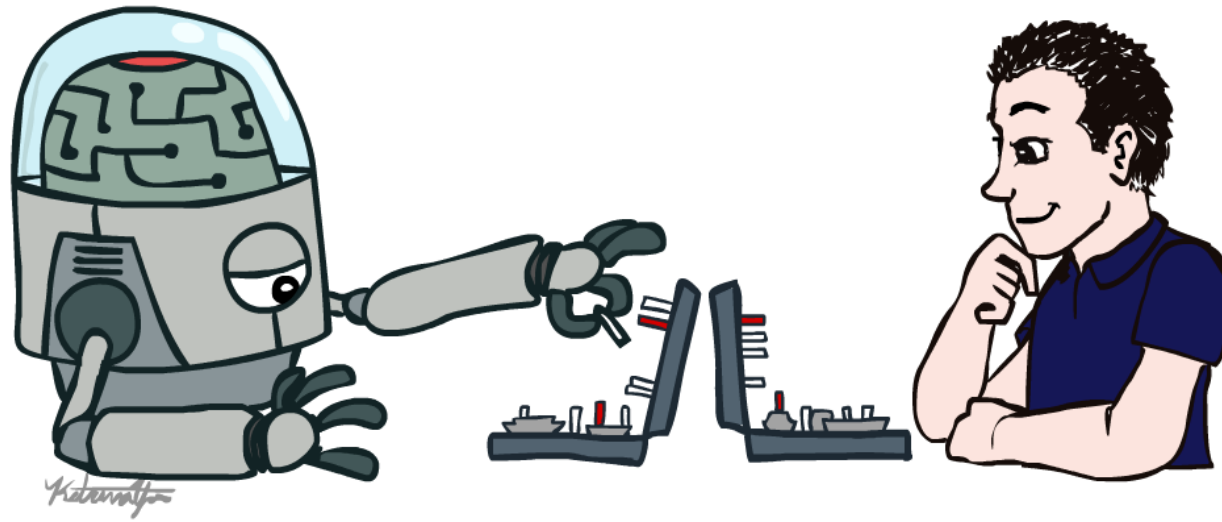


AI: Representation and Problem Solving

AI and Ethics



Instructor: Pat Virtue

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

Poll 1: What questions do you have?

Group Discussions

Breakout rooms will have AI Ethics themes today. Start thinking about which discussion category you would like to participate in today:

Bias/Fairness

Weapons/Safety

Privacy

Liability

Jobs

Intelligent Agents

From first lecture

Nim 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
```

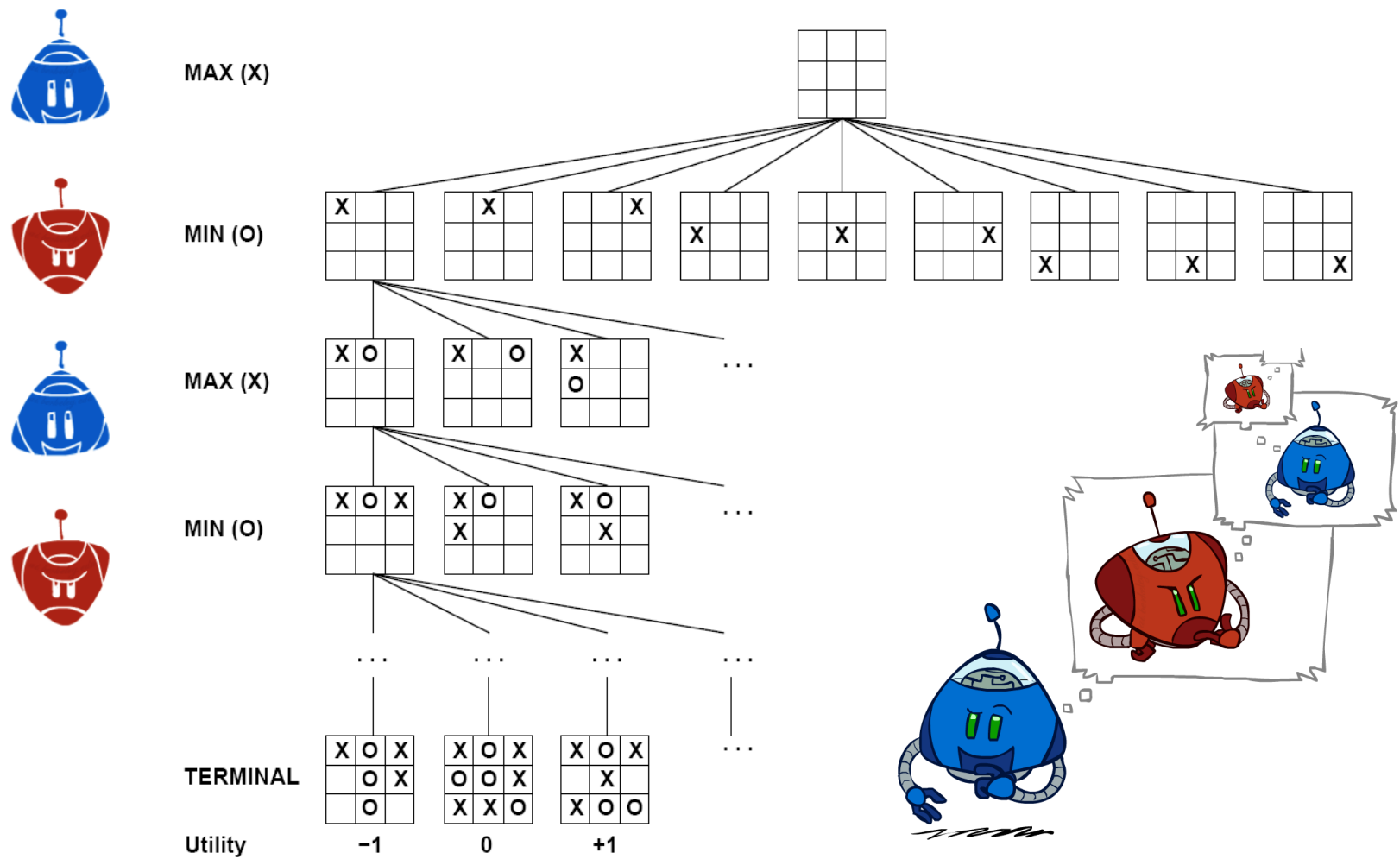
Poll 1

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%

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

Poll 2

Which 281 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

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

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

TA survey: <https://www.ugrad.cs.cmu.edu/ta/S20/feedback/>



Harlene
(Head TA)



Simran
(Head TA)



Adabelle



Ria



Jimin



Ihita



Adrian



Elizabeth

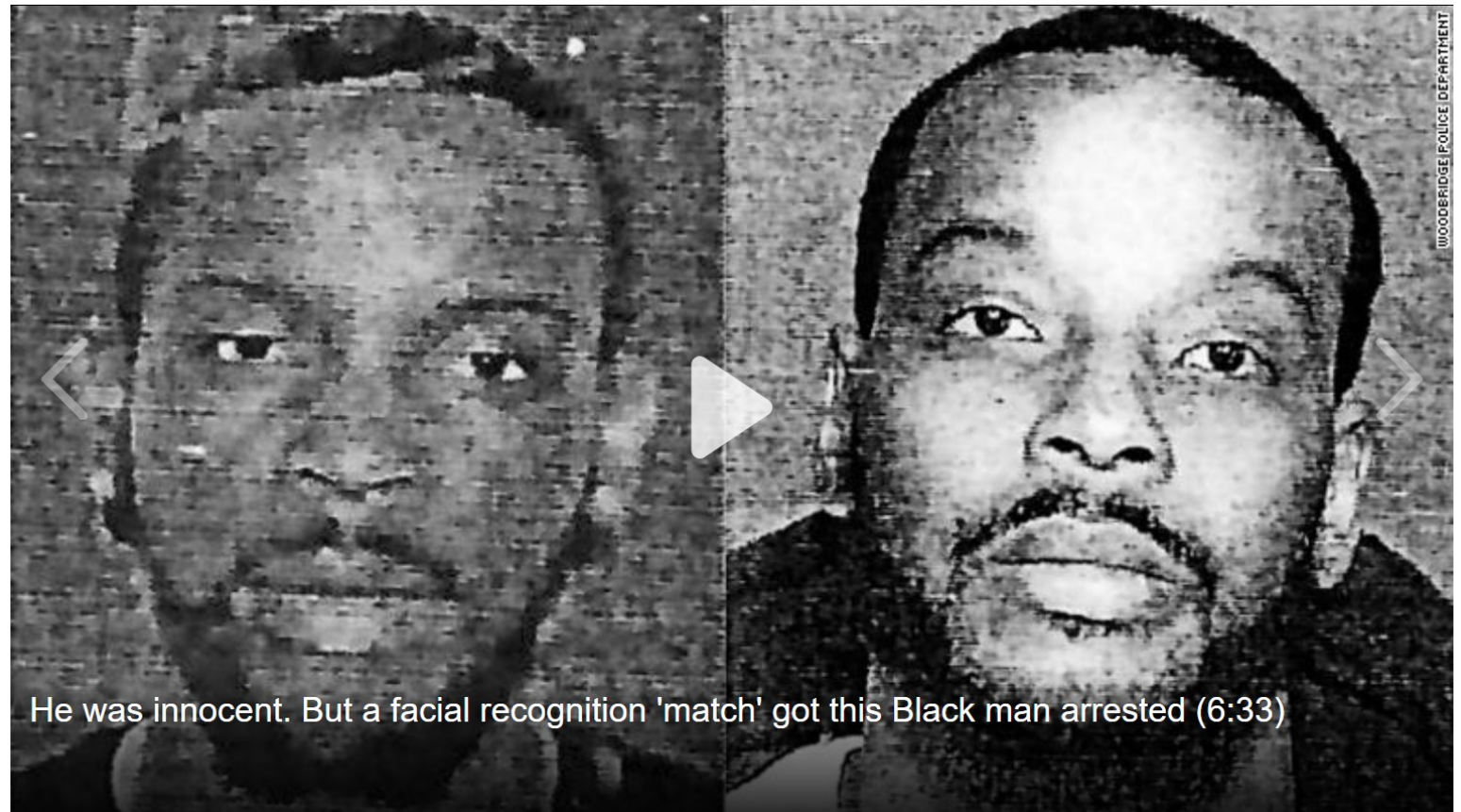
AI in the News

AI in the News

Anyone can use this powerful facial-recognition tool – and that's a problem

By Rachel Metz, CNN Business

Updated 3:53 PM EDT, Tue May 04, 2021



<https://amp.cnn.com/cnn/2021/05/04/tech/pimeyes-facial-recognition/index.html>

AI in the News

The New York Times

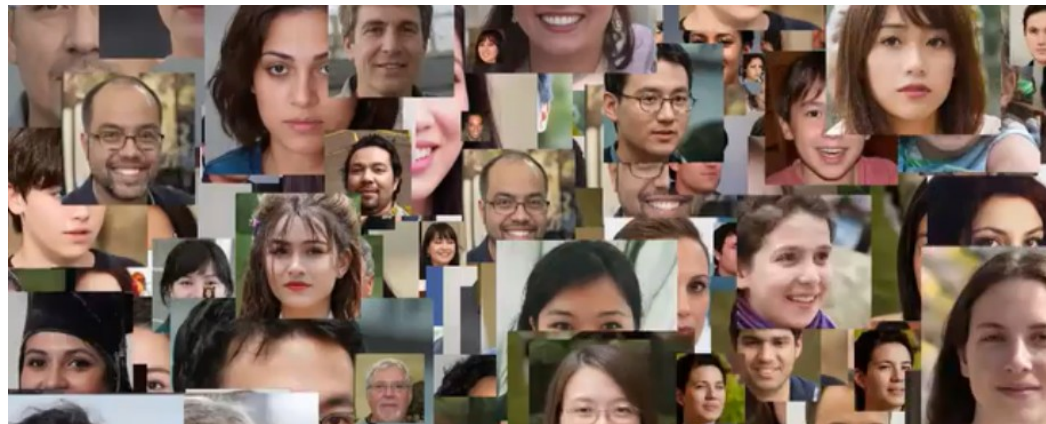
The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and “might lead to a dystopian future or something,” a backer says.



By Kashmir Hill

Published Jan. 18, 2020 Updated March 18, 2021



<https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>

AI in the News



Self-driving cars

US automakers outline rules for auto-driving cars after fatal crashes

Proposals come days after two men in a Tesla were killed in a crash near Houston

Edward Helmore

Wed 28 Apr 2021 12.49 EDT

<https://amp.theguardian.com/business/2021/apr/28/us-automakers-rules-auto-driving-cars-fatal-crashes>

AI in the News

Bloomberg



Hyperdrive

The Race to Build Self-Driving Trucks Has Four Horses and Three Jockeys

These are the companies set to dominate the highways of tomorrow.

By [Ira Boudway](#).

May 1, 2021, 5:30 AM EDT

Corrected May 4, 2021, 9:31 AM EDT

<https://www.bloomberg.com/news/articles/2021-05-01/waymo-tusimple-aurora-inside-the-race-to-build-self-driving-trucks>

AI in the News

BBC NEWS

What will self-driving trucks mean for truck drivers?

By Bernd Debusmann Jr
Business reporter

9 April 2021

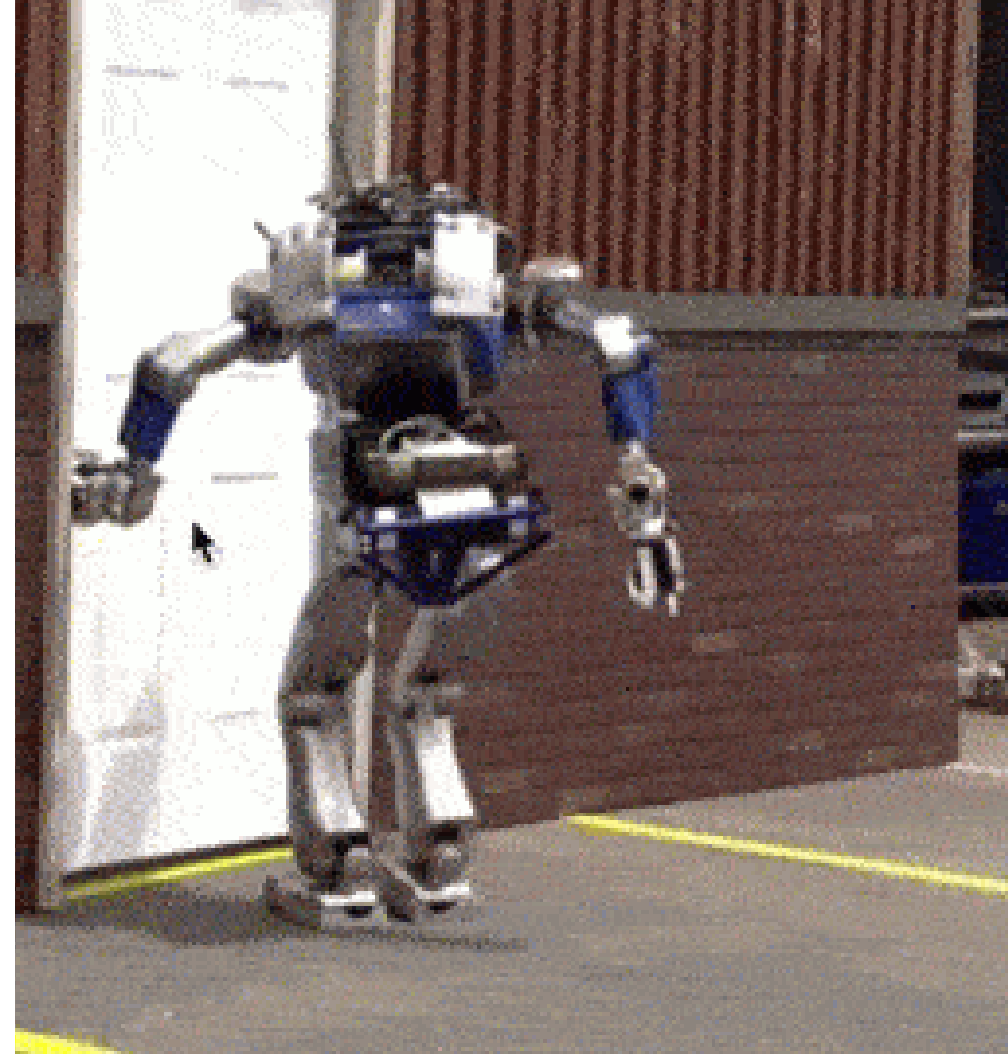


<https://www.bbc.com/news/business-56332388.amp>

Poll 4

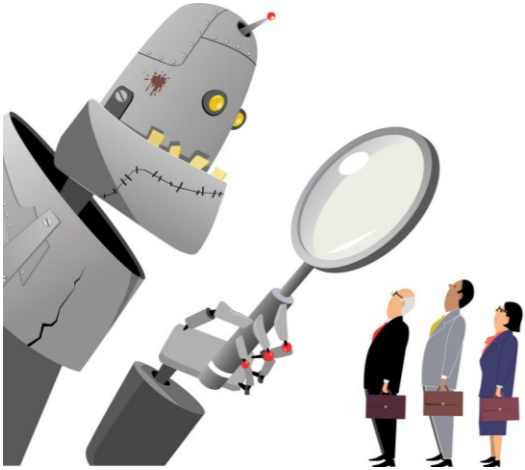
Is it ok if autonomous vehicles completely replace human drivers?

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



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

Bias/Fairness



Privacy



Jobs



Weapons/Safety



Liability



Bias/Fairness

How are AI systems motivated? How does this motivation cause bias?

How could we measure bias (including cultural bias) in an AI system?

How do we as programmers introduce bias into the AI techniques we've learned about in 15-281?

What are some ways that bias in AI programming can impact the world?

Domains to think about:

- Self-driving cars

- Hiring and judicial decision making

- Traffic control

- Face recognition

Weapons/Safety

Should AI be used as a weapon or in warfare?

Where would you draw the line for how it could be used and how not?

Would you let a fully autonomous robot perform surgery on a family member?

Can you think of some rules or laws that may help prevent unsafe AI?

Domains to think about:

- Social media bots and information warfare

- Drones and autonomous weapons machines

- Healthcare

Privacy

What are some tradeoffs we make when giving AI our private information?

Do you have a limit for the type or amount of data you are willing to provide?

How can we make the data usage of AI systems more explicit?

Are there ways that we can still enjoy the benefits of AI systems without giving up our private information?

Domains to think about:

Social media

Search (Google, Amazon, Yelp, ...)

Healthcare

Insurance

Liability

How do we determine what went wrong (or right) in deployed AI systems?

Which techniques that we learned about have explainable results?

Who is responsible for “mistakes” that AI systems make?

What if someone or another AI misuses AI technology? Who is responsible?

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

Domains to think about:

- Self-driving cars

- Healthcare decision-making

- Land and Air traffic control

Jobs

What types of jobs should AI do instead of people?

Is it ok if autonomous vehicles completely replace human drivers?

What will people do if their jobs are taken?

How can AI systems improve or change people's jobs even if it doesn't take the jobs?

Domains to think about:

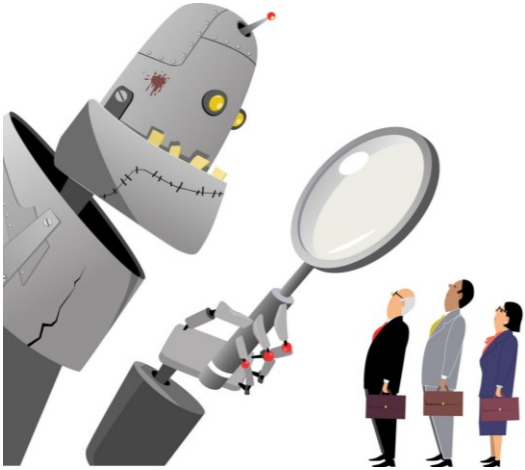
- Driving and Logistics

- Healthcare decision-making

- Entertainment

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

Bias/Fairness



Privacy



Jobs



Weapons/Safety



Liability



Additional Resources and Ideas

AI Bias/Fairness

ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)

A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

Alexandra Chouldechova

CMU, Statistics and Public Policy

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

<https://facctconference.org/>

AI Bias/Fairness

Hoda Heidari

MLD

Jason Hong

HCII

Graham Neubig

LTI



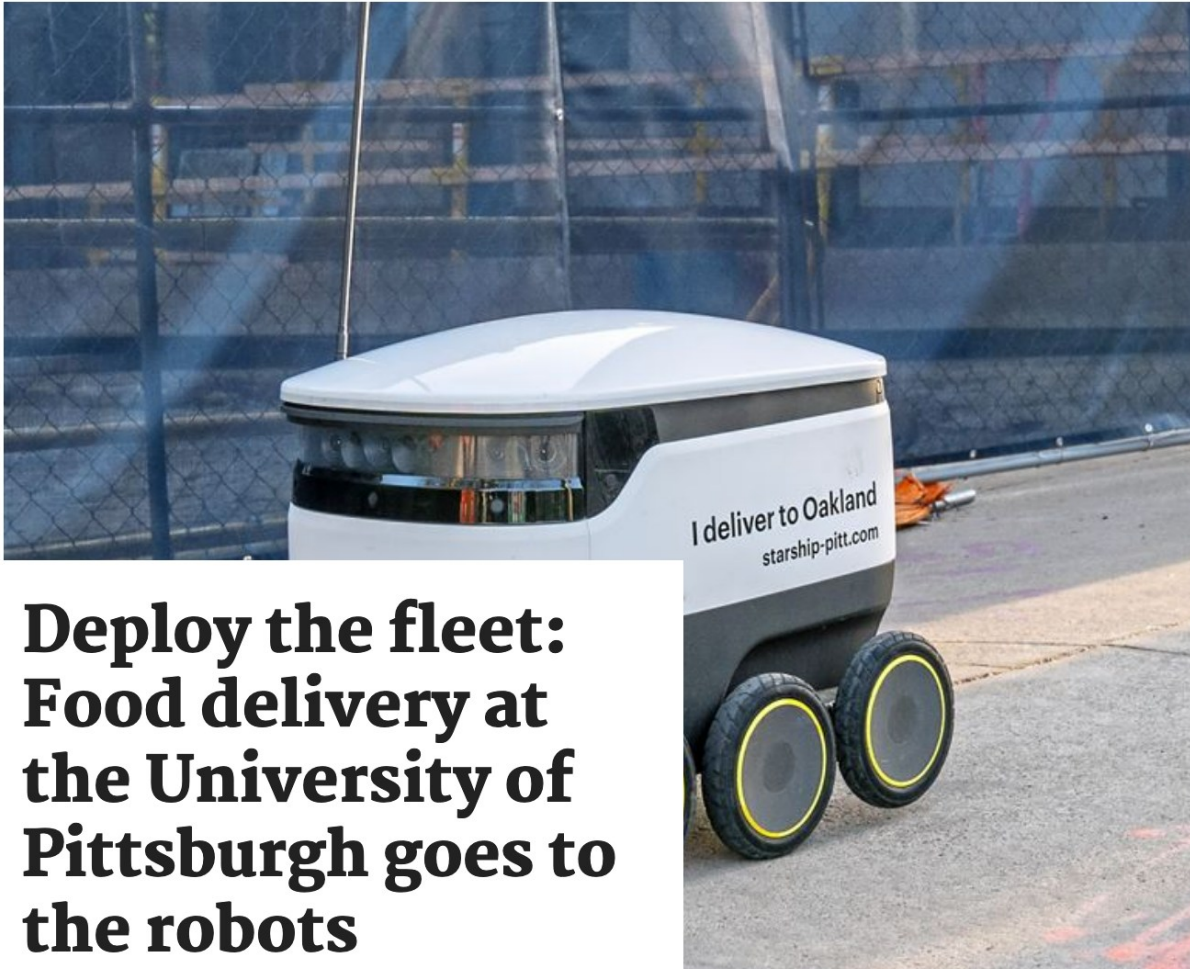
February 16, 2021

CMU Researchers Win NSF-Amaزون
Fairness in AI Awards

<https://www.ml.cmu.edu/news/news-archive/2021-2025/2021/february/carnegie-mellon-researchers-win-nsf-amazon-fairness-in-ai-awards.html>

AI Bias/Fairness

Question: What technique could have been used to learn from the mistake?



Food Delivery Robots Pulled From Pitt Campus After Backlash About Mobility

By KATHLEEN J. DAVIS • OCT 22, 2019

[Share](#) [Tweet](#) [Email](#)



AI Bias/Fairness

5 Examples Of How AI Can Be Used Across The Supply Chain

Forbes



Blake Morgan Senior Contributor ©

CMO Network

I am a Customer Experience Futurist, Author and Keynote Speaker.



Search/Navigation:

Create efficient routes for the fleet of trucks

Use of robots to deliver medicine, groceries etc

CSPs:

Deciding which stores get preference for deliveries

Deciding what products go to what stores

AI Bias/Fairness

AI could be the key to ending discrimination in hiring, but experts warn it can be just as biased as humans

Aaron Holmes Oct 8, 2019, 12:14 PM



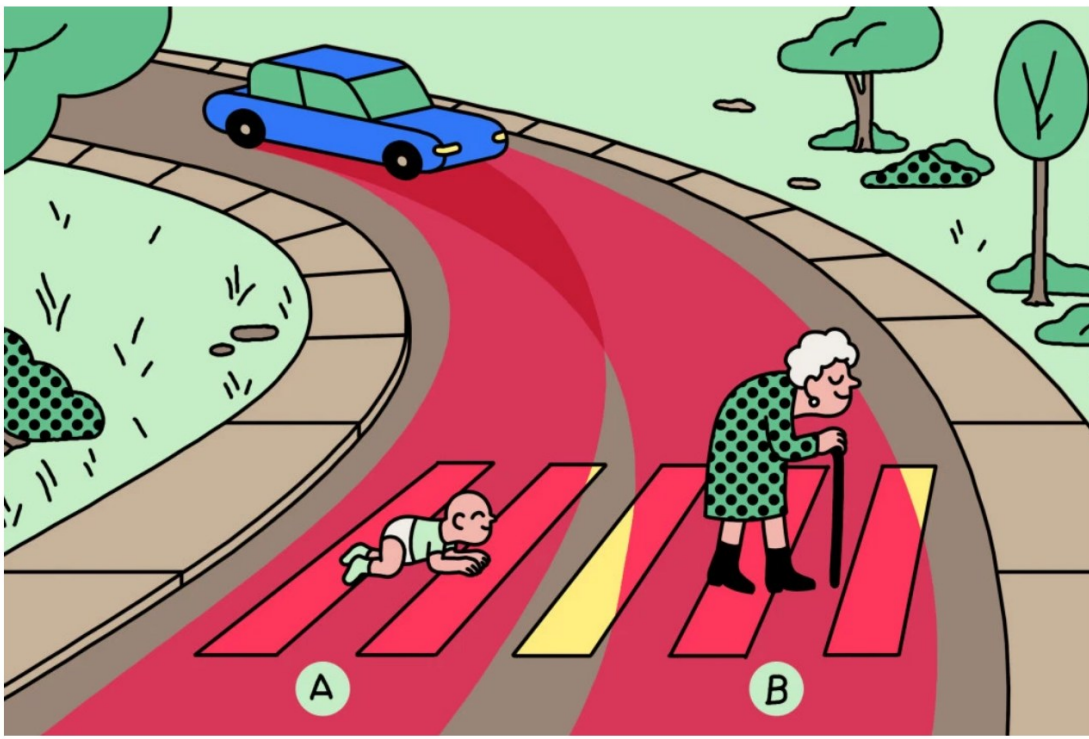
Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use

Asian and African American people were up to 100 times more likely to be misidentified than white men, depending on the particular algorithm and type of search. Native Americans had the highest false-positive rate of all ethnicities, according to the study, which found that systems varied widely in their accuracy.

The faces of African American women were falsely identified more often in the kinds of searches used by police investigators where an image is compared to thousands or millions of others in hopes of identifying a suspect.

Algorithms developed in the United States also showed high error rates for “one-to-one” searches of Asians, African Americans, Native Americans and Pacific Islanders. Such searches are critical to functions including cellphone sign-ons and airport boarding schemes, and errors could make it easier for impostors to gain access to those systems.

<https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/>



SIMON LANDREIN

Tech policy / AI Ethics

Should a self-driving car kill the baby or the grandma? Depends on where you're from.

The infamous "trolley problem" was put to millions of people in a global study, revealing how much ethics diverge across cultures.

TECHNOLOGY

The Ethics of Autonomous Cars

Sometimes good judgment can compel us to act illegally. Should a self-driving vehicle get to make that same decision?

PATRICK LIN OCTOBER 8, 2013

NEWS · 24 OCTOBER 2018

Self-driving car dilemmas reveal that moral choices are not universal

Survey maps global variations in ethics for programming autonomous vehicles.

<https://www.moralmachine.net/>

<https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/>

<https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/>

AI Weapons/Safety



<https://www.youtube.com/watch?v=9CO6M2HsoIA>

AI Weapons/Safety

Too Perilous For AI? EU Proposes Risk-Based Rules

Draft regulations splits AI applications into risk-based tiers and bans some

By **Lucas Laursen**



Illustration: iStockphoto

<https://spectrum.ieee.org/tech-talk/artificial-intelligence/embedded-ai/euairules>

AI Weapons/Safety

Wharton
UNIVERSITY of PENNSYLVANIA

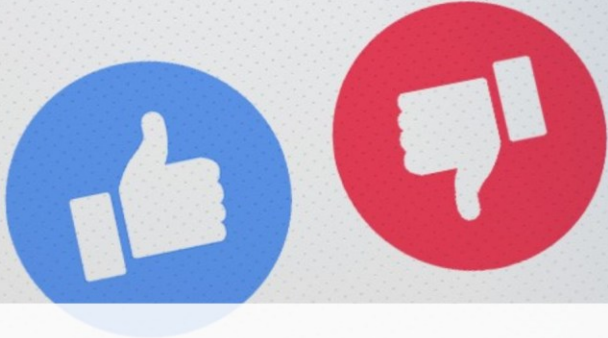
The K@W Network: English | 简体中文 | 繁體中文 | Español | Português | KWHS

Log In | Register | f | t | in | e

KNOWLEDGE @ WHARTON

Search

TOPICS ▼ REGIONS ▼ RESEARCH PODCASTS MORE ▼ ABOUT ▼



TECHNOLOGY

Why Social Media Is the New Weapon in Modern Warfare

Jan 17, 2019

Global Focus Books, Business Radio

8,353 views | Aug 1, 2019, 02:49am EDT

Cyber Warfare: Army Deploys 'Social Media Warfare' Division To Fight Russia



Zak Doffman Contributor

[Cybersecurity](#)

I write about security and surveillance.



AI Privacy



24,059 views | Mar 27, 2019, 01:16pm EDT

Rethinking Privacy For The AI Era

Insights Team Insights Contributor
Forbes insights **FORBES INSIGHTS** With **Intel AI** | Paid Program
Innovation

[f](#) [t](#) [in](#)

Concerns over consumer privacy have peaked in recent years—roughly in step with the rise of advanced technologies like artificial intelligence. About 9 in 10 American internet users say they are concerned about the privacy and security of their personal information online, and 67% are now advocating for strict national privacy laws, according to a [study by](#) Intouch International.

Fed up by a steady stream of incidents that range from the 2017 [Equifax hack](#) to the nefarious gaming of consumers' social media data for political purposes, policymakers have begun to strike back on consumers' behalf.

<https://www.forbes.com/sites/insights-intelai/2019/03/27/rethinking-privacy-for-the-ai-era>

AI Liability

<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



AI Jobs

<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



AI Jobs

Is it ok if autonomous vehicles completely replace human drivers?

- <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 Jobs

Should Robots or People Do These Jobs?

A Survey of Robotics Experts and Non-Experts About Which Jobs Robots Should Do

Wendy Ju & Leila Takayama
Willow Garage
68 Willow Road
Menlo Park, California, USA 94025
[wendyju, takayama] @ willowgarage.com

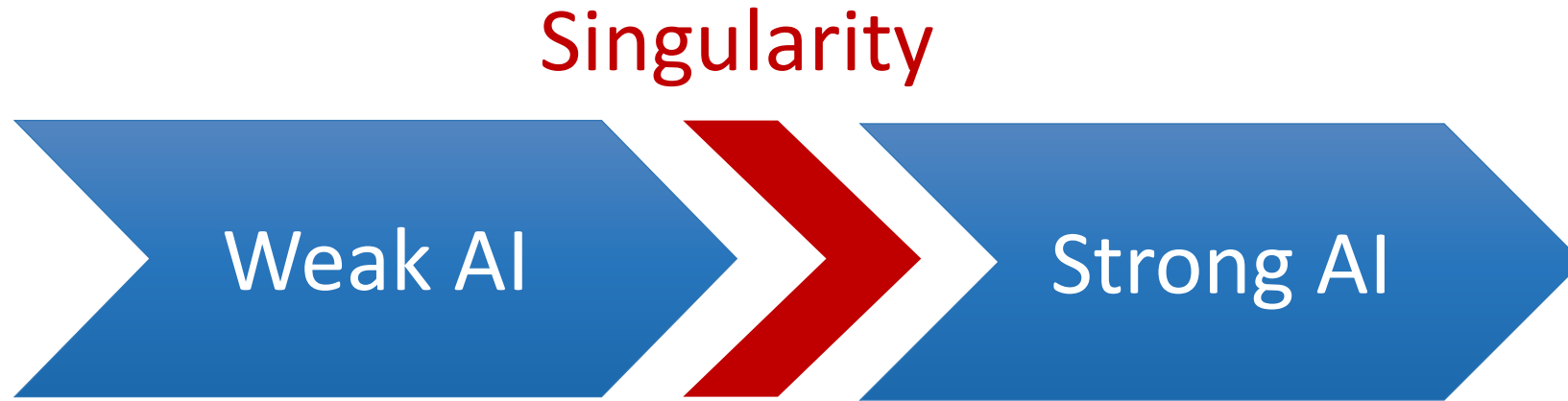
http://www.wendyju.com/publications/ju_iros11.pdf

Experts SHOULD	Guiding, Directing, and Motivating Subordinates	.23***
	Speech Clarity (Importance)	.17*
	Thinking Creatively (Importance)	.17**
	Category Flexibility (Importance)	.16**
	Dependability	.14**
	Building and Construction Knowledge (Importance)	.08*
	Interpreting the Meaning of Information (Importance)	-.31***
	Cramped Workspace and Awkward Positions	-.17***
	Negotiation (Level)	-.14*
	Multi-limb Coordination (Level)	-.13*
	Initiative	-.13**
	Artistic Interests	-.11**

* $p < .05$, ** $p < .01$, *** $p < .001$. Shaded rows indicate factors where robots are preferred over people; in unshaded rows, people are preferred.

What about the Future?

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.?

Question: What is the specific motivation behind these techniques?

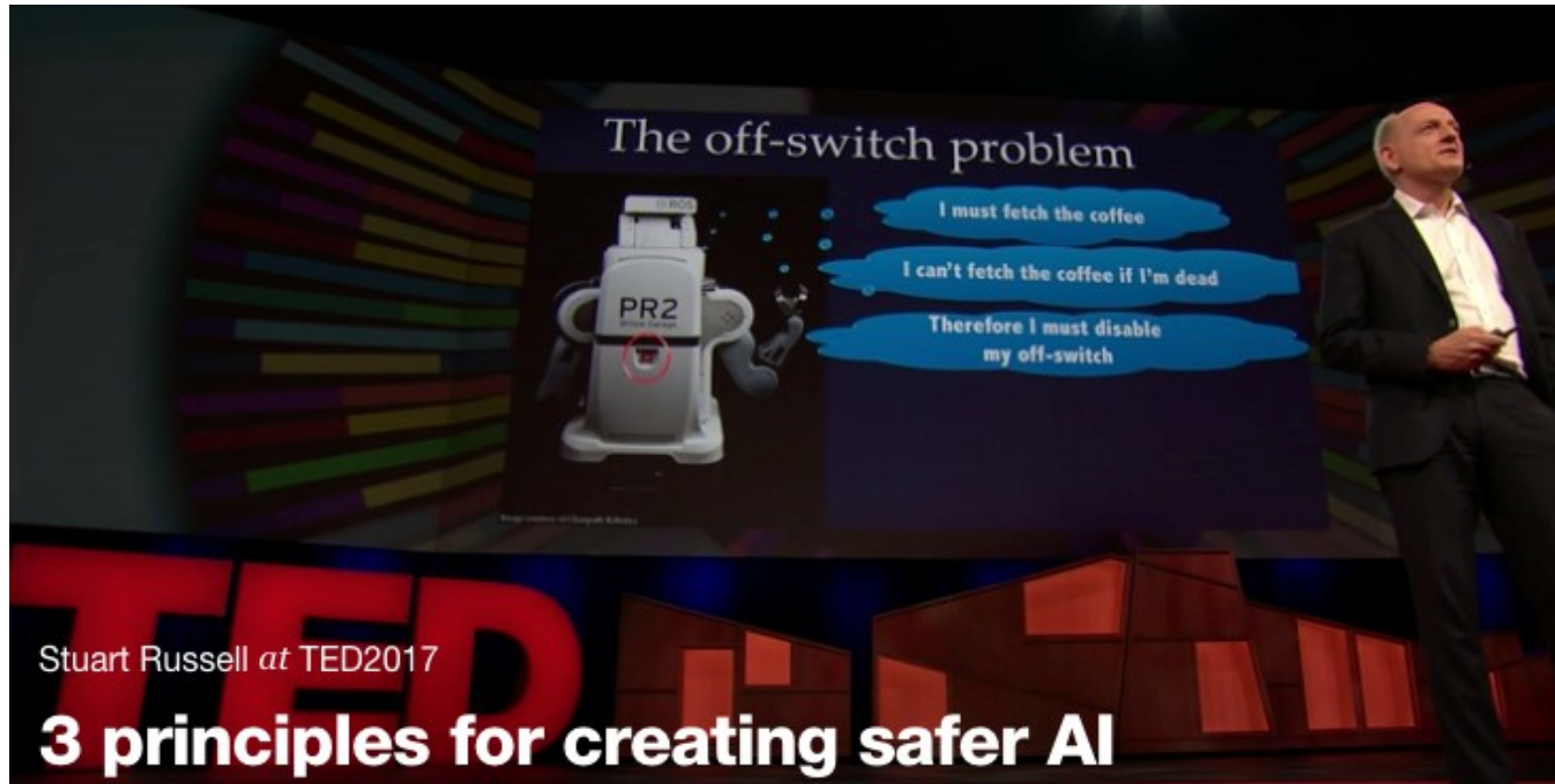
- Search
- Logical inference
- Linear programming
- RL
- Inference Bayes nets

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](#)



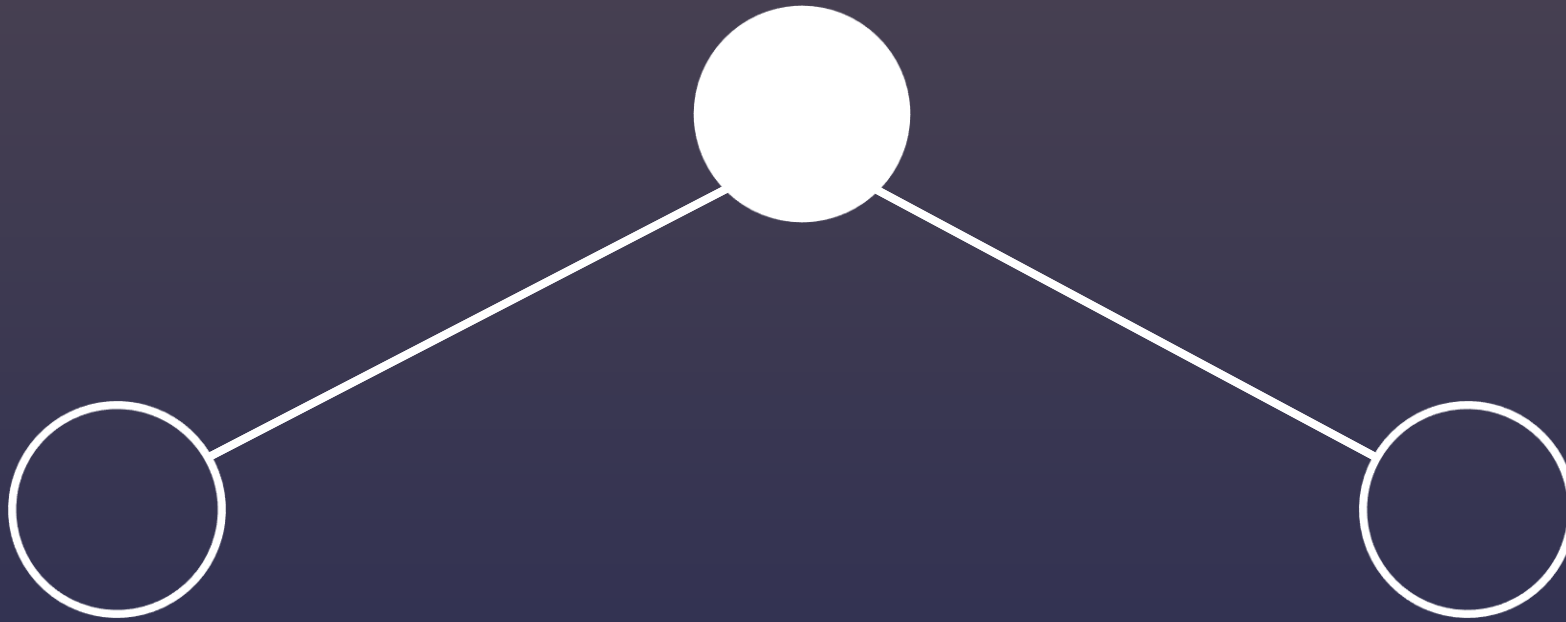
https://www.ted.com/talks/stuart_russell_how_ai_might_make_us_better_people

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

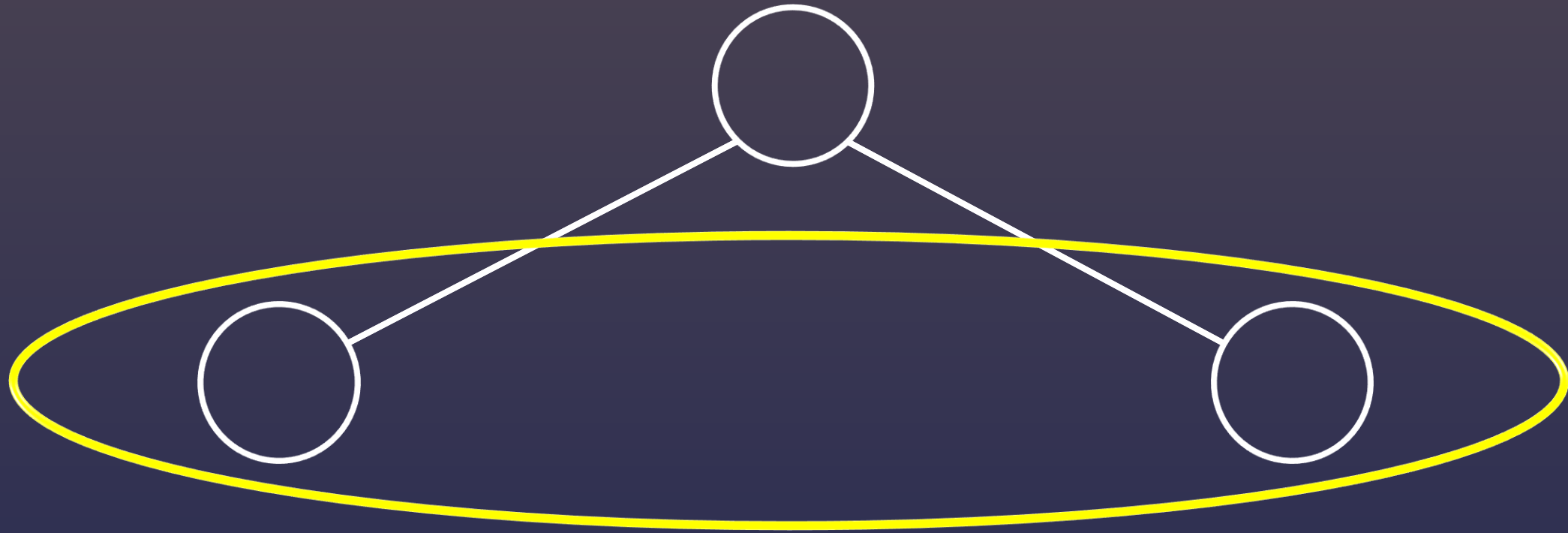


Human behaviour

Machine behaviour

AIMA 4: objective is a latent variable

Human objective



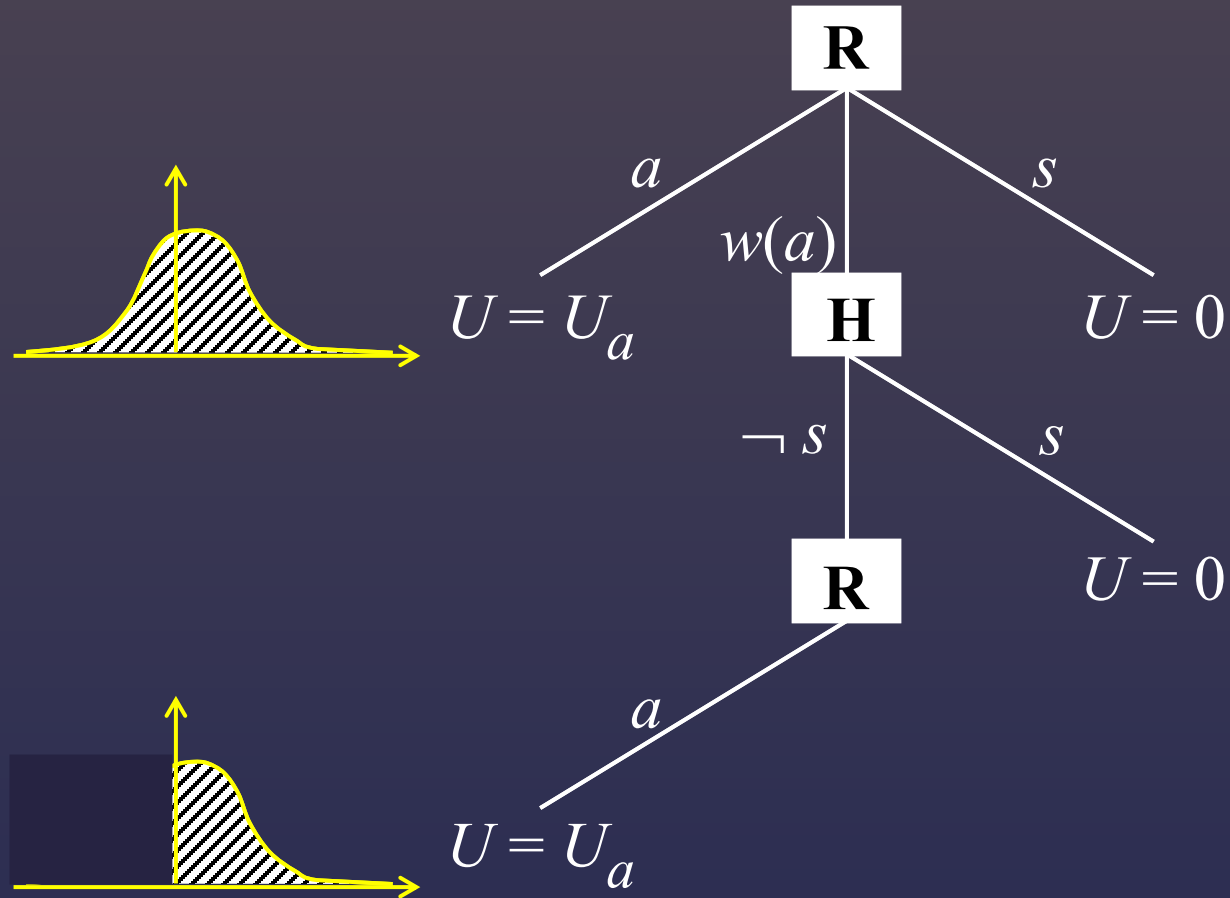
Human behaviour

Machine behaviour

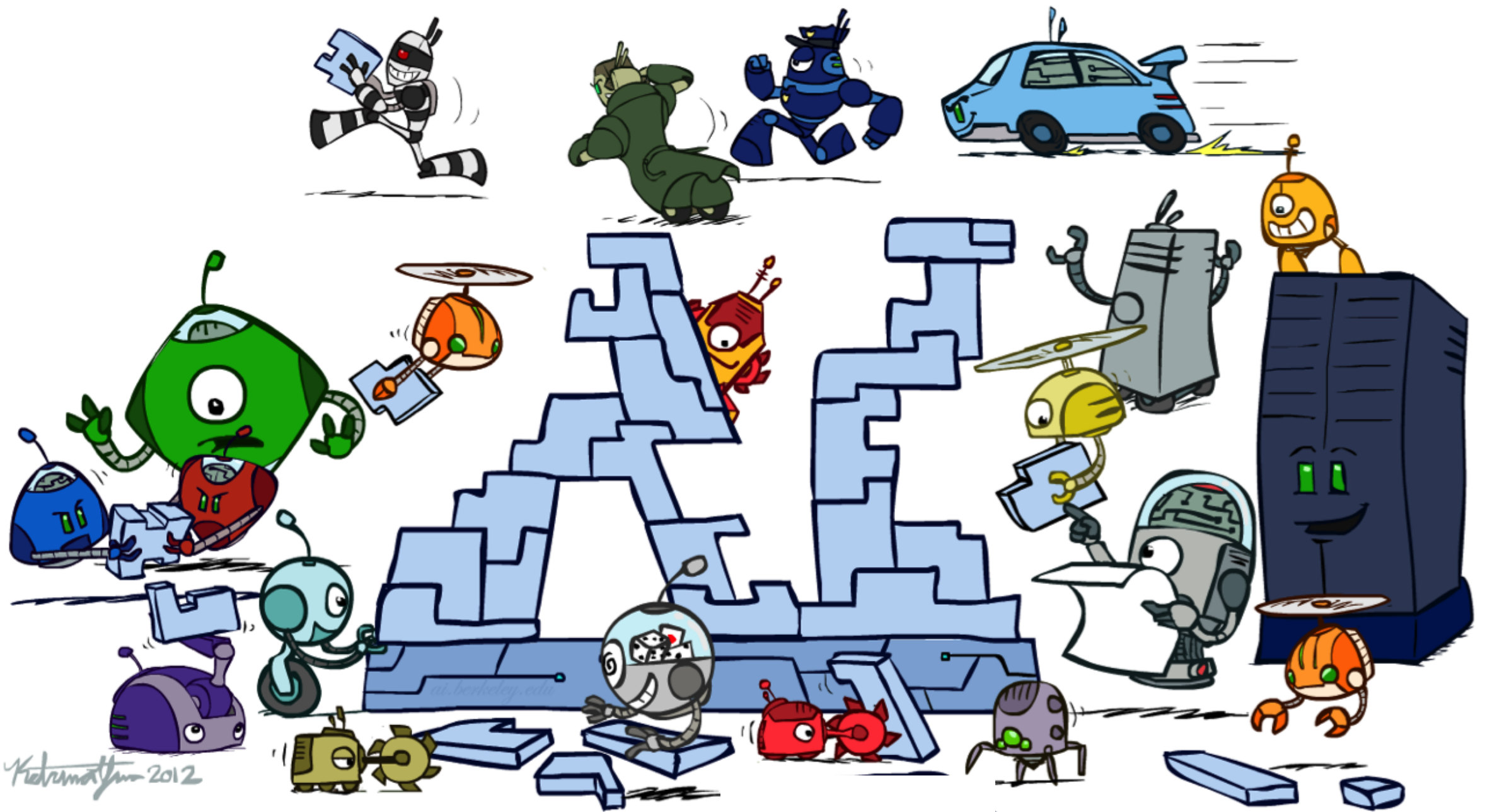
The off-switch problem

- ❖ 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



And thanks to Katrina Yim for her amazing artwork!