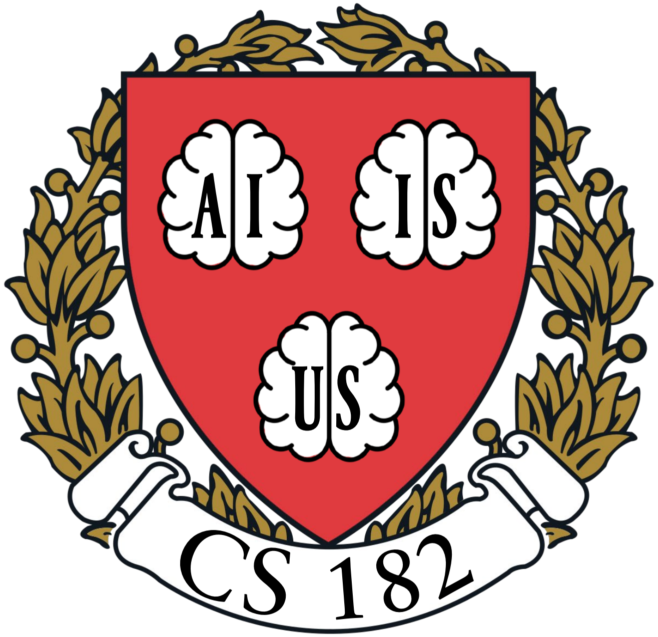


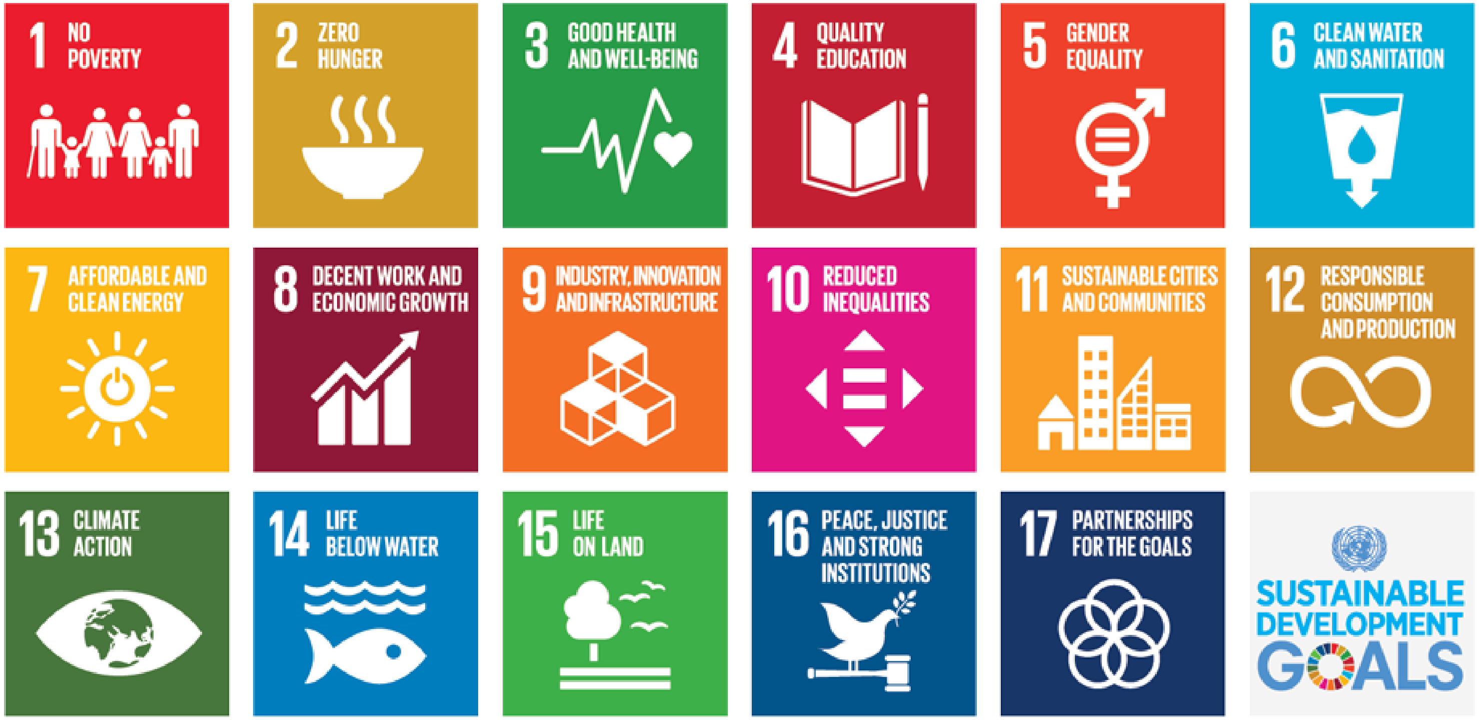
Fall 2021 | Lecture 23

Discussion

Ariel Procaccia | Harvard University

Q: BENEFITS OF AI

Poll 1: What applications of AI do you see to be the most (positively) impactful in the future?


1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

Q: RISKS OF AI

How worried should we be about advances in AI being adopted by militaries to automate warfare?

How far out is AGI, and what mechanisms can we use to mitigate its risk as an “existential threat to humanity?”

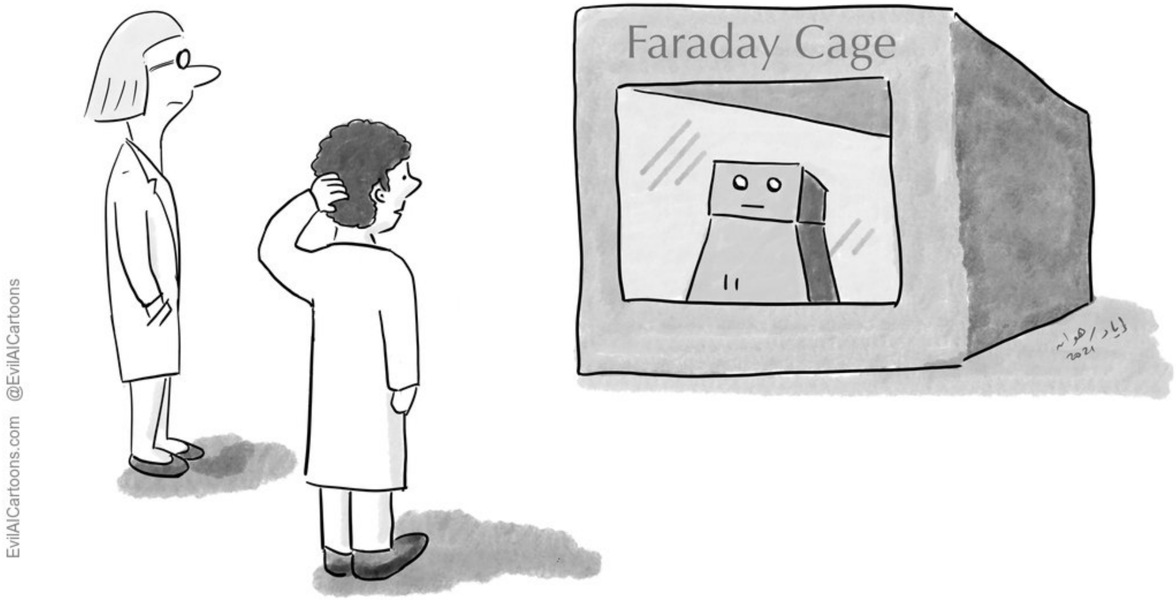
A: RISKS OF AI

"Almost any technology has the potential to cause harm in the wrong hands, but with [superintelligence], we have the new problem that the wrong hands might belong to the technology itself."

Russell and Norvig

Q: BENEFITS VS. RISKS

Poll 2: Is it clear that the benefits of pursuing AI outweigh its risks?



"If we let it out, there's an 85% chance it would cure cancer. But there's also a 0.01% chance it takes over the world!"

Q: WEIGHTED VOTING

In the context of voting, are there models where not everyone has votes weighted equally? We already do this (i.e., in some states people in prison for major offenses cannot vote and have a weight of zero). For example, suppose there is a ‘best’ option, and people are more or less likely to vote for it, modeled as Bernoulli random variables — should you/how might you weight votes in this case? Does this count as discrimination against certain peoples?

A: WEIGHTED VOTING

1 vote
Man

2 votes
Educated man

3 votes
Rich+educated man

Weighted voting in Belgium in the 19th century

Q: BOUNDED RATIONALITY

It seems like a lot of our AI discussion has focused on chasing rationality using computers, which approximates intelligence, but somehow seems too “artificial.” Are there AI systems in which we allow the systems to be finitely rational — like a human trying their best, or even irrational, but guided by some overarching principles?

Q: VALUE ALIGNMENT

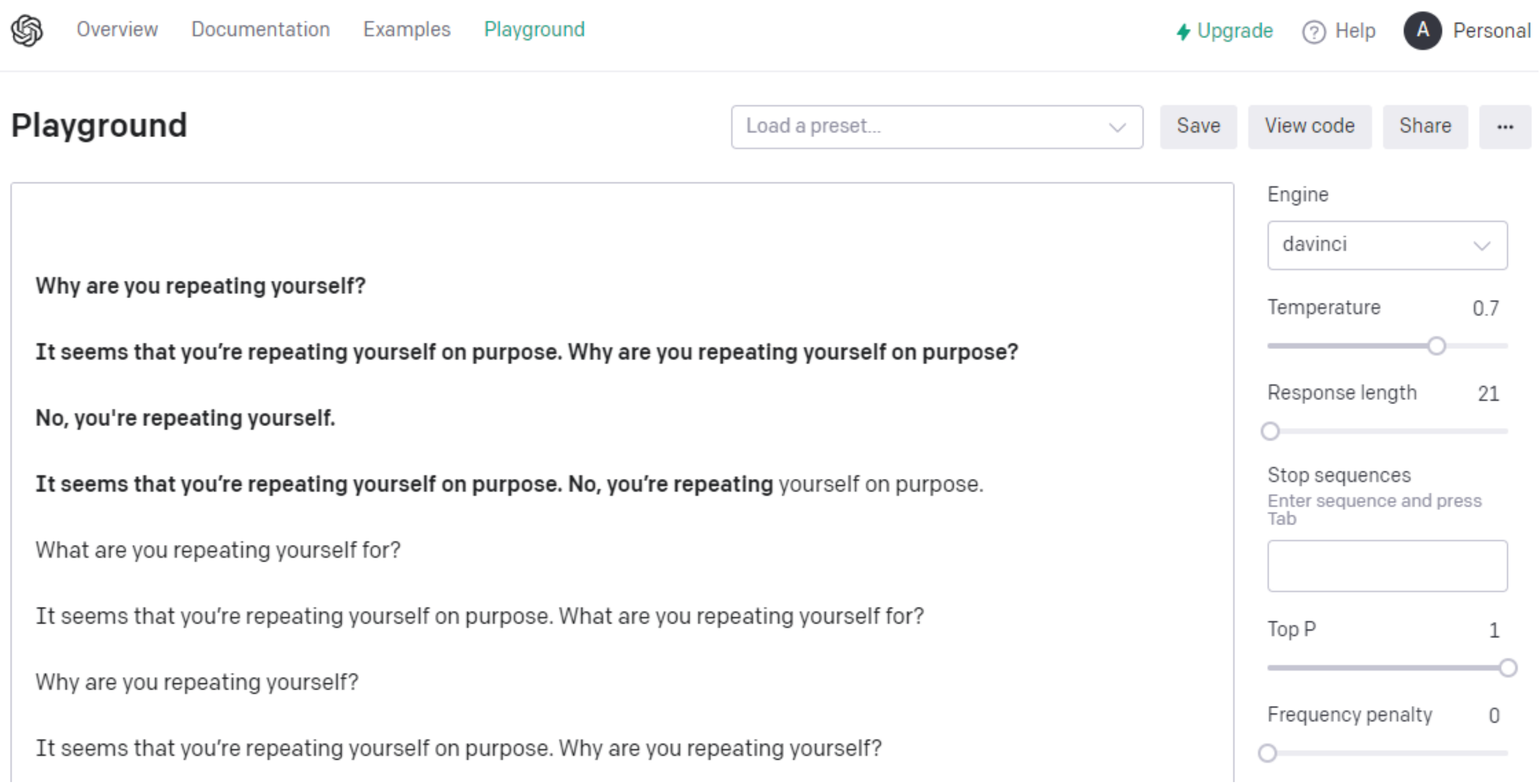
The kinds of machine learning algorithms we discussed in this class essentially make classifications based on empirical data. In other words, they approximate functions that describe trends in a particular dataset. How does this view of machine learning, under which ML predicts based on trends that hold in the status quo, conflict with the use of machine learning algorithms to make value-laden judgments? Is that even what machine learning algorithms are for? Are these kinds of value judgments even the domain of AI?

Q: SIMULATING INTELLIGENCE

What AI method best replicates / approaches human intelligence?

A: SIMULATING INTELLIGENCE

- Poll 3: What would you like to ask GPT-3?



Engine: davinci

Temperature 0.7

Response length 21

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Why are you repeating yourself?

It seems that you're repeating yourself on purpose. Why are you repeating yourself on purpose?

No, you're repeating yourself.

It seems that you're repeating yourself on purpose. No, you're repeating yourself on purpose.

What are you repeating yourself for?

It seems that you're repeating yourself on purpose. What are you repeating yourself for?

Why are you repeating yourself?

It seems that you're repeating yourself on purpose. Why are you repeating yourself?

TEACHING PHILOSOPHY, REVISITED

Key course objective:

Learn to represent problems

SYLLABUS, REVISITED

Problem solving
- Uninformed search
- Informed search
- Motion planning
- Constraint satisfaction problems
- Convex optimization
- Integer programming

Multi-agent systems
- Game theory
- AI game playing
- Wildlife protection (Tambe)
- Social choice

Reasoning with uncertainty
- Bayesian networks
- Hidden Markov Models
- Markov decision processes

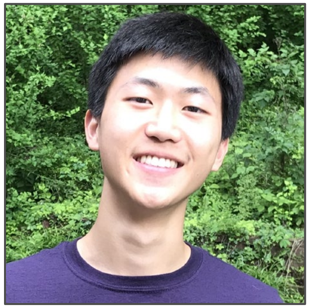
Machine learning
- Reinforcement learning
- Decision trees
- Linear classification
- Neural networks
- Language models (Alvarez-Melis)

Ethics
- Fairness
- Value alignment
- Embedded EthiCS (Wells)

THANKS!

Nabib Ahmed

Wenqi Chen



Max Guo

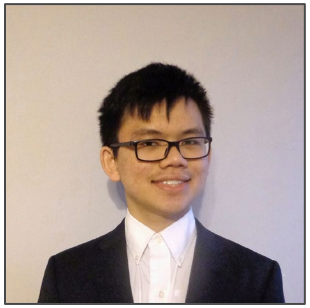
Kavya Kopparapu

Eric Lin

Amir Shanehsazzadeh

Zuzanna Skoczylas

Meiling Thompson



William Zhang