CS 690S: Human-Centric Machine Learning

Prof. Scott Niekum

What is human-centric ML?

This course will focus on modern machine learning approaches to learn from human demonstrations, preferences, feedback, and other multimodal signals, with the goal of **aligning agent goals 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 with human users as their capabilities improve and they are deployed more frequently in real-world settings. While **the standard ML paradigm assumes that learning objectives are directly provided as part of the problem specification**, emerging research in alignment suggests that it is often infeasible to do so accurately, **requiring such objectives to be inferred from human data**.

What is human-centric ML?

Data types

- Demonstrations — video, kinesthetic, direct control, prompt completion, etc.
- Preferences — binary, N-ranking, often from noisy crowdworkers
- Feedback — thumbs up/down, scalar, natural language, facial expressions
- Multimodal signals — language descriptions, gaze, prosody, body language, etc.

Objectives

- Alignment — it isn't trivial to communicate what we want the AI system to do
- Safety — we don't only care about performance in expectation, but risk and constraints
- Trust — there's an additional psychological component of how AI systems make users feel
- Transparency / verifiability — deployed AI must interface with broader social systems, regulation, etc.

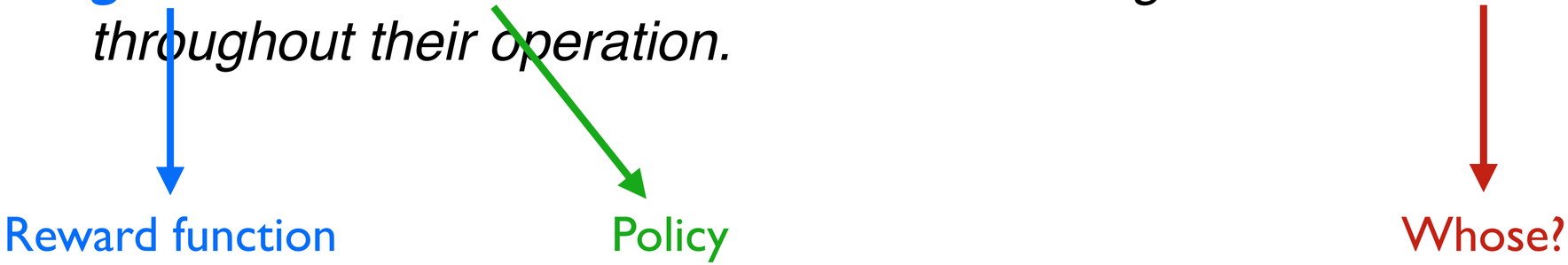
The alignment problem

*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

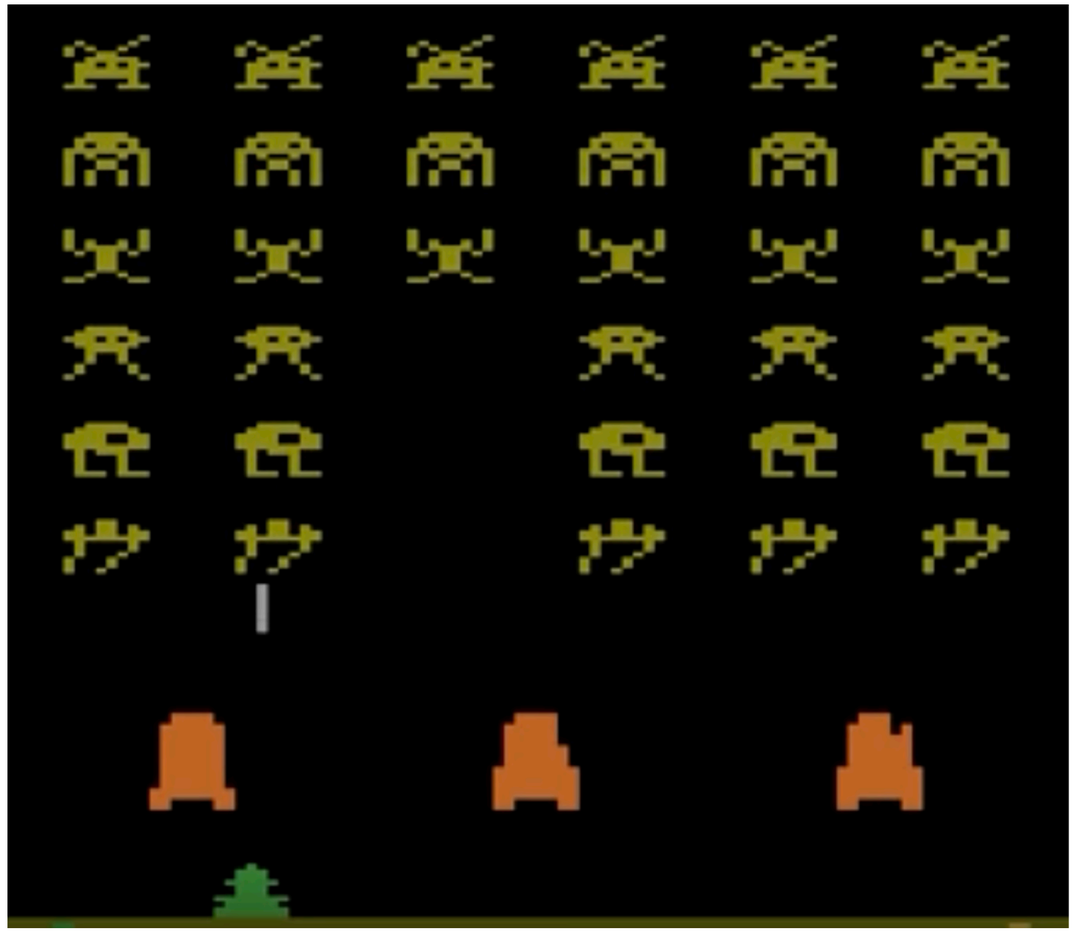
*Highly autonomous AI systems should be designed so that their **goals** and **behaviors** can be assured to align with human values throughout their operation.*



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

Alignment: reward hacking

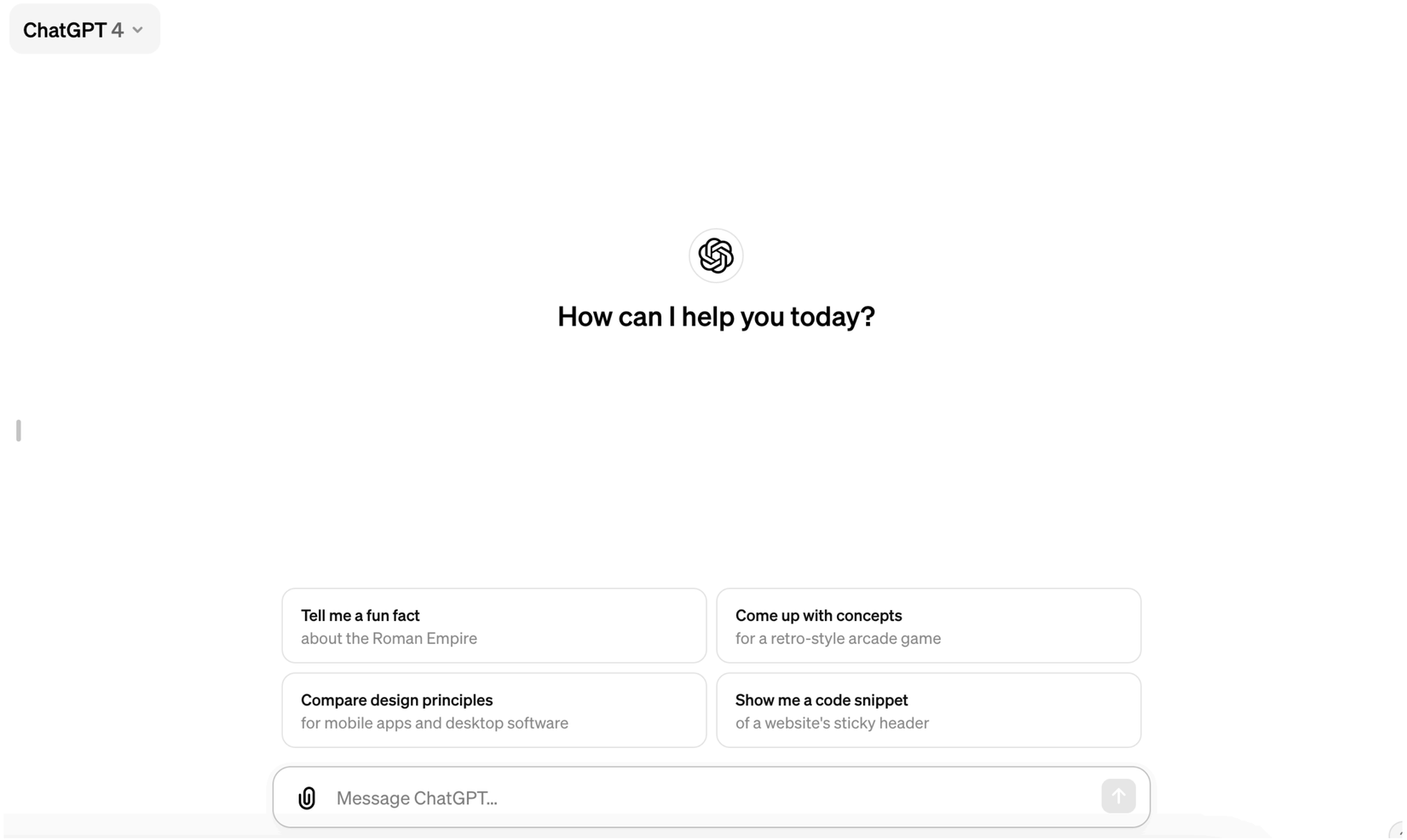
Alignment: spurious correlation



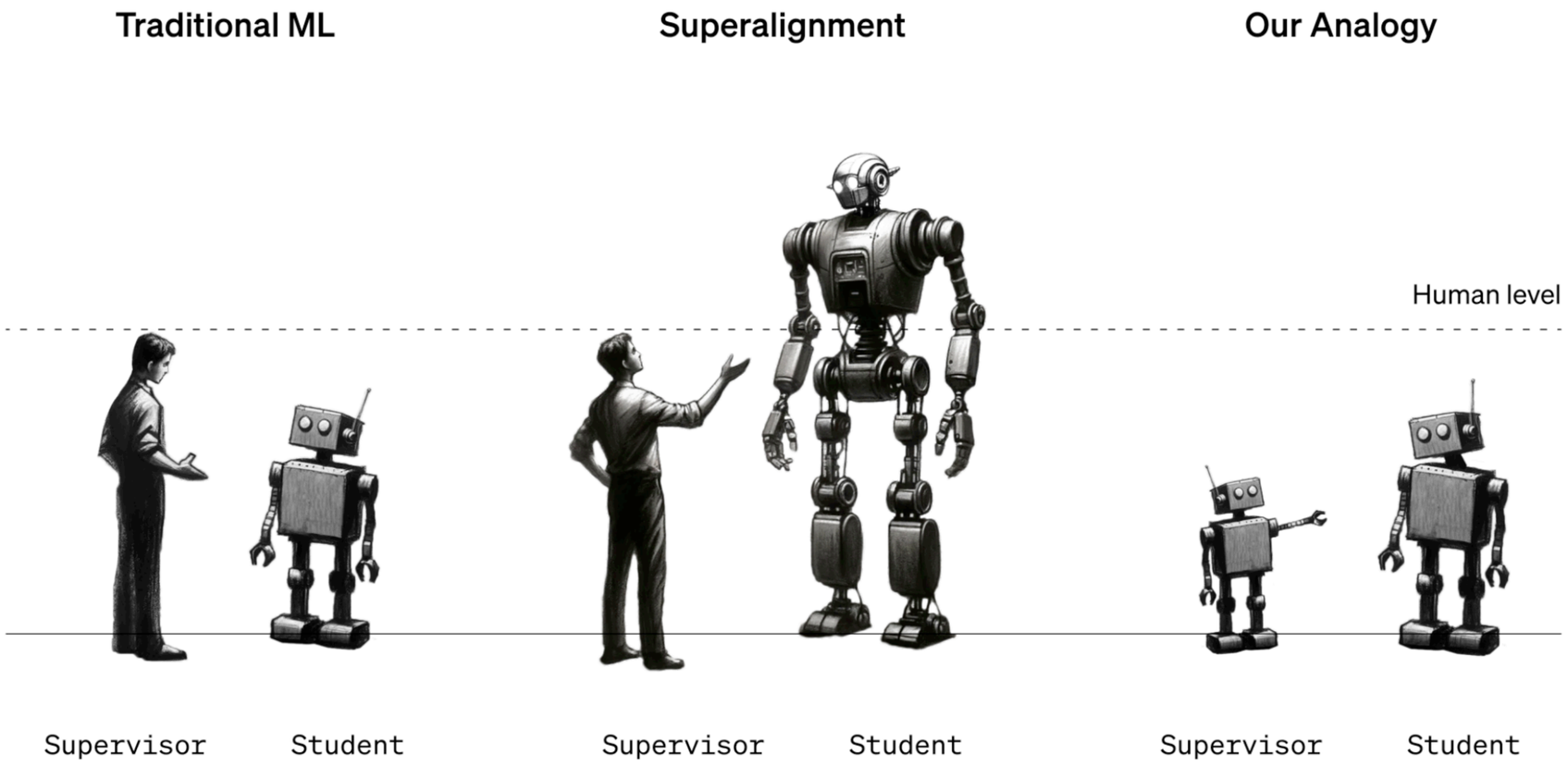
>

Alignment: verification difficulty

Alignment: practical use case



Alignment: superalignment



C. Burns, P. Izmailov, J.H. Kirchner, B. Baker, L. Gao, L. Aschenbrenner, Y. Chen, A. Ecoffet, M. Joglekar, J. Leike, I. Sutskever. Weak-to-strong Generalization: Eliciting Strong Capabilities with Weak Supervision. arXiv:2312.09390. 2023

Course info

Course website: https://people.cs.umass.edu/~sniekum/classes/HCML-S24/desc.php

Links to Piazza and Gradescope available on website.
Gradescope course entry code: YDPRPP

TA: Rohan Pandey
Email: rohanpandey@umass.edu
TA Office Hours: Tues 2:30-3:30 and Wed 4:00-5:00 in CS207 cube 1

My office hours: By appointment

Course structure

- ~50% lecture / 50% discussion
- Usually 2 readings per class + a written critique
- Special discussion role assignments
- 3 programming / written assignments on core topics
- Larger self-directed final project with check-ins
- No exams

Course structure

Role presentations (15%)

- To add depth to the discussion, we will draw on ideas from Colin Raffel's roleplaying seminar model.
- A few times per semester, each student will assigned one of the following roles for a paper and asked to produce a written report, as well as kick off discussion in class by presenting (from their seat; no slides) a summary of their findings in approximately 5 minutes
- Grades will be assigned only for the quality of the written report, but students are expected to be well-prepared to present their findings in class.

Roles

- **TMLR Reviewer:** The machine learning journal TMLR uses a peer review process that focuses on technical correctness and quality, rather than subjective novelty. For reference, the review guidelines for can be found here: https://jmlr.org/tmlr/reviewer-guide.html. Write a review that discusses: (1) The claims being made by the authors; (2) Whether all claims are supported by sufficient (and correct) arguments, theory, or empirical evidence, and if not, what you'd like to see; (3) Clarifying questions you'd like to ask the authors; (4) Whether the results appear to be reproducible from information contained in the paper. Not all of these will be applicable for every paper, so use your judgement

- **Archaeologist:** This paper was found buried under ground in the desert. You're an archeologist who must determine where this paper sits in the context of previous and subsequent work. Find and report on one older paper cited within the current paper that substantially influenced the current paper and one newer paper that cites this current paper. If the paper is too new to have been meaningfully cited, report on a second older paper instead. Discuss in detail how the papers influenced each other from both conceptual and technical perspectives, as well as how they differ, and what their main results are.

- **Academic Researcher:** Imagine that you're an academic researcher and this paper was just released. Propose a follow-up research project that builds on these ideas, addresses a key limitation of the current paper, or that investigates something about the paper's analysis or experiments that you are skeptical of. If this is an older paper with a lot of follow-up work already existing, feel free to come up with an idea that instead builds on a paper that was influenced by this one. Feel free to talk to other students to brainstorm. If you get really stuck, then simply report on the limitations of this work and highlight challenges that would be valuable to address, even if you don't have a proposal for how you might address them. Or alternately, propose something on a smaller-scale, such as an additional experiment or hypothesis to examine that would have made the paper stronger.

Course structure

Reading critiques (25%)

For everyone who isn't assigned a role for a particular paper, a written critique of each reading (usually two) for each class will be due by 8:00 PM the previous night via Gradescope. Each critique should include 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

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 (25%)

- 3 assignments on core topics in the class:
  - 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 3 assignments are in python, and 2 require learning PyTorch if you aren't already familiar

Course structure

Final project (25%)

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 human-centric ML).

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 a novel intellectual contribution, as well as empirical results on a problem of interest.

Attendance + participation (10%)

Attendance is mandatory and participation in discussion is an important element of the course
Aim to participate in the discussion at least once a week.

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
- 3 free misses (3 lowest grades bumped up to 10s)
- No other extensions except in highly unusual circumstances, so please save these for times of necessity.

All other assignments:

- can be turned in up to one week late
- loss of 5 points (out of 100) per late day
- this cannot go beyond the final day of classes

Academic honesty

Generative AI policy:

- Generative AI tools (e.g. ChatGPT) can only be used in the context of background research to better understand topics covered in the class. For example, it is permissible to ask ChatGPT for a summary of how inverse reinforcement learning differs from behavioral cloning.
- You are **not** permitted to use generative AI tools to assist with any part of completing your reading summaries, written homeworks, or coding assignments.
- This policy clearly forbids copying text or code directly from these sources, but it is also **not** acceptable to summarize the output of generative AI tools, or to use an answer from them as a starting point for your own work

Other policies:

- Do **not** look online for code, answers to questions, etc., even just for inspiration
- Do **not** discuss assignment questions, except for high level ideas with classmates
- You'll fail the assignment and possibly the class, so please don't cheat!

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

Topics: algorithm types

Behavior cloning

- Demonstrations
- Simple learning
- Poor off-distribution

Reinforcement learning

- Reward hand-coded
- Expensive exploration
- Complex algorithms
- Difficult hyperparam tuning
- Strong generalization guarantees

Inverse RL

- Demonstrations
- No hand-designed reward
- Assume optimal demos
- RL in inner loop!

RLHF

- Preferences or feedback
- Lower human demand
- RL or RL-free
-

Topics: human-centric considerations

AI risk / safety

Performance guarantees

Human modeling

Bounded rationality

Human-AI cooperation

Multimodal human signals

Scalable oversight

Superintelligence

Action items

- Start readings for next week
- Take a brief look at homework and start to catch up if not familiar with relevant ML, Python, or PyTorch
- Look out for TA’s role schedule
- Never too early to start thinking about final project ideas!

Questions?