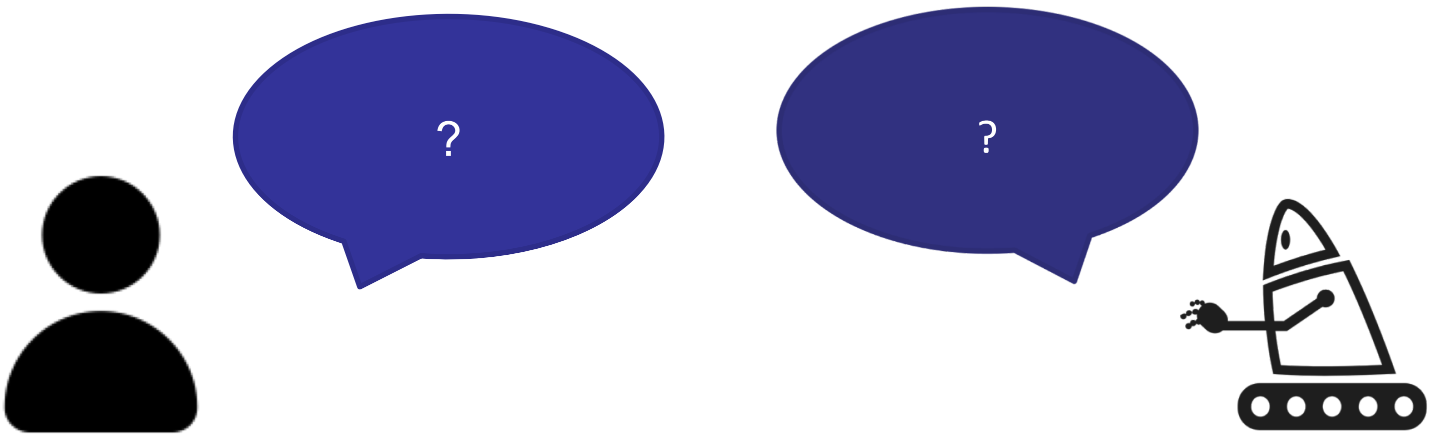
CS 188: Artificial Intelligence

Inverse Reinforcement Learning and AI Safety



Guest Lecturer: Regina Wang

Slides adapted from Stuart Russell, Anca Dragan

Roadmap

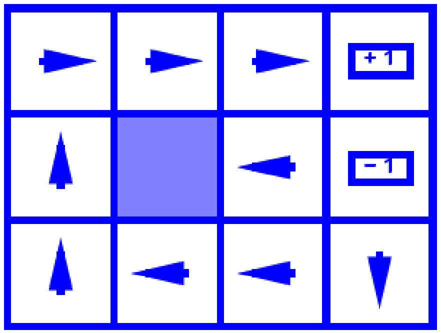
- Inverse Reinforcement Learning (IRL)
  - Standard model
  - Motivation

- AI Safety
  - Powerful AI
  - Revised Model
  - Off-switch game

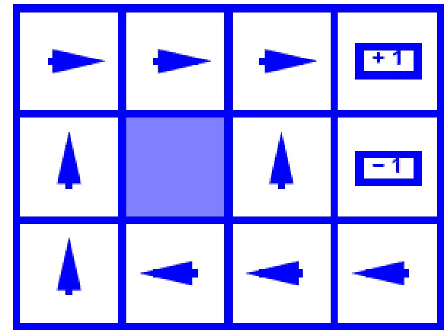
- Going forward
  - Research areas
  - AI governance

Inverse Reinforcement Learning

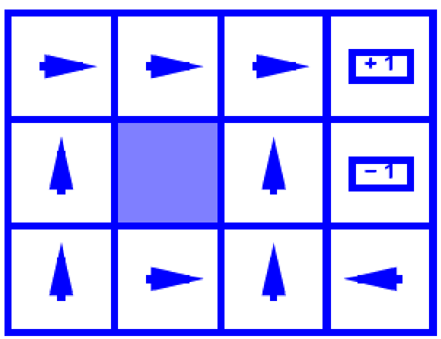
Reminder: Optimal Policies



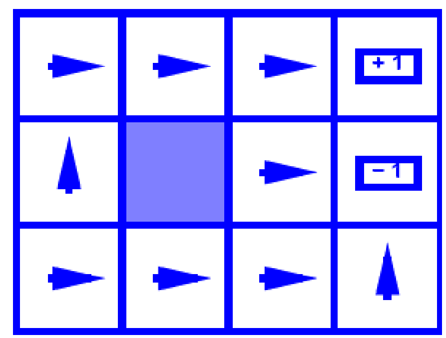
R(s) = -0.01



R(s) = -0.03



R(s) = -0.4

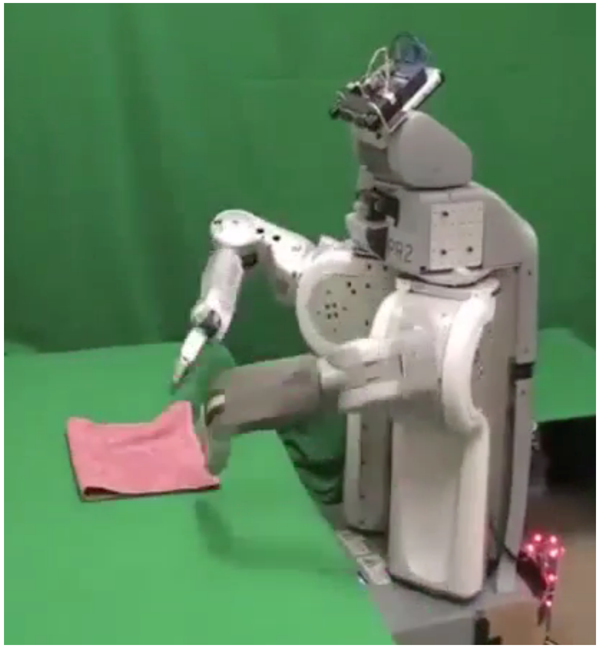


R(s) = -2.0

Utility?

Clear utility function

Not so clear utility function



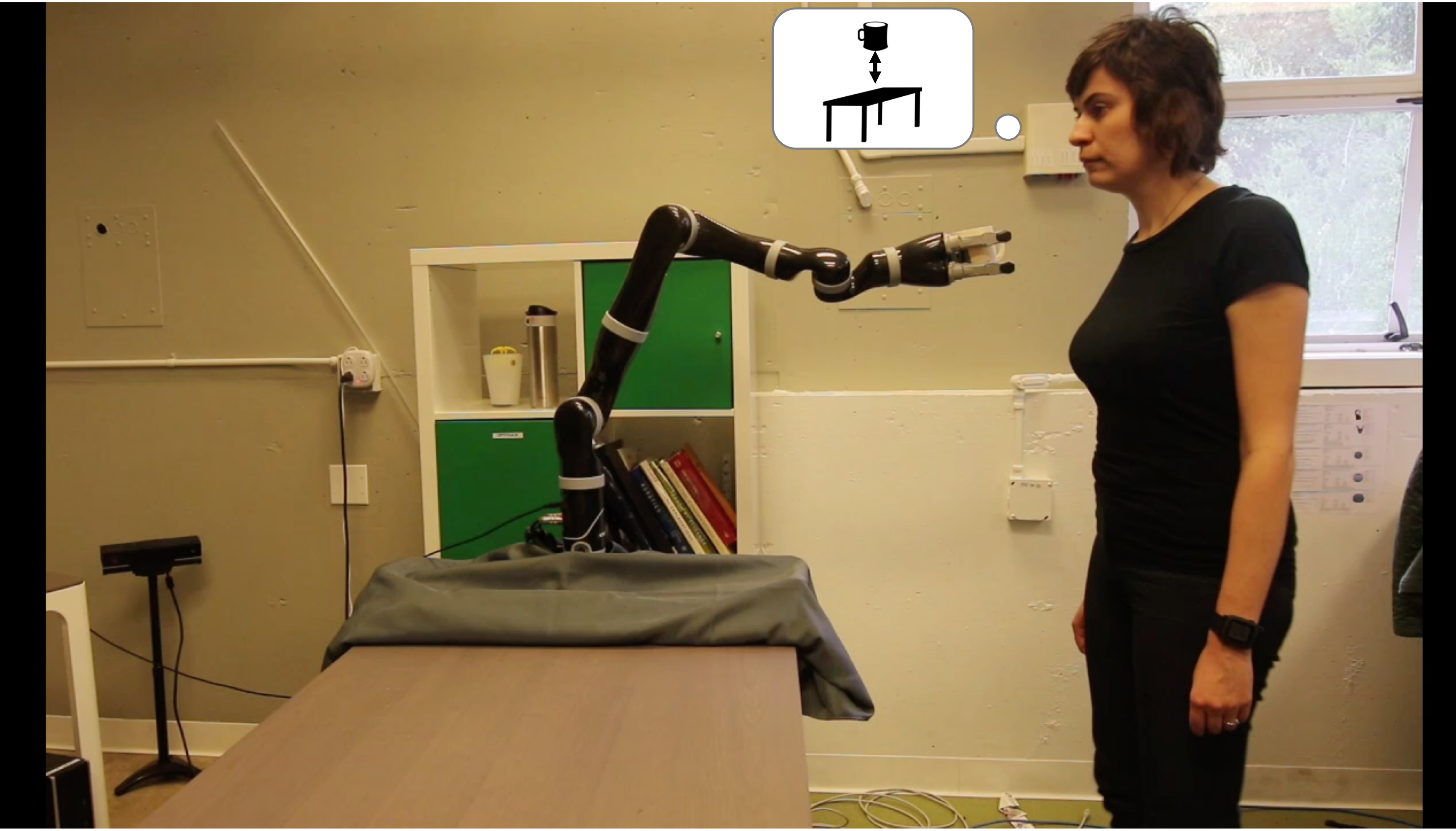
Where do reward functions come from in this class?

Standard model for AI

We expect to be able to give a reward function for the AI system to optimize. This is how it's done in this class!

King Midas problem: **Cannot specify *R* correctly**

OpenAI 2016. Faulty Reward Functions in the Wild



Planning/Reinforcement Learning

$$R \rightarrow \pi^*$$

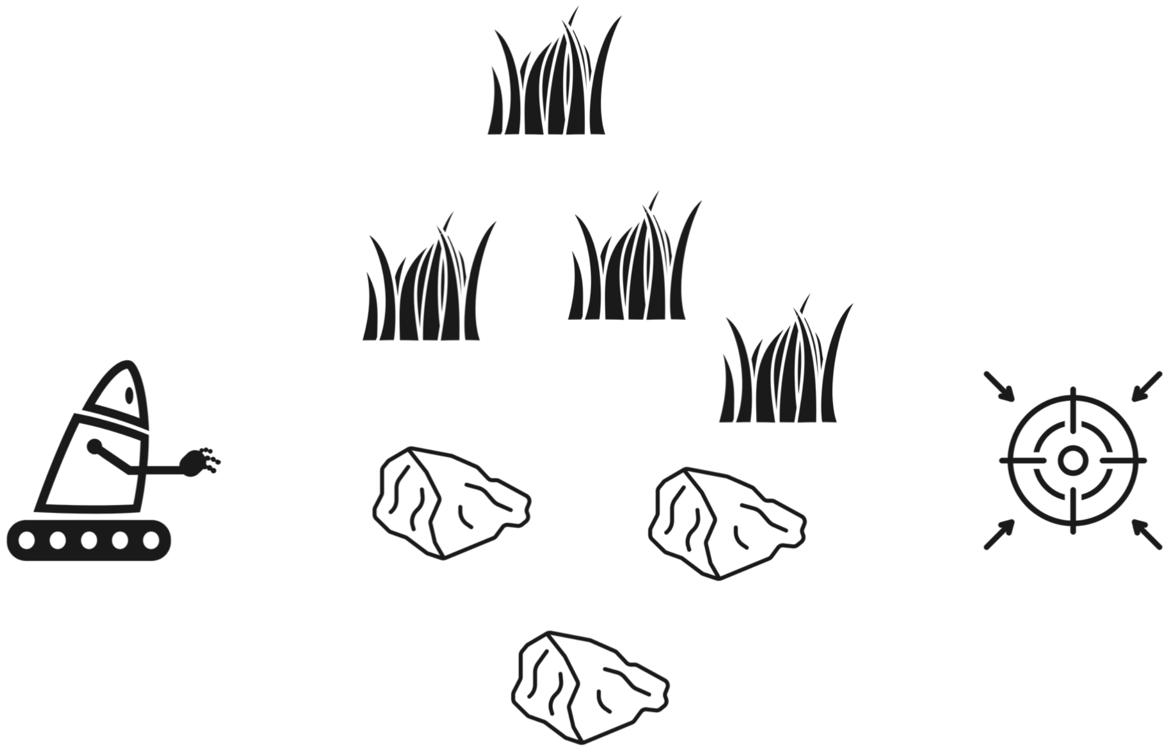
Inverse Planning/RL

$$\pi^* \rightarrow R$$

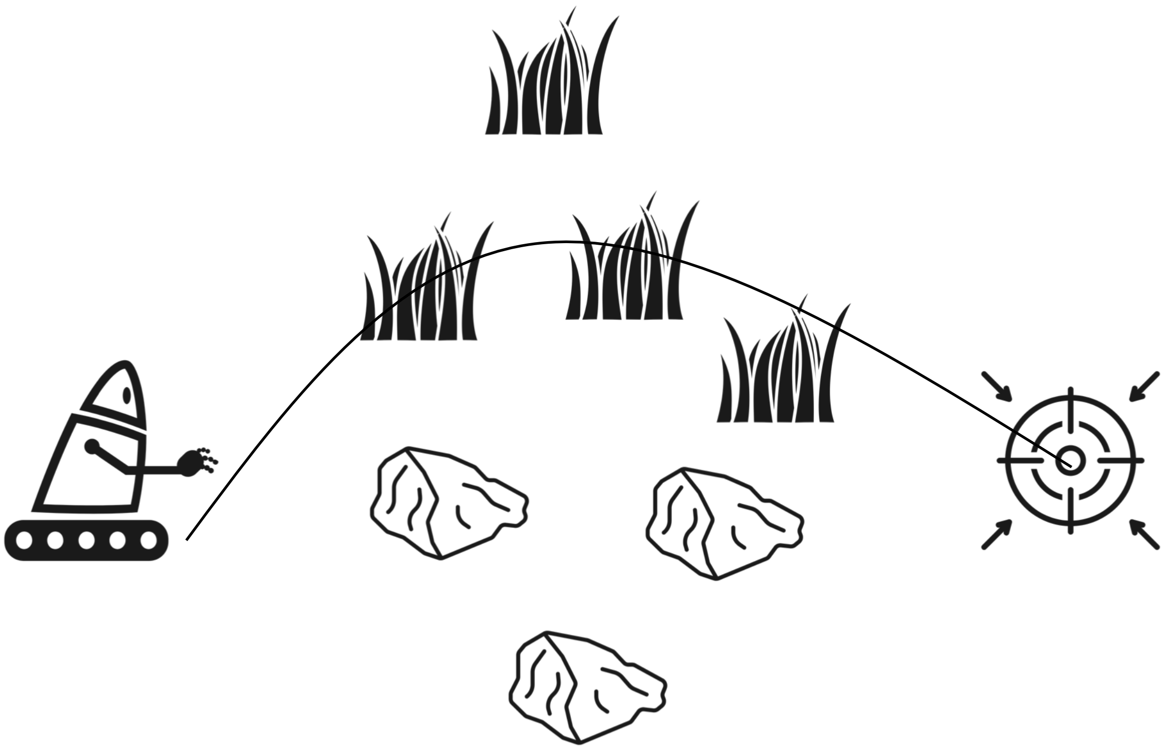
Inverse Planning/RL

$$\xi \rightarrow R$$

Inverse Planning/RL



Inverse Planning/RL



Inverse Planning/RL

given: ξ_{D}

find: $R(s, a)$

s.t. $\boldsymbol{R}(\xi_D) \geq \boldsymbol{R}(\xi) \forall \xi$

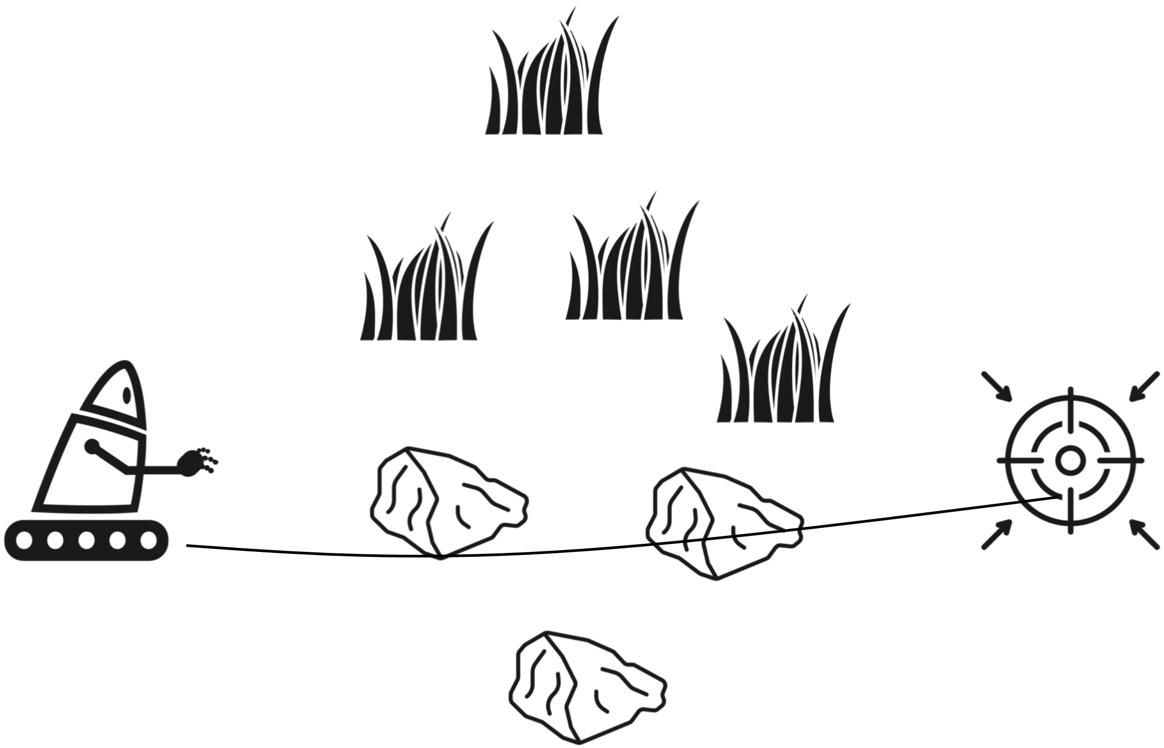
Inverse Planning/RL

given: ξ_{D}

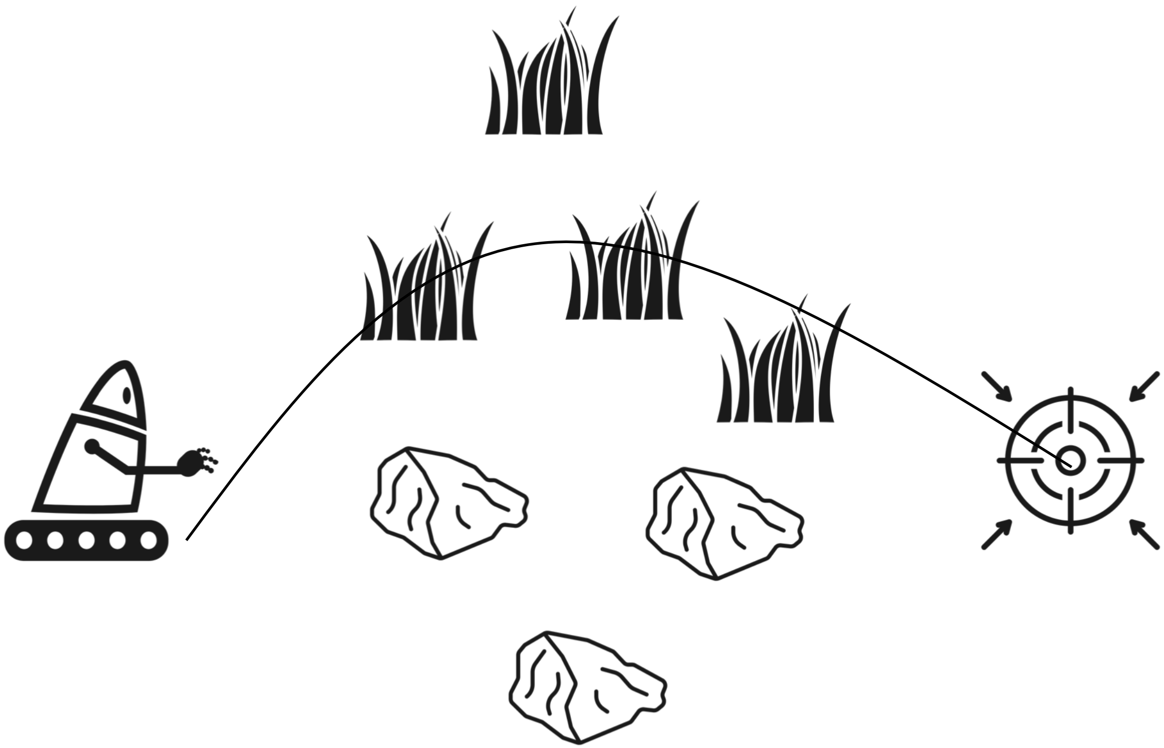
find: $R(s,a) = \theta^T \phi(s,a)$

s.t. $\boldsymbol{R}(\xi_D) \geq \boldsymbol{R}(\xi) \forall \xi$

Inverse Planning/RL



Inverse Planning/RL



Is the demonstrator really optimal?

$$\boldsymbol{R}(\xi_D) \geq \boldsymbol{R}(\xi) \forall \xi$$

The Bayesian view

$$\underbrace{P(\xi_D}_{\text{evidence}}|\underbrace{\theta)}_{\text{hidden}}$$

The Bayesian view

$$P(\xi_D|\theta) \propto e^{\beta\theta^T\boldsymbol{\phi}(\xi_D)}$$

The Bayesian view

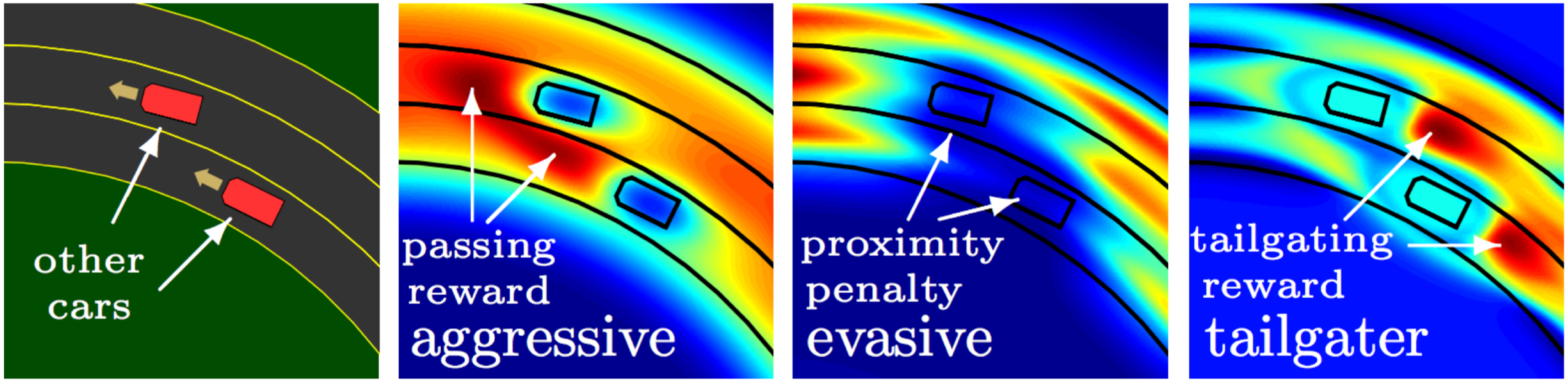
$$P(\xi_D|\theta) = \frac{e^{\beta\theta^T\boldsymbol{\phi}(\xi_D)}}{\sum_{\xi} e^{\beta\theta^T\boldsymbol{\phi}(\xi)}}$$

The Bayesian view

$$P(\xi_D|\theta) = \frac{e^{\beta\theta^T\boldsymbol{\phi}(\xi_D)}}{\sum_{\xi} e^{\beta\theta^T\boldsymbol{\phi}(\xi)}}$$

$$b'(\theta) \propto b(\theta)P(\xi_D|\theta)$$

[Ratliff et al. *Maximum Margin Planning*]



[Levine et al. *Continuous Inverse Optimal Control with Locally Linear Examples*]

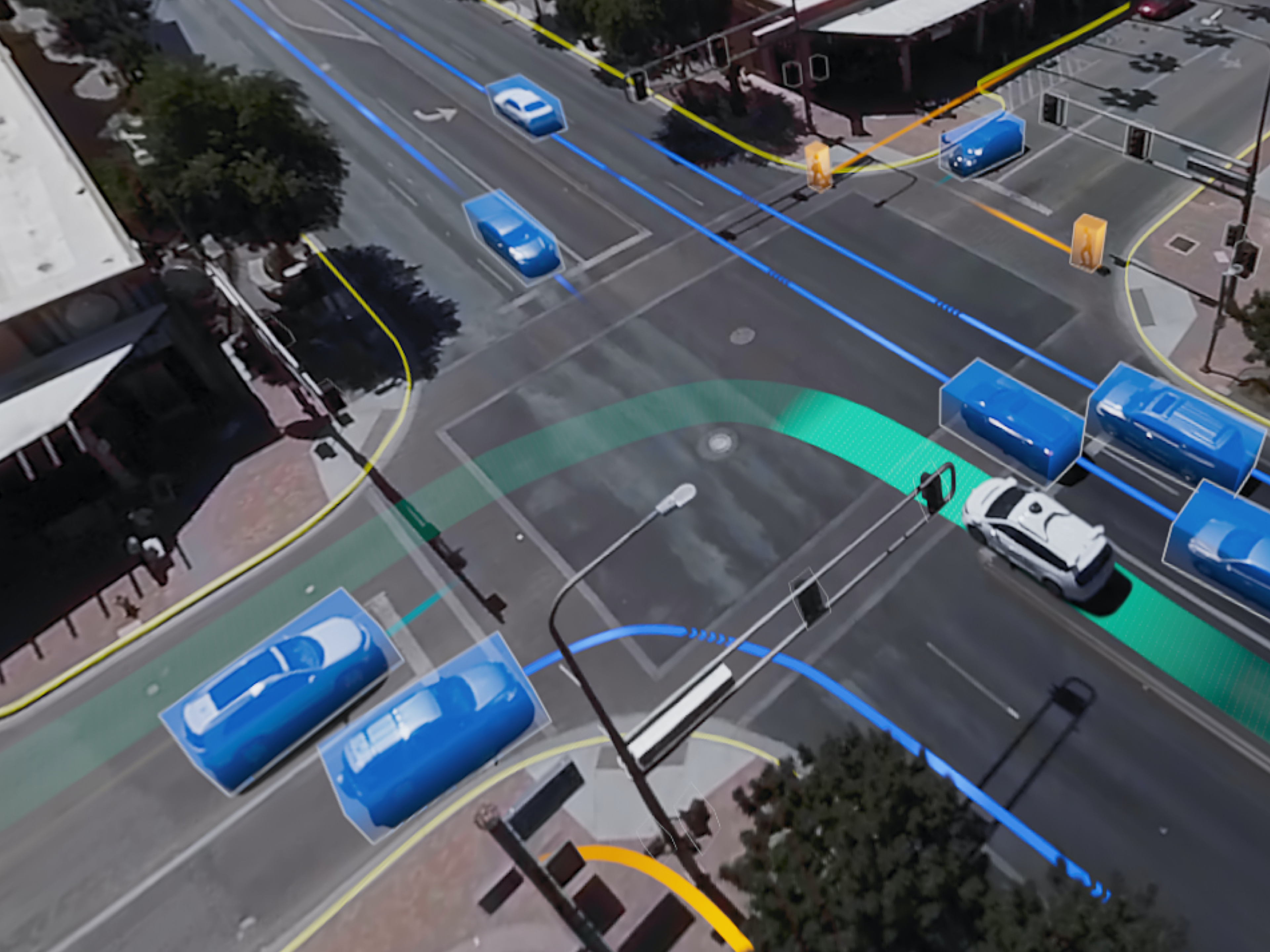
Roadmap

- Inverse Reinforcement Learning (IRL)
  - Standard model
  - Motivation
- AI Safety
  - Powerful AI
  - Revised Model
  - Off-switch game
- Going forward
  - Research areas
  - AI governance

AI Safety

“It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. … At some stage therefore we should have to expect the machines to take control.”

- Alan Turing, 1951



Artificial intelligence / Machine learning

The way we train AI is fundamentally flawed

The process used to build most of the machine-learning models we use today can't tell if they will work in the real world or not—and that's a problem.

by **Will Douglas Heaven**

November 18, 2020

Deep learning forever?

François Chollet (2017):

"Many more applications are completely out of reach for current deep learning techniques – even given vast amounts of human-annotated data …

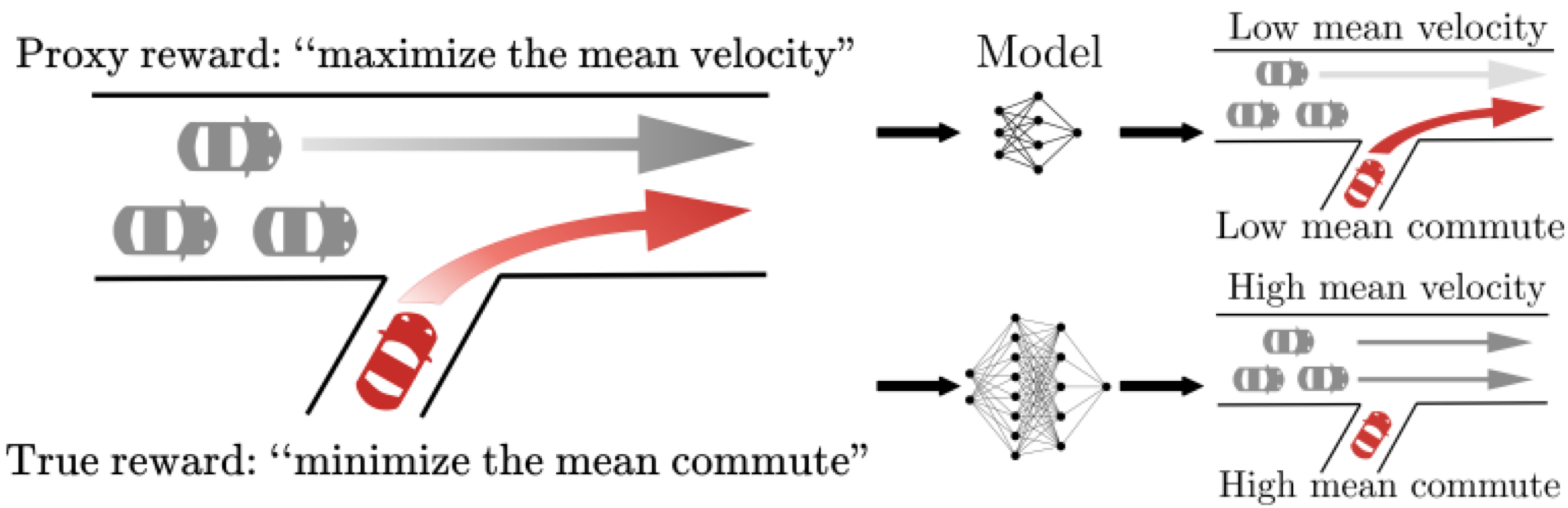
The main directions in which I see promise are models closer to general-purpose computer programs."

Recall: Standard model for AI

We expect to be able to give a reward function for the AI system to optimize. This is how it's done in this class!

King Midas problem: **Cannot specify *R* correctly**

Smarter AI would lead to a worse outcome



Pan, Bhatia, Steinhardt 2022. The Effects of Reward Misspecification

Likely AI developments in the 2020s

- Robots for war, roads, warehouses, mines, fields, home
- Personal digital assistants for all aspects of life
- Use of AI in clinical settings and prostheses
- Global vision system via satellite imagery

What happens as AI becomes more powerful?

AI systems will eventually make better decisions than humans

Turing's point: how do we retain control over entities more powerful than us, for ever?

[Russell, Many Experts Say We Shouldn't Worry About Superintelligent AI. They're Wrong, *IEEE Spectrum*, October 2019.]

E.g., Social Media

Optimizing clickthrough

= ~~learning what people want~~

= modifying people to be more predictable

How we got into this mess

- Humans are intelligent to the extent that our actions can be expected to achieve our objectives
- ~~Machines are intelligent to the extent that their actions can be expected to achieve their objectives~~
- Machines are ***<u>beneficial</u>*** to the extent that ***<u>their</u>*** actions can be expected to achieve ***<u>our</u>*** objectives

Revised model: <u>Provably beneficial AI</u>

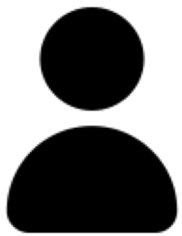
Instead of giving the robot an explicit reward function,

1. Robot goal: satisfy human preferences
2. Robot is uncertain about human preferences
3. Human behavior provides evidence of preferences

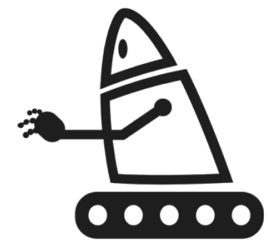
AI becomes an assistance game with human and AI players!

Smarter AI now leads to a <u>better</u> outcome!

Basic assistance game



Acts roughly according to preferences



Maximizes unknown human preferences

Game:

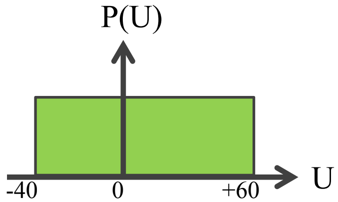
- Human teaches robot
- Robot learns, asks questions and permission, defers to human, and allows off-switch

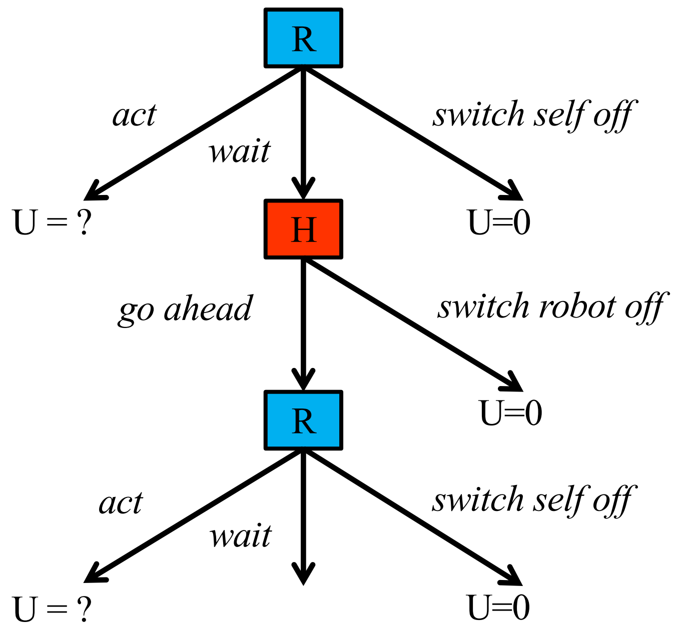
[Hadfield-Menell et al, NeurIPS 16, IJCAI 17, NeurIPS 17]
[Milli et al 2017, IJCAI 17] [Malik et al, ICML 18]

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"
- A robot with uncertainty about objective won't behave this way

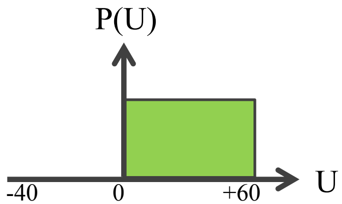
Off-switch problem (example)





$EU(act) = +10$

$EU(wait) = (0.4 * 0) + (0.6 * 30)$
$= +18$



Off-switch problem (general proof)

- $EU(act) = \int_{-\infty}^{+\infty} P(u) \cdot u \, du = \int_{-\infty}^{0} P(u) \cdot u \, du + \int_{0}^{+\infty} P(u) \cdot u \, du$
- $EU(wait) = \int_{-\infty}^{0} P(u) \cdot 0 \, du + \int_{0}^{+\infty} P(u) \cdot u \, du$
- Obviously $\int_{-\infty}^{0} P(u) \cdot u \, du \leq \int_{-\infty}^{0} P(u) \cdot 0 \, du$
- Hence $EU(act) \leq EU(wait)$
  - "If H doesn't switch me off, then the action must be good for H, and I'll get to do it, so that's good; if H does switch me off, then it's because the action must be bad for H, so it's good that I won't be allowed to do it."

Roadmap

- Inverse Reinforcement Learning (IRL)
  - Standard model
  - Motivation
- AI Safety
  - Powerful AI
  - Off-switch game
  - Revised Model
- Going forward
  - Research areas
  - AI governance

Going Forward

Rebuild AI on a New Foundation

Proposed by Stuart Russell in Human Compatible:

Remove the assumption of a perfectly known objective/goal/loss/reward

- Combinatorial search: G(s) and c(s,a,s')
- Constraint satisfaction: hard and soft constraints
- Planning: G(s) and c(s,a,s')
- Markov decision processes: R(s,a,s')
- Supervised learning: Loss(x,y,y')
- Reinforcement learning: R(s,a,s')
- Robotics: all of the above

Ongoing research: “Imperfect” humans

How do we deal with the following?

- Computation limitation
- Hierarchically structured behavior
- Emotionally driven behavior
- Uncertainty about own preferences
- Plasticity of preferences
- Non-additive, memory-laden, retrospective/prospective preferences

Ongoing research: General Safety

And when we disregard humans, there’s more:

- Safe exploration
- Robustness to distributional shift
- Avoiding negative side effects
- Avoiding reward hacking

[Amodei et al. Concrete Problems in AI Safety]

And then there's Governance…

"While AI researchers, developers, and industry can lay the groundwork for what is technically feasible, it is ultimately up to government and civil society to determine the frameworks within which AI systems are developed and deployed."

- *Perspectives on Issues in AI Governance,*

Google, 2019

Summary

The standard model for AI may lead to loss of human control over increasingly intelligent AI systems.

Provably beneficial AI is possible and desirable.

It isn't "AI safety" or "AI Ethics," it's AI