Reinforcement Learning & Learning to Promote Learning

Emma Brunskill
Stanford University
AI to Automate Humans

Will Robots Replace Human Drivers, Doctors and Other Workers?

Robots could take over 38% of U.S. jobs within about 15 years

Will robots replace workers by 2030?
Artificial Intelligence to Amplify People
Reinforcement Learning
Action (turn left)
Image (pixel colors) → Score → Action (turn left)
Reinforcement Learning

Observation \rightarrow Policy: Map Observations \rightarrow Actions \rightarrow Action

Reward →

*Policy*: Map Observations → Actions

*Goal*: Choose actions to maximize expected rewards
But Don’t Know How World Works!

Policy: Map Observations $\rightarrow$ Actions

Goal: Choose actions to maximize expected rewards
Reinforcement Learning Progress
Real Potential: Humans & AI
Reinforcement Learning with and for People

Goal: Choose actions to maximize expected rewards

Policy: Map Observations $\rightarrow$ Actions
Reinforcement Learning with and for People

Policy: Map Observations $\rightarrow$ Actions
Goal: Choose actions to maximize student outcomes

Student's answer

Math exercise

Pass exam
Cheap to try things, or Simulate ≠ High stakes Hard to model
Reinforcement Learning & Learning to Promote Learning

Making better decisions by
1) Learning from past experience
2) Having humans help machines
A Classrooms

Avg Score: 95
A Classrooms  
Avg Score: 95

B Classrooms  
Avg Score: 92
What should we do for a new student?

A Classrooms

Avg Score: 95

B Classrooms

Avg Score: 92
Comes Up in Many Domains: e.g. Equipment Maintenance Scheduling
Comes Up in Many Domains: e.g. Patient Treatment Ordering
Core Aspect of Intelligent Behavior

Data about past decisions & outcomes

How best to act in the future?

Image: https://upload.wikimedia.org/wikipedia/commons/f/f0/DARPA_Big_Data.jpg
Challenge: Counterfactual Reasoning

A Classrooms
Avg Score: 95

B Classrooms
Avg Score: 92

B Classrooms
Avg Score: ????
Challenge: Generalization to Untried Policies

A Classrooms
Avg Score: 95

B Classrooms
Avg Score: 92

B Classrooms
Avg Score: ????
Policy: Player state → level
Goal: Maximize engagement
Old data: ~11,000 students
Statistical Predictive model (e.g. Predict if student will get next level correct)

Image: https://upload.wikimedia.org/wikipedia/commons/f/f0/DARPA_Big_Data.jpg
Use Models as a Simulator

Predictive statistical model of player behavior

Observation → Reward → Action

Model of player engagement

Goal: Choose actions to maximize expected rewards
Goal: Choose actions to maximize expected rewards
Alternative: Reweigh Old Experience to Look Like New Policy

- No statistical predictive model assumptions

**Behavior Policy**

**Student outcomes**

**New Policy**

**Student outcomes**
Alternative: Reweigh Old Experience to Look Like New Policy

- No statistical predictive model assumptions

- Unbiased* estimate of new policy’s performance

*Under mild assumptions
We used to find a policy with 30% higher engagement (Mandel et al. 2014)
When Making Many Decisions...

Fraction Identification Tutor

Let's name fractions using number lines!

1. Brittany bought a watermelon to share with three of her friends. Each of the watermelon pieces were equal-sized. Brittany ate 1/4 of the watermelon. Use the number line to show how much of the watermelon Brittany ate.

Number of sections: 4
Towards Better Estimates of New Policies

- Trade bias and variance
- New methods to combine models & direct evaluation (Guo, Thomas, B 2017; Thomas and B 2016)
Towards Using Old Data to Confidently Identify Better Policies for Future Use

Data about past decisions & outcomes

Image: https://upload.wikimedia.org/wikipedia/commons/f/f0/DARPA_Big_Data.jpg

How best to act in the future?
Reinforcement Learning & Learning to Promote Learning

• Making better decisions by
  • Learning from past experience
  • Having humans help machines
Histogram Tutor

DESCRIPTIONS AND HISTOGRAMS (1/3 points)
The price of airline tickets varies over time. The following is a histogram that could describe the distribution of airplane ticket prices. Select the best option for each of the questions below.

The x-axis should be labeled as

- Time
- Ticket Price
- Frequency ✖
- Distribution
Continually Improving Tutoring System

Correct/Wrong

At end, post test
Improving Across Many Students
Over Time Tutoring System Stopped Giving Some Problems to Students
System Self-Diagnosed that Problems Weren’t Helping Student Learning

DESCRIPTIONS AND HISTOGRAMS (1/3 points)

The price of airline tickets varies over time. The following is a histogram that could describe the distribution of airplane ticket prices. Select the best option for each of the questions below.

The x-axis should be labeled as

- Time
- Ticket Price
- Frequency ×
- Distribution

View Unit in Studio
Humans are Invention Machines

New actions

New sensors
Goal: Choose actions to maximize expected rewards
Goal: Choose actions to maximize expected rewards
Human in the Loop Reinforcement Learning

Goal: Choose actions to maximize expected rewards
Goal: Choose actions to maximize expected rewards
Chrissy loves exploring outdoors. Yesterday, she saw a herd of 12 elk being chased by a pack of 8 wolves. How many animals in total did Chrissy see while she was exploring?

'animals' needs to be the total of all important parts.
Chrissy loves exploring outdoors. Yesterday, she saw a herd of 12 elk being chased by a pack of 8 wolves. How many animals in total did Chrissy see while she was exploring?

New actions = new hints

Learning where to ask for new hints

People helping computers to teach people
People Helping Computers to Teach People Tasks

Doroudi, Kamar, Brunskill, Horvitz, CHI 2016
Reinforcement Learning &
Learning to
Promote Learning

• Making better
decisions by
  – Learning from past experience
  – Having humans help machines
Thanks to

and Karan Goel, Travis Mandel, Yun-En Liu, NSF, ONR, Microsoft, Google, Yahoo & IES
Reinforcement Learning & Learning to Promote Learning

- Making better decisions by
  - Learning from past experience
  - Having humans help machines