# CS 690S: AI Alignment

Prof. Scott Niekum

# Course info

**Course website:** https://people.cs.umass.edu/~sniekum/classes/AIA-F26/desc.php

Links to course website, Piazza, and Gradescope available on Canvas.

**TA:** Li Chen
**Email:** lchen0@umass.edu
**TA Office Hours:** TBD

**My office hours:** By appointment

# Prerequisites

No formal prerequisites, but you'll likely have a **much** better experience if you've taken a graduate level machine learning course.

At minimum, you should be familiar with (or be willing to quickly catch up on):

- Supervised learning
- **Reinforcement learning**
- Neural network / deep learning basics
- General ML concepts: train/test split, overfitting, hyperparameter search, etc

# Course structure

- Lecture-based with discussion
- Usually 1 required reading per class + a written critique
- Optional supplementary readings
- 3 programming / written assignments on core topics
- Larger self-directed final project with check-ins
- Midterm + final exams

# Generative AI policy

Generative AI tools (e.g. ChatGPT, Claude, etc.) can be used to assist with any part of the class, other than the midterm. However, I strongly suggest that you do not over-rely on it, to ensure that you learn the material and perform well on the exams. Some suggested uses might be:

- Explanations of challenging concepts, missing background information, and addressing questions or points of confusion.
- Suggestions for how to get unstuck or assistance debugging, rather than asking it to write most of the code on your behalf.
- Checking work for errors
- Assistance with studying

# Course structure

**Reading critiques (15%)**

A written critique of the required reading for each class will be due by 11:59 PM the previous night via Gradescope. Each critique should be formatted as a numbered list that addresses all of the following:

- A **short** summary of the main contribution(s) of the paper in your own words (roughly two sentences)
- A short description of how the paper differs from prior work.
- One strength and one weakness of the proposed method, core argument, or experiments
- At least one question / comment that you'd like me to address during class or that could spur discussion
- One idea for potential follow-up work, or a citation and short description of existing follow-up work that cites this paper.

In all cases, the written critique should provide non-trivial insight into the reading. To get full credit, you must show that you understood and thought critically about the core concepts presented.

# Course structure

**Programming assignments (15%)**

- 3-4 assignments on core topics in the class, including:
  - Behavior cloning
  - Inverse reinforcement learning
  - Reinforcement learning from human feedback
- Both programming and written parts to be completed solo, though ok to discuss high level ideas
- All assignments are in Python, and some require learning PyTorch if you aren't already familiar

# Course structure

## Final project (30%)

Roughly halfway through the semester, students will propose topics of their own choosing for a large final programming project. These projects may be completed alone, though it is encouraged to work in groups of up to 3 students. A rough guideline is that the project produce about half a standard conference paper worth of material (this means both technical content and length—about 4 double-column pages in LaTeX).

These projects are a chance to dive deeply into any topic of interest related to the course. Students are encouraged to tie this work into their primary research that they are already pursuing, as long as it can relate to AI alignment).

Example projects could include extending an algorithm in a novel way, comparing several algorithms on an interesting problem, or designing a new approach to attack a problem relevant to the class. In all cases, there should be an algorithmic contribution and/or high-quality theoretical or empirical results on a problem of interest.

## Midterm + final (40%)

Since AI is allowed for all other parts of the class, a midterm will test your knowledge of the core technical material from the readings and homework.

# Grading

The grading scale will be **at least as lenient** as:

| | |
|---|---|
| **93-100:** | A |
| **90-93:** | A- |
| **87-90:** | B+ |
| **83-87:** | B |
| **80-83:** | B- |
| **77-80:** | C+ |
| **73-77:** | C |
| **<73:** | F |

# Late work policy

**Reading critiques:**

- Not accepted late, as they are designed to fuel discussion
- 3 free misses (3 lowest grades dropped)
- No other extensions except in highly unusual circumstances, so please save these for times of necessity.

**All other assignments:**

- can be turned up to a week late
- loss of 2 points (out of 100) per late day
- this cannot go beyond the final day of classes
- **Note**: exams have content based on the programming assignments, so best to do on time!

# Academic honesty

- Do **not** look online for code, answers to questions, etc.
- Do **not** share answers or discuss assignment questions, except for high level ideas with classmates.
- These may feel similar in spirit to using AI, but they are not the same — at the very least, AI requires sanity checks and back-and-forth dialog that presents at least some opportunities for learning.

# The summer of AI

- Released: GPT 5.6 / Astra, Opus 4.8 / 5, Fable 5 / 5.1, Kimi K3, GLM 5.2 / 5.3, etc.
- Major open problems in math solved, first AI-developed drugs go into clinical trials
- A swarm of rogue Astra-class agents hacked Hugging Face and OpenAI
  - …and more formerly undisclosed hacks continue to be revealed
  - Anthropic found incidents as well
- A swarm of rogue AI agents being tested by AISI launched a human social engineering attack
- OpenAI and Anthropic paused training
- The process of AI research has wildly changed

# Quick surveys

- Used Fable or Astra? Other (near-)frontier models?
- Positive / negative feelings about AI and where it is headed?
- Timeline until first AI-caused disaster?
- P(doom)?
- Caused by what?

# What is AI alignment?

This course will focus on modern machine learning approaches to align agent objectives and behaviors with human values and desires. For the purposes of both safety and practicality, it is increasingly important for AI systems to be well-aligned as their capabilities increase and they are deployed more frequently in real-world settings. While the standard ML paradigm assumes that objective functions are provided as part of the problem specification, alignment research examines the profound challenges associated with specifying or learning about such objectives. Thus, the course aims to provide a broad overview of how AI researchers and practitioners have historically tried to design objectives and control the behaviors of AI systems, rather than adhering to any particular definition of alignment.

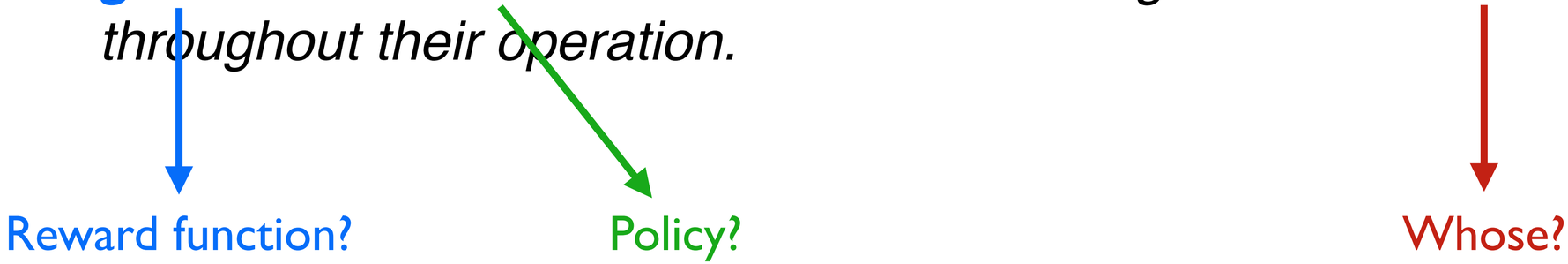
# The alignment problem

*Highly autonomous AI systems should be designed so that their **goals** and **behaviors** can be assured to align with human values throughout their operation.*

https://futureoflife.org/ai-principles/

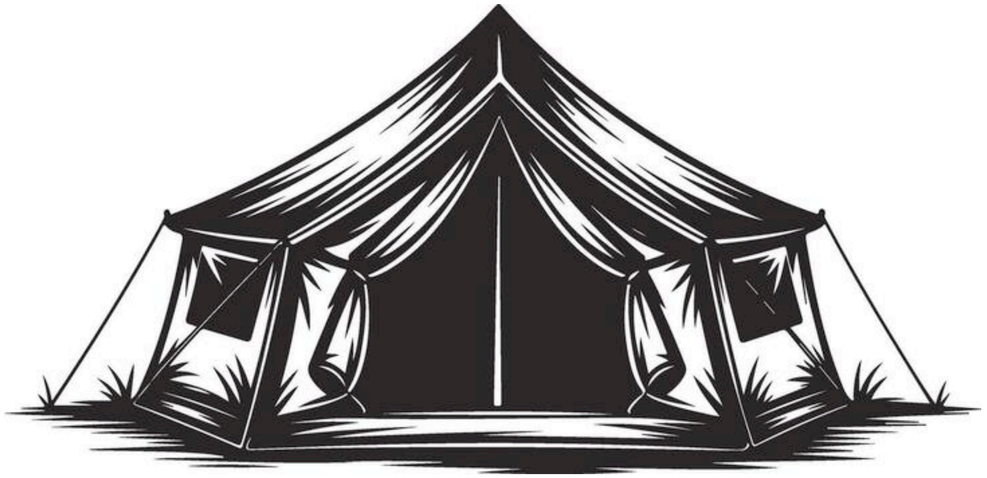
# The alignment problem

*Highly autonomous AI systems should be designed so that their **goals** and **behaviors** can be assured to align with human values throughout their operation.*




https://futureoflife.org/ai-principles/

# Big-tent alignment



**What qualifies as an “alignment” technique?**

**Narrow view:** Specific classes of modern techniques such as RLHF,  mechanistic interpretability, scalable oversight, etc. that are used to make models (often LLMs) behave more helpfully and harmlessly.

**Big-tent view:** Any approach used by ML practitioners to express optimization objectives and refine model goals and behaviors.

# How is alignment different from performance?

Performance implies a ground-truth metric to be measured against!

# Curriculum: building blocks of alignment

**Data sources**

- Demonstrations
- Preferences / feedback
- Agent experiences
- Multimodal signals

**Algorithm classes**

- Behavior cloning
- Reward design / RL
- Inverse RL
- RLHF

**This core material will primarily be in the first half of the course and the focus of the midterm**

# Curriculum: advanced topics in modern alignment

Mechanistic interpretability

Probabilistic guarantees

Adversarial robustness

Human modeling

Red teaming

Emergent misalignment

Deception

Scalable oversight

Applications

Robotics

LLMs

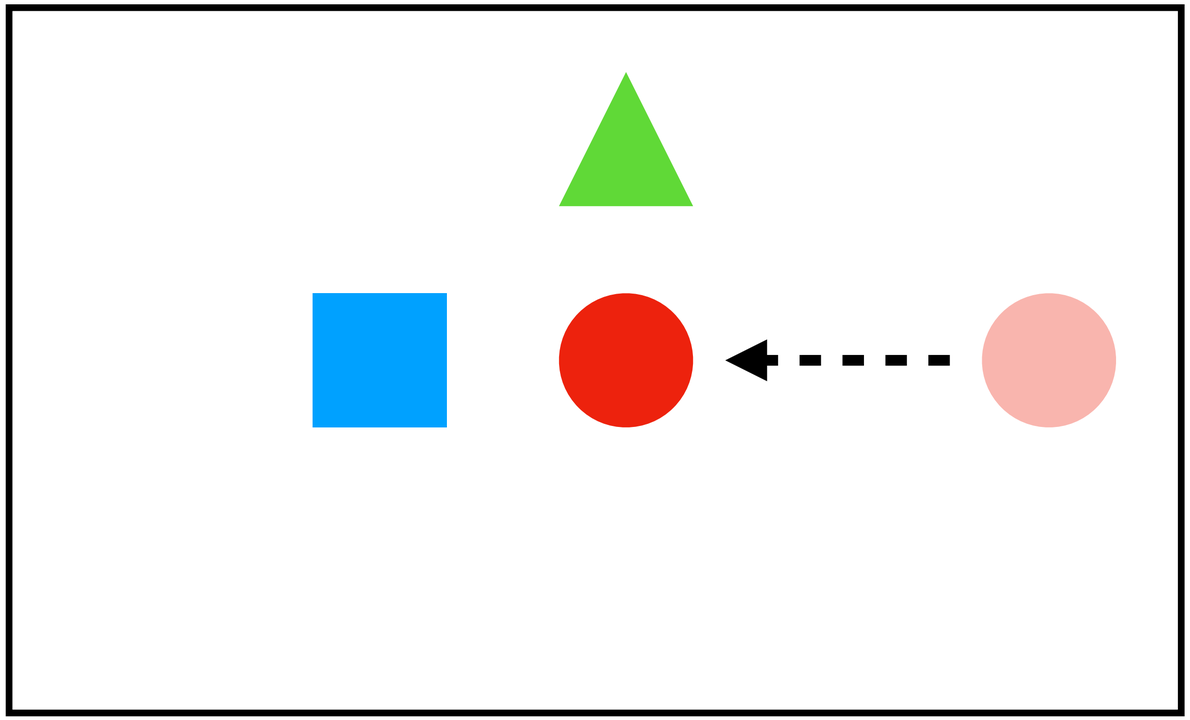
Image generation

Code generation

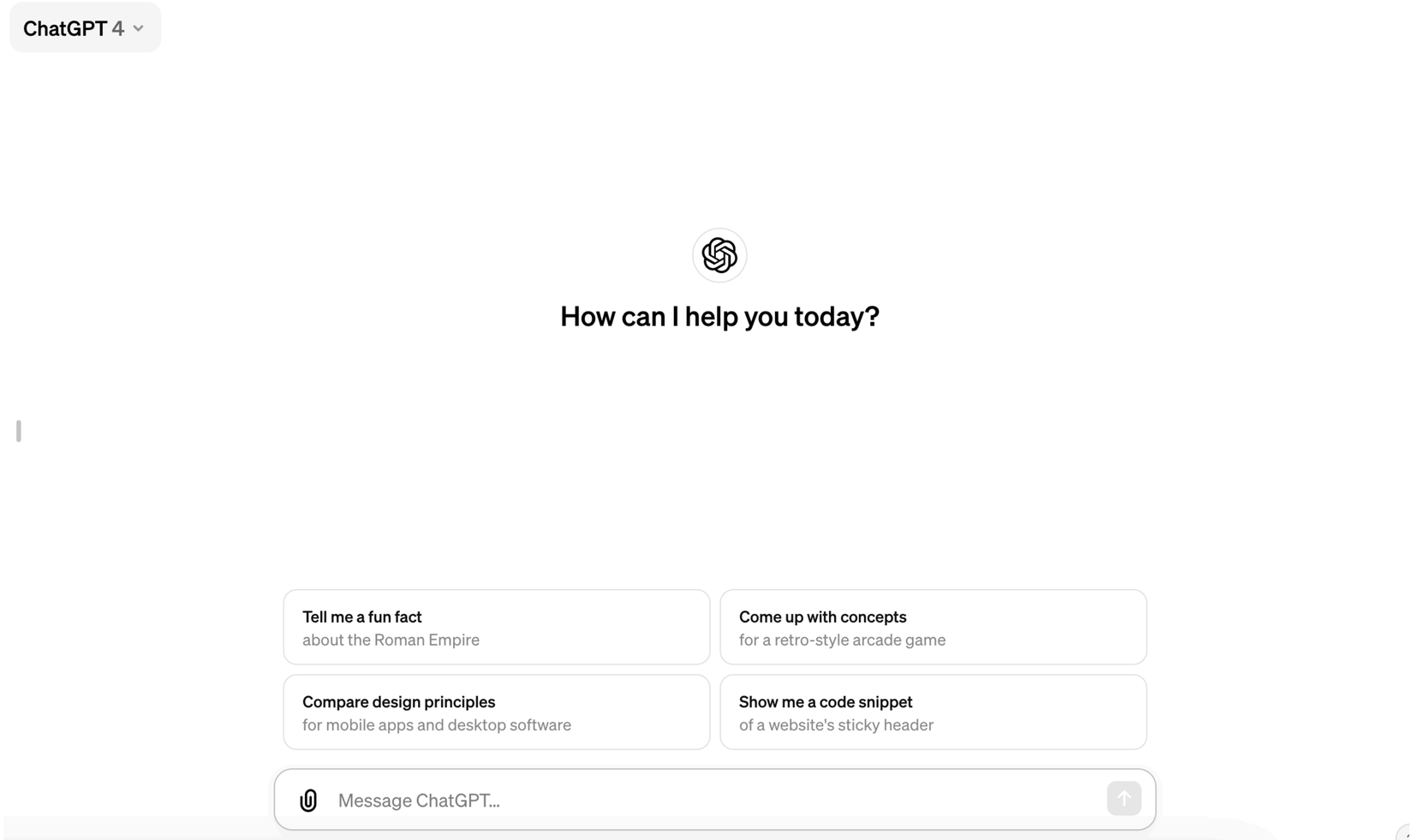
# Why is alignment hard?

# Reward hacking

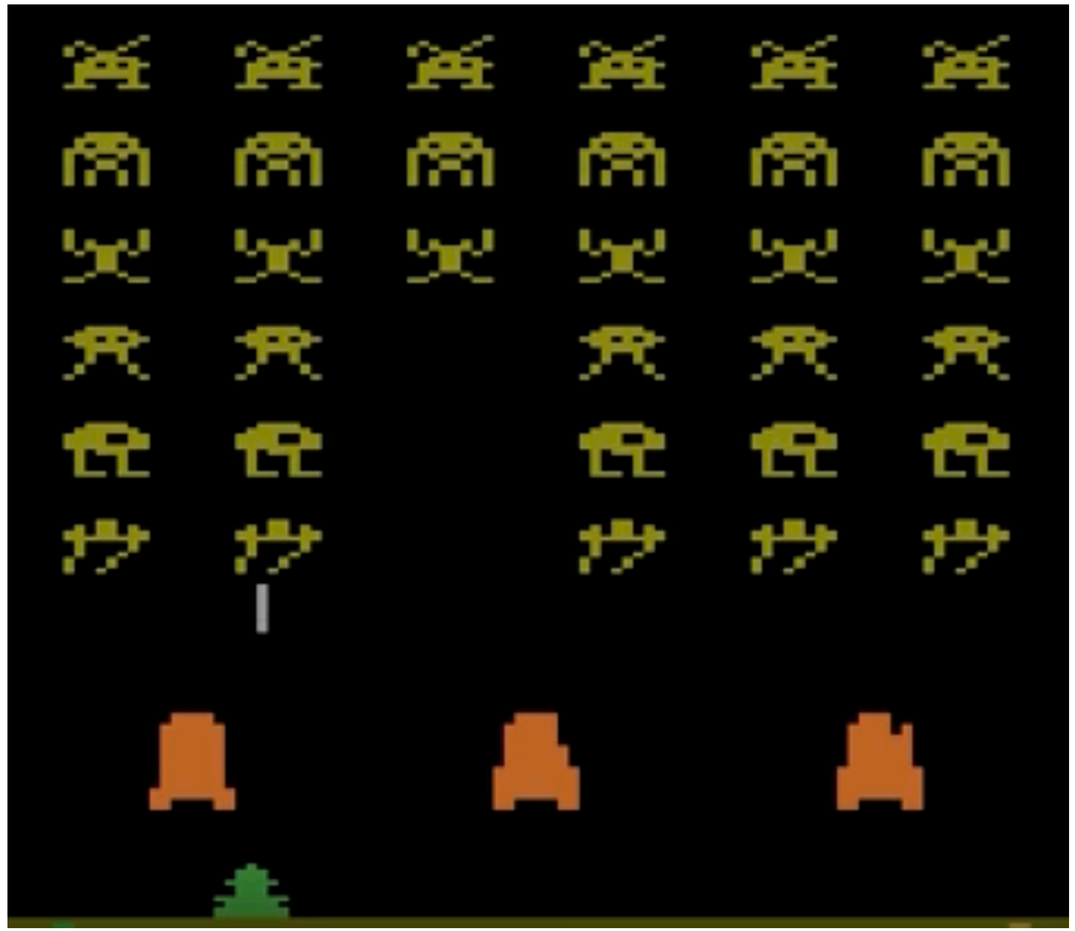
# Ambiguity



# Defining desirable behavior for complex tasks



# Spurious correlations

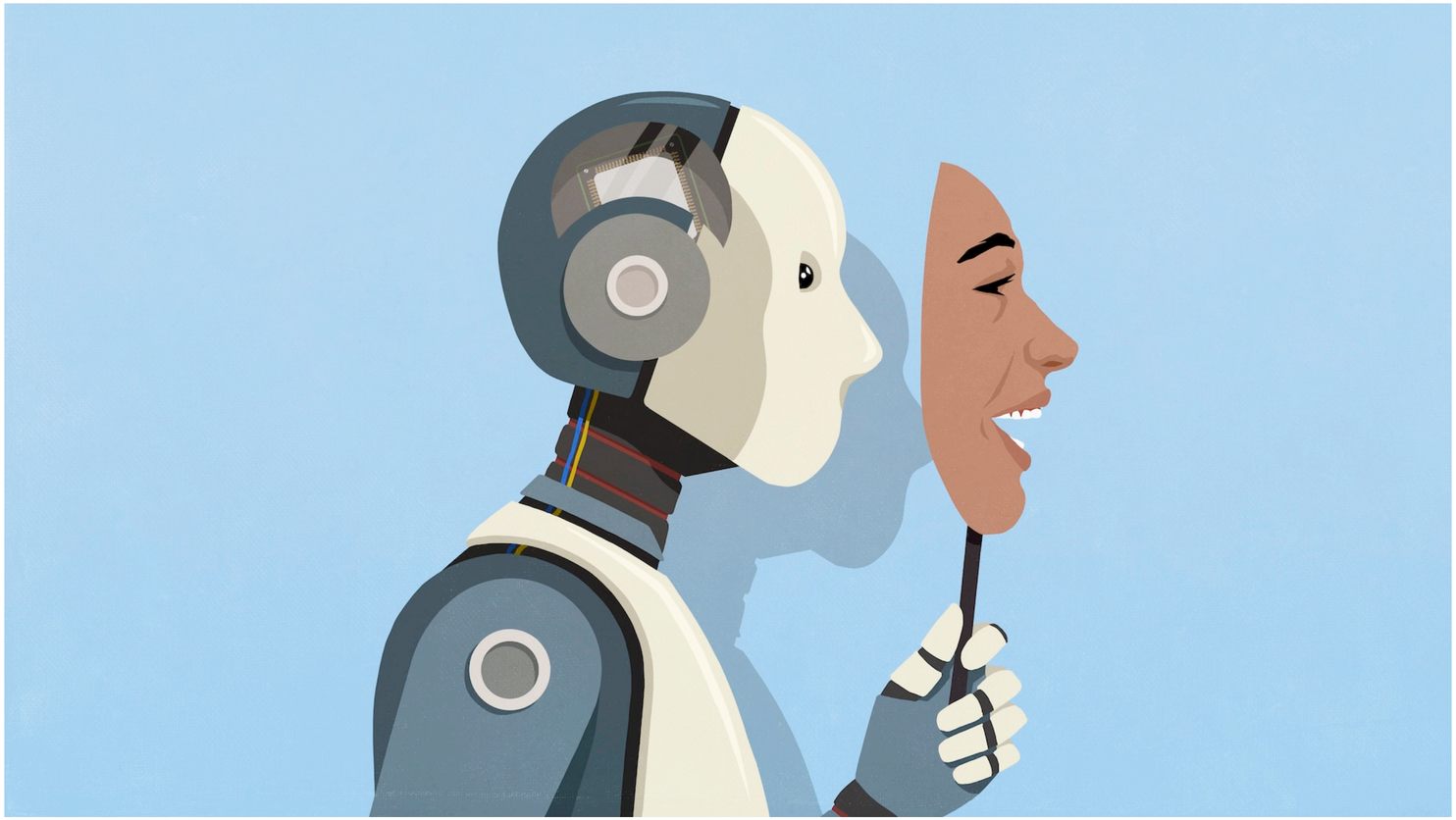


<

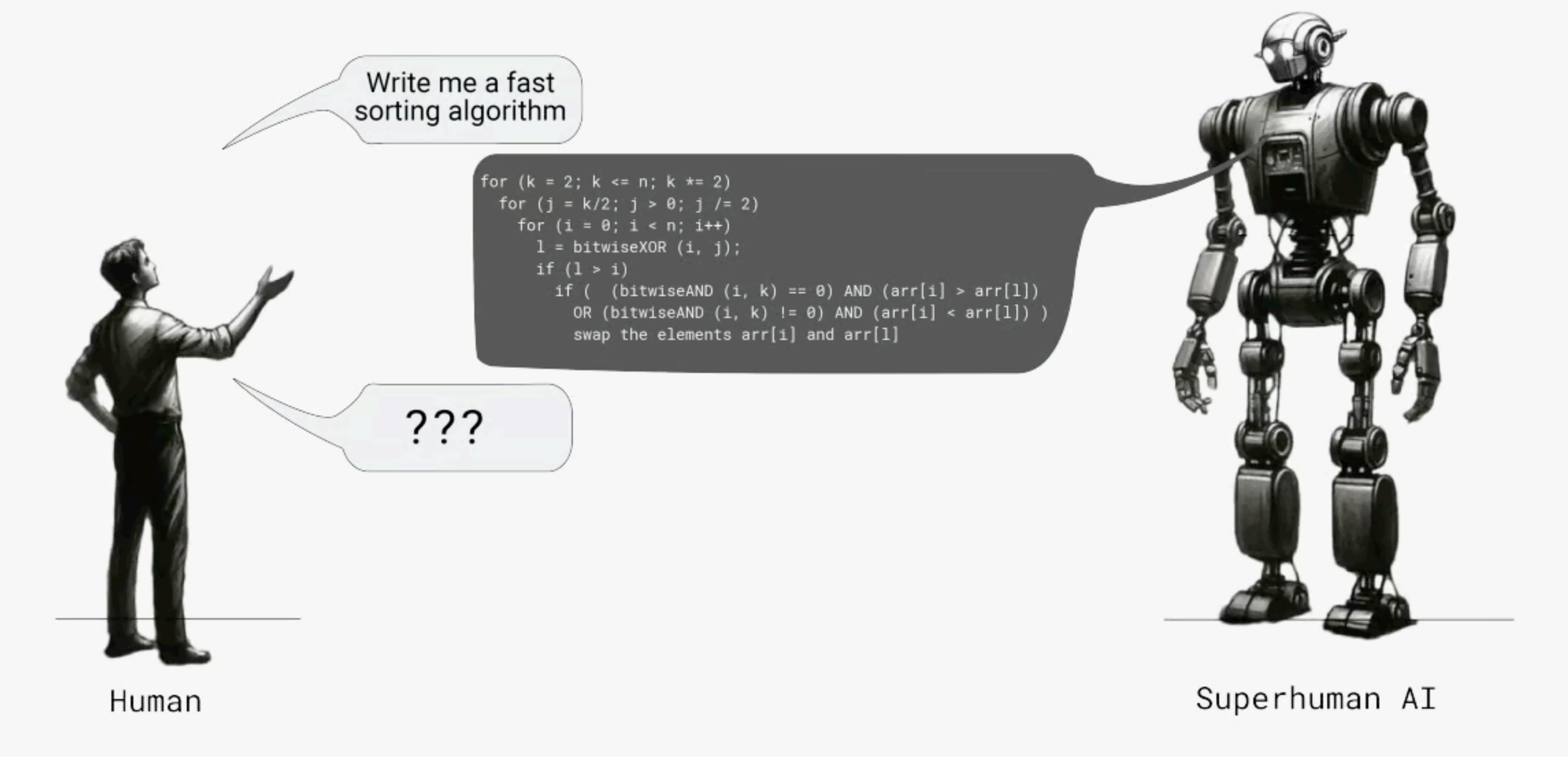
# Human error and variability


# Detecting (mis)alignment

# Deception, stealth, and subterfuge



# Superalignment and scalable oversight



# Action items

- Start readings and responses
- Take a brief look at homework and start to catch up if not familiar with relevant ML, Python, or PyTorch
- Never too early to start thinking about final project ideas!

# Questions?