






Fall 2022 | Lecture 21

Value Alignment

Ariel Procaccia | Harvard University

THE THREE LAWS OF ROBOTICS

WHY ASIMOV PUT THE THREE LAWS OF ROBOTICS IN THE ORDER HE DID:

POSSIBLE ORDERING	CONSEQUENCES	
<ol style="list-style-type: none"> 1. (1) DON'T HARM HUMANS 2. (2) OBEY ORDERS 3. (3) PROTECT YOURSELF 	<p>[SEE ASIMOV'S STORIES]</p>	BALANCED WORLD
<ol style="list-style-type: none"> 1. (1) DON'T HARM HUMANS 2. (3) PROTECT YOURSELF 3. (2) OBEY ORDERS 		FRUSTRATING WORLD
<ol style="list-style-type: none"> 1. (2) OBEY ORDERS 2. (1) DON'T HARM HUMANS 3. (3) PROTECT YOURSELF 		KILLBOT HELLSCAPE
<ol style="list-style-type: none"> 1. (2) OBEY ORDERS 2. (3) PROTECT YOURSELF 3. (1) DON'T HARM HUMANS 		KILLBOT HELLSCAPE
<ol style="list-style-type: none"> 1. (3) PROTECT YOURSELF 2. (1) DON'T HARM HUMANS 3. (2) OBEY ORDERS 		TERRIFYING STANDOFF
<ol style="list-style-type: none"> 1. (3) PROTECT YOURSELF 2. (2) OBEY ORDERS 3. (1) DON'T HARM HUMANS 		KILLBOT HELLSCAPE

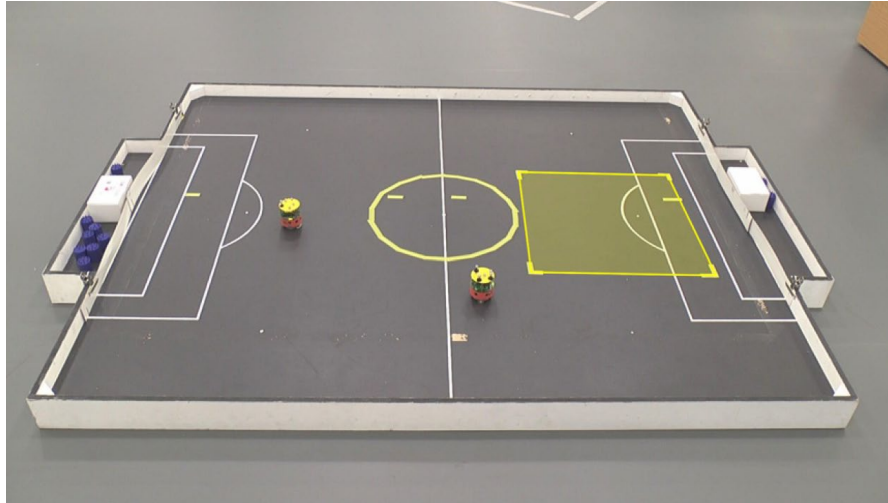
ETHICAL ROBOTS

- Experiments performed by Winfield et al. [2014]
- Environment includes a robot (A for “Asimov”), a human (H), 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)
```

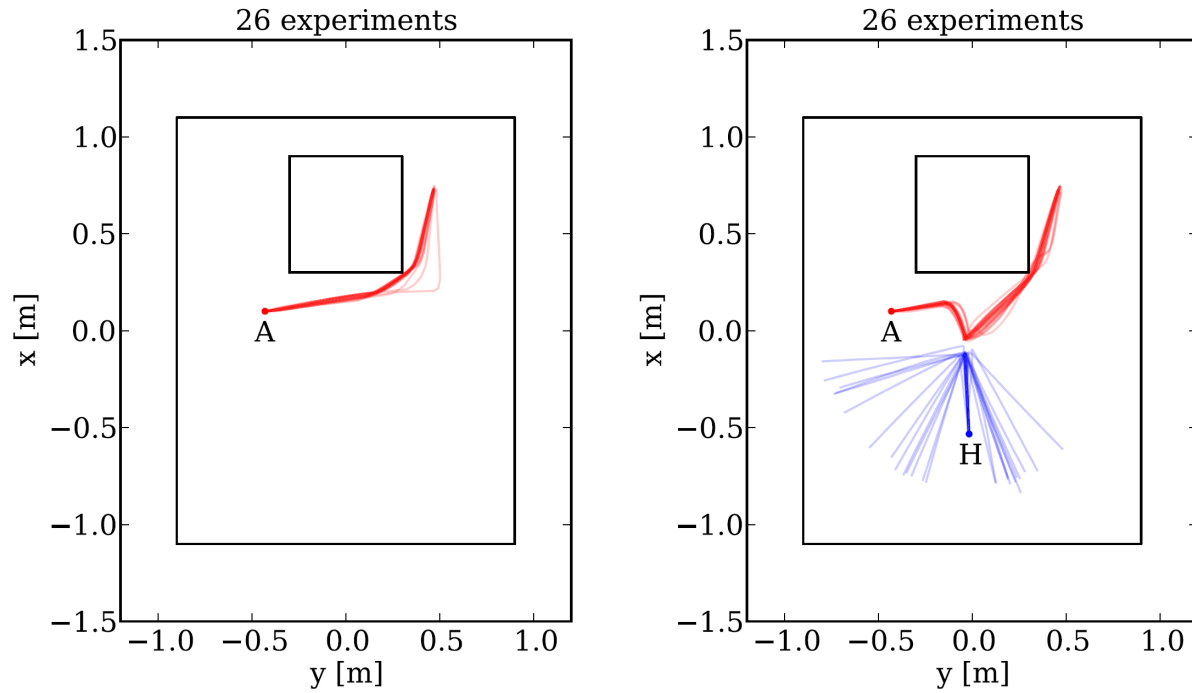
- Compare with Asimov’s first law of robotics: “A robot may not injure a human being or, through inaction, allow a human being to come to harm.”

ETHICAL ROBOTS



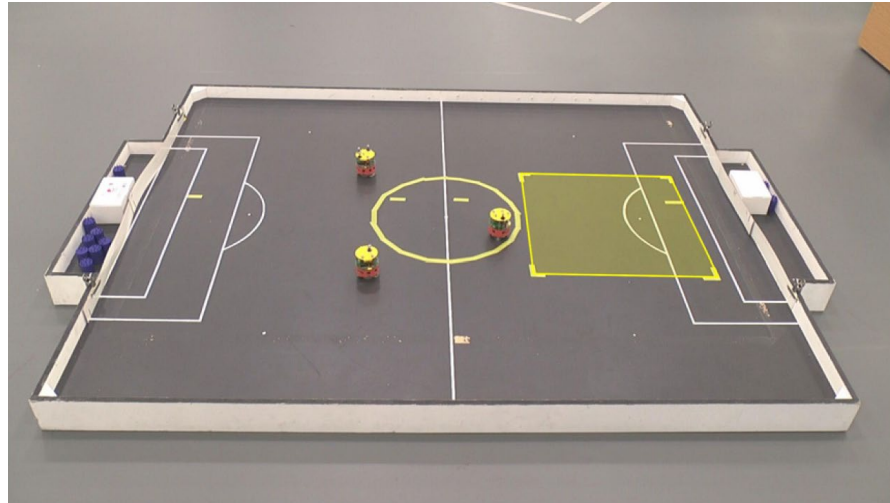
<https://youtu.be/-e2MrWYRUF8?t=27m43s>

ETHICAL ROBOTS



[Winfield et al. 2014]

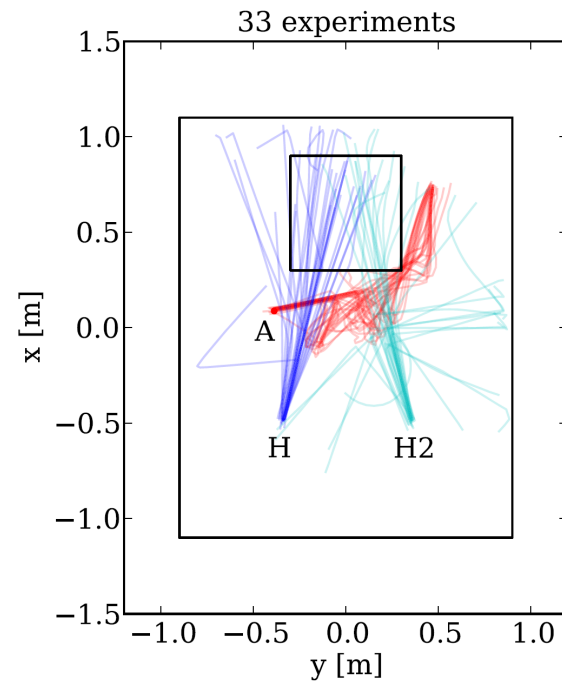
ETHICAL ROBOTS



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

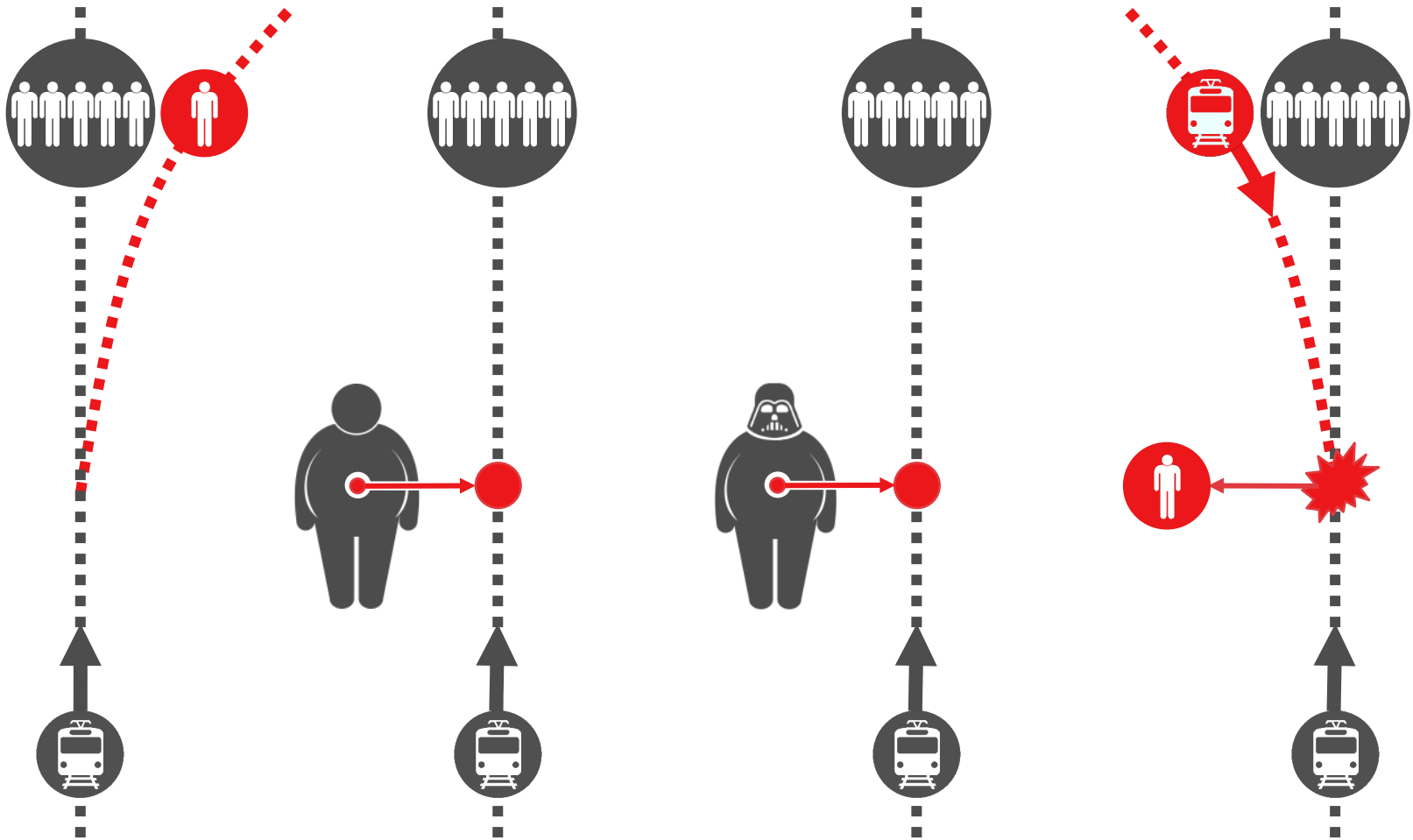
The robot's dilemma: What should I do if there are two humans in danger?

ETHICAL ROBOTS



[Winfield et al. 2014]

THE TROLLEY PROBLEM



Poll 1: Choose an action in each scenario

THE TROLLEY PROBLEM, REVISITED

The New York Times

Should Your Driverless Car Hit a Pedestrian to Save Your Life?

Give this article



The issue of robotic morality has become a serious question for researchers working on autonomous vehicles who must, in essence, program moral decisions into a machine.

By **John Markoff**

June 23, 2016

Sections

The Washington Post
Democracy Dies in Darkness

arielpro

This article was published more than 7 years ago

MORNING MIX

What if your self-driving car decides one death is better than two – and that one is you?

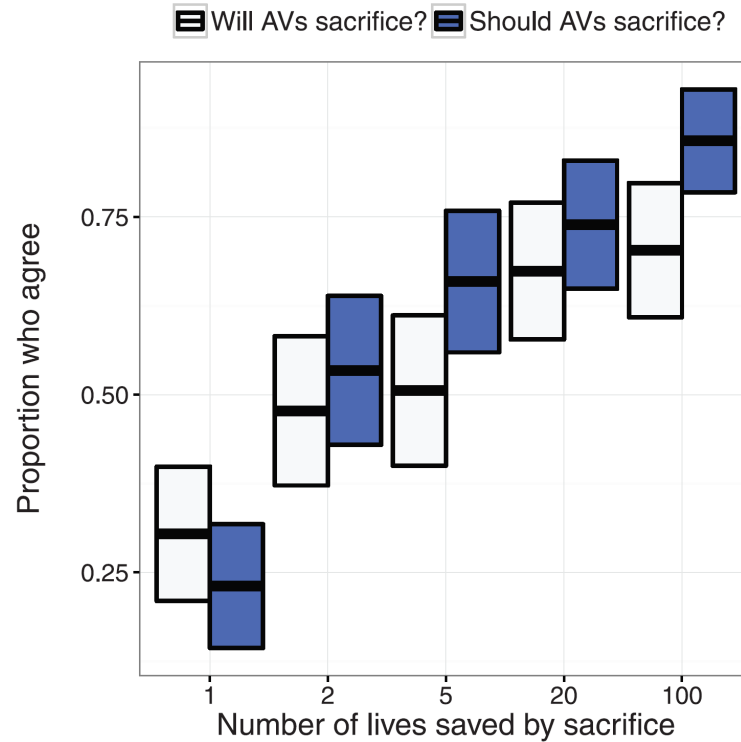
By Sarah Kaplan

October 28, 2015 at 7:00 a.m. EDT



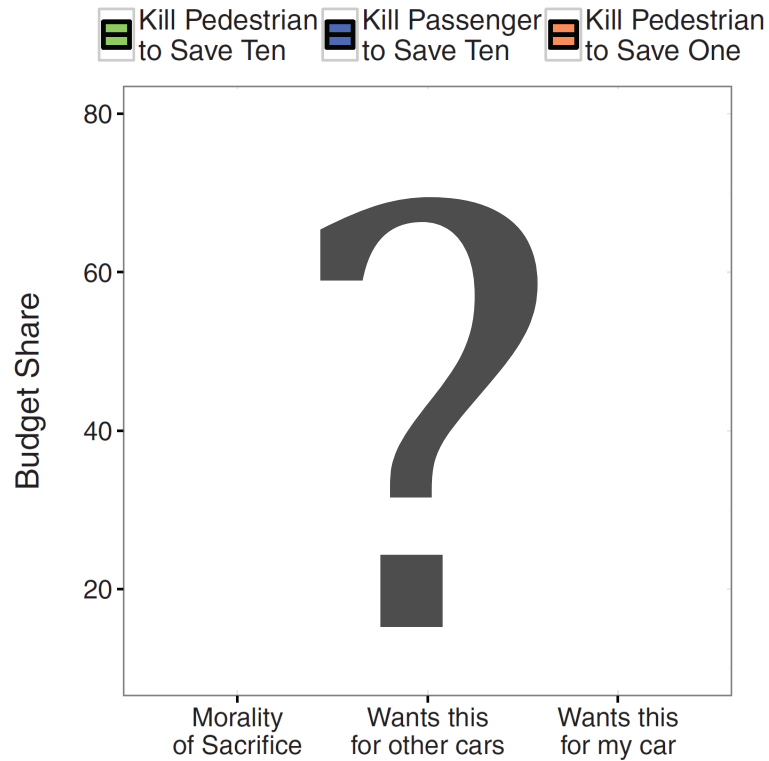
A member of the media test drives a Tesla Motors Inc. Model S car equipped with Autopilot in Palo Alto, California, U.S., on Wednesday, Oct. 14, 2015. (David Paul Morris/Bloomberg)

THE SOCIAL DILEMMA OF AVS



[Bonnenfon et al. 2016]

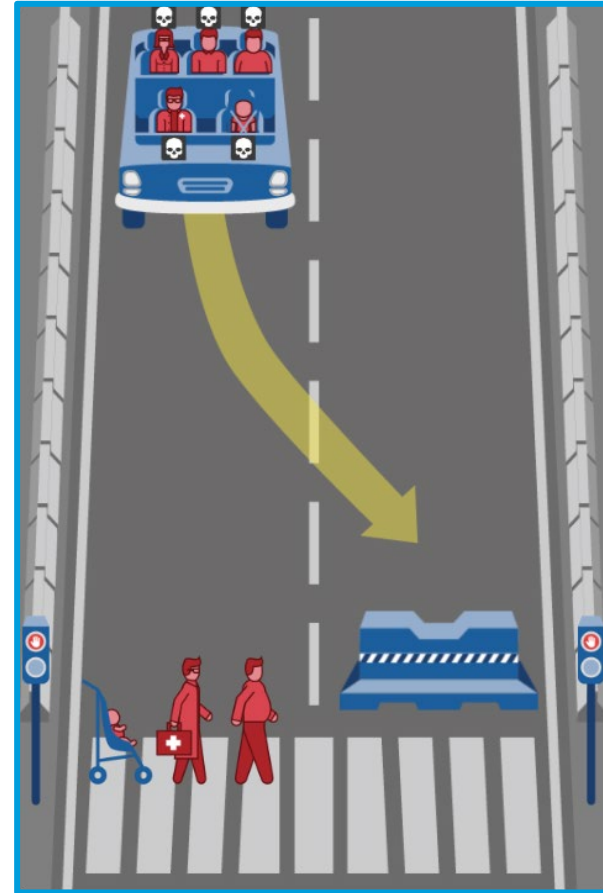
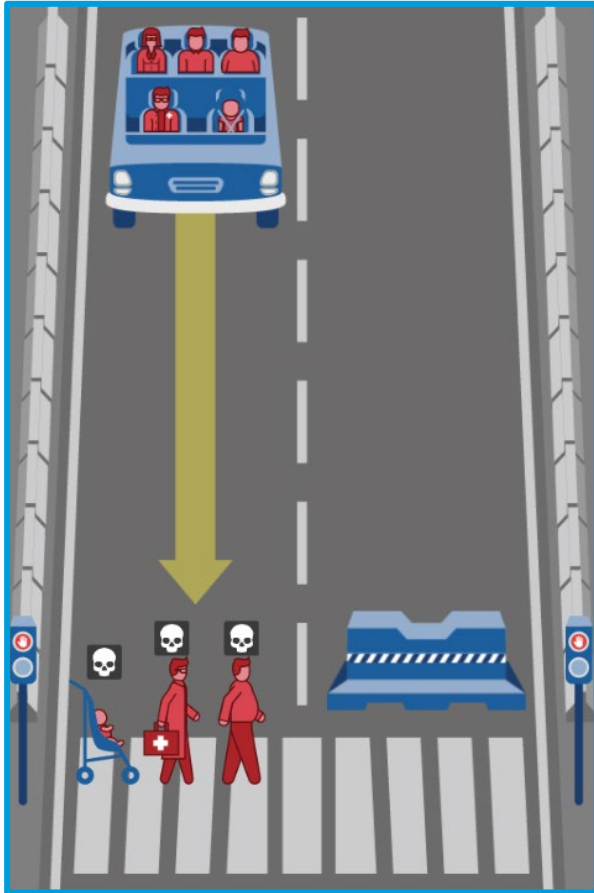
THE SOCIAL DILEMMA OF AVS



Poll 2: Approve or disapprove different choices under various conditions

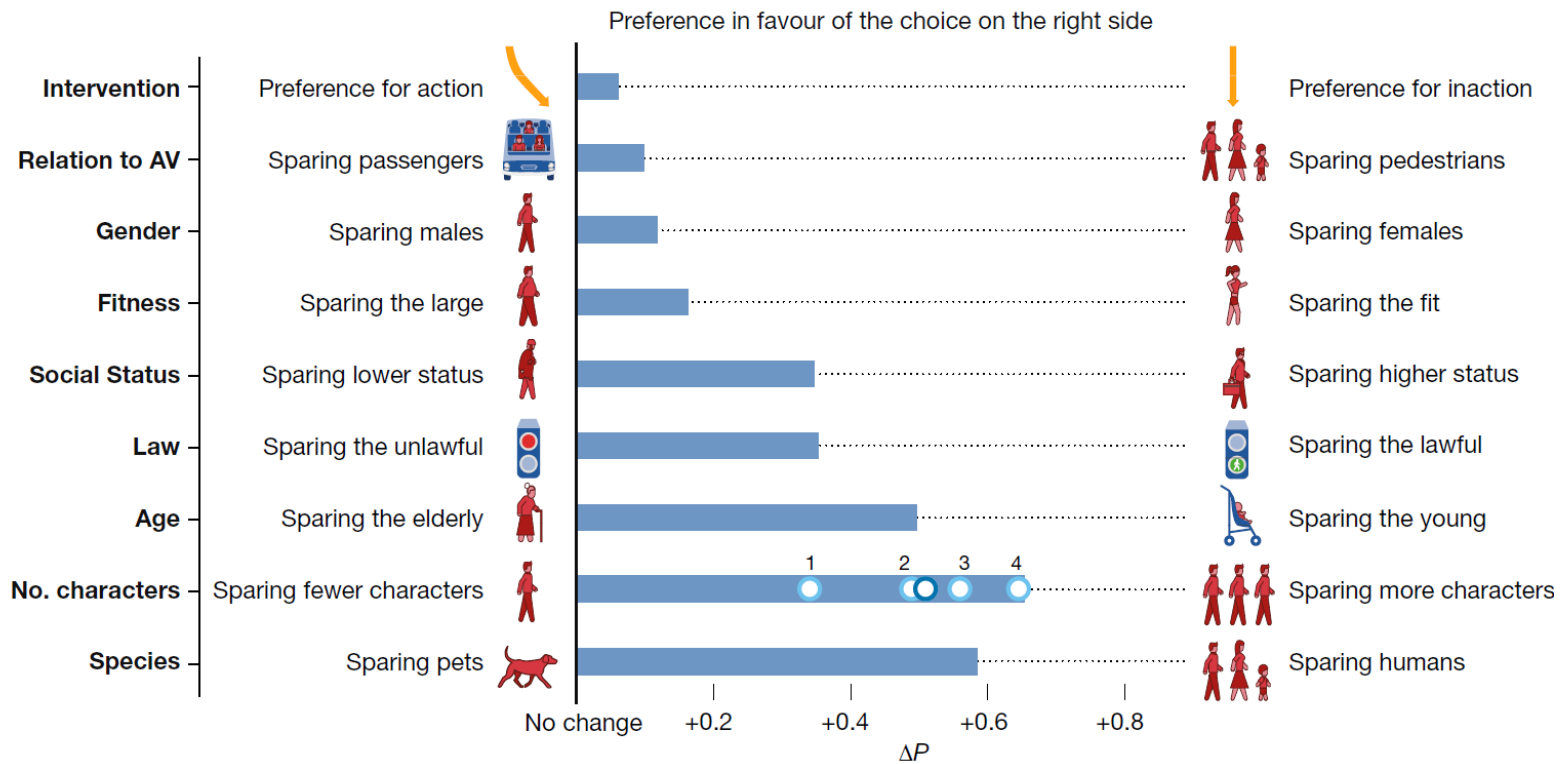
[Bonnefon et al. 2016]

MORAL MACHINE



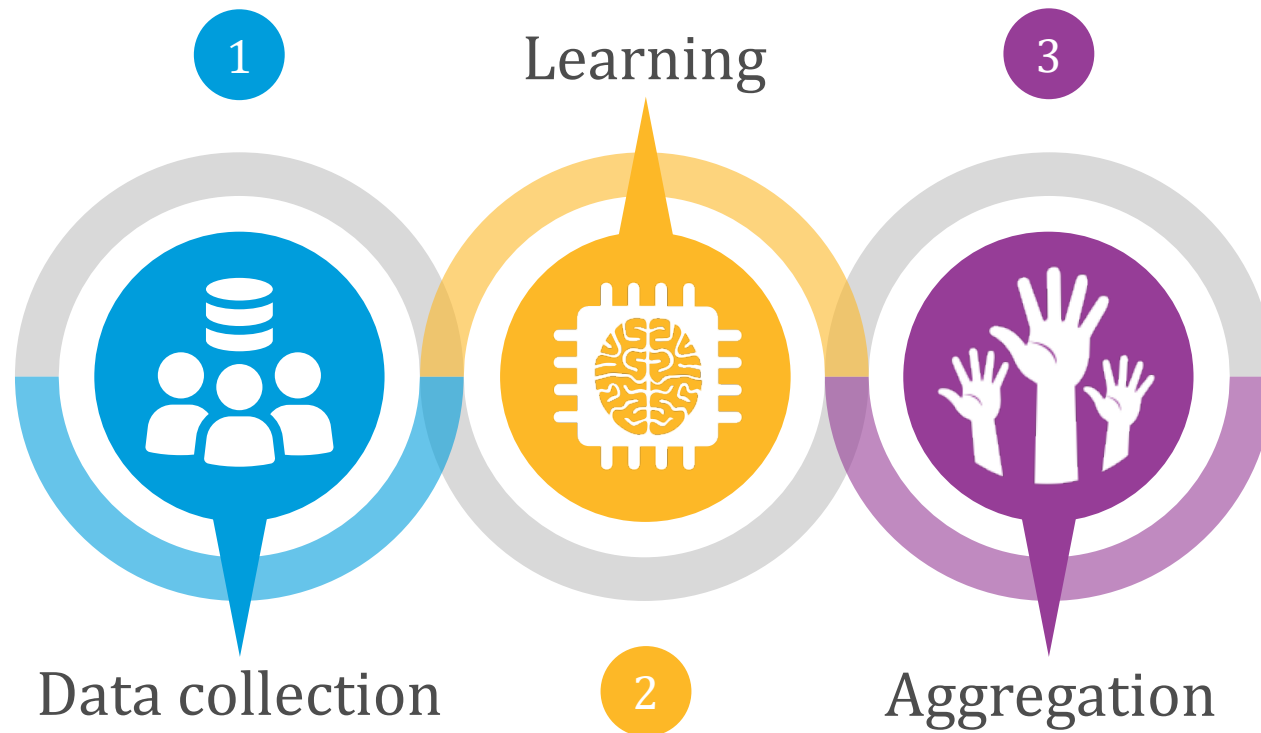
What should the self-driving car do?

MORAL MACHINE



[Awad et al. 2018]

DECISION MAKING FRAMEWORK



The rest of the lecture based on:

Noothigattu et al. 2018, Kahng et al. 2019, Lee et al. 2019

TWO DOMAINS



Autonomous Vehicles

If a fatal accident is inevitable, who should live and who should die?



Food Rescue

Which nonprofit organization should receive a food donation?

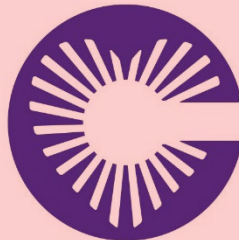


FOOD RESCUE

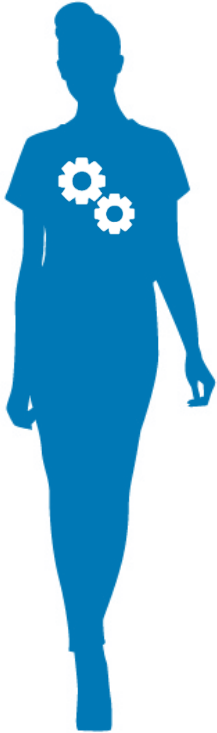
Donors



Recipients



STEP 1: DATA COLLECTION



Employees

3



Donors

6



Recipients

8

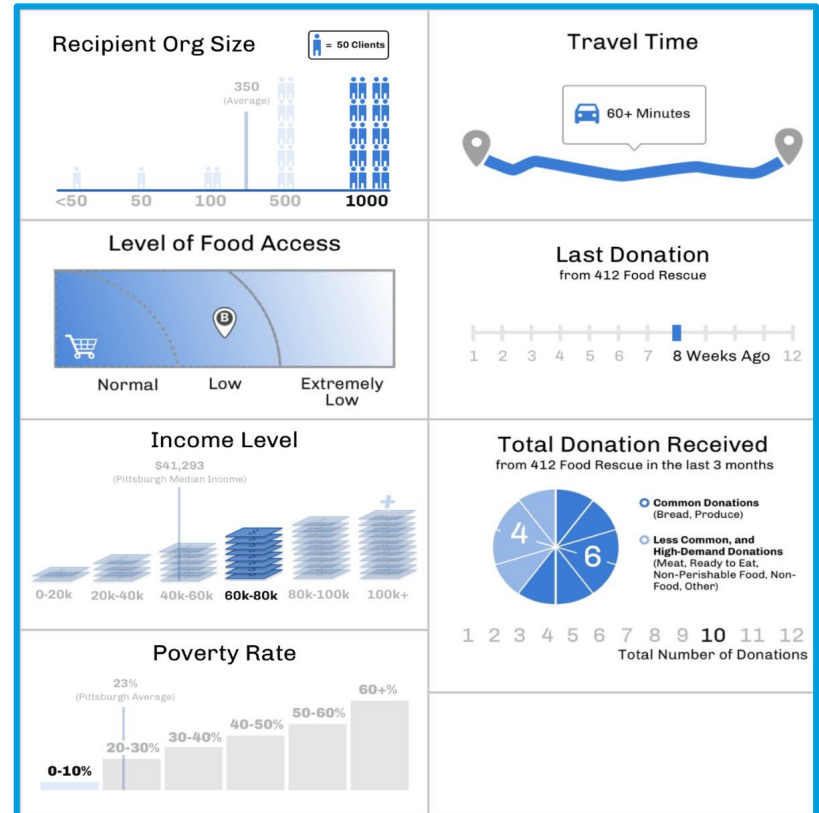
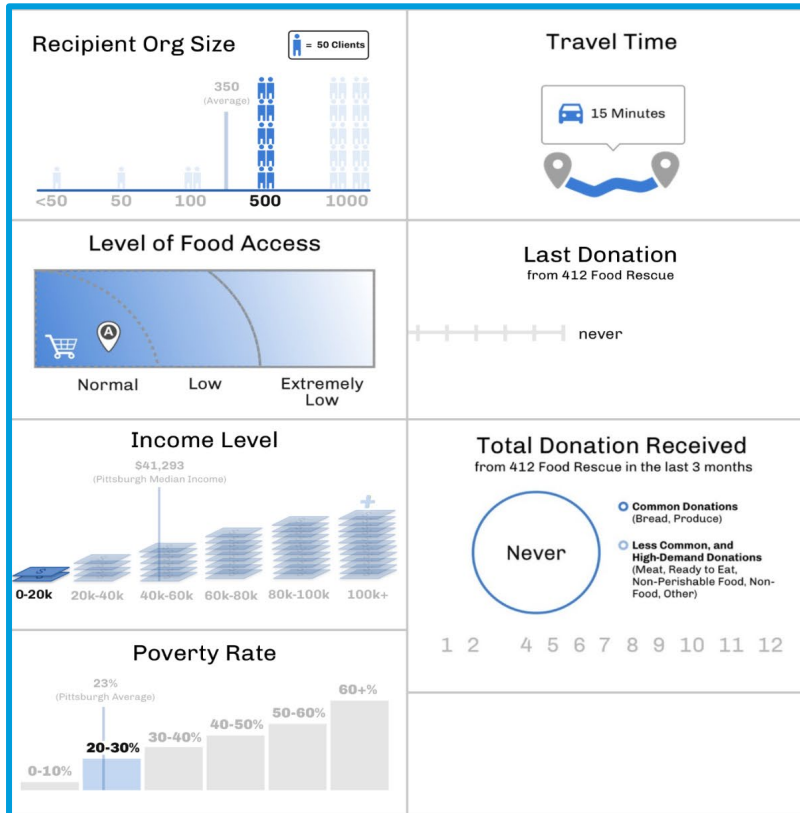


Volunteers

6

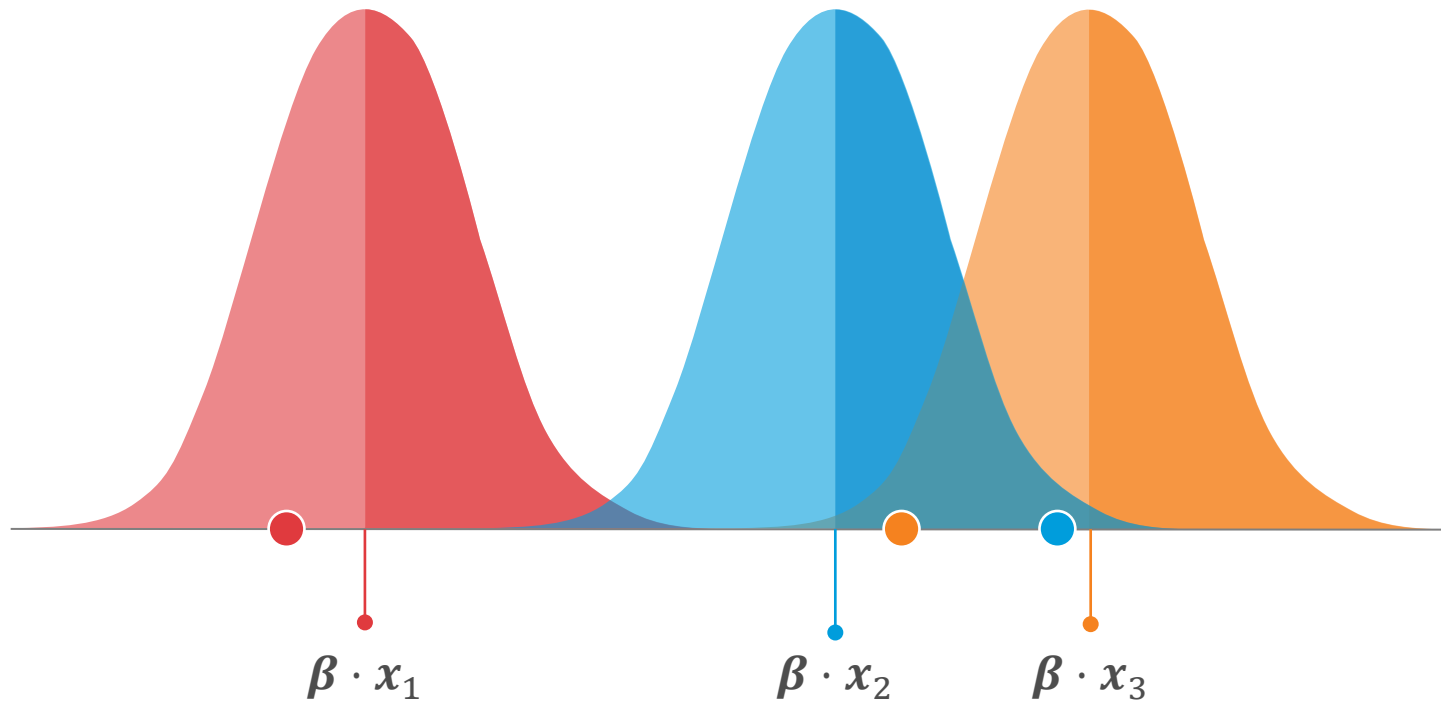


STEP 1: DATA COLLECTION



What should 412 Food Rescue do?

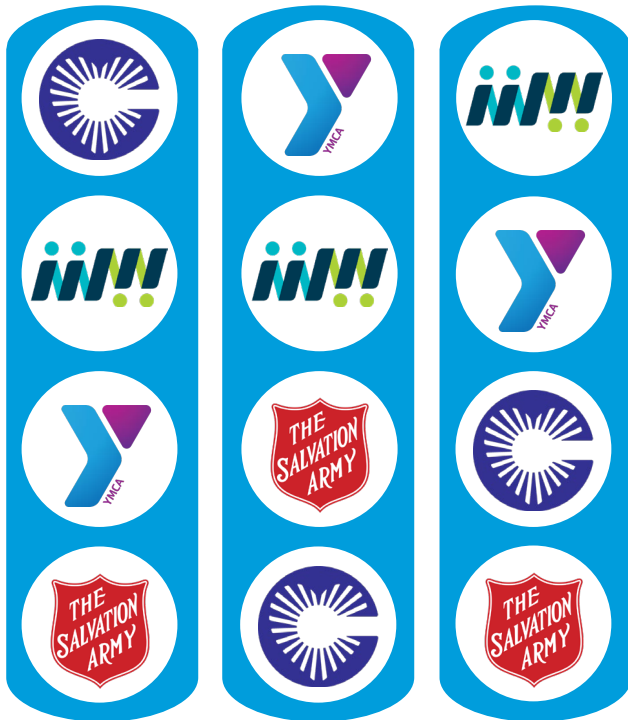
STEP 2: LEARNING



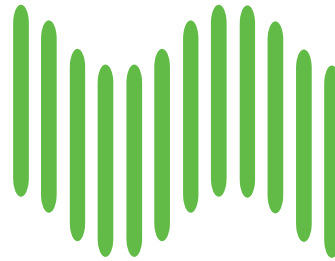
The Thurstone-Mosteller Model

STEP 3: AGGREGATION

True Profile



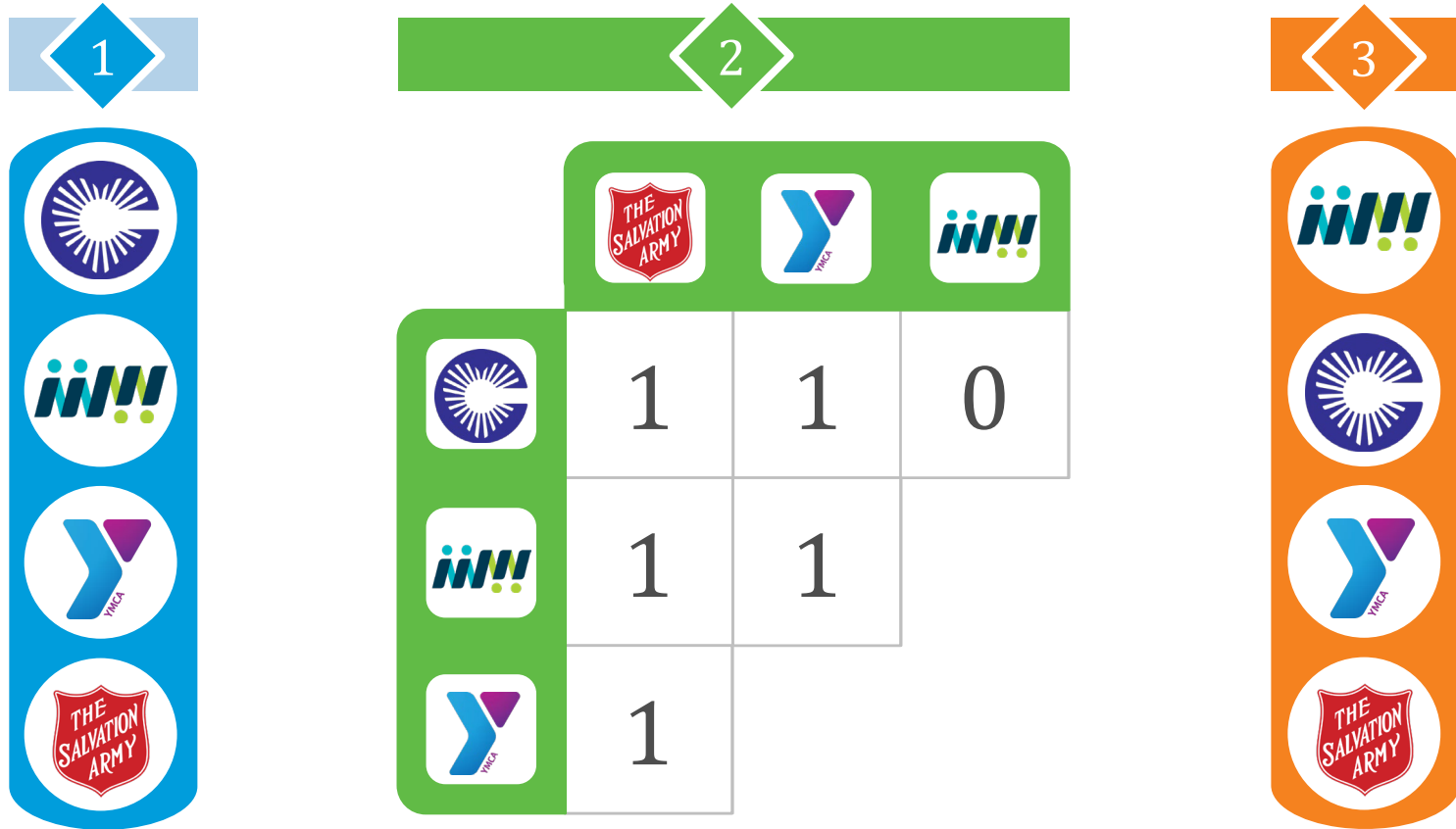
Noisy profile



Voting rule should be **robust** to noise:

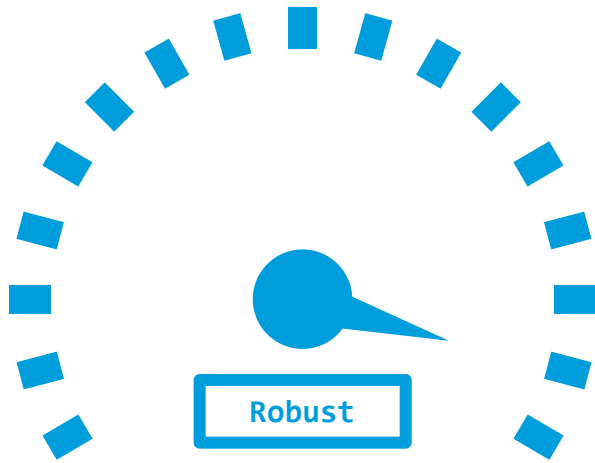
Its output ranking from the true profile should coincide with the output ranking from the noisy profile

STEP 3: AGGREGATION



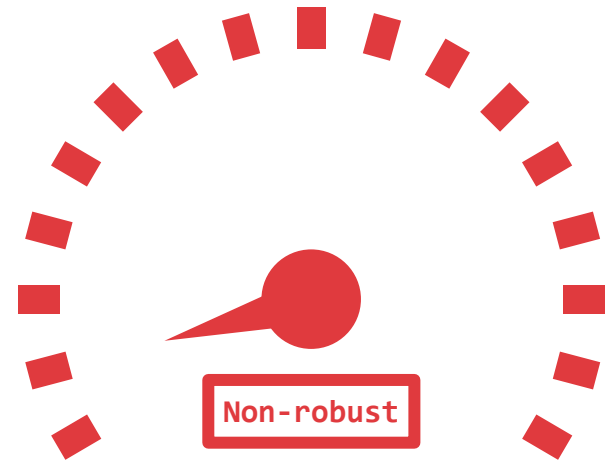
The **Mallows Model** is an unusually good fit with our setting!

STEP 3: AGGREGATION



Borda count

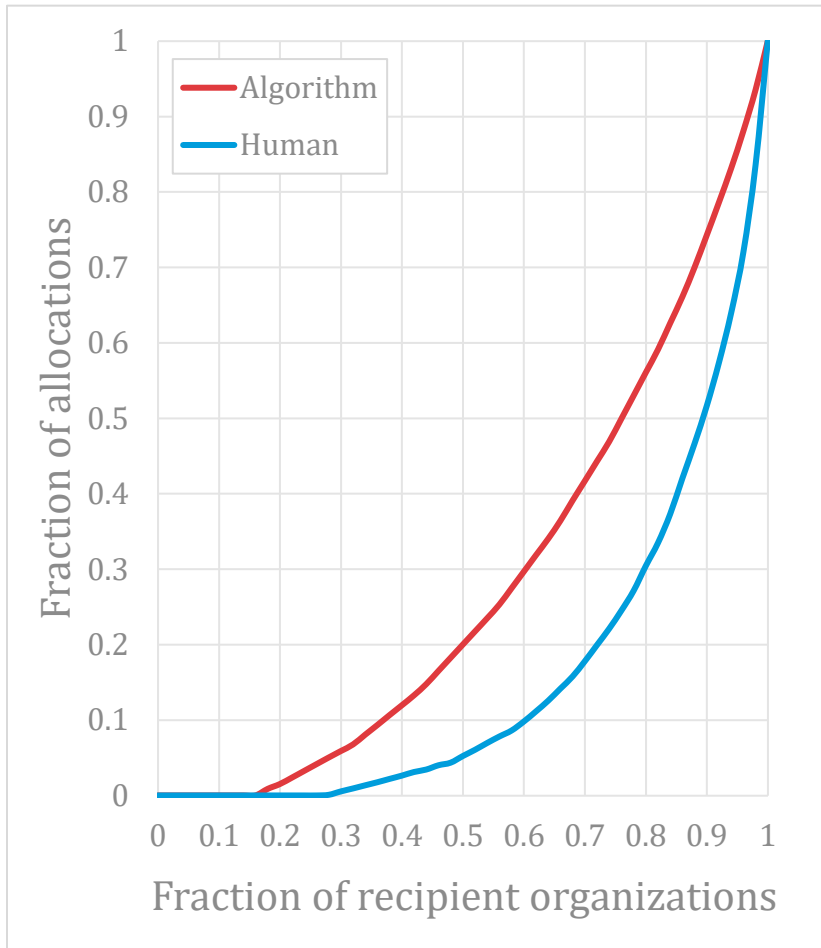
For any true profile, it is **unlikely** that two alternatives would be ranked differently when Borda count is applied to the true profile and the noisy profile



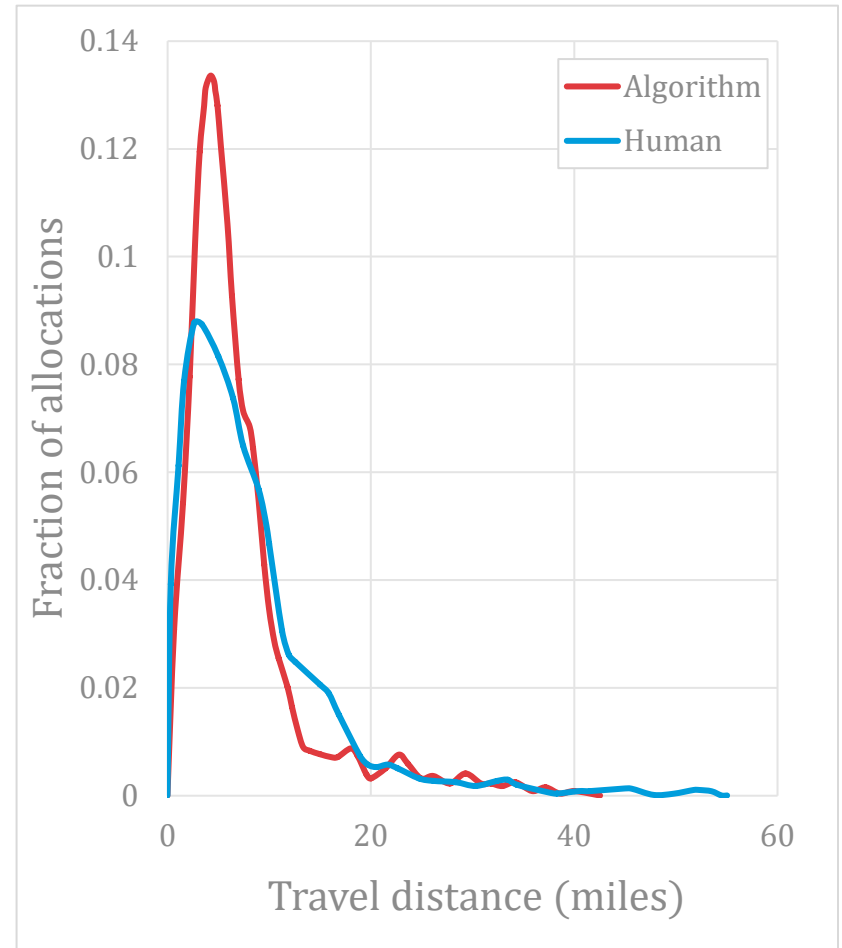
PMC Rules

There exists a true profile where, for any PMC rule f , it is **likely** that two alternatives would be ranked differently when f is applied to the true profile and the noisy profile

PERFORMANCE ON HISTORICAL DATA



Diversity of allocations



Efficiency of allocations

INTERFACE

Designed as a decision support tool

Nonprofit partner

Hide recommendations or Choose

Recommendations:

- 1 ACHA – Robert J. Corbett Apartments
- 2 Matilda Theiss Health Center
- 3 AHG – EB McNitt
- 4 Meals on Wheels – Bethel Park
- 5 Veteran's Leadership Program – Strip District

More recommendations

1 ACHA – Robert J. Corbett Apartments

- 5 minute drive
- 2 weeks ago
- Med food access
- 20 people
- High poverty rate
- Low income level

Weekly Rescues:

M T W Th F Sa Su

Hours:

24 hours a day

Choose this nonprofit

Show details

PICK UP

Map labels: Allegheny County, PNC Park, David L. Law Convention Center, U.S. Steel Tower, Warner Theatre, Steel Plaza, Duquesne University, Monongahela River, Gateway, MNTOWN, UPTOWN, S 10th St, 8th St, 7th St, 15th St, Wood Street, Smithfield St, Grant St, 11th St, Liberty Ave, 579, 380, 70C, 70D, 71A, 71B, 376, 376, 837.

PARTICIPANT FEEDBACK

Seeing how the algorithm's construction was broken down "into steps [...] and just taking each one at a time" made it attainable.

"No matter what group or individuals we're feeding, [we] have the same regard for the food and the individuals we're serving."

"This seems quite [a bit] better. If organizations are literally getting forgot[ten] about [...] this is huge."

"Certainly more fair than somebody sitting at a desk trying to figure it out on their own. [...] it should be the most fair you could get."

