

# **671 - PRINCIPLES OF AI**

Lara J. Martin (she/they)

TA: Aydin Ayanzadeh (he)

8/31/2023 - Introduction

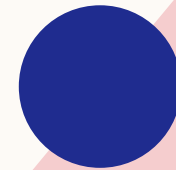
# WHAT TO EXPECT TODAY

Introduction to Lara & Aydin

Course-Specific Junk

Getting to know you

What is AI?



# WHO IS LARA?

[laramar@umbc.edu](mailto:laramar@umbc.edu), [laramartin.net](http://laramartin.net)

- Applied NLP, Neurosymbolic methods
- CS & Linguistics BS @ Rutgers → MLT @ CMU  
→ HCC PhD @ GT → CIFellows @ UPenn →  
Assistant Prof. @ UMBC



[https://upload.wikimedia.org/wikipedia/commons/a/a4/Map\\_of\\_USA\\_with\\_state\\_and\\_territory\\_names\\_2.png](https://upload.wikimedia.org/wikipedia/commons/a/a4/Map_of_USA_with_state_and_territory_names_2.png)

# WHO IS AYDIN?

[aydina1@umbc.edu](mailto:aydina1@umbc.edu)

# HOW TO

- Blackboard → submissions & grades
- Course Website (Linked to from Blackboard), but also:
  - <https://redirect.cs.umbc.edu/courses/graduate/671/fall23/>
  - <https://laramartin.net/Principles-of-AI/>
- Slack (also on Blackboard)
  - JOIN NOW
  - [redacted]

# TEXTBOOK




Artificial Intelligence: A Modern Approach  
By Stuart Russell & Peter Norvig  
4<sup>th</sup> Edition (lavender)


Digital copy through Blackboard

Might be able to get some of the  
information from 3<sup>rd</sup> edition, but the  
chapter numbers won't match


# TEXTBOOK


## Course Content

☐  Course Website/Syllabus


☐  My Textbooks & Course Resources

☐  Class Slack

☐  Assignments

☐  Request Help (RT)

## CMSC 671 Principles of Artificial Intelligence (01.4879) FA2023

| Item Info   | Notes   |
|---|---|
|  | <b>Artificial Intelligence: A Modern Approach, 4th Edition</b><br>ISBN: 9780134671932 By: Stuart Russell; Peter Norvig<br><input checked="" type="checkbox"/> Opted In Course Materials Initiative (CMI)<br>Required<br>⌚ The last day to opt out is 09/13/2023.<br><div>Want to opt-out?</div> <div>Read Now</div> |

# OFFICE HOURS

- Lara: after class on Tuesdays & Fridays at 11am-12pm in ITE 216
  - <https://calendly.com/laramar/schedule>
- Aydin: Wednesdays at 2-4:30pm in ITE 334



# LEARNING OBJECTIVES

- Predict the behavior of different search algorithms (**HW1**)
- Construct and query a knowledge base using first-order logic (**HW2**)
- Define decision making problems, and implement agents that can solve them (**HW3**)
- Apply probabilistic reasoning to problems with uncertainty (**HW4**)
- Compare and contrast AI methods to determine an appropriate method for a given problem (**Midterm**)
- Reflect on the societal impacts of the AI methods and applications discussed in class (**Class Knowledge Checks**)
- Develop and run AI experiments to work towards solving modern problems (**Final Project**)

|                        |     |
|------------------------|-----|
| Class Knowledge Checks | 10% |
| Paper Presentation     | 5%  |
| Homework               | 40% |
| Midterm                | 20% |
| Final Project          | 25% |

# POLICIES

Late days: Each student is allowed 5 late days (no excuses needed, no points off) – for HW only

Collaboration: pairs allowed on HW, groups of 3-5 on final project, can discuss w/ others for paper presentation & class knowledge checks

# ACADEMIC INTEGRITY

- If you feel the need to cheat, come to me or Aydin first
- If you cheat or plagiarize, you...
  - aren't learning anything
  - wasting money paying for tuition
  - will get an F on the assignment (at the very least)
- More details on course website



# **WHAT ABOUT CHATGPT?**

# WHAT IS CHATGPT?

---

## GPT-4 Technical Report

---

OpenAI\*

### Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

### 1 Introduction

This technical report presents GPT-4, a large multimodal model capable of processing image and

## 2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.<sup>2</sup> We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

## 3 Predictable Scaling

A large focus of the GPT-4 project was building a deep learning stack that scales predictably. The primary reason is that for very large training runs like GPT-4, it is not feasible to do extensive model-specific tuning. To address this, we developed infrastructure and optimization methods that have very predictable behavior across multiple scales. These improvements allowed us to reliably predict some aspects of the performance of GPT-4 from smaller models trained using  $1,000\times$  –  $10,000\times$  less compute.

### 3.1 Loss Prediction

The final loss of properly-trained large language models is thought to be well approximated by power laws in the amount of compute used to train the model [41, 42, 2, 14, 15].

To verify the scalability of our optimization infrastructure, we predicted GPT-4's final loss on our internal codebase (not part of the training set) by fitting a scaling law with an irreducible loss term (as in Henighan et al. [15]):  $L(C) = aC^b + c$ , from models trained using the same methodology but using at most 10,000x less compute than GPT-4. This prediction was made shortly after the run

# KNOWN ISSUES

- Bad reproducibility
- Copyright issues
- Can't explain what it's doing
- Can't remember things long term
- Confident bullshitter

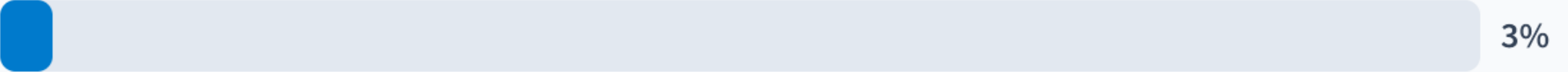
What program/department are you in?





## Why are you taking this course?

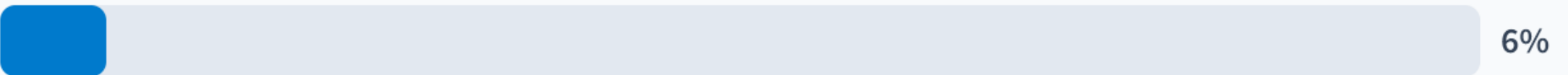
Requirement



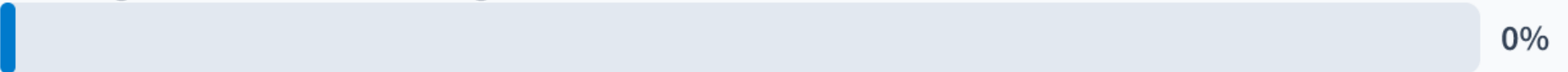
Interested in AI



Interested in Dr. Martin's work



Nothing else seemed interesting



## Have you taken an AI course in undergrad?

Yes

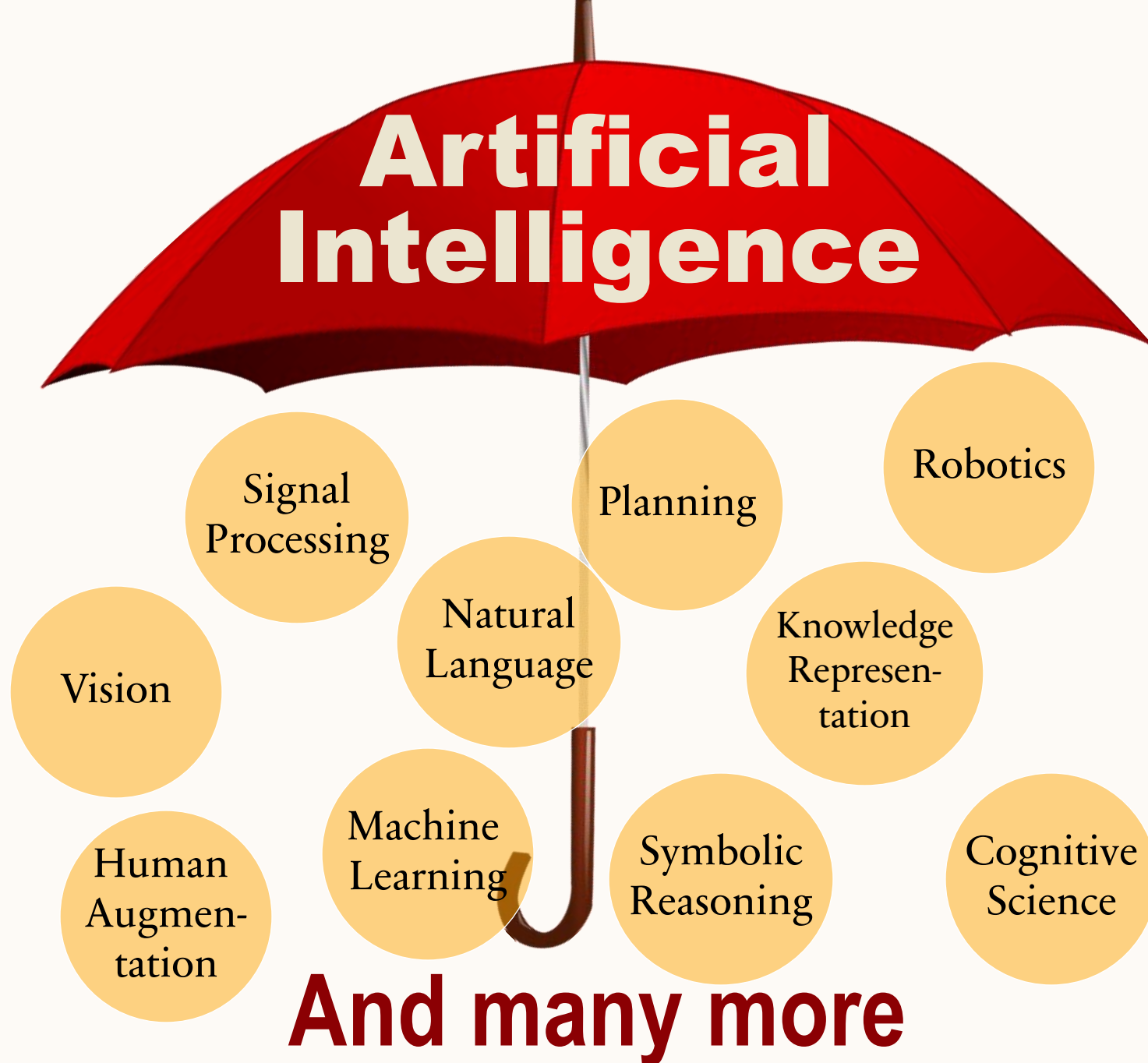


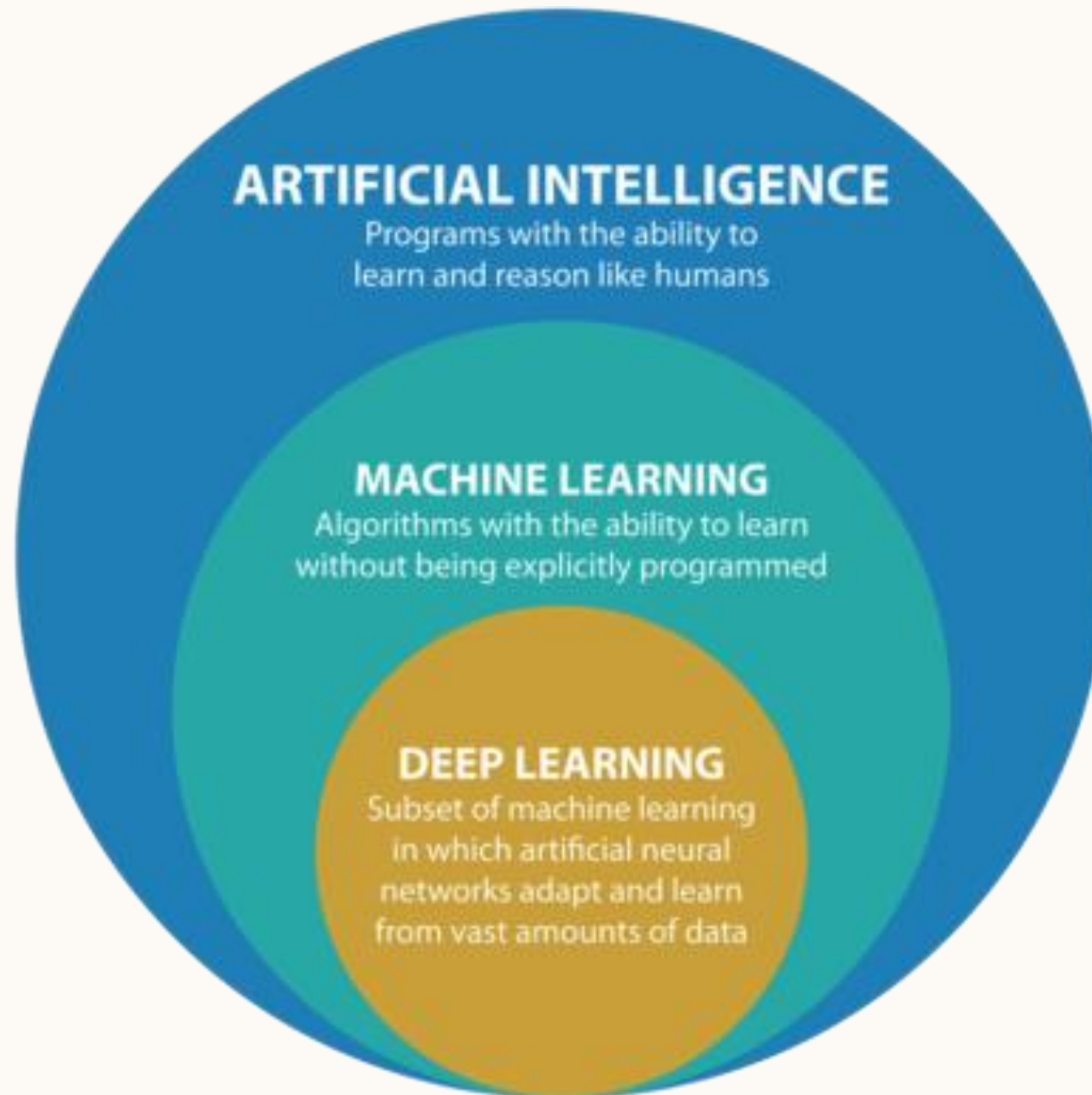
No





# WHAT IS AI?





# GOALS OF AI

- **Represent** and **store** knowledge
- **Retrieve** and **reason** about knowledge
- **Behave intelligently** in complex environments
- **Learn** from environment and interactions
- **Develop** interesting and useful applications
- **Interact** with people, agents, and environment

# WHY AI?

## Engineering

- To get machines to do a wider variety of useful things
  - Understand spoken natural language
  - Recognize individual people in visual scenes
  - Find the best travel plan for your vacation

## Cognitive Science

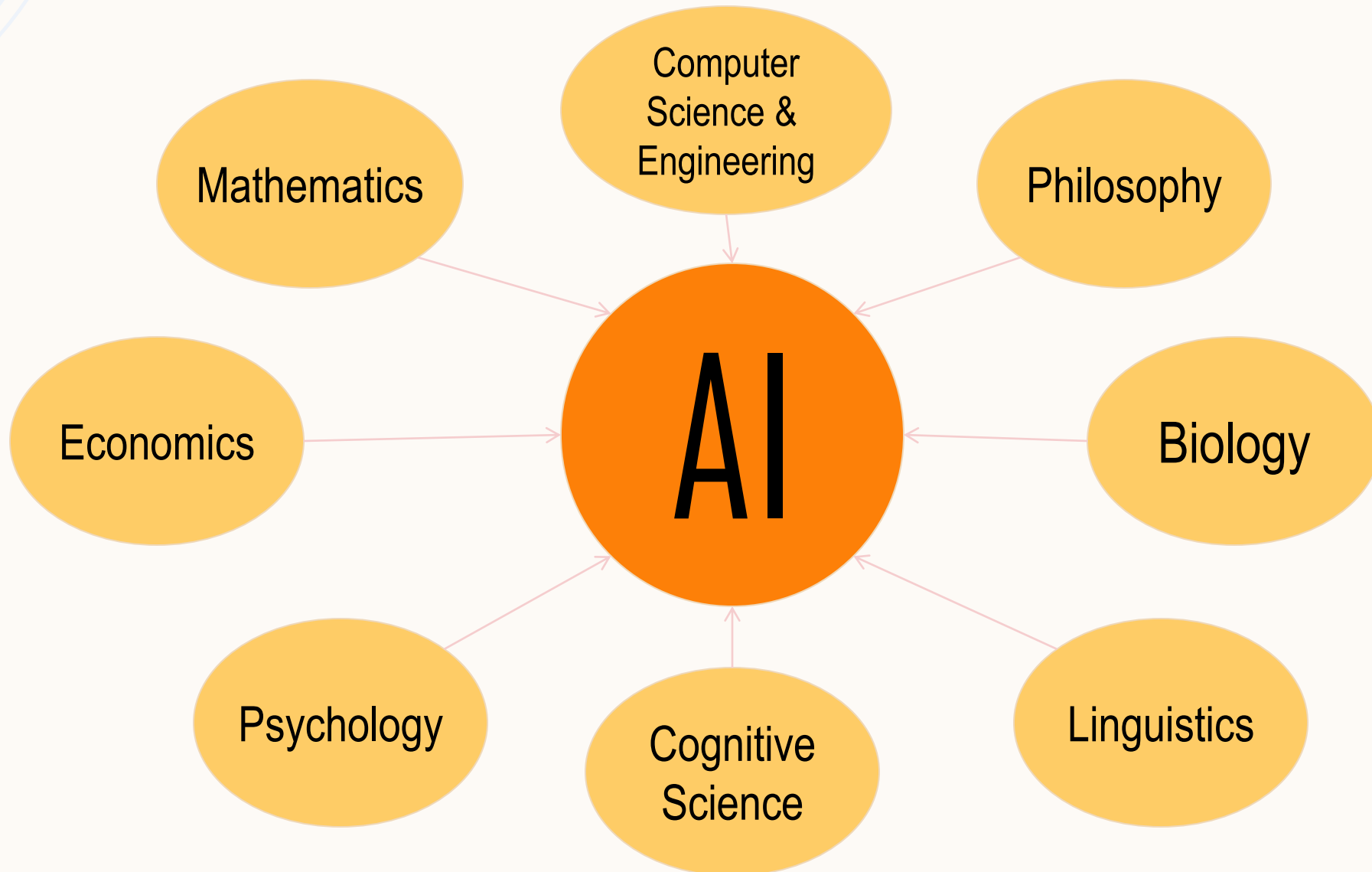
- Help understand how natural minds work
  - Visual perception, memory, learning, language, etc.

## Philosophy

- As a way to explore interesting (and important) philosophical questions



# FOUNDATIONS OF AI



# WHAT ARE SOME EXAMPLES OF AI IN YOUR DAILY LIFE?

Chatbots/ChatGPT

→ Siri

Auto complete

Recommendation → Netflix

spam

Traffic prediction

Weather prediction

Face recognition

Stock market

Navigation

Speech-to-text & text-to-speech

Self-driving cars

Text-to-image → TikTok

filters

Deep fakes

Game AI

Protein folding

Gesture recognition

Handwriting recognition  
& OCR (optical character  
recognition)

Roomba

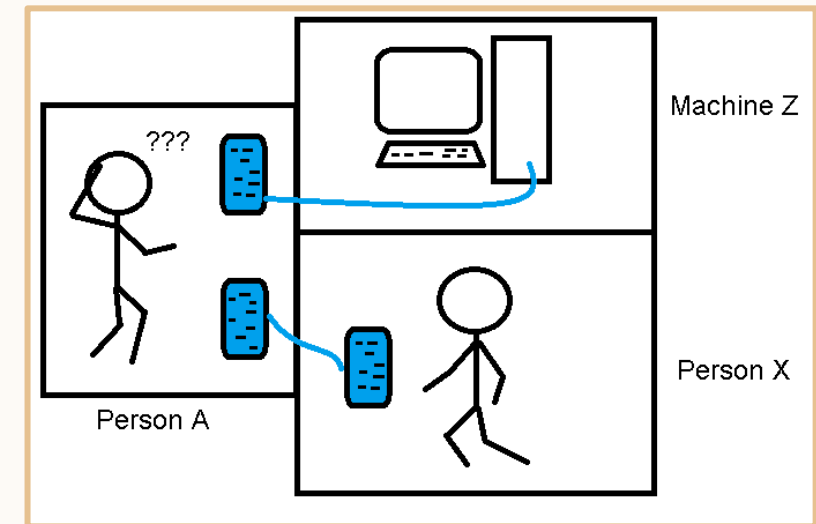
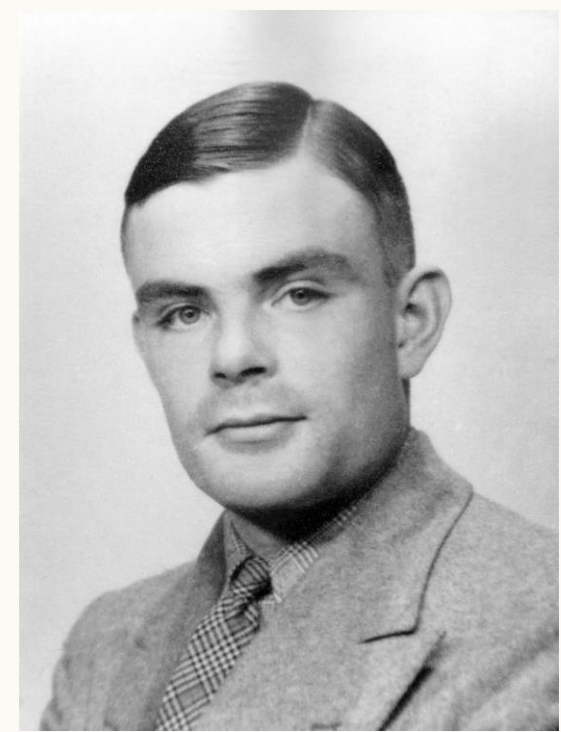
Cockpit AI



# **HOW DO WE KNOW IF IT'S WORKING?**

# TURING TEST

- Three rooms:
- 1 person, 1 computer, and 1 interrogator
  - The interrogator can communicate with the other two
  - The interrogator tries to decide which is the person
  - Both try to convince the interrogator they are the person
- If the machine succeeds, the machine can think
- ...Right?



Slide by Dr. Cynthia Matuzek

Image: [filipinofreethinkers.org/2012/06/23/turings-tremendous-talent-and-trenchant-test/turing-test](https://www.filipinofreethinkers.org/2012/06/23/turings-tremendous-talent-and-trenchant-test/turing-test)

<https://cdn.britannica.com/81/191581-050-8C0A8CD3/Alan-Turing.jpg>

ARTIFICIAL INTELLIGENCE

# Google Engineer Claims AI Chatbot Is Sentient: Why That Matters

Is it possible for an artificial intelligence to be sentient?

By Leonardo De Cosmo on July 12, 2022



# FOR NEXT CLASS

- Join the Slack
- Find the textbook
- Read chapter 2