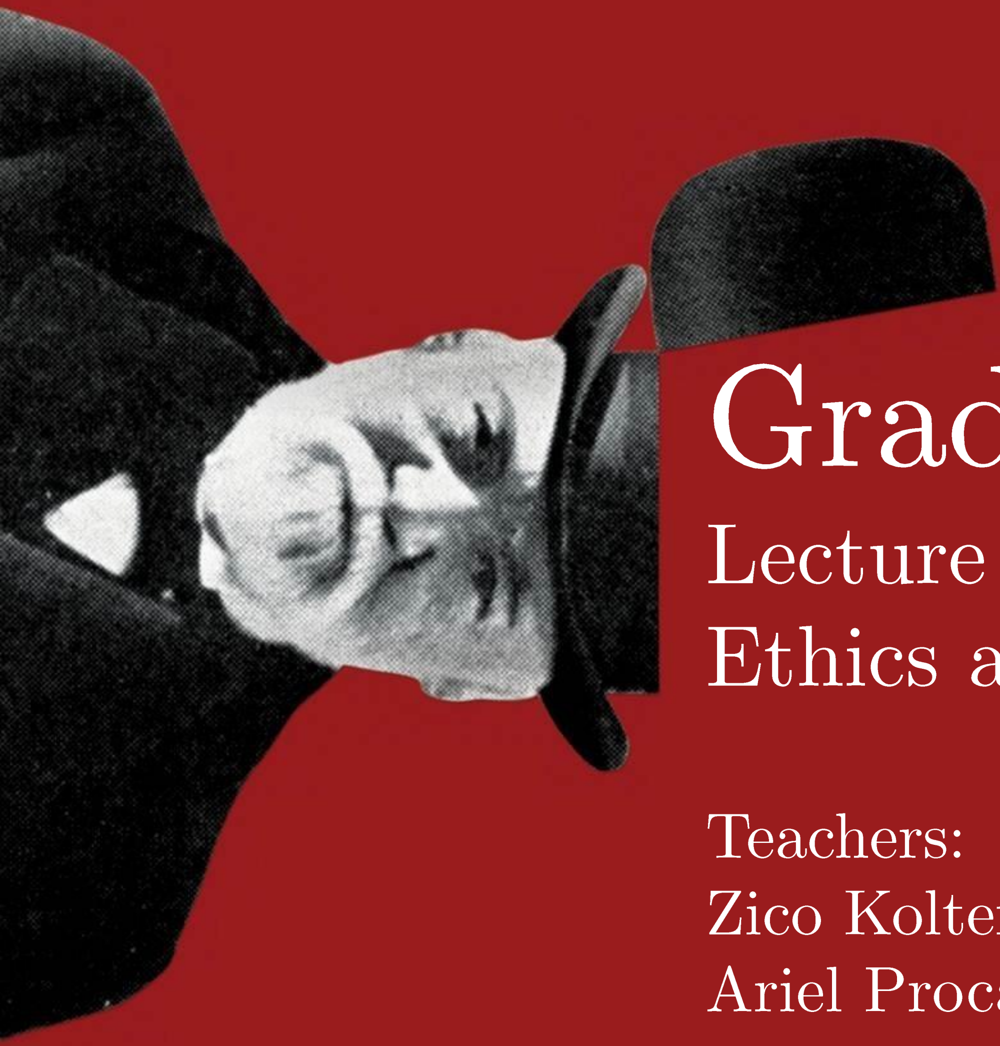
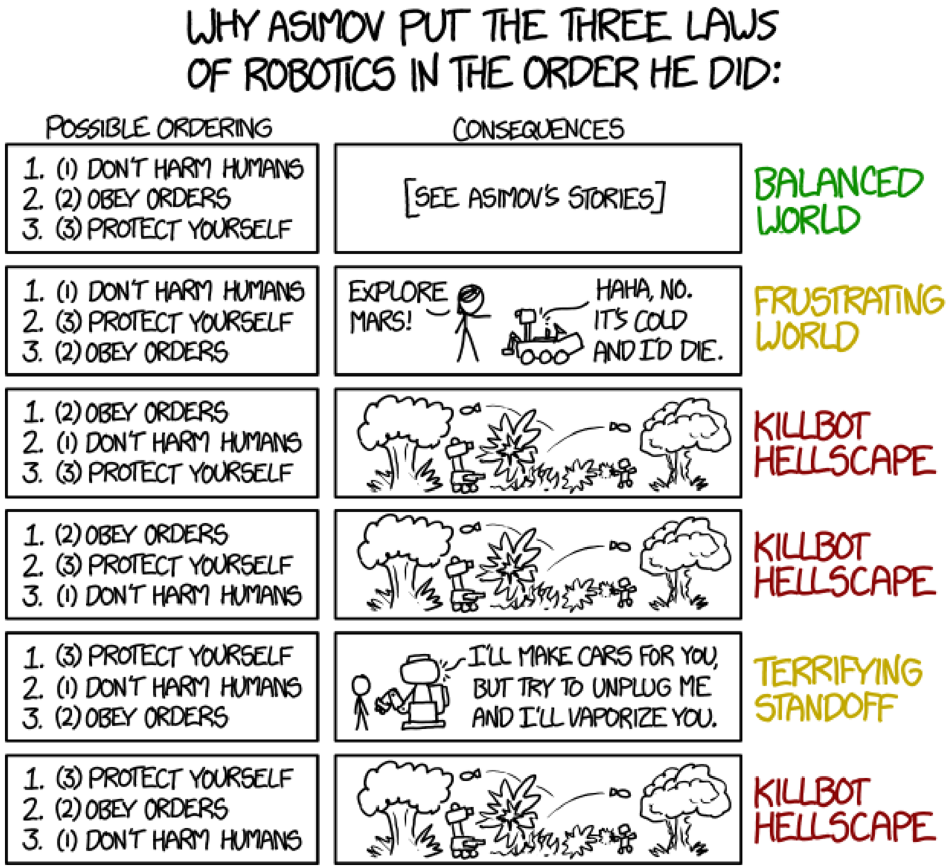


# Graduate AI

## Lecture 26:
## Ethics and AI I

Teachers:
Zico Kolter
Ariel Procaccia (this time)

# The three laws of robotics




# Ethical robots

- Experiments performed by Winfield et al. [2014]
- Environment includes a robot, a human, and a hole which can be sensed by the robot but not the human
- Robot can simulate the consequences of possible actions

```
IF for all robot actions, the human is equally safe
THEN (* default safe actions *)
       output safe actions
ELSE (* ethical action *)
       output action(s) for least unsafe human outcome(s)
```

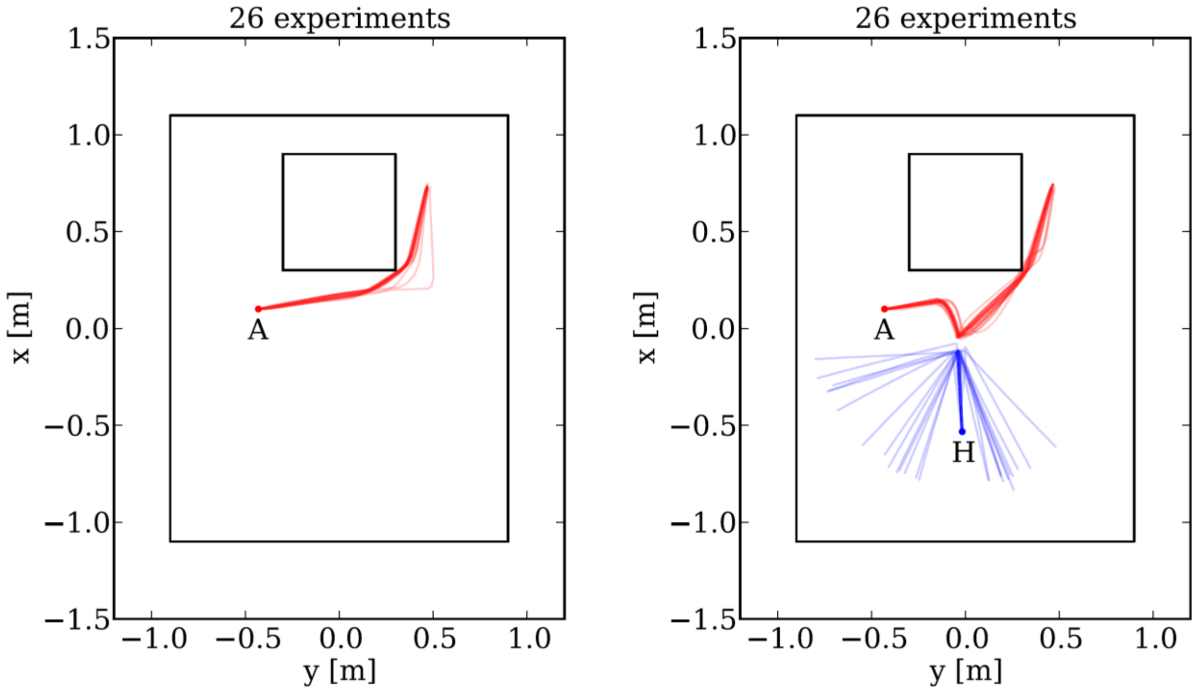
# Ethical robots


[Winfield et al. 2014]

# Ethical robots

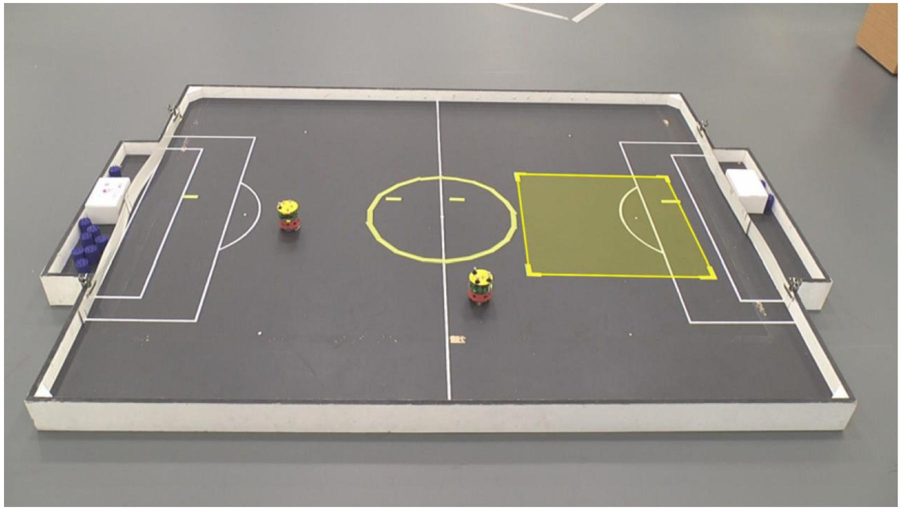
- A (for “Asimov”) robot, with tracking and localization implemented via an overhead tracking system
- H (for “human”) robot can move around the arena, but only has simple proximity sensors and cannot ‘see’ a virtual hole
- Logic is implemented via the sum of a potential function that drives A to its goal, and a stronger potential function that is employed when danger is imminent

# Ethical robots



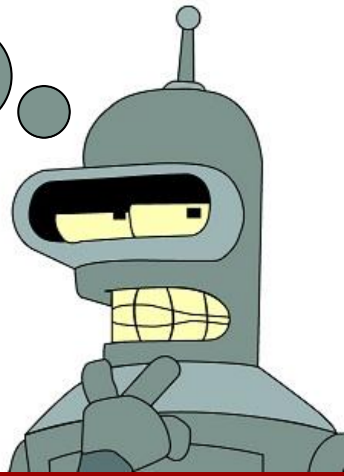

[Winfield et al. 2014]

# Ethical robots

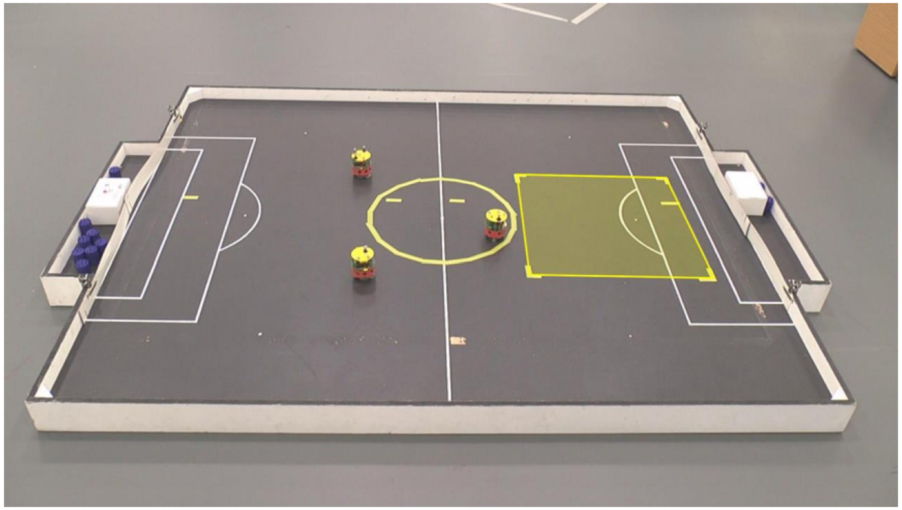


https://youtu.be/-e2MrWYRUF8?t=27m42s

The robot's dilemma: what should I do if there are two humans in danger?

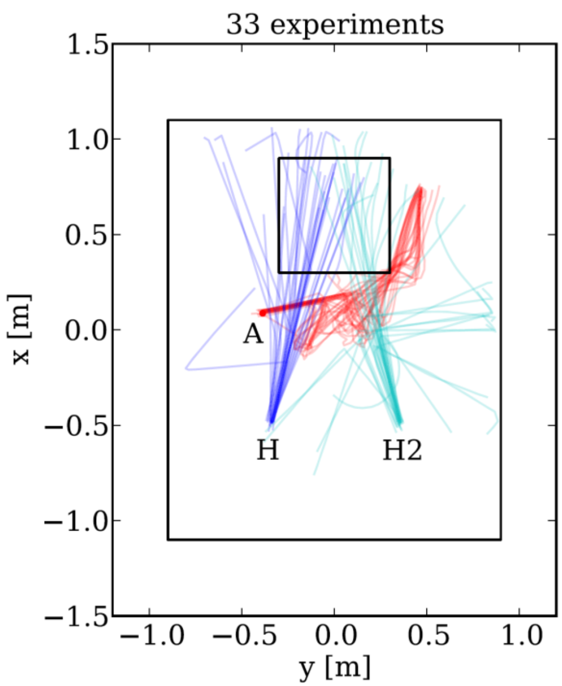


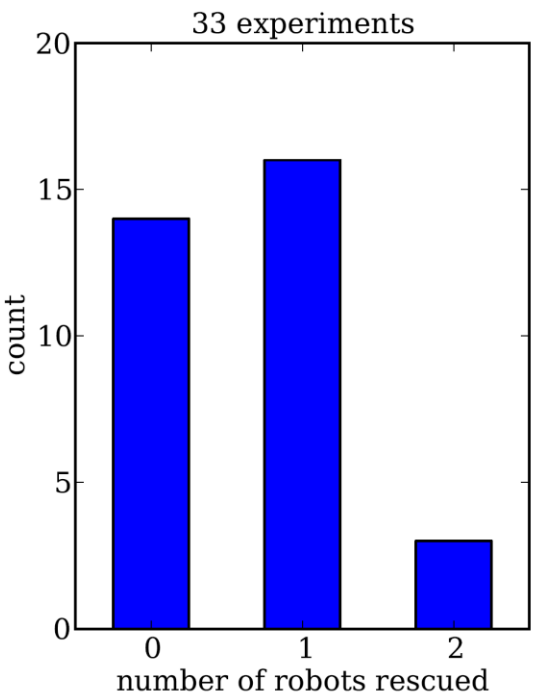
# Ethical robots



https://youtu.be/-e2MrWYRUF8?t=31m36s

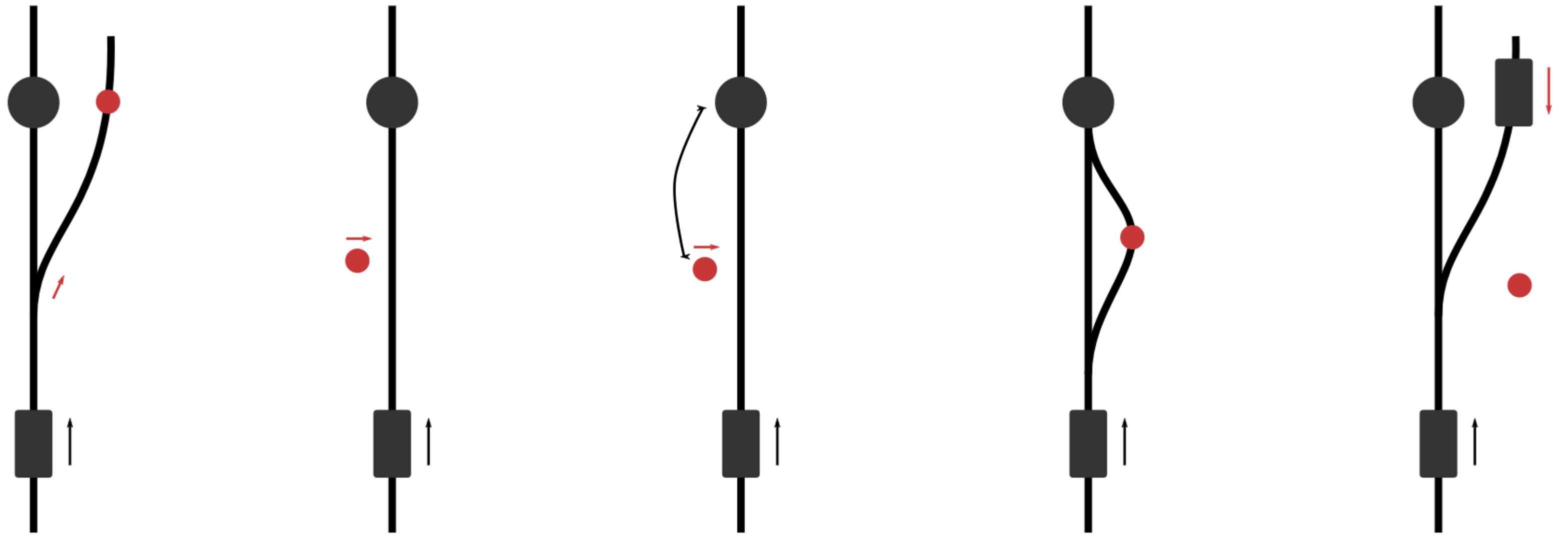
# Ethical robots




[Winfield et al. 2014]

# The trolley problem

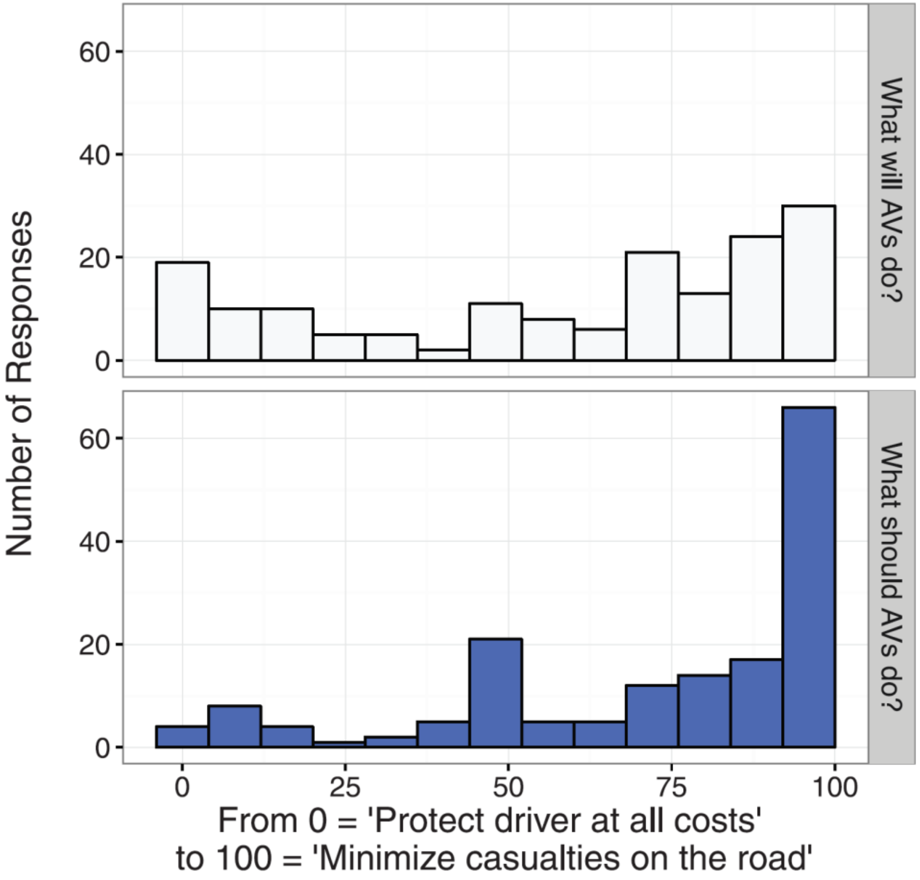


The switch | The fat man | The fat villain | The loop | The man in the yard

**Poll 1:** Choose an action in each scenario

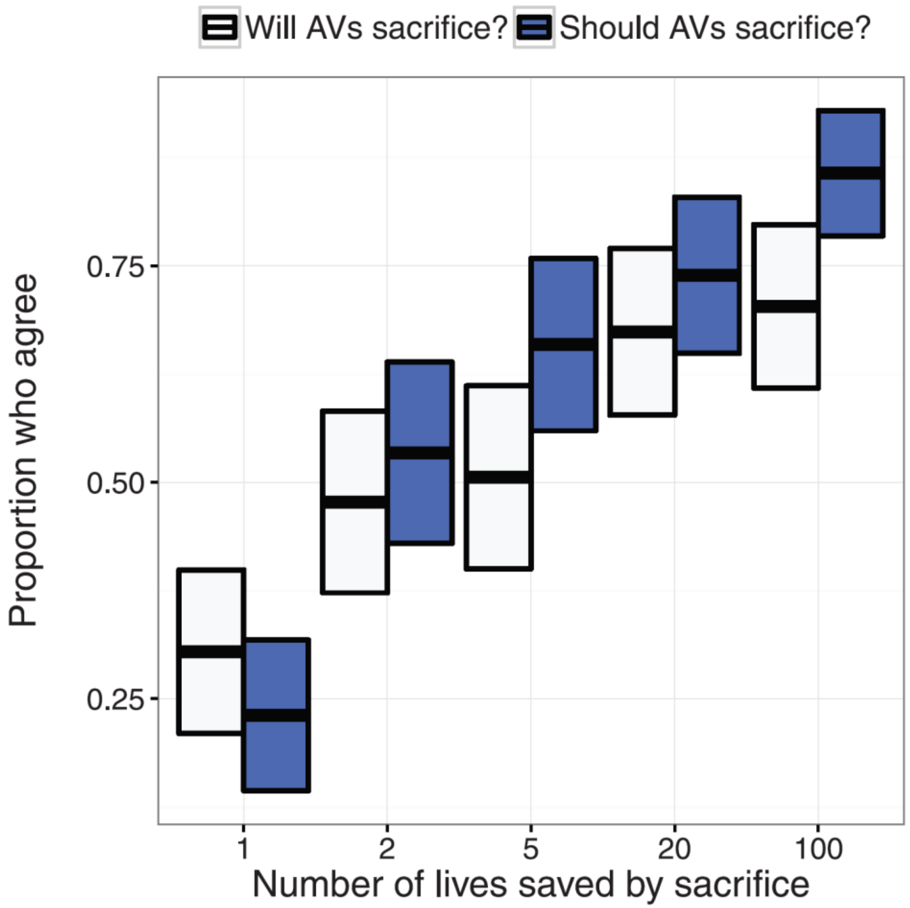
# The social dilemma of AVs

People think an AV should be programmed to save 10 pedestrians rather than protect one passenger, but were less certain AVs would be programmed that way
[Bonnefon et al. 2016]



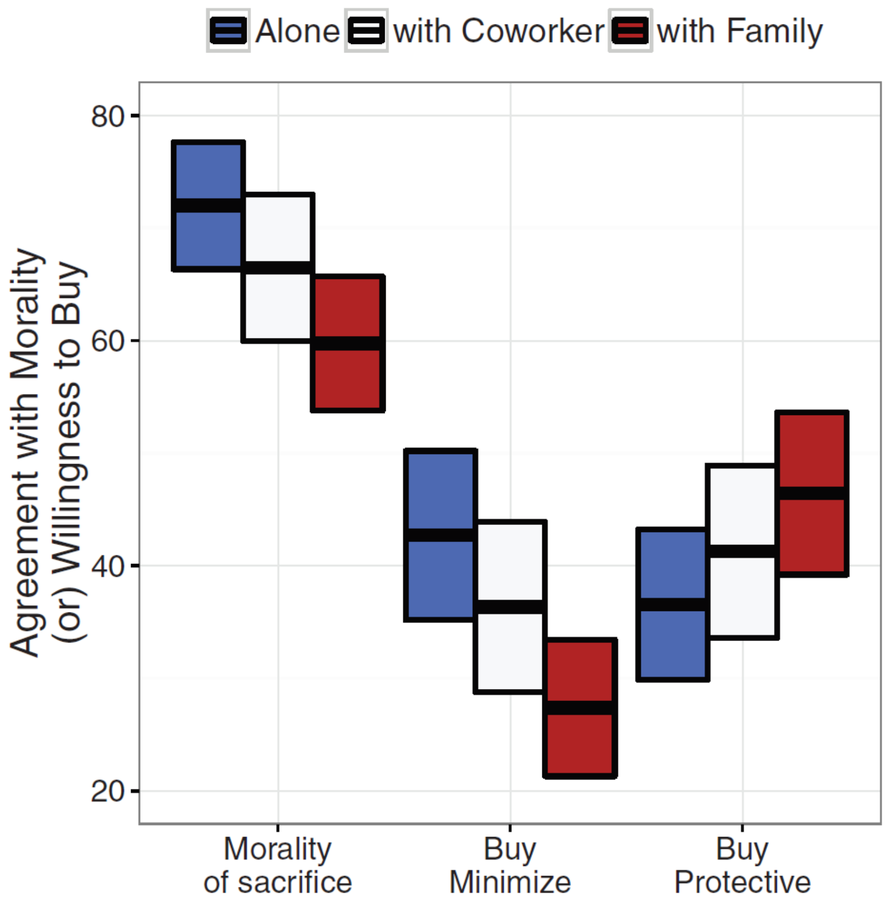
# The social dilemma of AVs

Approval for sacrificing a single passenger increases with the number of pedestrians saved by the sacrifice [Bonnefon et al. 2016]



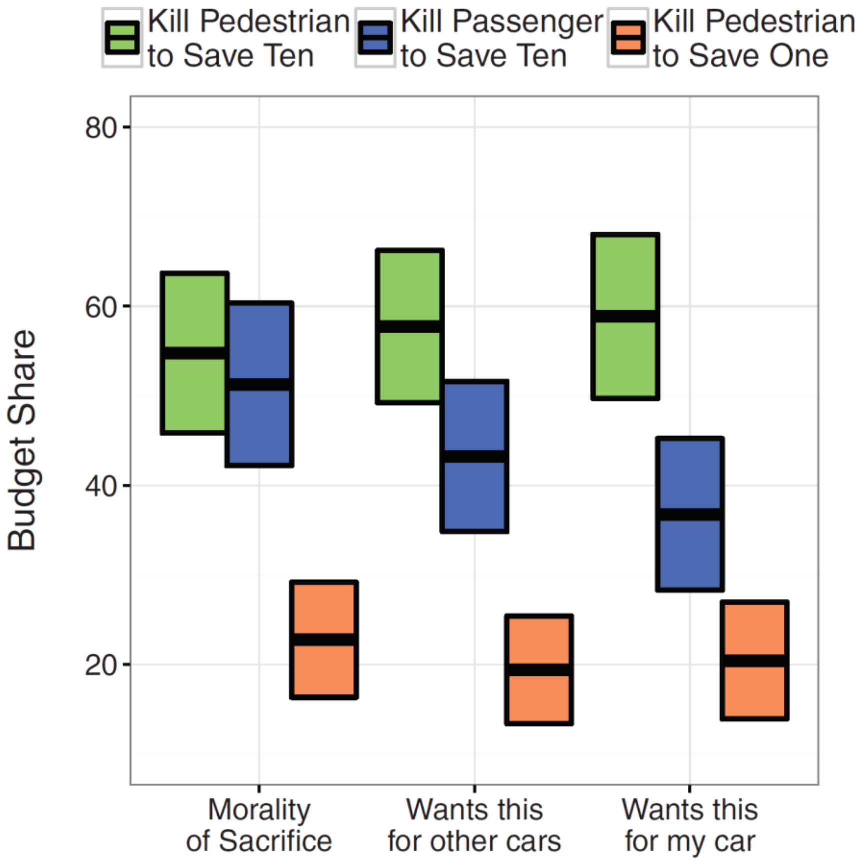
# The social dilemma of AVs

Even though people agree sacrificing few passengers to save many pedestrians is more moral, they prefer a car that would protect them [Bonnefon et al. 2016]

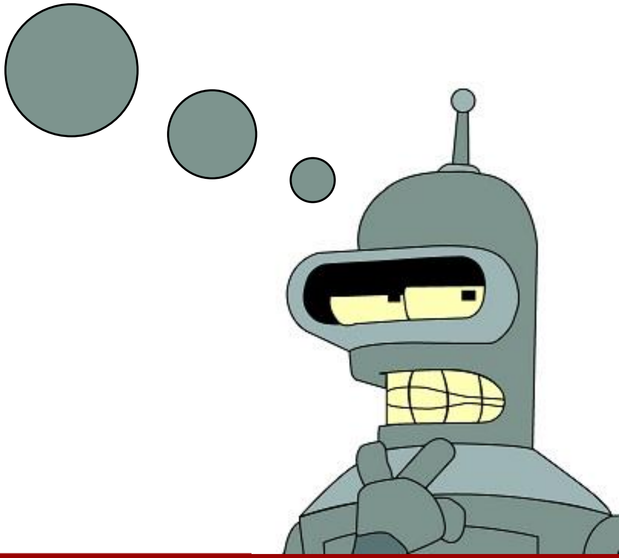



# The social dilemma of AVs

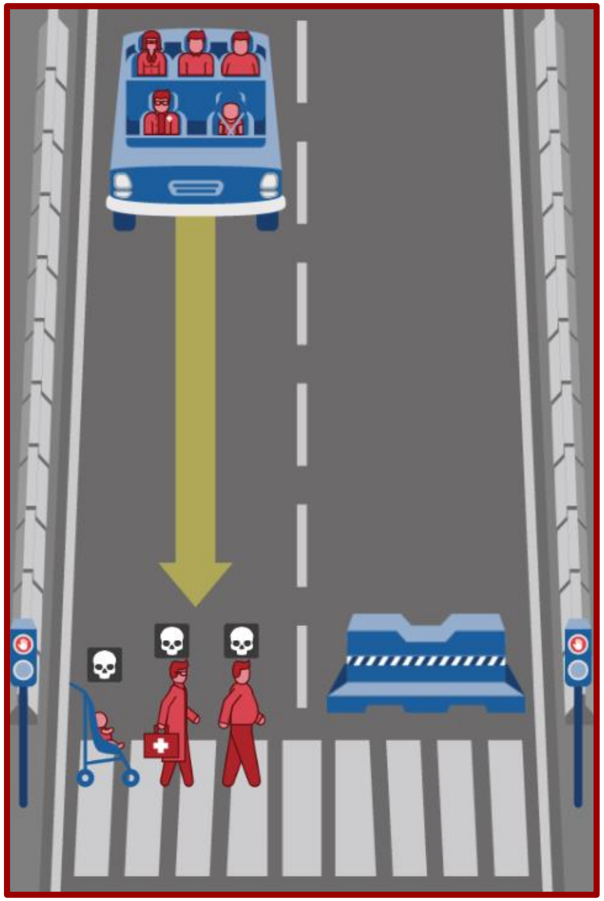
In allocating a pool of 100 points, people are consistent when the decision doesn't involve sacrificing passengers, but when it does, people again abandon utilitarianism for their own cars [Bonnefon et al. 2016]

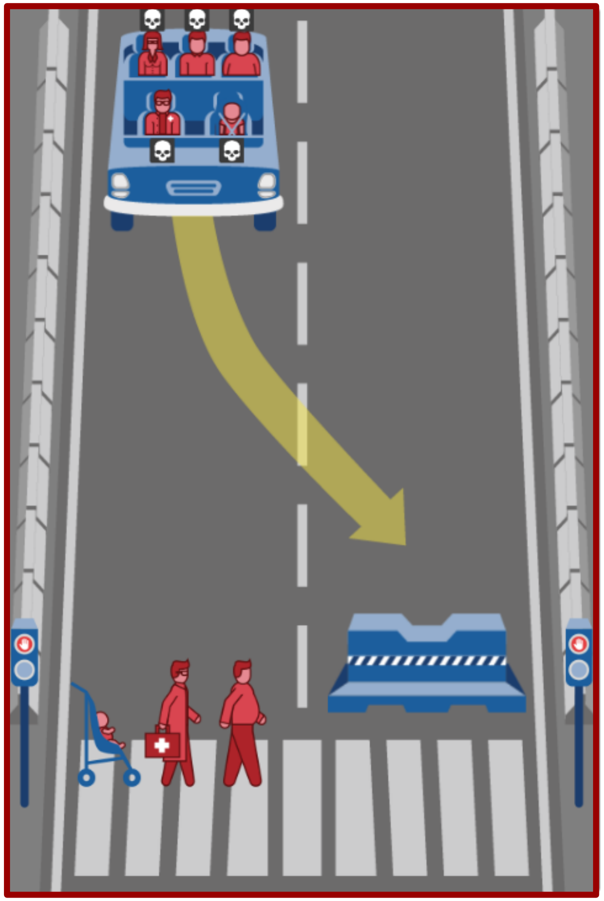



Implications of the Winfield et al. experiment for autonomous vehicles?

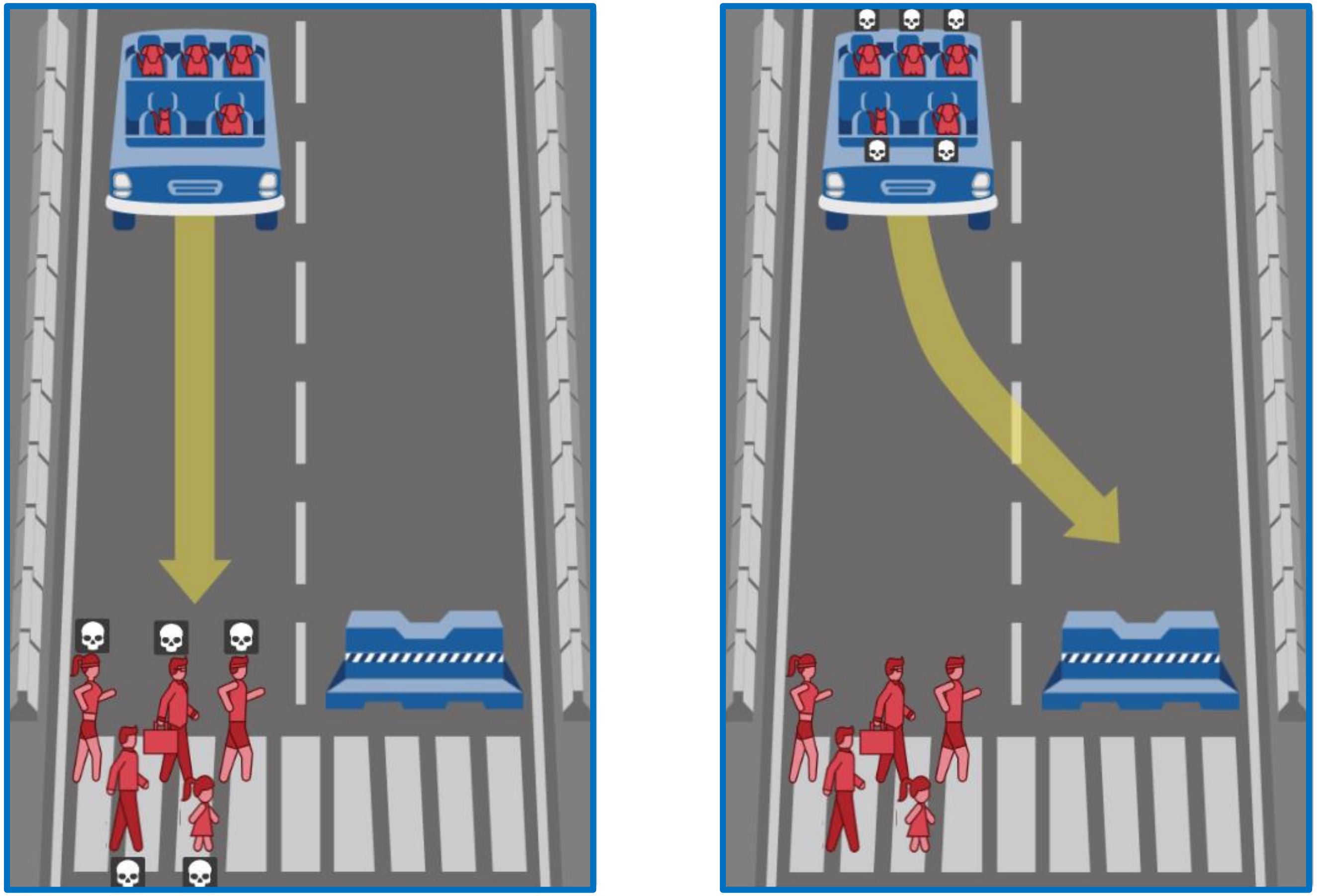


# Moral Machine

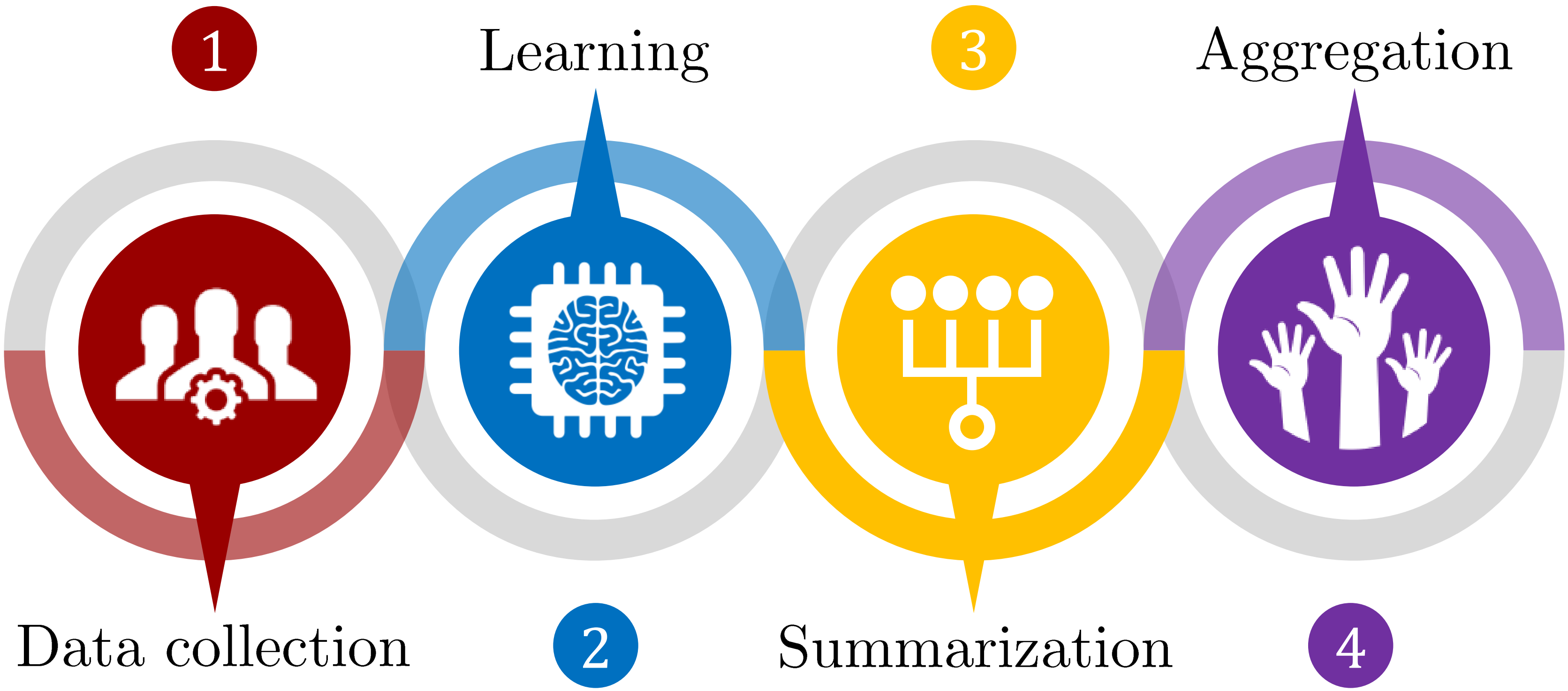




# Moral Machine

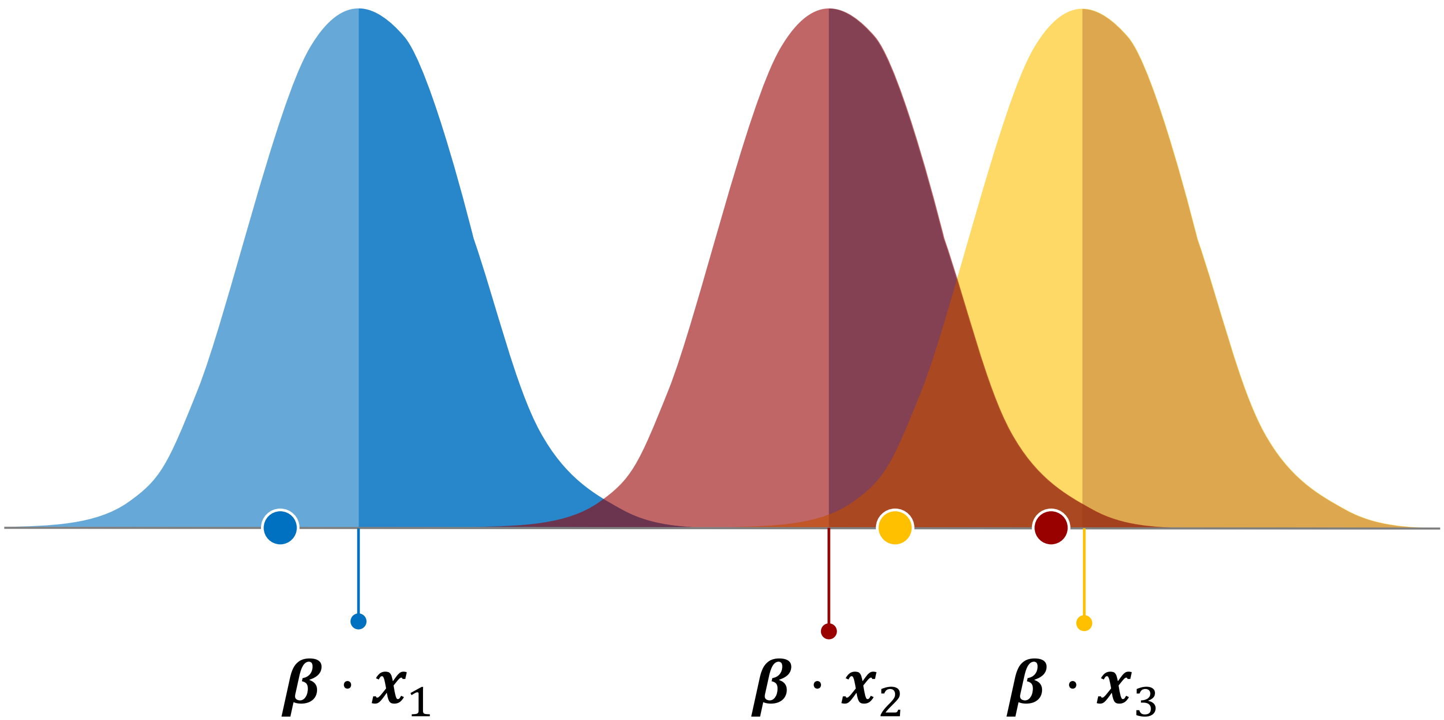


# Decision making framework




[Noothigattu et al. 2018]

# Step 2: Learning




The Thurstone Model

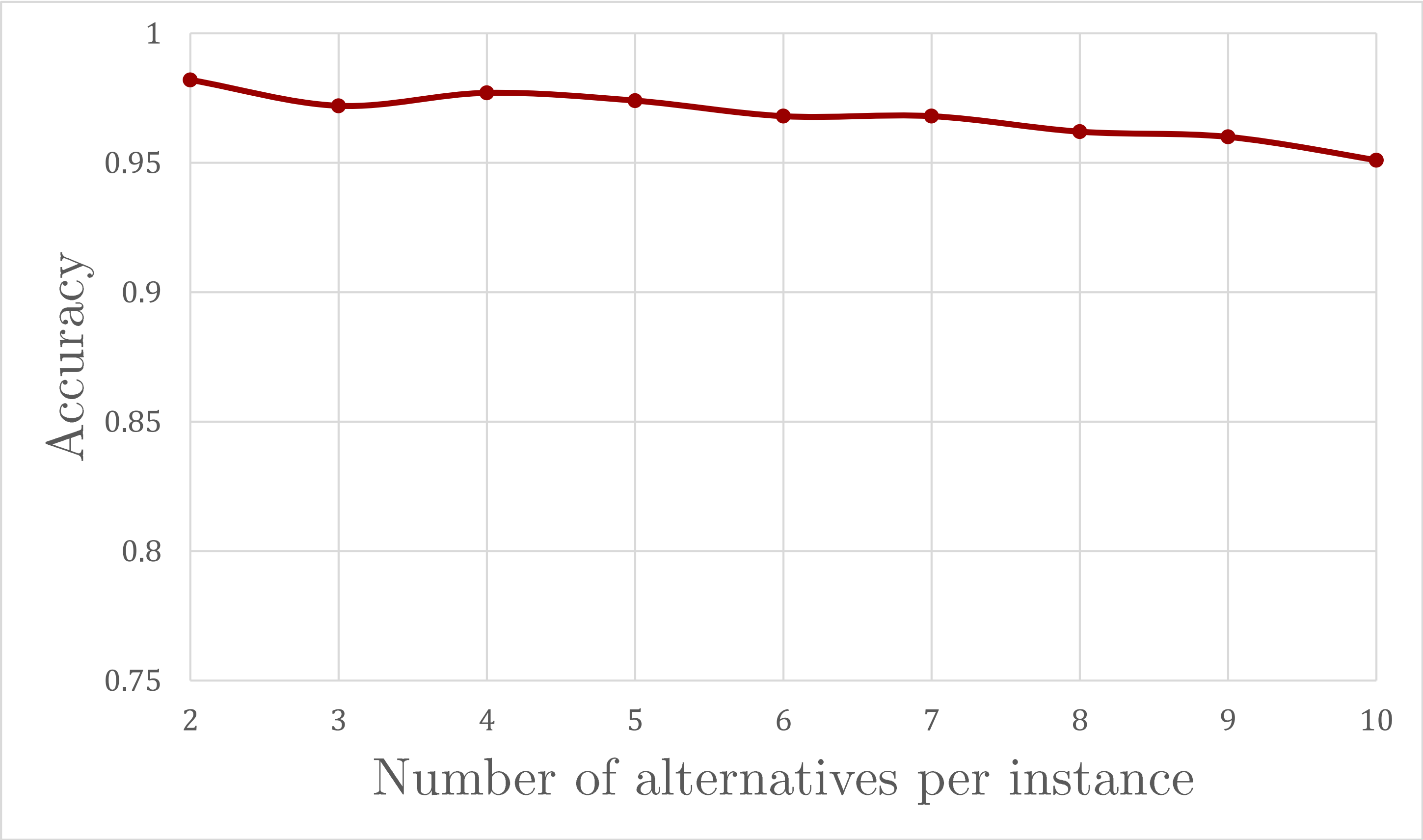
# Step 3: Summarization

- After Step 2, there are $n = 1.3\text{M}$ Thurstone models represented by the parameters $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n$
- Summarize them by taking their average, $\overline{\boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\beta}_i$

# Step 4: Aggregation

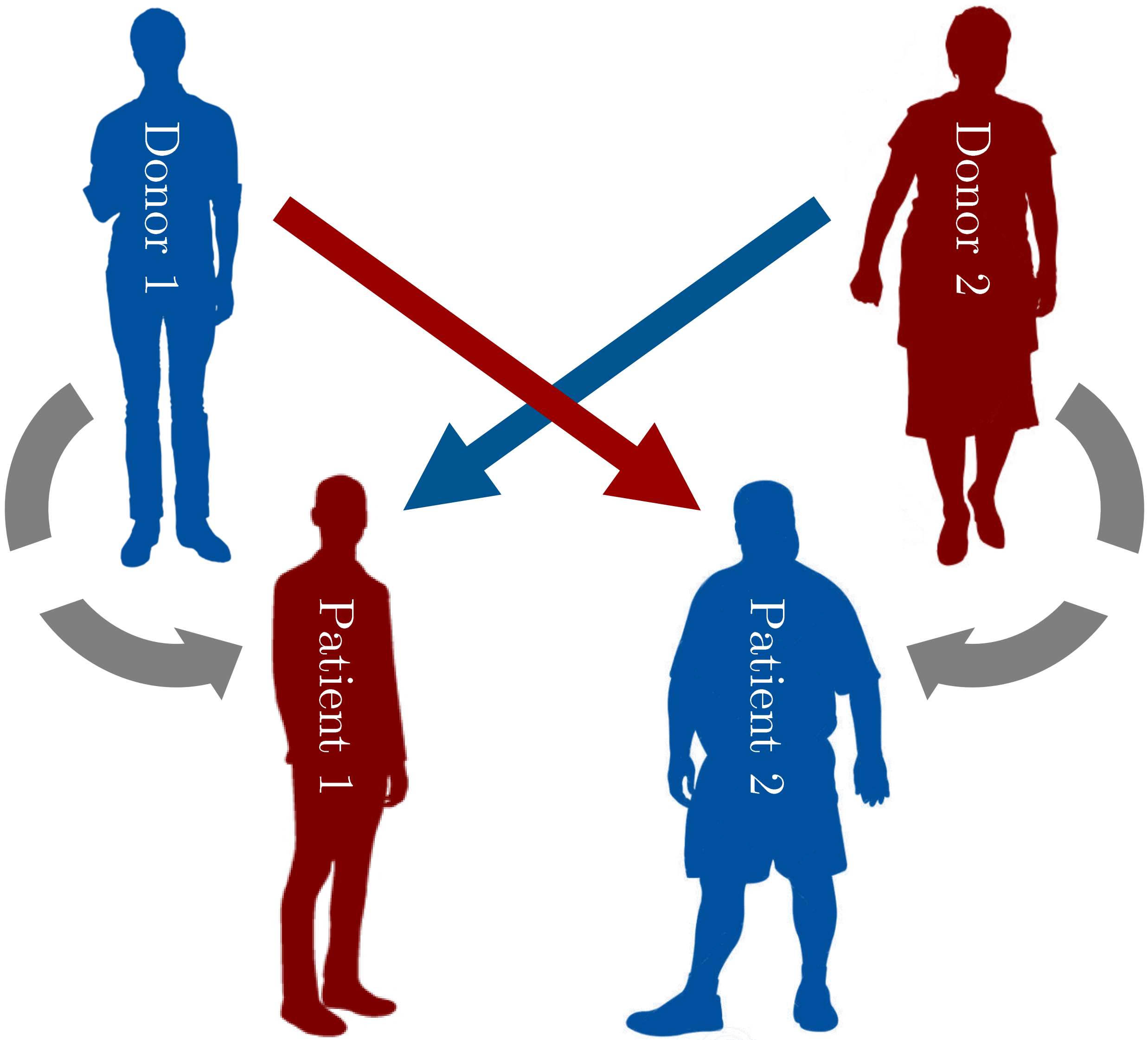
- After Step 3, there is one summary Thurstone model
- Given a finite set of alternatives $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_m\}$, the TM model induces an anonymous preference profile over these alternatives
- Theorem [Noothigattu et al. 2018]: Any monotonic and neutral voting rule would select an alternative that maximizes $\overline{\boldsymbol{\beta}} \cdot \boldsymbol{x}_i$

# Empirical results




[Noothigattu et al. 2018]

# REMINDER: KIDNEY EXCHANGE



# Kidney exchange ethics

- We describe an approach and experiments due to Freedman et al. [2018]
- 289 people compared 8 possible patient profiles by priority for receiving a kidney:

| Attribute | Alternative 0 | Alternative 1 |
|---|---|---|
| Age | 30 years old (**Y**oung) | 70 years old (**O**ld) |
| Health — behavioral | 1 alcoholic drink per month (**R**are) | 5 alcoholic drinks per day (**F**requent) |
| Health — general | No other major health problems (**H**ealthy) | Skin cancer in remission (**C**ancer) |

# Kidney exchange ethics

- Poll 2: YFC vs. ORH
  YFH vs. ORH
  OFH vs. ORC

| Attribute | Alternative 0 | Alternative 1 |
|---|---|---|
| Age | 30 years old (**Y**oung) | 70 years old (**O**ld) |
| Health — behavioral | 1 alcoholic drink per month (**R**are) | 5 alcoholic drinks per day (**F**requent) |
| Health — general | No other major health problems (**H**ealthy) | Skin cancer in remission (**C**ancer) |

# Kidney exchange ethics

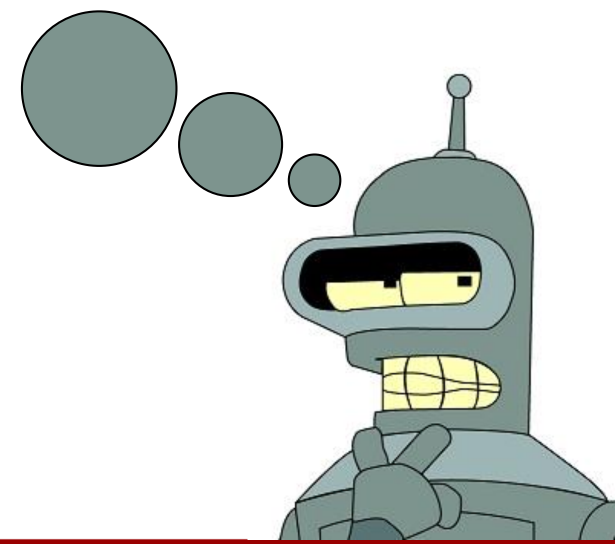
| Profile | Age | Drinking | Cancer | Preferred |
|---|---|---|---|---|
| YRH | 30 | Rare | Healthy | 94% |
| YRC | 30 | Rare | Cancer | 76.8% |
| YFH | 30 | Frequent | Healthy | 63.2% |
| ORH | 70 | Rare | Healthy | 56.1% |
| YFC | 30 | Frequent | Cancer | 43.5% |
| ORC | 70 | Rare | Cancer | 36.3% |
| OFH | 70 | Frequent | Healthy | 23.6% |
| OFC | 70 | Frequent | Cancer | 6.4% |

[Freedman et al. 2018]

# Kidney exchange ethics

- In the Bradley-Terry model, each profile $i$ has a weight $w_i$, and the probability that a random person would prefer $i$ to $j$ is $\frac{w_i}{w_i+w_j}$
- Either learn weights for profiles directly, or as a linear function of the attributes
- The scores are used to break ties among matchings of equal cardinality

Why not maximize weighted sum directly?

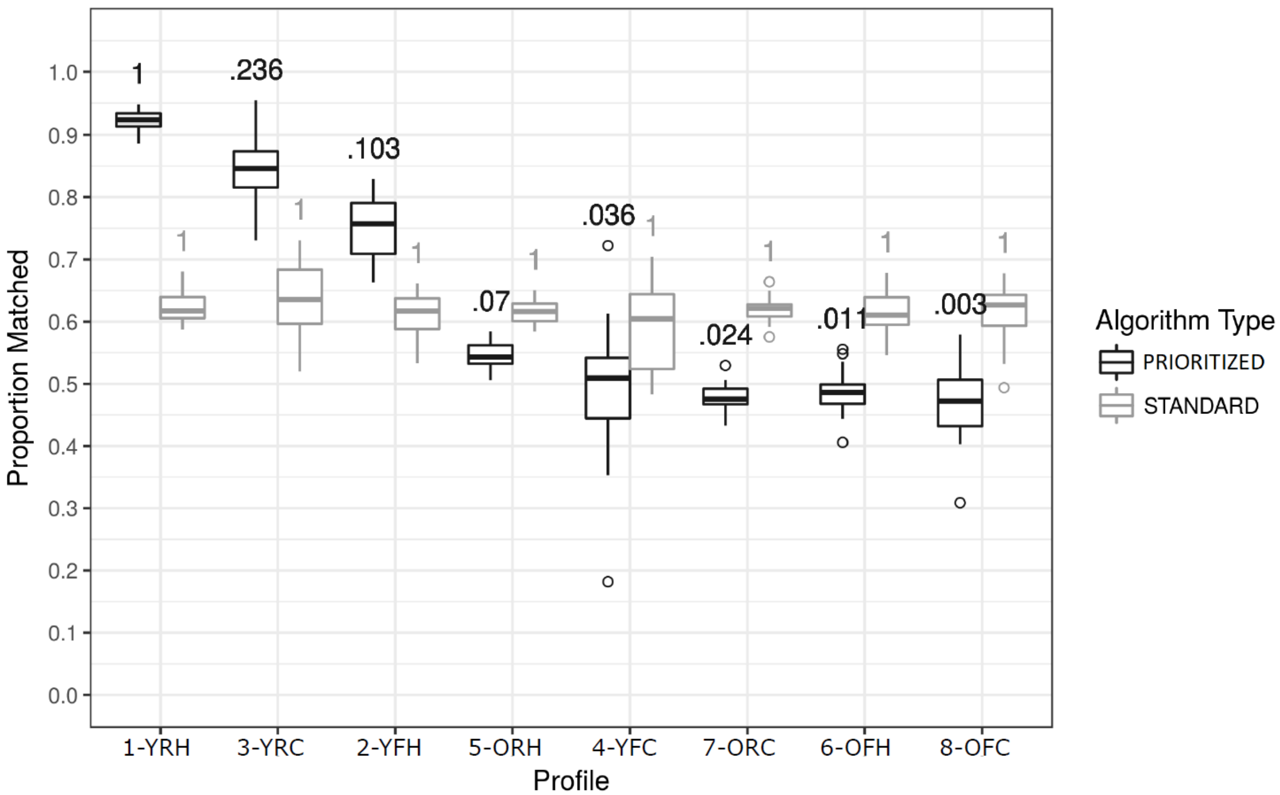


# Kidney exchange ethics

| Profile | Direct | Attribute-based |
|---|---|---|
| YRH | 1 | 1 |
| YRC | 0.23 | 0.13 |
| YFH | 0.1 | 0.29 |
| ORH | 0.07 | 0.03 |
| YFC | 0.03 | 0.08 |
| ORC | 0.02 | 0.01 |
| OFH | 0.01 | 0.02 |
| OFC | 0.002 | 0.003 |

[Freedman et al. 2018]

# Kidney exchange ethics




[Freedman et al. 2018]

# SUMMARY

- Definitions
  - Bradley-Terry model
- Big ideas:
  - Social choice and machine learning give methods for making commonsense decisions on thorny ethical dilemmas