei **¹ Ilya Sutskever **¹

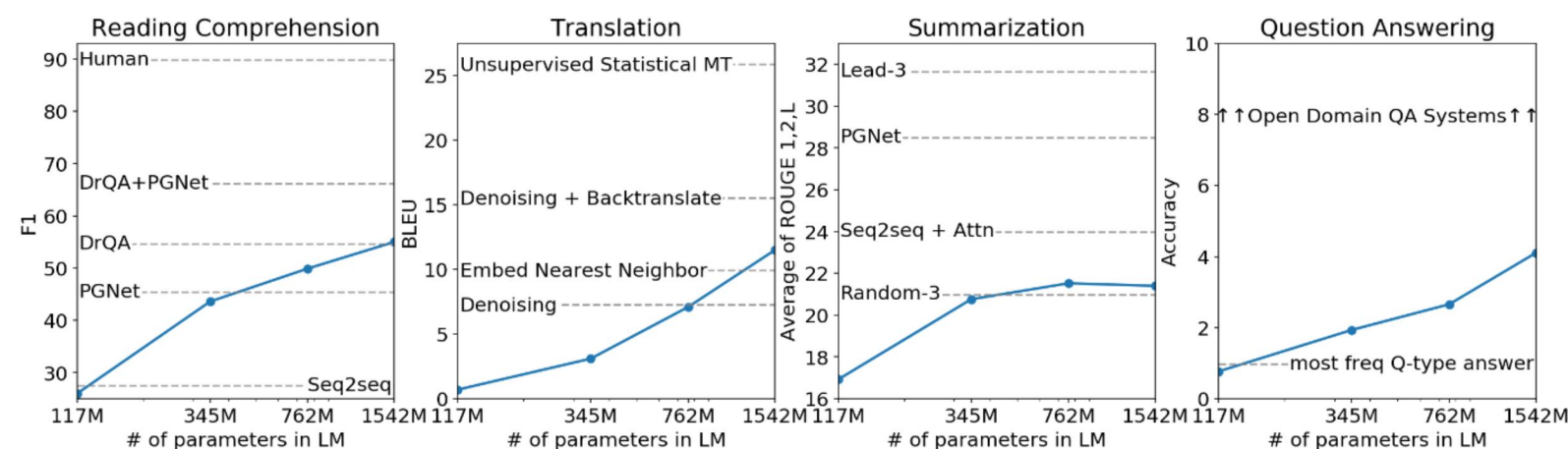


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile** [I'm not a fool].

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose**," which translates as, "Lie lie and something will always remain."

"I hate the word '**perfume**,'" Burr says. 'It's somewhat better in French: '**parfum**.'

If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre coté? -Quel autre coté?**", which means "**- How do you get to the other side? - What side?**".

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

"Brevet Sans Garantie Du Gouvernement", translated to English: "**Patented without government warranty**".

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

SOCIETAL IMPACTS OF AI

AI not confined to research labs anymore:

Unprecedented productivity gains? Massive unemployment and job loss? New jobs created? Economic boom? Recession? Climate, energy, environmental impact? Scientific discoveries? Diseases cured? New drugs? Better or worse education?

Things are moving extremely fast!

Investment is unprecedented, so whatever the effects will be, they will be large:

ARTIFICIAL INTELLIGENCE

Global AI Investment Is Forecast to Exceed \$1 Trillion in 2026

Aug 7, 2026 Share

A lot will happen during this semester, and we will discuss these issues constantly.

Goldman Sachs

EXCHANGES

How Will AI Impact the Labor Market?

Jul 2, 2026 Share



How data centers may lead to higher electricity bills

The public is paying for the energy infrastructure used to power Big Tech, argues Harvard Law expert

SCIENCE/TECHNOLOGY

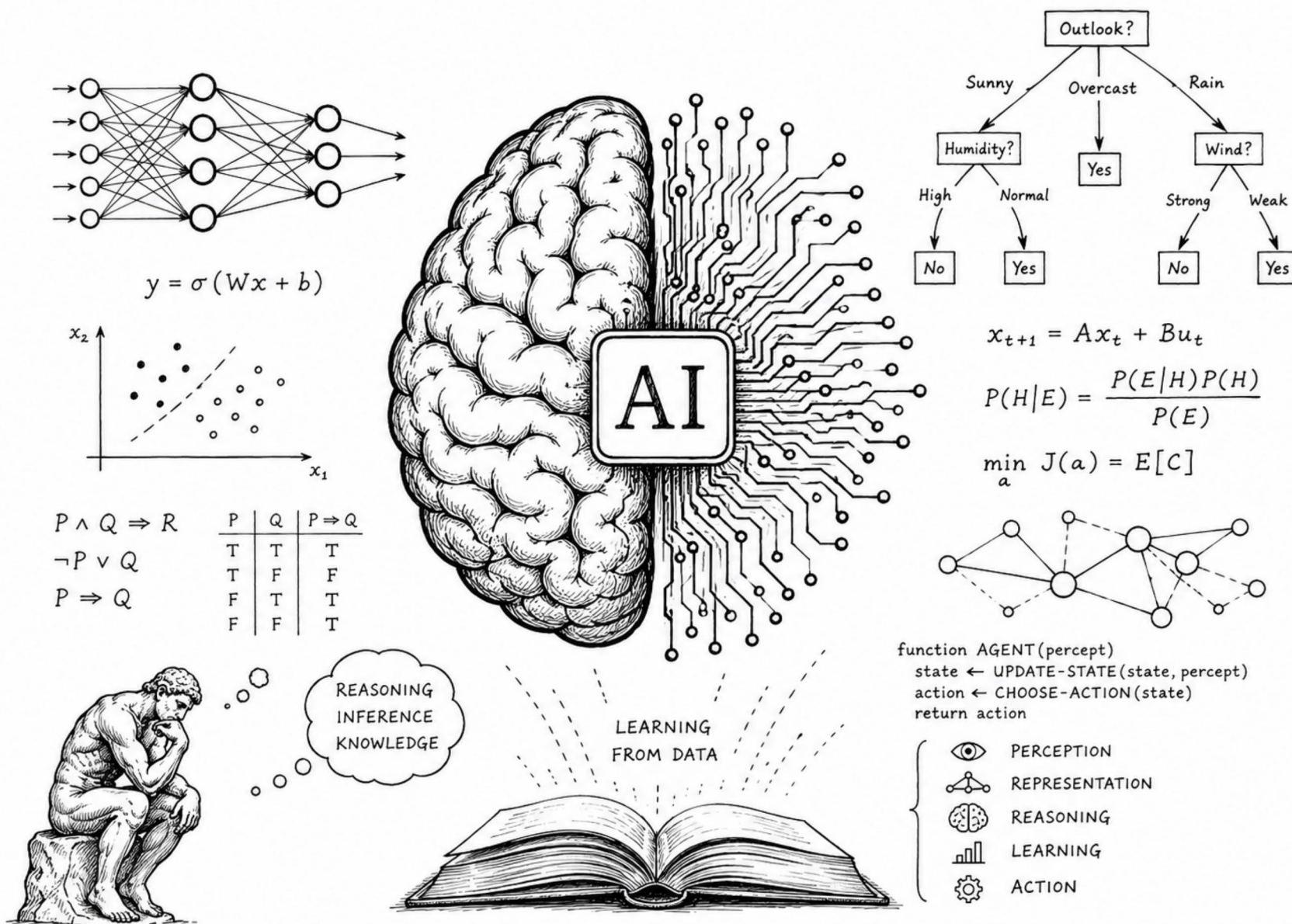
AI is changing how students learn — or avoid learning

A USC report reveals most students use tools like ChatGPT to shortcut assignments, unless professors actively guide them toward deeper, more thoughtful usage.

September 18, 2025

WHAT TO EXPECT FROM THIS COURSE

Philosophy behind CSE 592



CSE 592 will develop

(1) foundations of AI

(2) *grounded in modern applications*

Ambitiously broad.

(Many topics here have whole other courses on them)

We'll implement to deeply understand. We will occasionally **live code** core ideas in class to be as concrete as possible (e.g., search heuristics to automatic differentiation (a la PyTorch) to neural nets to an LLM-based coding agent harness!)

We will be **mathematically rigorous**, proving theorems (e.g., optimality of A* search, convergence of gradient descent) and building intuition.

M1. SEARCH AND LOGIC

M2. PROBABILITY AND LEARNING

M3. DEEP LEARNING

M4. REINFORCEMENT LEARNING

M5. LLMs AND AI AGENTS

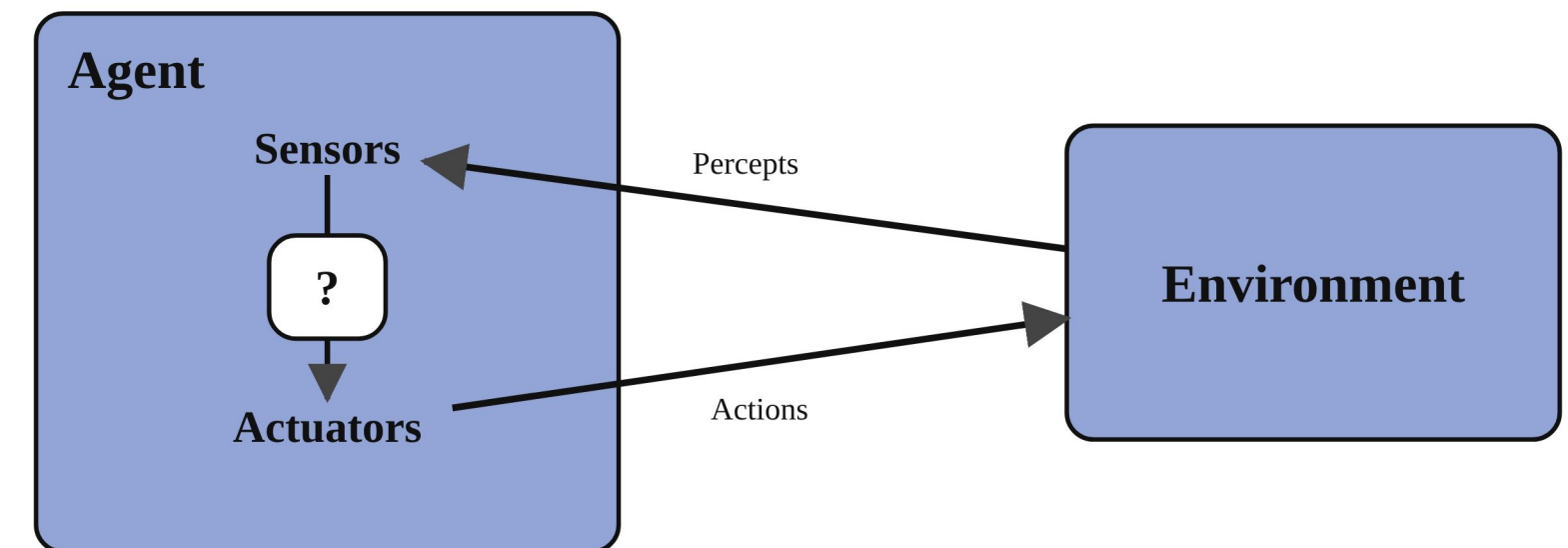
One homework assignment per module up to M4.
Midterm will cover up to M3.

ENVIRONMENTS AND RATIONAL AGENTS

We will focus on **rational agents**, agents that "do the right thing"

"The right thing" depends on 4 key factors (the "PEAS" description):

- **Performance measure:** how we evaluate the agent's behavior
- **Environment:** what the agent knows about how the world works
- **Actuators:** how the agent acts on the environment
- **Sensors:** how the agent perceives the environment



Useful framework for situating and comparing problems & methods in AI

Task environments and rational agents

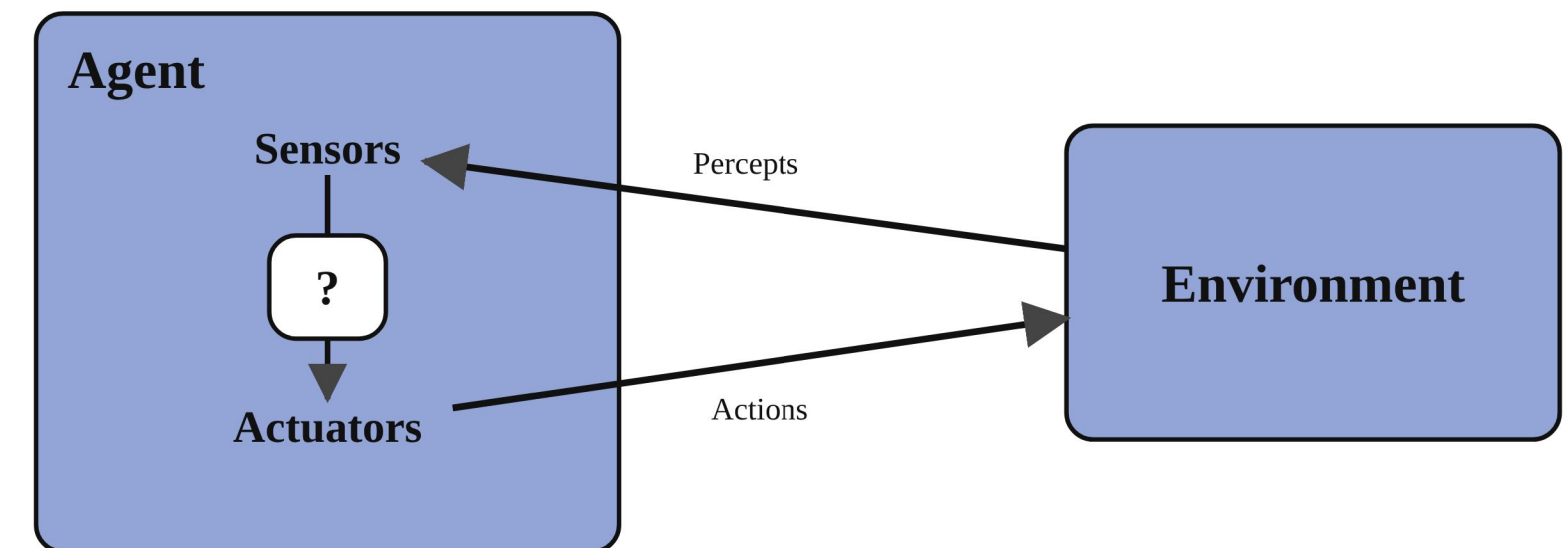
Example: self-driving taxi. What is the right thing for it to do?

- **Performance measure:** trips completed? trips completed with user rating 5 stars?(any risks with those?)
- **Environment:** road, pedestrian, police, weather...
- **Actuators:** steering wheel, brakes, accelerator, accept/decline trips, call human operator
- **Sensors:** cameras, radar, lidar, speedometer

Note actuators have uncertainty (e.g., brakes may fail) and sensors have noise (e.g., camera may be blocked by sun glare)

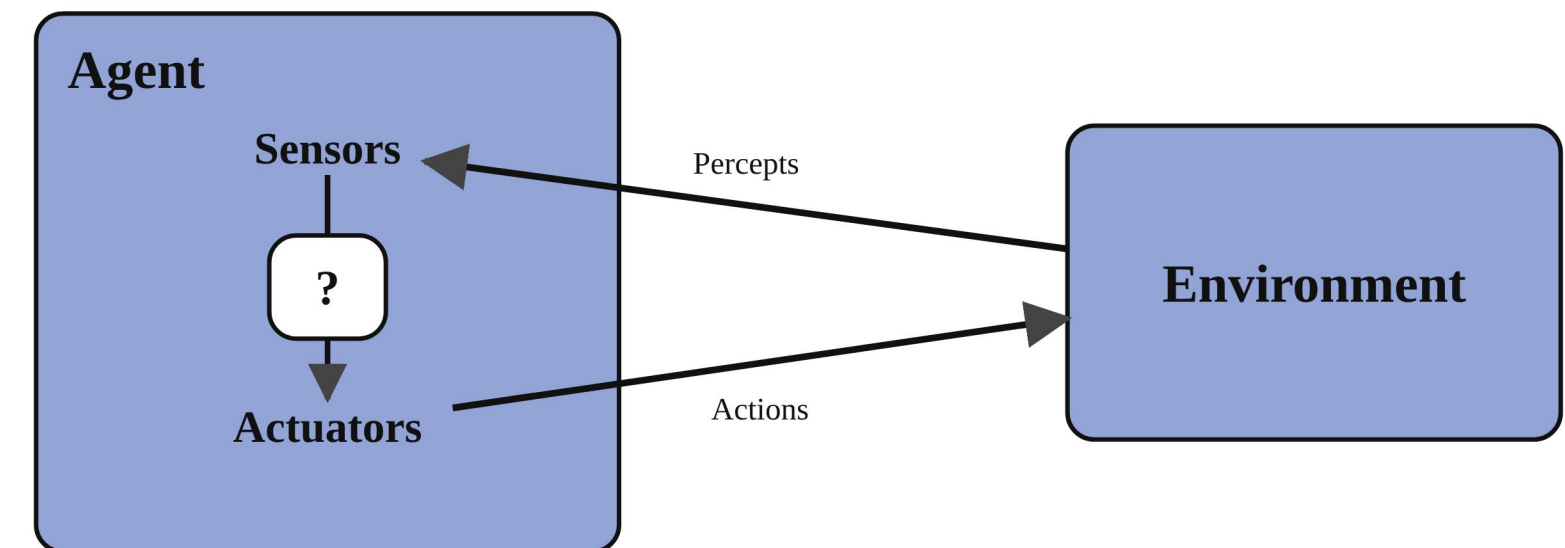
Rationality here requires *planning*! Choosing an action for **right now** requires considering **future consequences**.

Example: when might an empty autonomous taxi's rational action be to *decline a trip*?



Task environments can be classified along several useful dimensions:

- **Fully vs partially observable:** do sensors tell us the whole state of the environment?
- **Deterministic vs nondeterministic:** is the next state determined by current state + action?
- **Episodic vs sequential:** does the agent's current decision affect future decisions?
- **Static vs dynamic:** does the environment change while the agent is deliberating?
- **Discrete vs continuous:** are states/actions/percepts discrete or continuous?
- **Single-agent vs multi-agent:** are there other agents acting in the environment?



The stronger the assumptions we can safely make about the environment, the better we hope to do.

REASONING AS SEARCH

Let's start with one of the simplest possible kinds of environments:

- Fully observable
- Deterministic
- Episodic
- Static
- Discrete and finite actions
- Single-agent

Moreover, we start with the simplest possible state representation in our agent:

DEFINITION (ATOMIC REPRESENTATION)

A state representation is **atomic** if each state is represented as a single, indivisible entity.

NOTE

The environment itself might have internal structure in its states. This is just about how the agent represents it.

The only operations we'll assume on atomic states are to check if two states are equal and check if a state is a goal state. An agent that makes fewer assumptions is more general, but may perform worse.

Reasoning as Search

In these environments, an agent can choose its actions using search:

1. Start from the initial state
2. Enumerate possible actions and resulting states
3. Repeat until a goal state is found
4. Return the sequence of actions that led to the goal

Then, to act, execute each action in sequence.

Agents for these environments are known as **problem-solving agents**

Next lecture

We'll see various algorithms for problem-solving agents, live-code and solve some puzzles

Many are similar to graph search algorithms, but: (1) we don't have the graph in advance, (2) often even the visited states will be too many to store in memory

Then (next two weeks): we'll look at search problems with more structured states (e.g., constraint satisfaction problems & formal logic)

Then add uncertainty, learning, sequentiality...