

# Announcements

## Assignments

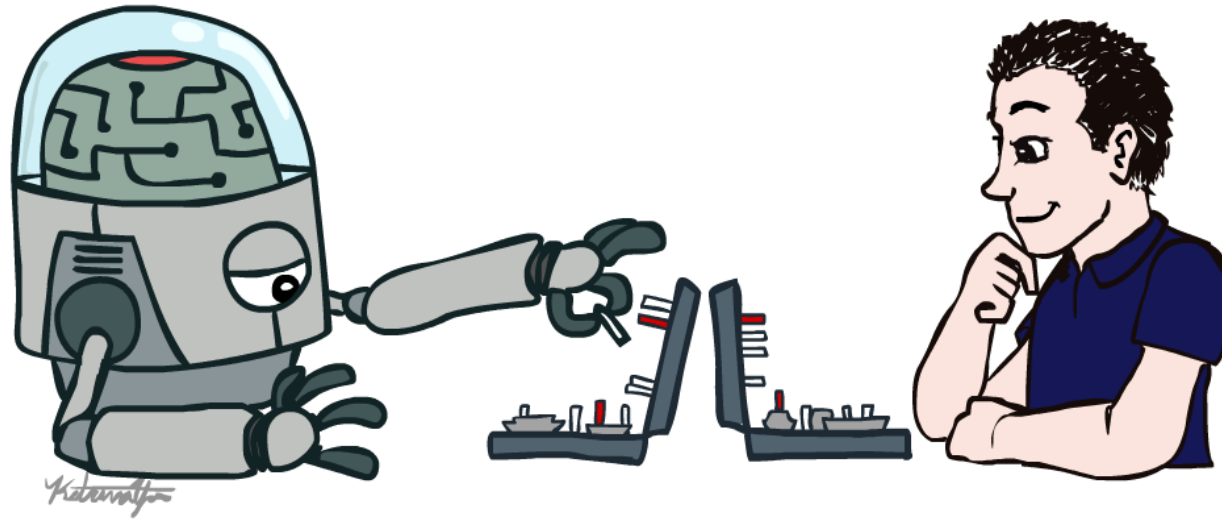
- HW12
  - Due tomorrow, 12/6, 10 pm

## Final Exam

- Thur, 12/12, 1-4 pm, GHC 4401
- See Piazza for details
- Recitation → Review session, Fri, 3-4:20 pm, GHC 4401
- Practice exam and learning objectives coming soon!
  - Practice exam solution session: Wed, 12/11, 12-2pm, GHC 4215

# AI: Representation and Problem Solving

## Human-Compatible AI



Instructors: Pat Virtue & Fei Fang

Slide credits: CMU AI and <http://ai.berkeley.edu>

# Intelligent Agents

From first lecture

## Candy Grab

- A. 11 pieces on the table
- B. Take turns taking either 1 or 2 pieces
- C. Person that takes the last piece wins!

```
class Agent
    function getAction(state)
        return action
```

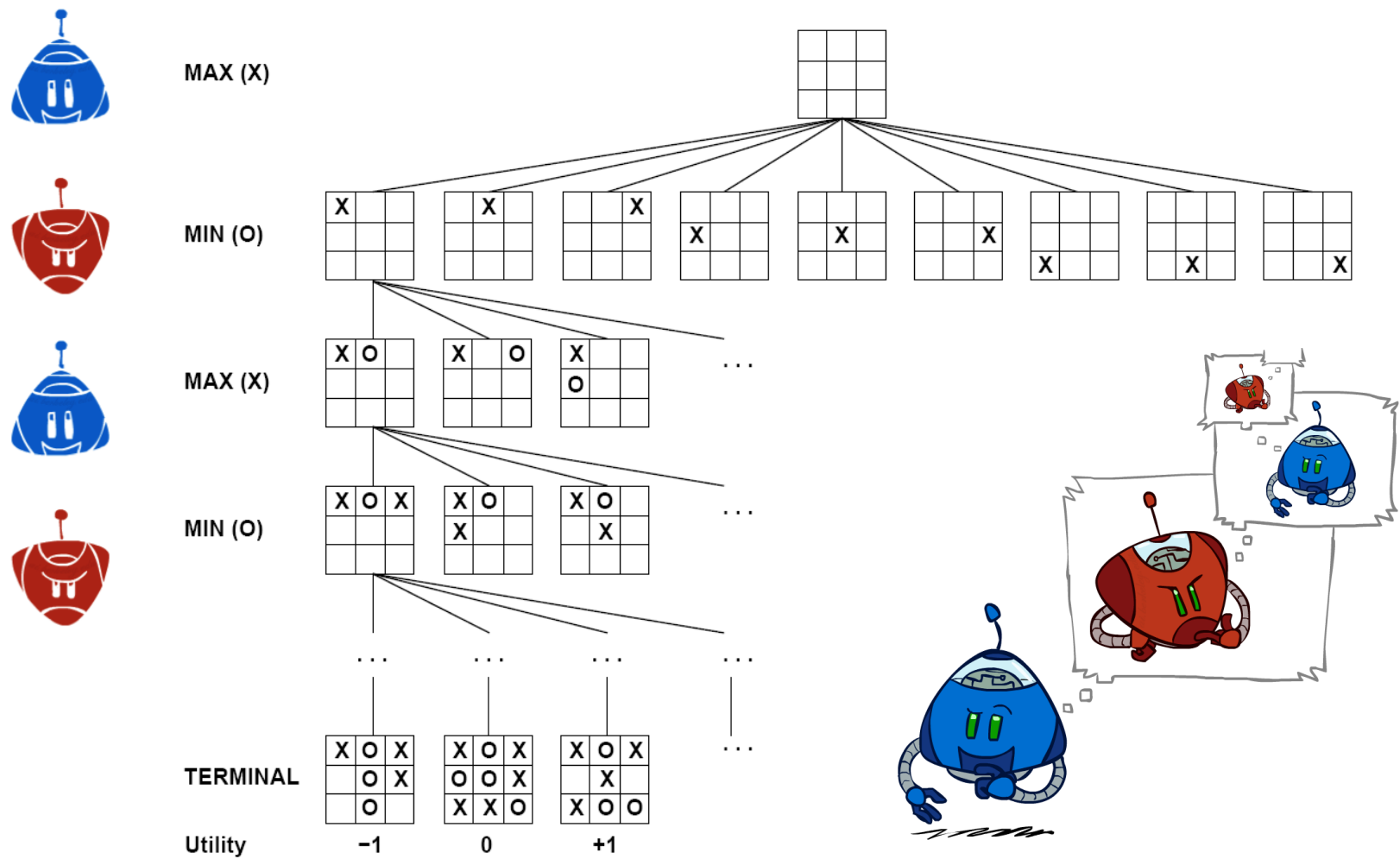
# Piazza Poll 1

## Three “Intelligent” Agents

Which agent code is the most “intelligent”?

# Games – Three “Intelligent” Agents

## A: Search



# Games – Three “Intelligent” Agents

## B: Encode the pattern

```
function getAction( numPiecesAvailable )  
  
    if numPiecesAvailable % 3 == 2  
        return 2  
    else  
        return 1
```

```
10's value:Win  
9's value:Lose  
8's value:Win  
7's value:Win  
6's value:Lose  
5's value:Win  
4's value:Win  
3's value:Lose  
2's value:Win  
1's value:Win  
0's value:Lose
```

# Games – Three “Intelligent” Agents

C: Record statistics of winning positions

| Pieces Available | Take 1 | Take 2 |
|------------------|--------|--------|
| 2                | 0%     | 100%   |
| 3                | 2%     | 0%     |
| 4                | 75%    | 2%     |
| 5                | 4%     | 68%    |
| 6                | 5%     | 6%     |
| 7                | 60%    | 5%     |

# Piazza Poll 1

## Three “Intelligent” Agents

Which agent code is the most “intelligent”?

- A. Search
- B. Encode multiple of 3 pattern
- C. Keep stats on winning positions

# Piazza Poll 2

281

Which ~~381~~ technique is the most intelligent?

- A. Search
- B. Logical inference
- C. Numeric optimization
- D. Q-learning
- E. Approximate Q-learning
- F. Exact inference Bayes nets
- G. Approximate inference Bayes nets

# Value of Information

## Ghostbusters

Given:

$$P(G = g_{top}) = 0.5$$

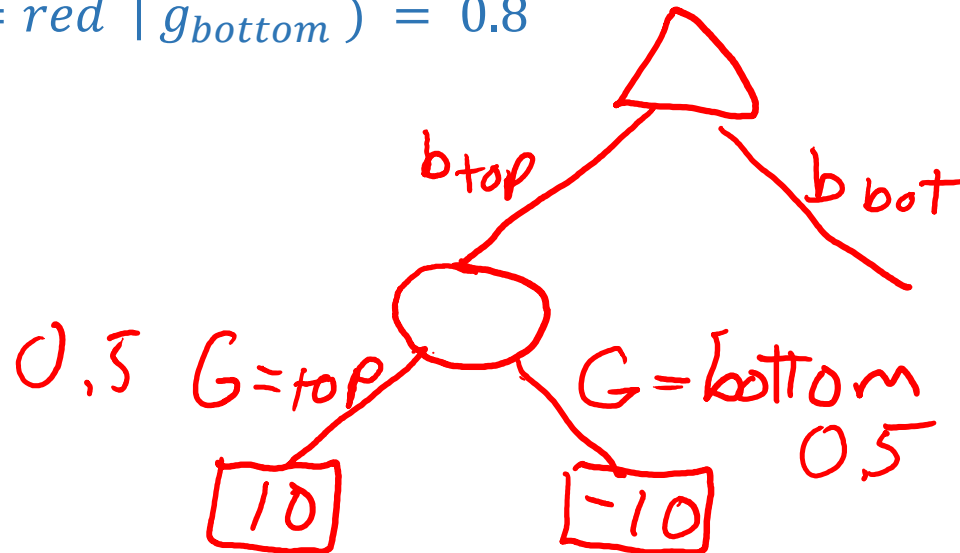
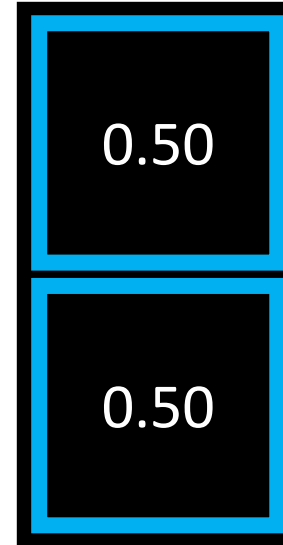
$$P(G = g_{bottom}) = 0.5$$

$$P(S_{top} = red \mid g_{top}) = 0.8$$

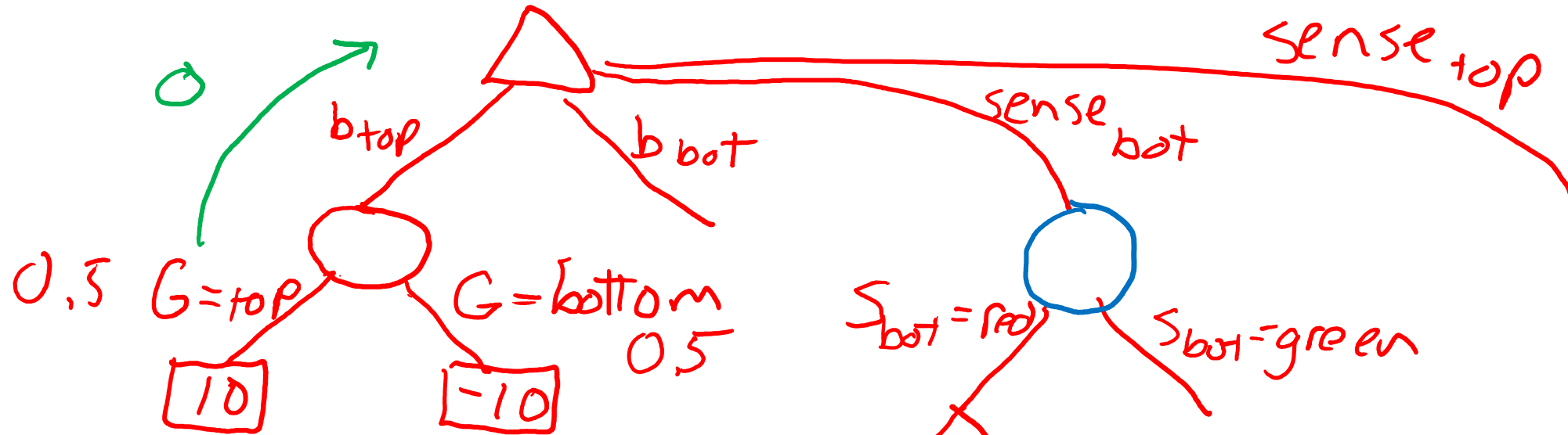
$$P(S_{top} = red \mid g_{bottom}) = 0.4$$

$$P(S_{bottom} = red \mid g_{top}) = 0.4$$

$$P(S_{bottom} = red \mid g_{bottom}) = 0.8$$



# Value of Information



$$P(G = g_{top}) = 0.5$$

$$P(G = g_{bottom}) = 0.5$$

$$P(S_{top} = red \mid g_{top}) = 0.8$$

$$P(S_{top} = red \mid g_{bottom}) = 0.4$$

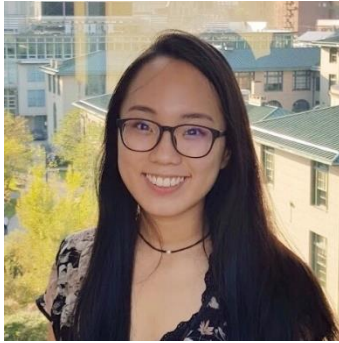
$$P(S_{bottom} = red \mid g_{top}) = 0.4$$

$$P(S_{bottom} = red \mid g_{bottom}) = 0.8$$

$$P(S_{bot} = red) = \sum_G P(G, S_{bot}) =$$

FCE: <https://cmu.smartevals.com/>

TA survey: <https://www.ugrad.cs.cmu.edu/ta/F19/feedback/>



Angela  
Yang



Claire Wang  
(Head TA)



George  
Brown



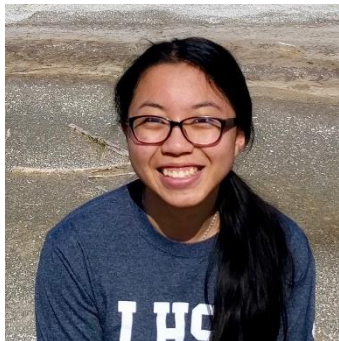
Michelle  
Ma



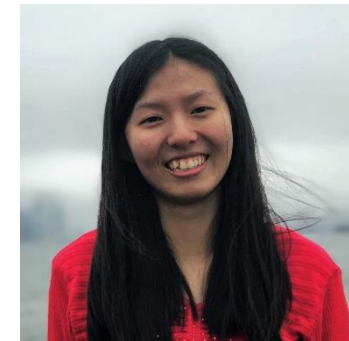
Sean  
Pereira



Chakara (Tian)  
Owarang



Tina  
Wu



Vicky  
Zeng

# AI in the News

NEWSLETTERS • ARTIFICIAL INTELLIGENCE

## Could Artificial Intelligence Save Us From Itself?

By [Jeremy Kahn](#) November 26, 2019



<https://fortune.com/2019/11/26/ai-is-the-problem-and-the-solution/>

# AI in the News

<https://www.forbes.com/sites/grantfreeland/2019/12/02/maximize-the-promise-and-minimize-the-perils-of-artificial-intelligence-ai/amp/>



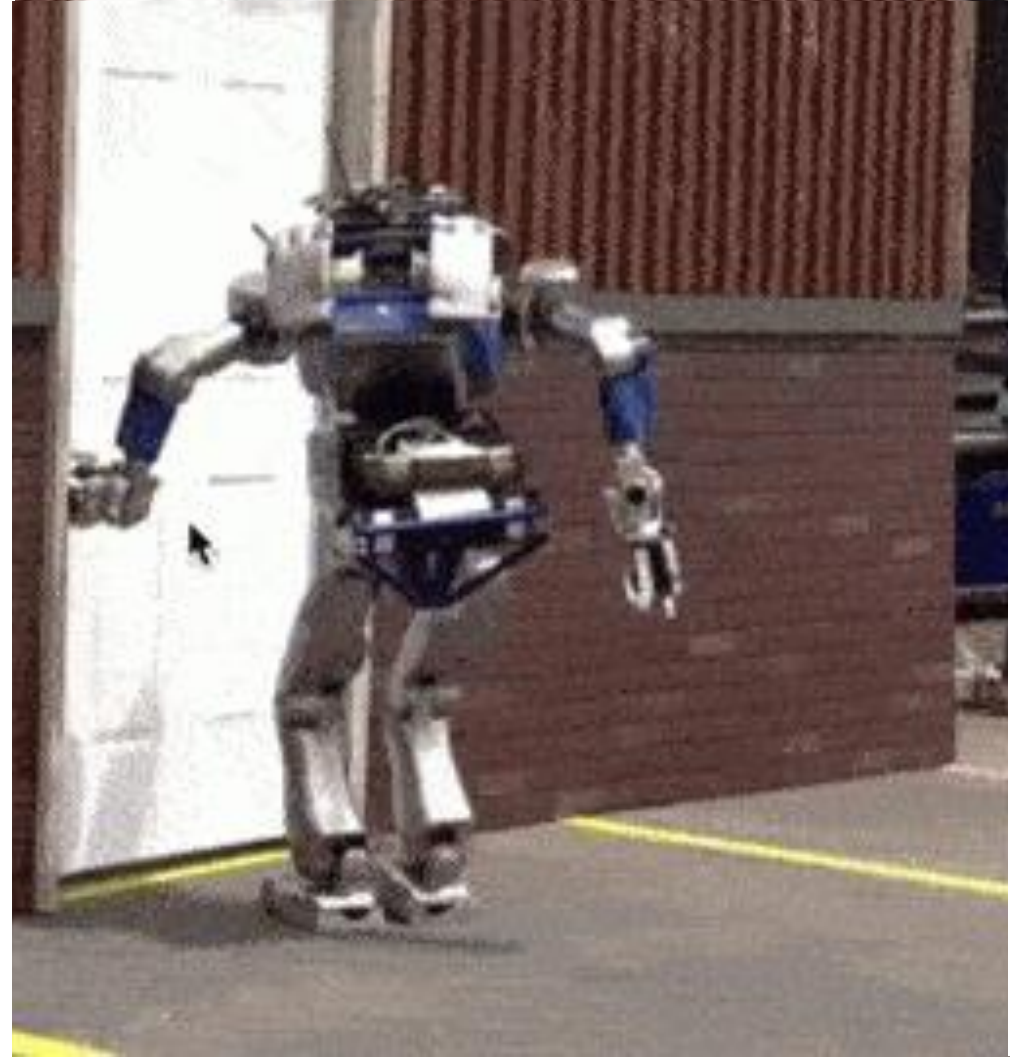
1,749 views | Dec 2, 2019, 7:20 am

## Maximize The Promise And Minimize The Perils Of Artificial Intelligence (AI)

**Grant Freeland** Contributor ⓘ  
Leadership Strategy

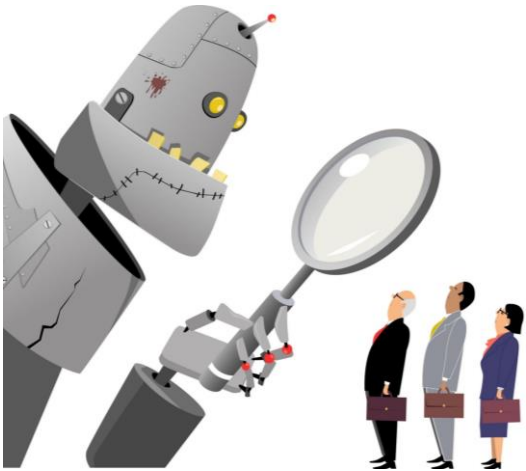


# Should We Worry about Today's A.I.?



# Should We Worry about Today's A.I.?

## Bias



Images:

<https://medium.com/@turalt/ai-isnt-biased-we-are-b74ec94d1698>

# AI Bias



Alexandra Chouldechova

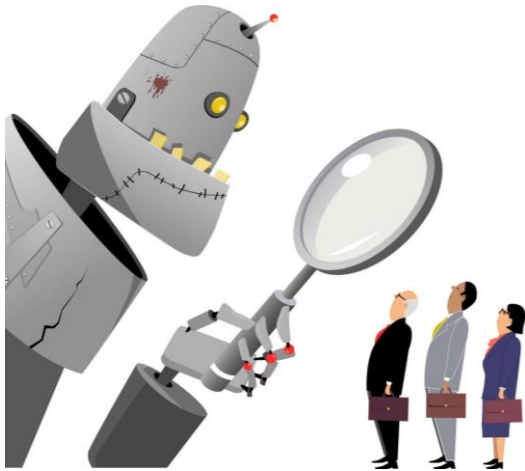
CMU, Statistics and Public Policy

<http://www.contrib.andrew.cmu.edu/~achoulde/>

<https://fatconference.org/>

# Should We Worry about Today's A.I.?

## Bias



## Weapons



Images:

<https://medium.com/@turalt/ai-isnt-biased-we-are-b74ec94d1698>

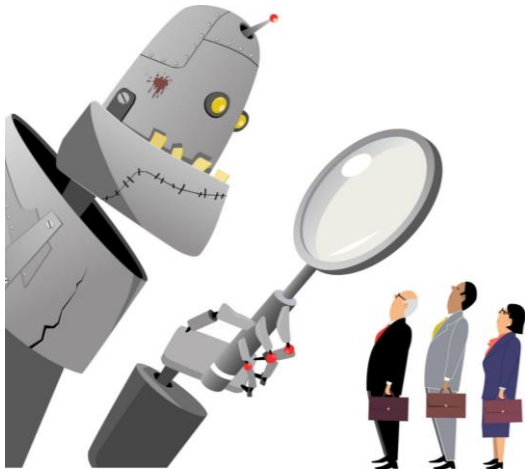
<http://futureoflife.org/2016/09/20/podcast-what-is-nuclear-risk/>

# AI Weapons



# Should We Worry about Today's A.I.?

## Bias



## Weapons



## Liability



Images:

<https://medium.com/@turalt/ai-isnt-biased-we-are-b74ec94d1698>

<http://futureoflife.org/2016/09/20/podcast-what-is-nuclear-risk/>

<https://electrek.co/2016/09/25/tesla-model-s-crashes-into-gym-driver-claims-autonomous-acceleration-tesla-says-drivers-fault/>

# Piazza Poll 3

## AI Explainability

Which of the following techniques have explainable results?

(SELECT ALL THAT APPLY)

- A. Search
- B. Logical inference
- C. Numeric optimization
- D. Q-learning
- E. Approximate Q-learning
- F. Exact inference Bayes nets
- G. Approximate inference Bayes nets
- H. Deep learning

# AI in the News

<https://undark.org/2019/12/04/black-box-artificial-intelligence/>

NEWS

## Unpacking the Black Box in Artificial Intelligence for Medicine

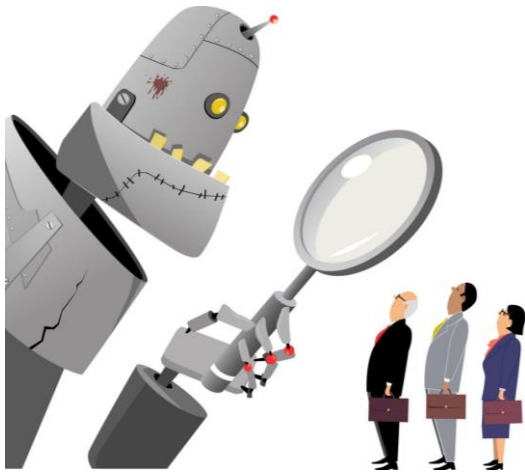
*Deep learning will radically change aspects of our medical care. How well do we need to understand how AI tools work?*

Visual: Yuichiro Chino / Getty Images



# Should We Worry about Today's A.I.?

## Bias



## Weapons



## Liability



## Jobs



Images:

<https://medium.com/@turalt/ai-isnt-biased-we-are-b74ec94d1698>

<http://futureoflife.org/2016/09/20/podcast-what-is-nuclear-risk/>

<https://electrek.co/2016/09/25/tesla-model-s-crashes-into-gym-driver-claims-autonomous-acceleration-tesla-says-drivers-fault/>

<http://ot.to/>

# AI in the News

<https://www.supplychaindigital.com/technology/ups-invests-autonomous-driving-firm-tusimple>

SUPPLY CHAIN

LOGISTICS

TECHNOLOGY

## UPS invests in autonomous driving firm TuSimple

By SEAN GALEA-PACE • Aug 15, 2019, 10:20AM



## Piazza Poll 4

Is it ok if autonomous vehicles completely replace human drivers?

# Piazza Poll 4

Is it ok if autonomous vehicles completely replace human drivers?

## AI in the News

- <https://www.brookings.edu/research/what-jobs-are-affected-by-ai-better-paid-better-educated-workers-face-the-most-exposure/>
- <https://www.vox.com/platform/amp/policy-and-politics/2019/12/3/20965464/2020-presidential-candidates-jobs-automation-ai>
- <https://www.irishtimes.com/business/technology/short-window-to-stop-ai-taking-control-of-society-warns-ex-google-employee-1.4104535>

# AI Challenge: Humans

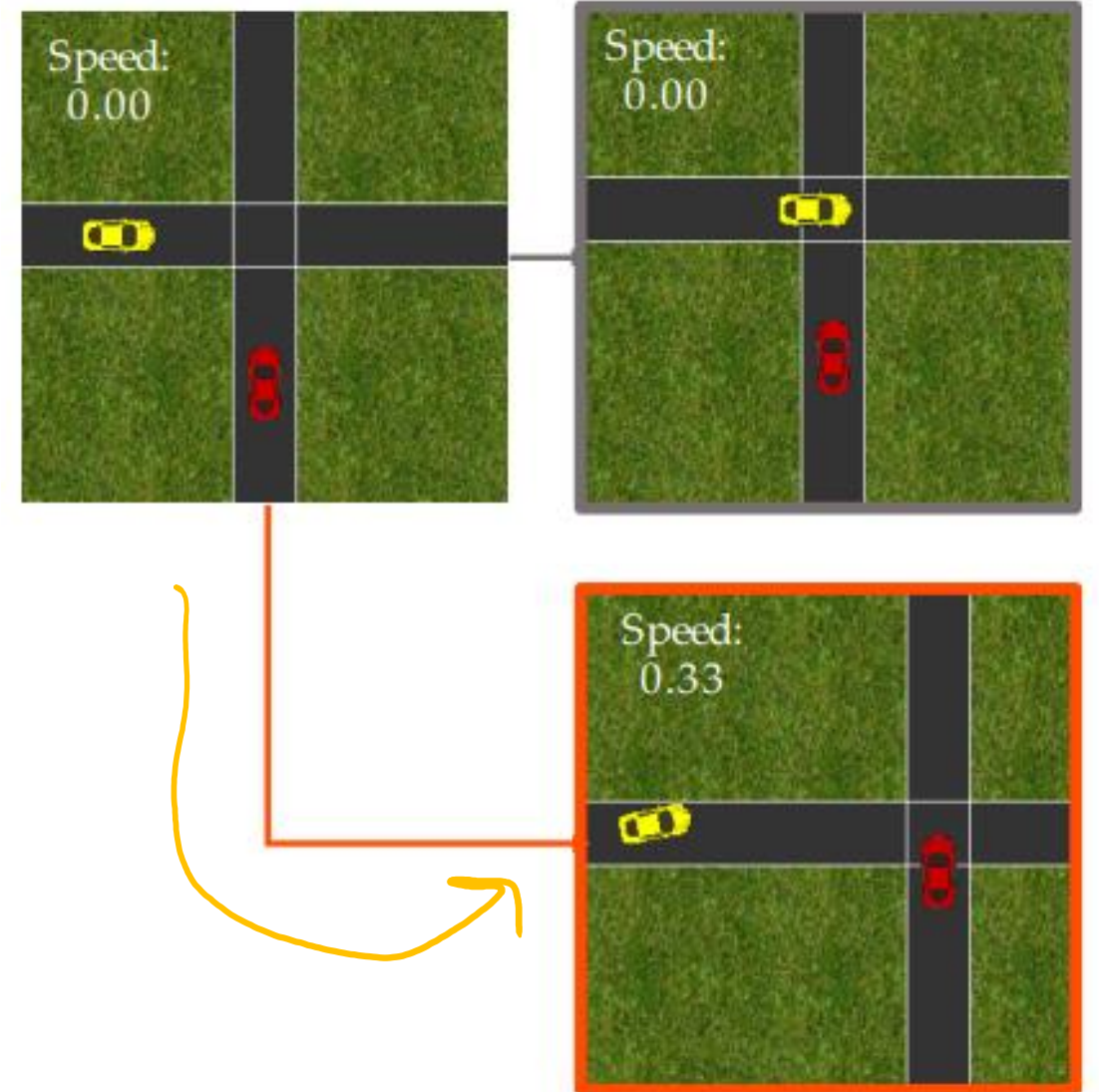
Handing humans a drink



Anca Dragan  
UC Berkeley, EECS  
CMU PhD

# AI Challenge: Humans

## Driving with humans



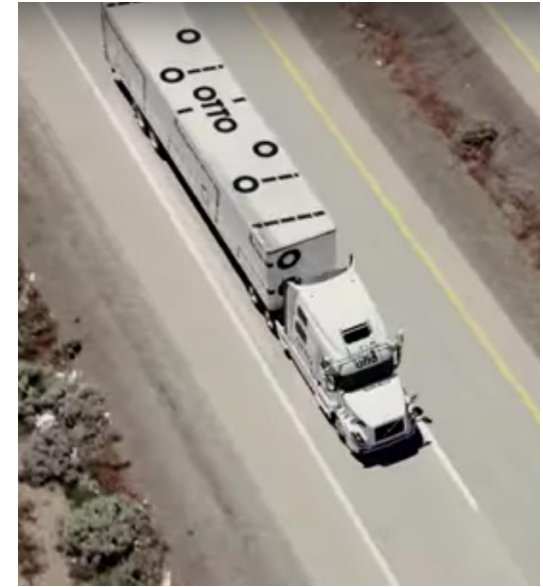
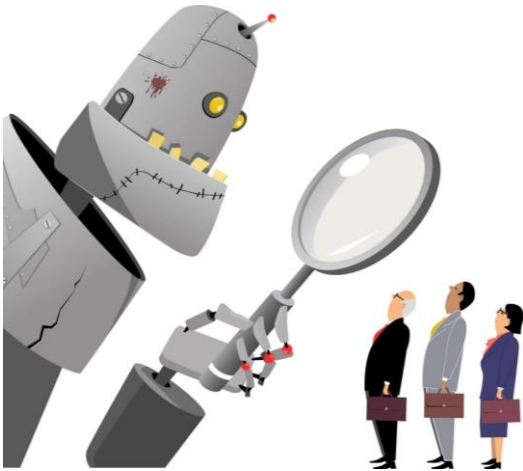
Dorsa Sadigh  
Stanford, CS

<https://stanford-iliad.github.io/pdfs/publications/sadigh2016planning.pdf>

## Piazza Poll 5

Once autonomous vehicles are readily available, should it be illegal for humans to drive?

# Narrow A.I.



Images:

<https://medium.com/@turalt/ai-isnt-biased-we-are-b74ec94d1698>

<http://futureoflife.org/2016/09/20/podcast-what-is-nuclear-risk/>

<https://electrek.co/2016/09/25/tesla-model-s-crashes-into-gym-driver-claims-autonomous-acceleration-tesla-says-drivers-fault/>

<http://ot.to/>

AI in the News

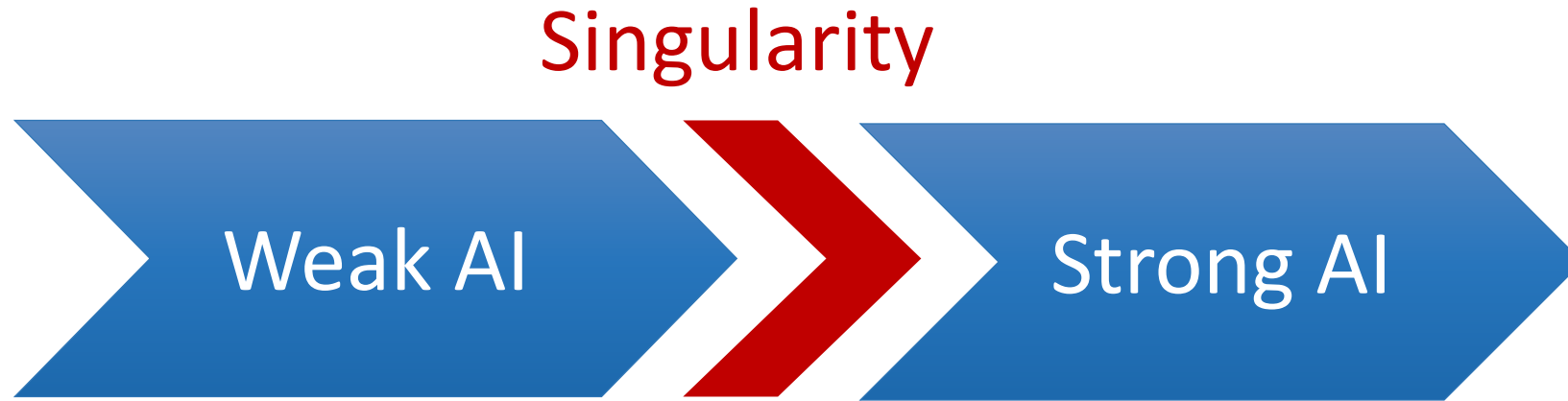
Nov. 5, 2017

**The New York Times**

**Building A.I. That Can Build A.I.**

<https://mobile.nytimes.com/2017/11/05/technology/machine-learning-artificial-intelligence-ai.html>

# Should we worry about future A.I.?



- Narrow AI
- Limited number of applications

- Artificial General Intelligence (AGI)
- Recursive self-improvement
- Beyond human control

# Should we worry about future A.I.?

What motivates agents?

Candy grab

Ana: "taking 2 makes it 8"

Bob: "taking 1 makes it 7"

Ana: "taking 2 makes it 5"

Bob: "taking 2 makes it 3"

Ana: "taking 1 makes it 2"

Bob: "taking 2 makes it 0"

I WIN!



# Should we worry about future A.I.?

Question: What is the specific motivation behind these techniques?

- Search
- Logical inference
- Linear programming
- RL
- Inference Bayes nets

# Should we worry about future A.I.?

Question: What motivation could cause problems?

# Should we worry about future A.I.?

Stuart Russell, UC Berkeley [Center for Human-Compatible AI](#)



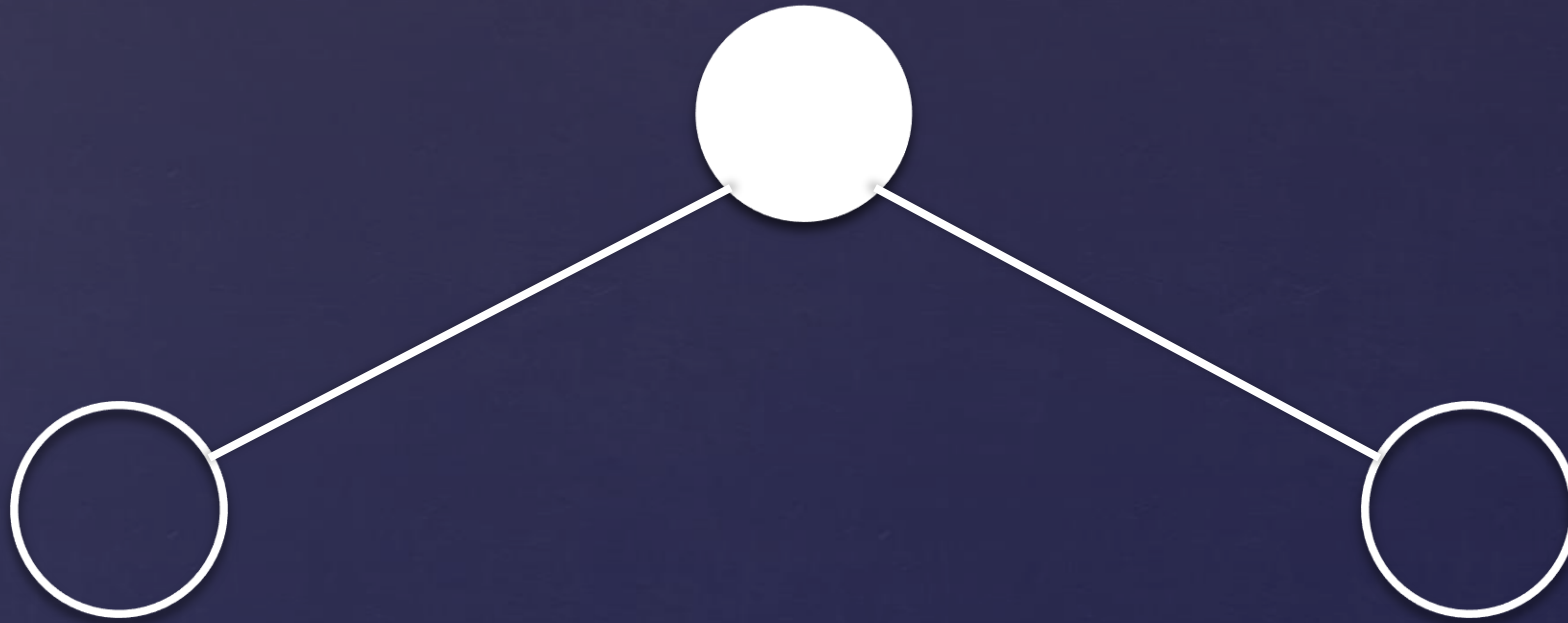
[https://www.ted.com/talks/stuart\\_russell\\_how\\_ai\\_might\\_make\\_us\\_better\\_people](https://www.ted.com/talks/stuart_russell_how_ai_might_make_us_better_people)

# Three simple ideas

1. The robot's only objective is to maximize the realization of human values
2. The robot is initially uncertain about what those values are
3. The best source of information about human values is human behavior

# AIMA 1,2,3: objective given to machine

Human objective

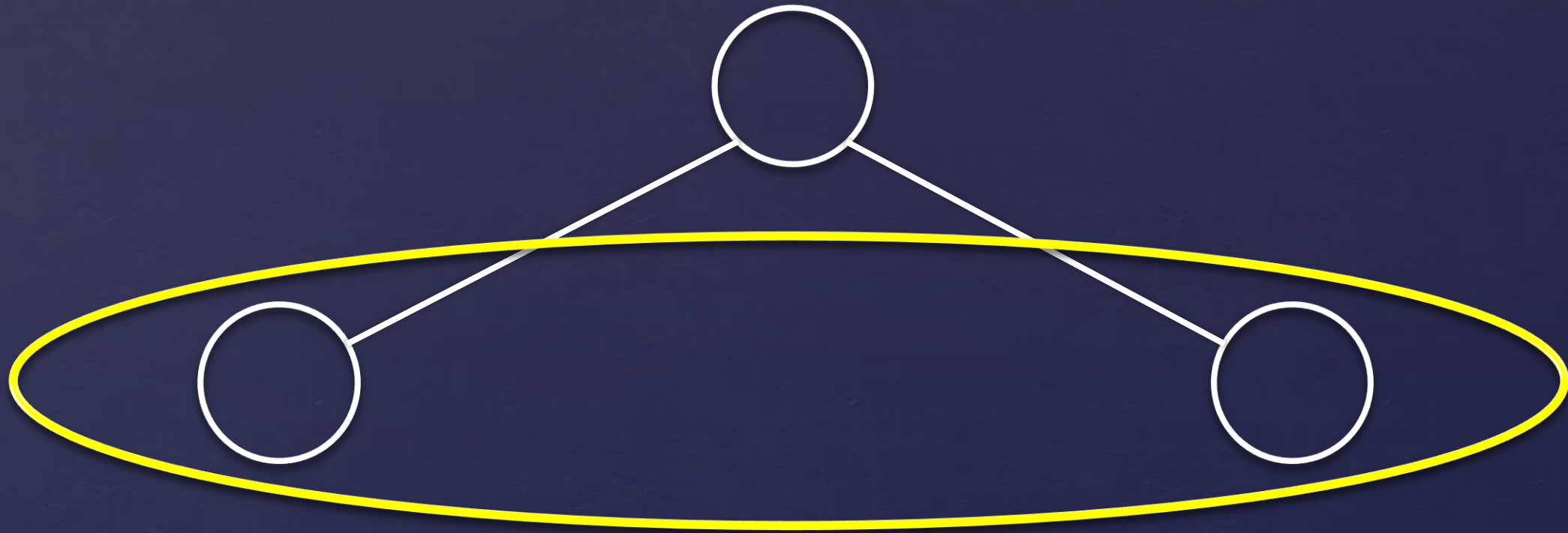


Human behaviour

Machine behaviour

# AIMA 4: objective is a latent variable

Human objective



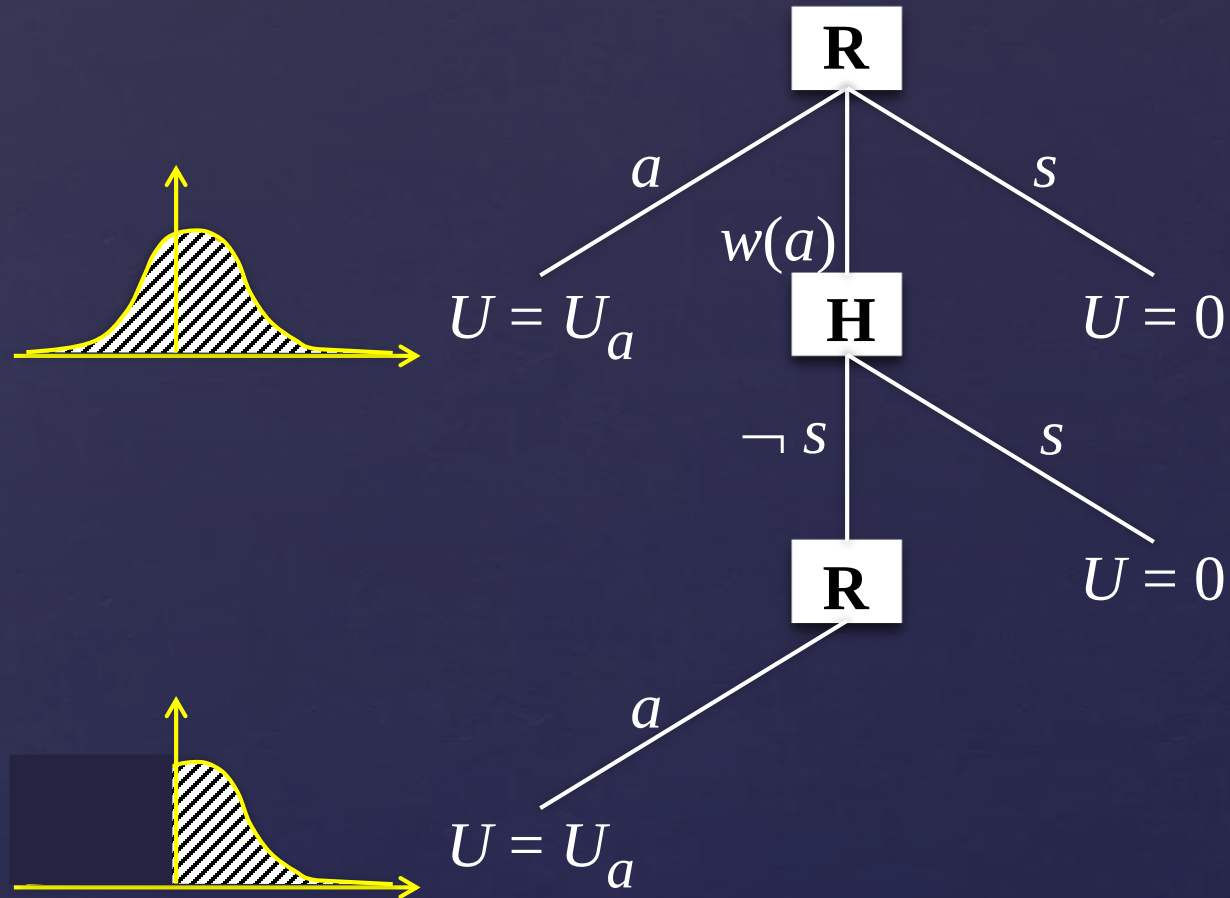
Human behaviour

Machine behaviour

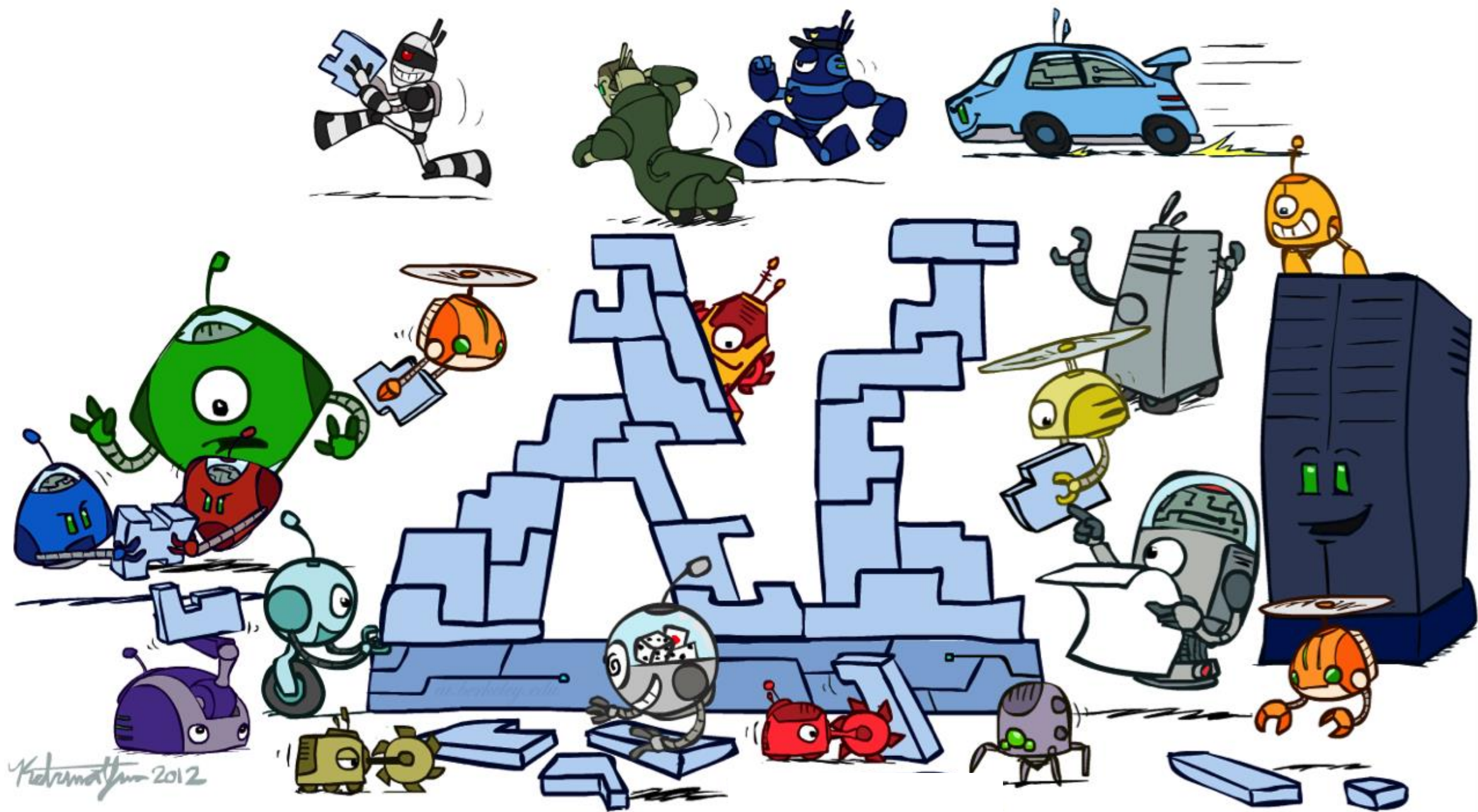
# The off-switch problem

- ❖ A robot, given an objective, has an incentive to disable its own off-switch  
(You can't fetch the coffee if you're dead)
- ❖ How can we prevent this?
- ❖ Answer: robot must allow for *uncertainty* about the true human objective
  - ❖ The human will only switch off the robot if that leads to better outcomes for the true human objective
  - ❖ Theorem: it's *in the robot's interest* to allow it
  - ❖ Theorem: Such a robot is *provably beneficial*

# Off-switch model



$w(a)$  preferred to  $a$  or  $s$



# Thanks to Our Course Staff!!

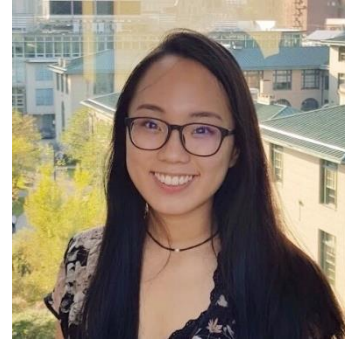
Administrative  
Assistant

Liv Zane

Videographer

John Lombardo

Teaching Assistants



Angela  
Yang



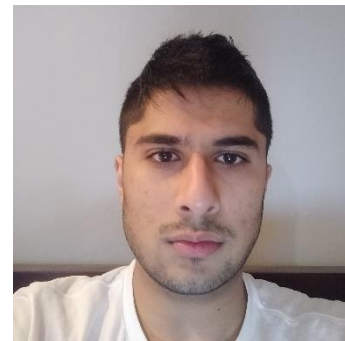
Claire Wang  
(Head TA)



George  
Brown



Michelle  
Ma



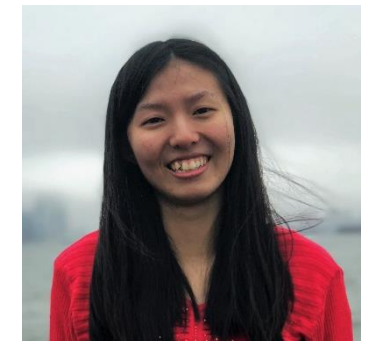
Sean  
Pereira



Chakara (Tian)  
Owarang



Tina  
Wu



Vicky  
Zeng