Reinforcement Learning and Reward

Emma Brunskill
CS234
Week 10
Winter 2022

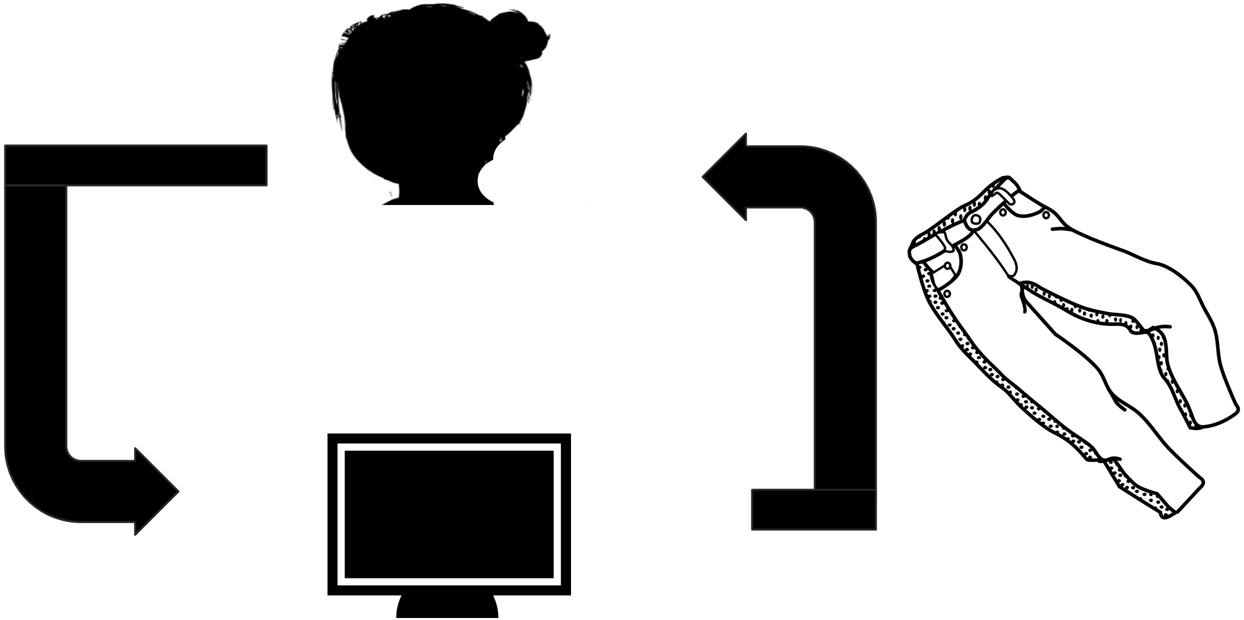
Where We Are

- Last: Learning from historical data
- Now: Reinforcement Learning in the Wild
    - Rewards, alignment
    - Using RL in applications

Plan for today

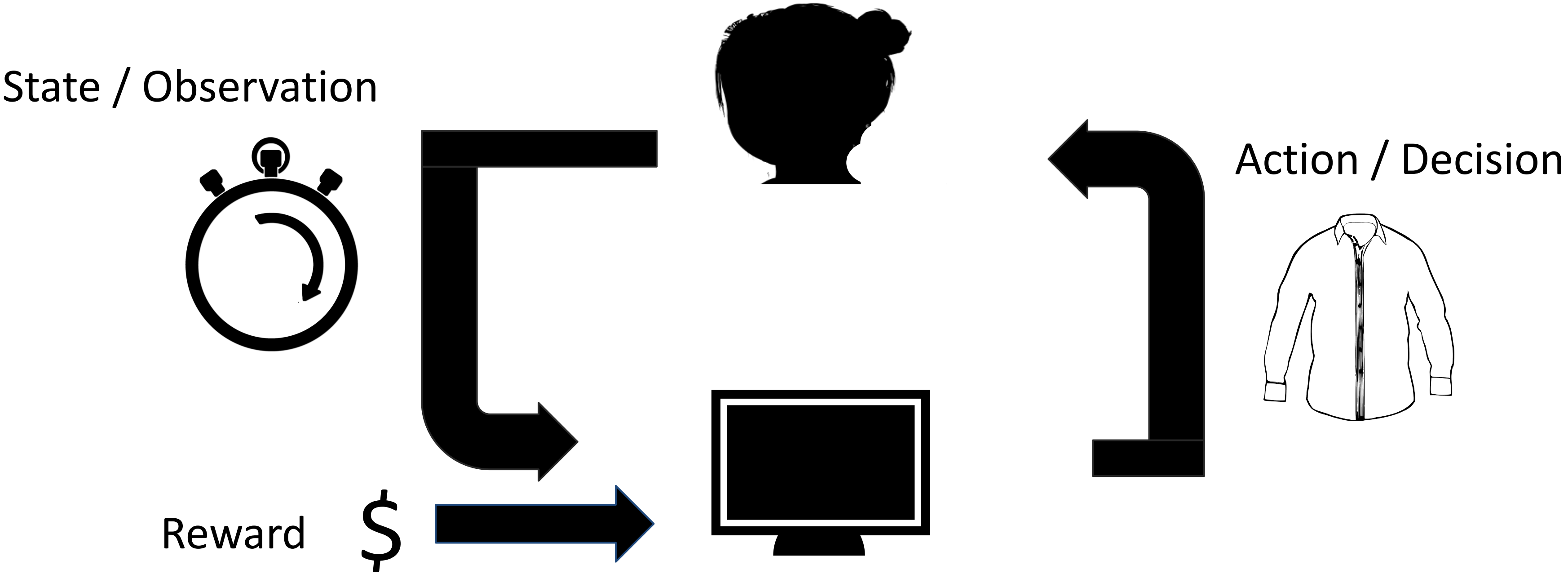
- Reward in RL
- Panel

Reinforcement Learning



Pants image by OpenClipart-Vectors from Pixabay

Decision Policy



State / Observation

Action / Decision

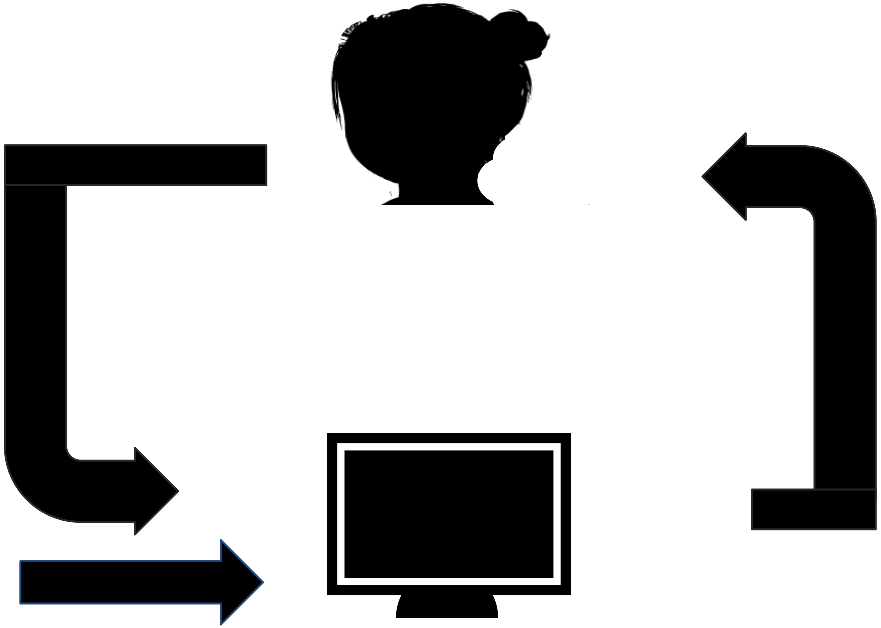
Reward $

(Decision) Policy: if observe this then do that

Example: If looked at blouse for 10 sec Then show another blouse

stopwatch modified from image by Katarzyna Tyl from Pixabay

Advertising Example



State / Observation:

View time

Action / Decision

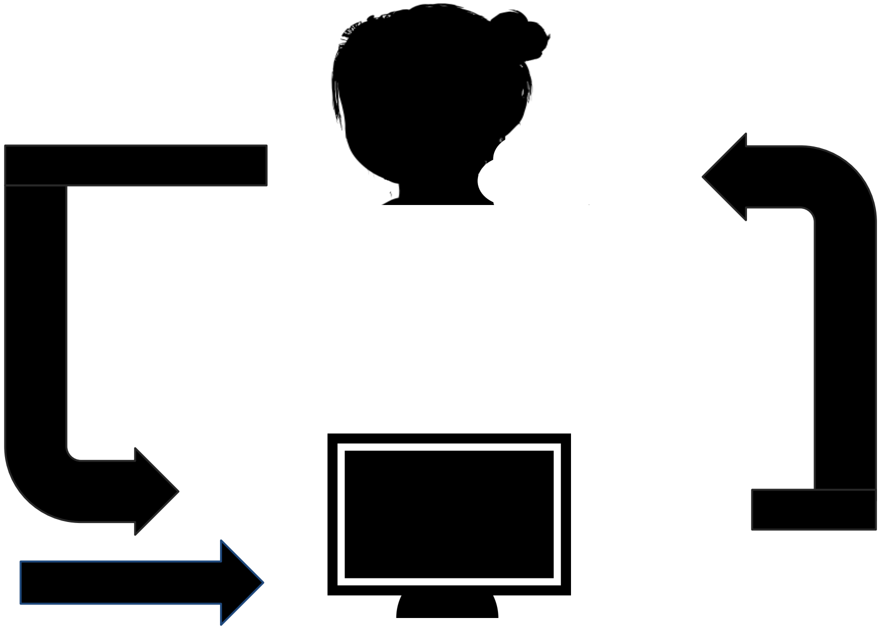
Choose web ad

Reward: Click on ad

Robot Learning to Unload Dishwasher

State / Observation:

Camera image of kitchen



Action / Decision

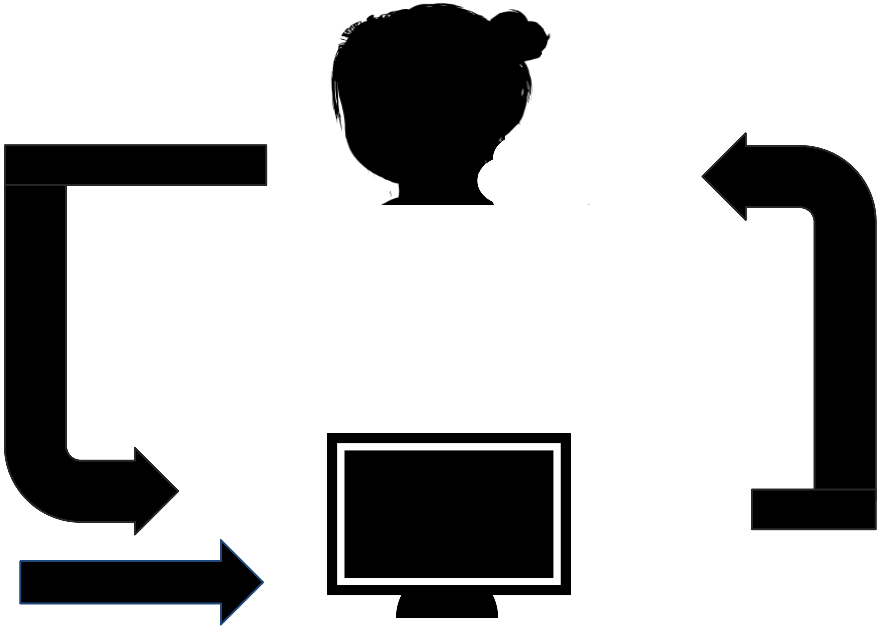
Move robot joint

Reward:
If all dishes in dishwasher +1
Else 0

Blood Pressure Management

State / Observation:

Blood pressure
Gender
Location



Action / Decision

Suggest exercise or meditation

Reward:
If in healthy range: +1
If use medication: -0.05
-

Beyond Expected Reward

- In this class focused on expected scalar reward
- In many real settings
  - Distribution of outcomes (distributional RL, conditional value at risk, …)
  - Multiple-objective (high reward and low cost and …)
  - Constrained maximization (safety, fairness, …)

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last — a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO

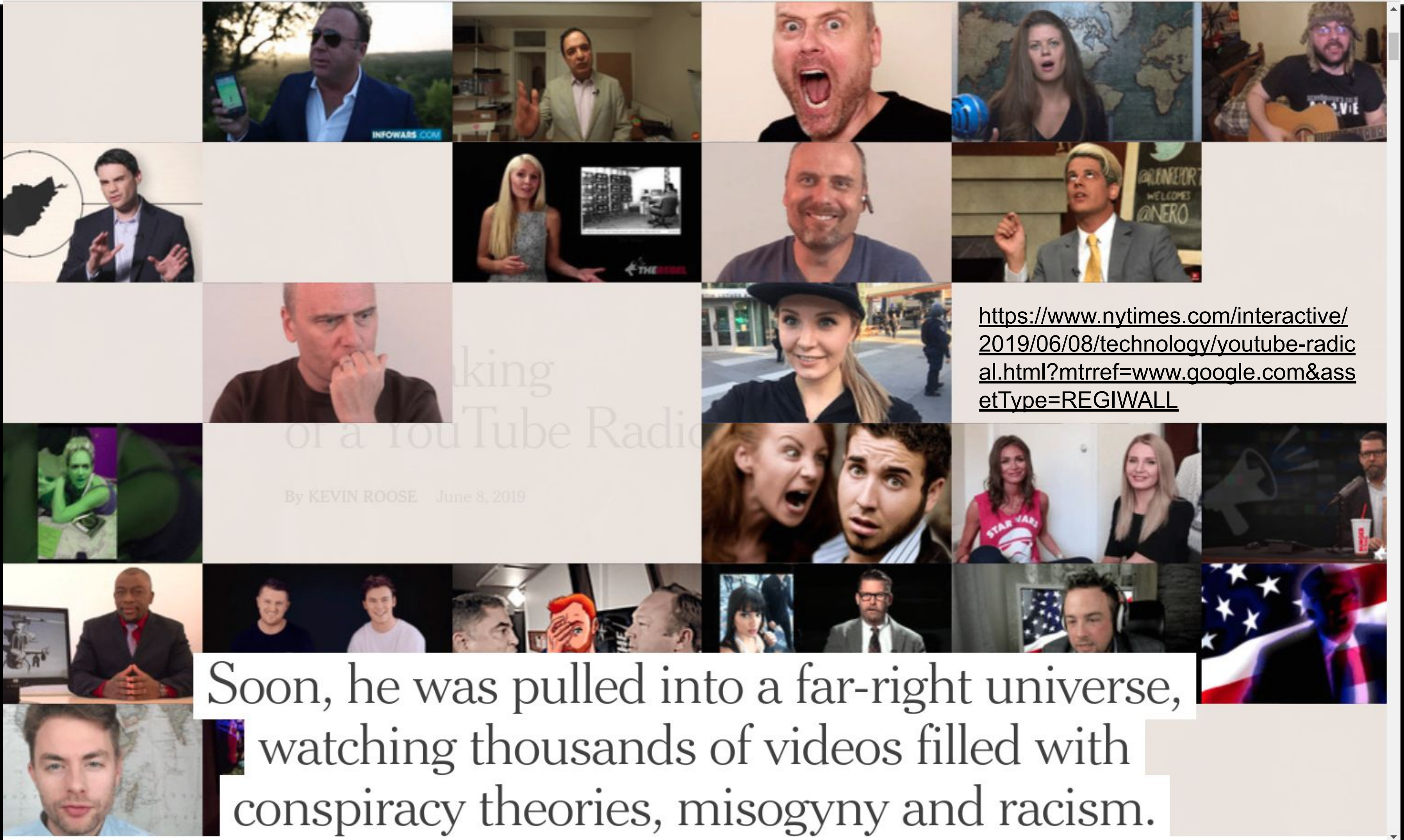
Recall Example During My 1st Lecture: AI Teacher

- Student initially does not know addition (easier) nor subtraction (harder)
- Teaching agent can provide activities about addition or subtraction
- Agent gets rewarded for student performance:
  - +1 if student gets problem right,
  - -1 if get problem wrong
- (Think/Discuss) What type of policy would a RL agent learn? Is this what the human designer of this system would likely want?

The New York Times

TECHNOLOGY

Caleb Cain was a college dropout looking for direction. He turned to YouTube.



aking

of a YouTube Radic

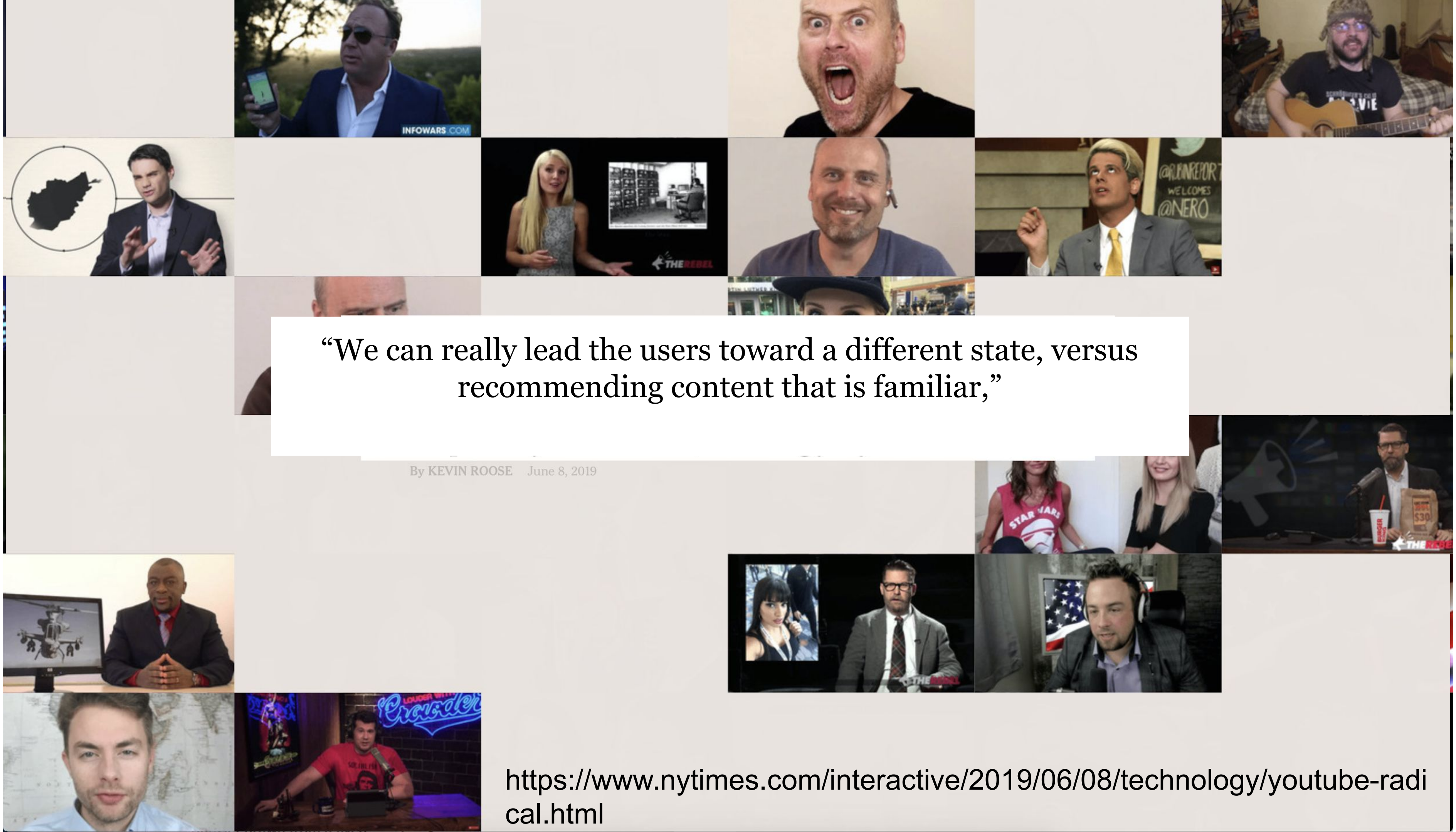
By KEVIN ROOSE June 8, 2019

Soon, he was pulled into a far-right universe, watching thousands of videos filled with conspiracy theories, misogyny and racism.

https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html?mtrref=www.google.com&assetType=REGIWALL

- In last 2 years have been trying out using reinforcement learning
- "… designed to maximize users' engagement over time by predicting which recommendations would expand their tastes and get them to watch not just one more video but many more."

https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

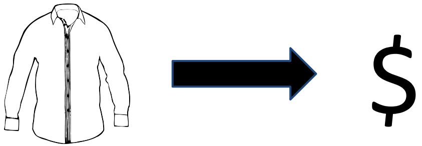


"We can really lead the users toward a different state, versus recommending content that is familiar,"

By KEVIN ROOSE June 8, 2019

https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

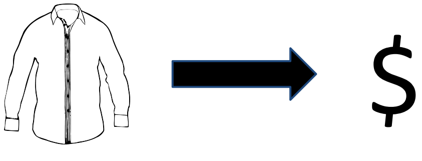
Supervised Learning



$

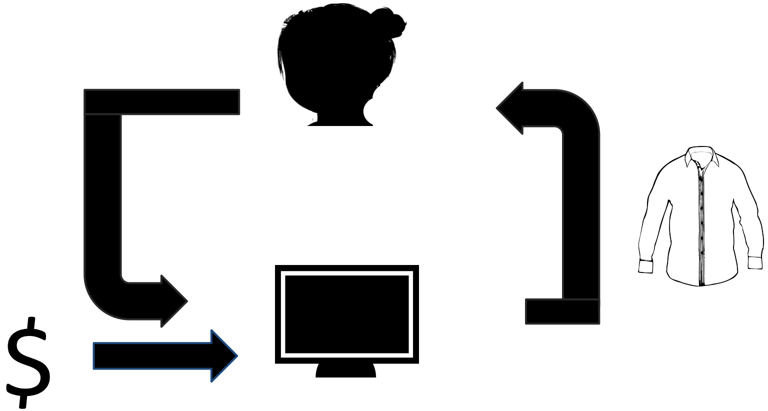
Recommend things people already like*

Supervised Learning



Recommend things people already like*

Reinforcement Learning

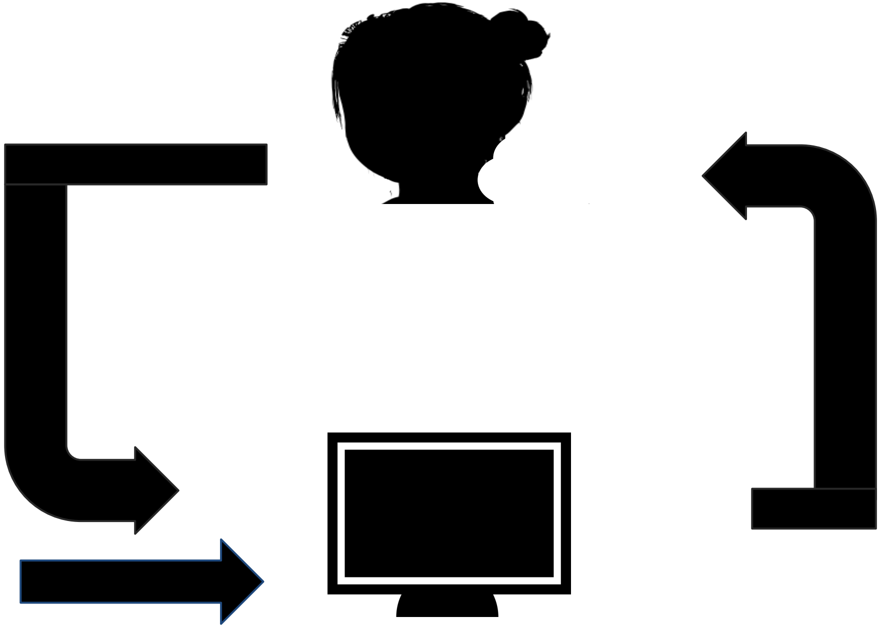


Provide recommendations so people will *(potentially change into people who)* buy more

Reinforcement Learning is Trying to Change (the State of) the World

State / Observation:

Blood pressure
Gender
Location

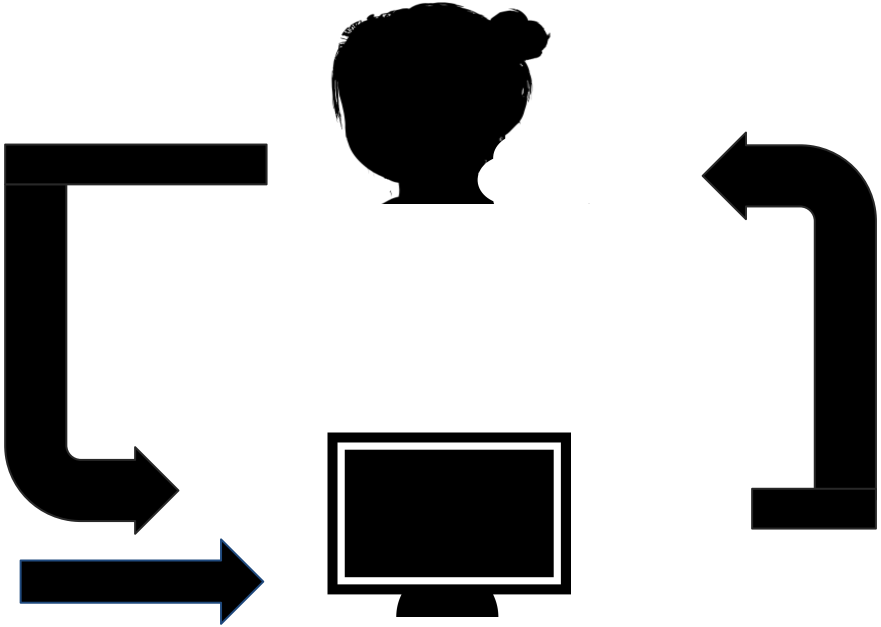


Action / Decision

Suggest exercise or meditation

Reward:
If in healthy range: +1
If use medication: -0.05

Reinforcement Learning is Trying to Change (the State of) the World



State / Observation:

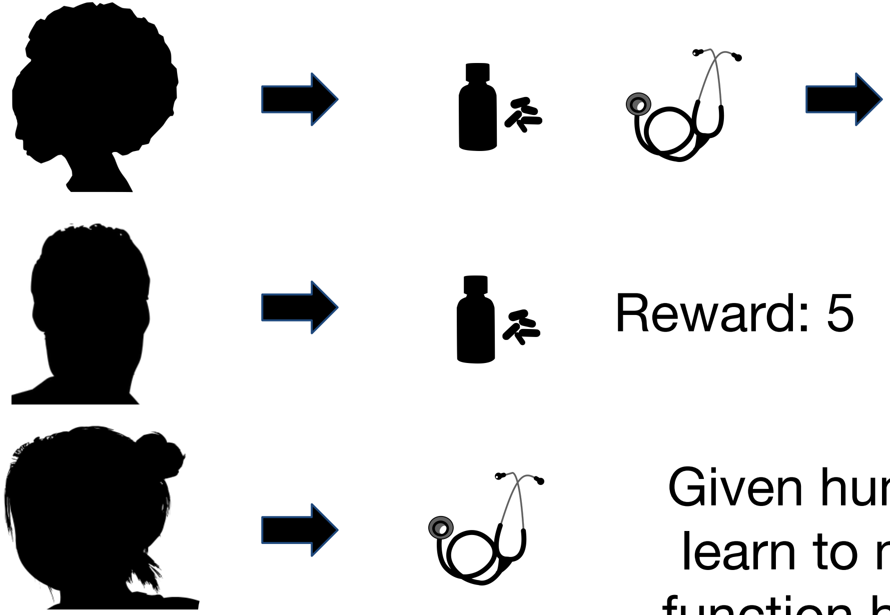
Blood pressure
Gender
Location

Action / Decision

Suggest exercise or meditation

What is the Reward?

One Idea: Learn the Rewards of People



Reinforcement Learning

Reward: 92

Multi-armed Bandits

Reward: 5

Imitation Learning & Inverse RL

Given human expert decisions, learn to mimic or learn reward function humans are optimizing

Value Alignment

- How can we ensure RL agent is optimizing for our desired rewards?
- Stuart Russell (recent general audience book on this broad topic is Human Compatible: AI and the Problem of Control)
- Anca Dragan, Smitha Milli, Dylan Hadfield-Menell, and others

Rest of Today: Panel in RL

-