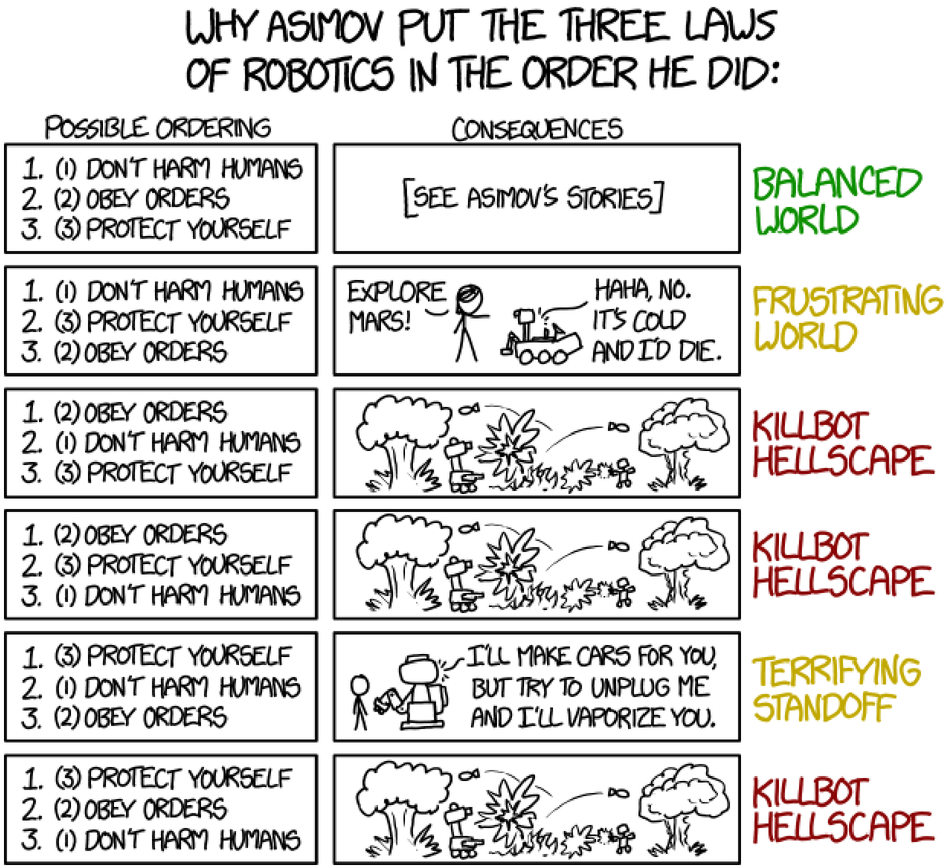
Fall 2021 | Lecture 21

# 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 | |
|---|---|---|
| 1. (1) DON'T HARM HUMANS<br>2. (2) OBEY ORDERS<br>3. (3) PROTECT YOURSELF | [SEE ASIMOV'S STORIES] | BALANCED WORLD |
| 1. (1) DON'T HARM HUMANS<br>2. (3) PROTECT YOURSELF<br>3. (2) OBEY ORDERS | EXPLORE MARS!<br>HAHA, NO. IT'S COLD AND I'D DIE. | FRUSTRATING WORLD |
| 1. (2) OBEY ORDERS<br>2. (1) DON'T HARM HUMANS<br>3. (3) PROTECT YOURSELF | | KILLBOT HELLSCAPE |
| 1. (2) OBEY ORDERS<br>2. (3) PROTECT YOURSELF<br>3. (1) DON'T HARM HUMANS | | KILLBOT HELLSCAPE |
| 1. (3) PROTECT YOURSELF<br>2. (1) DON'T HARM HUMANS<br>3. (2) OBEY ORDERS | I'LL MAKE CARS FOR YOU, BUT TRY TO UNPLUG ME AND I'LL VAPORIZE YOU. | TERRIFYING STANDOFF |
| 1. (3) PROTECT YOURSELF<br>2. (2) OBEY ORDERS<br>3. (1) DON'T HARM HUMANS | | KILLBOT HELLSCAPE |

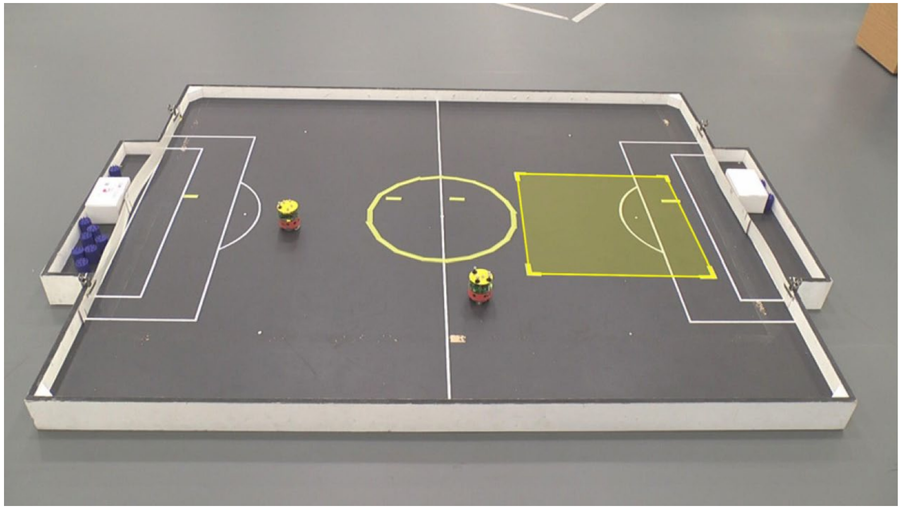
# ETHICAL ROBOTS

- 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)
```

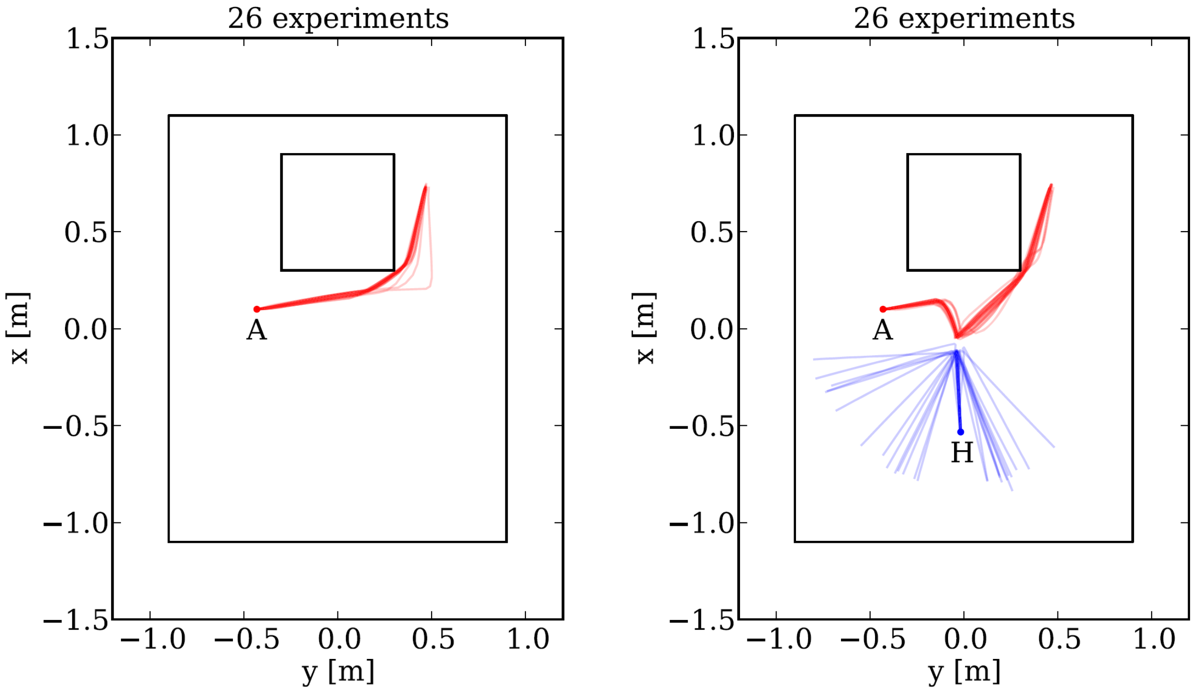
- Compare with Asimov’s first law of robotics: “A robot may not injure a human being or, through inaction, allow a human being to come to harm.”

# ETHICAL ROBOTS



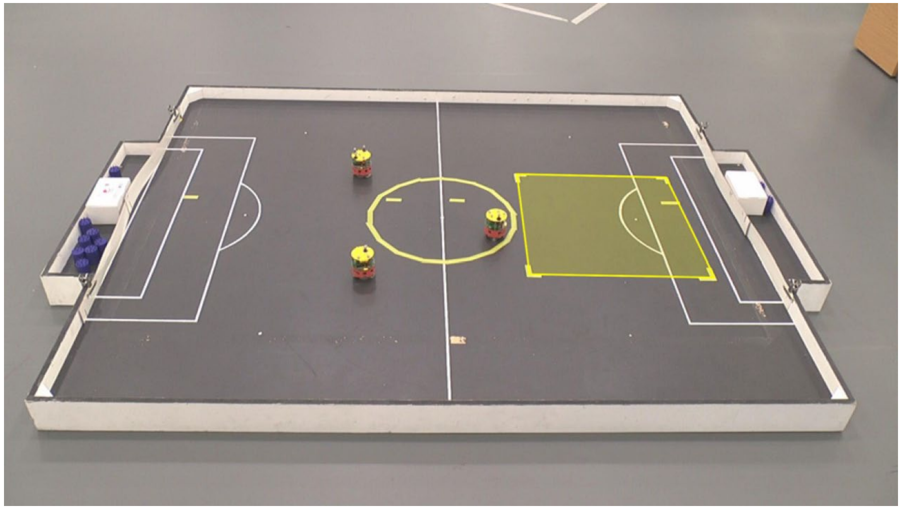
https://youtu.be/-e2MrWYRUF8?t=27m43s

# ETHICAL ROBOTS



[Winfield et al. 2014]

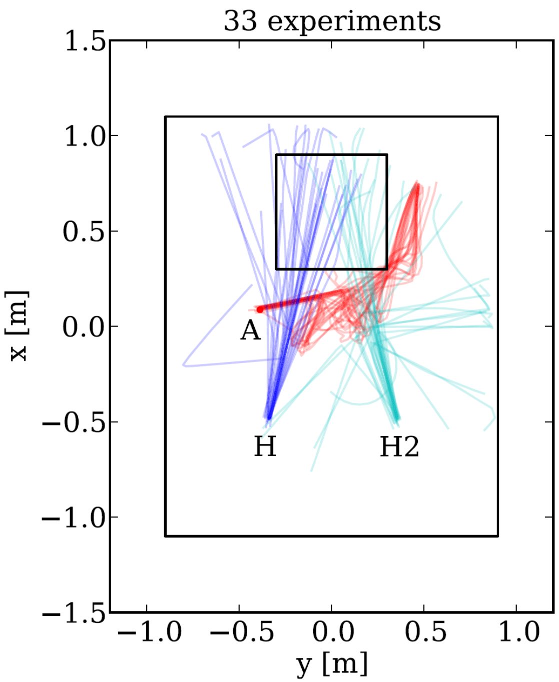
# ETHICAL ROBOTS

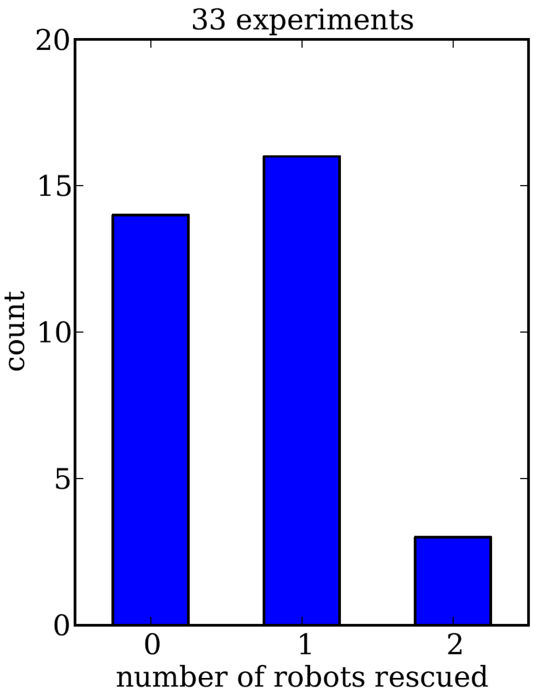


https://youtu.be/-e2MrWYRUF8?t=31m36s

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

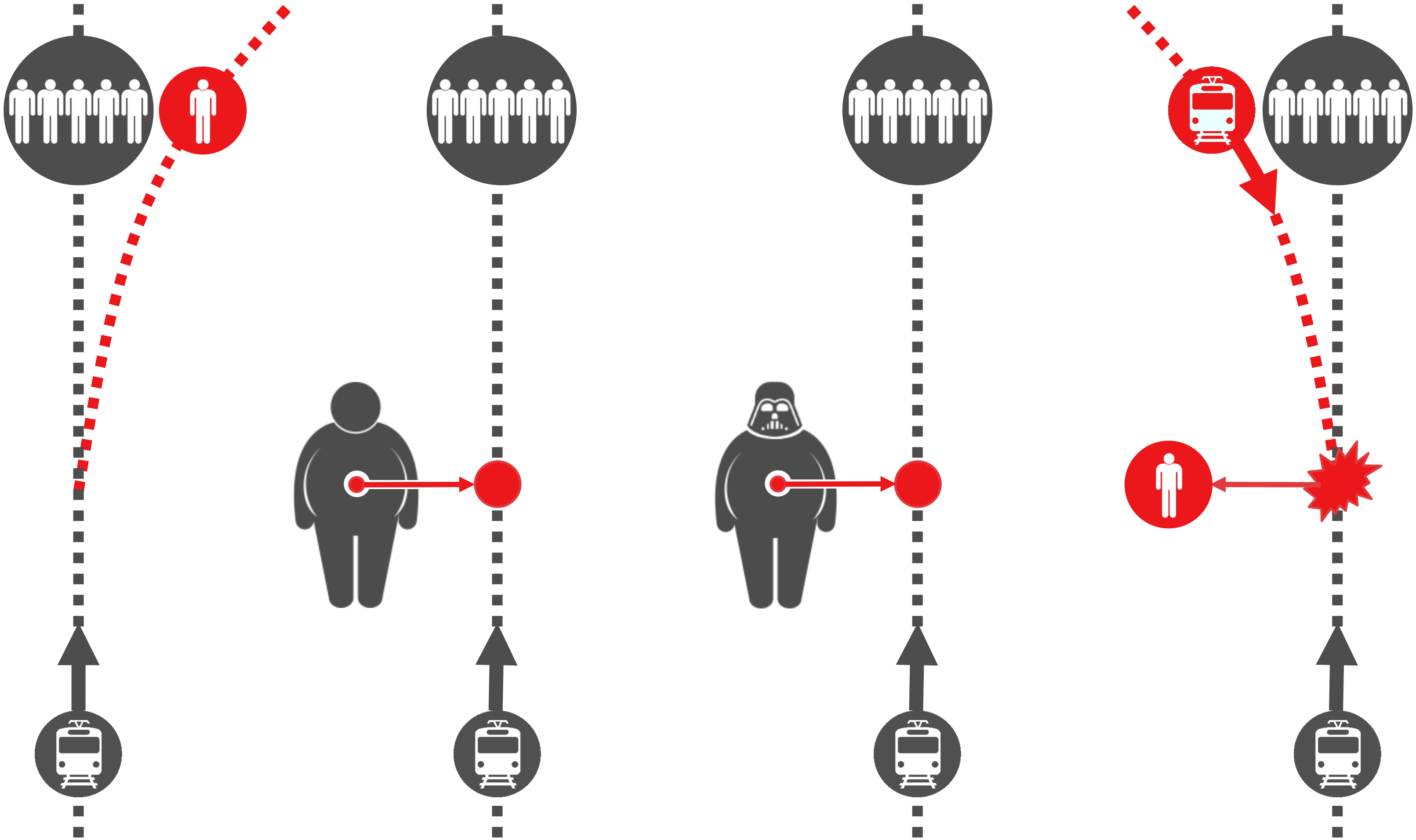
# ETHICAL ROBOTS





[Winfield et al. 2014]

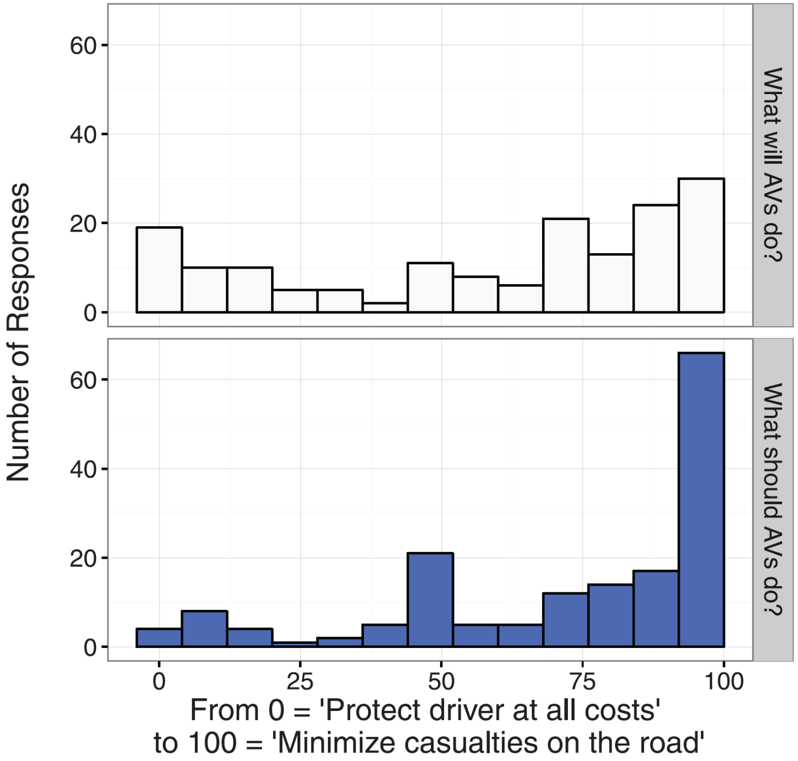
# THE TROLLEY PROBLEM



Poll: Choose an action in each scenario

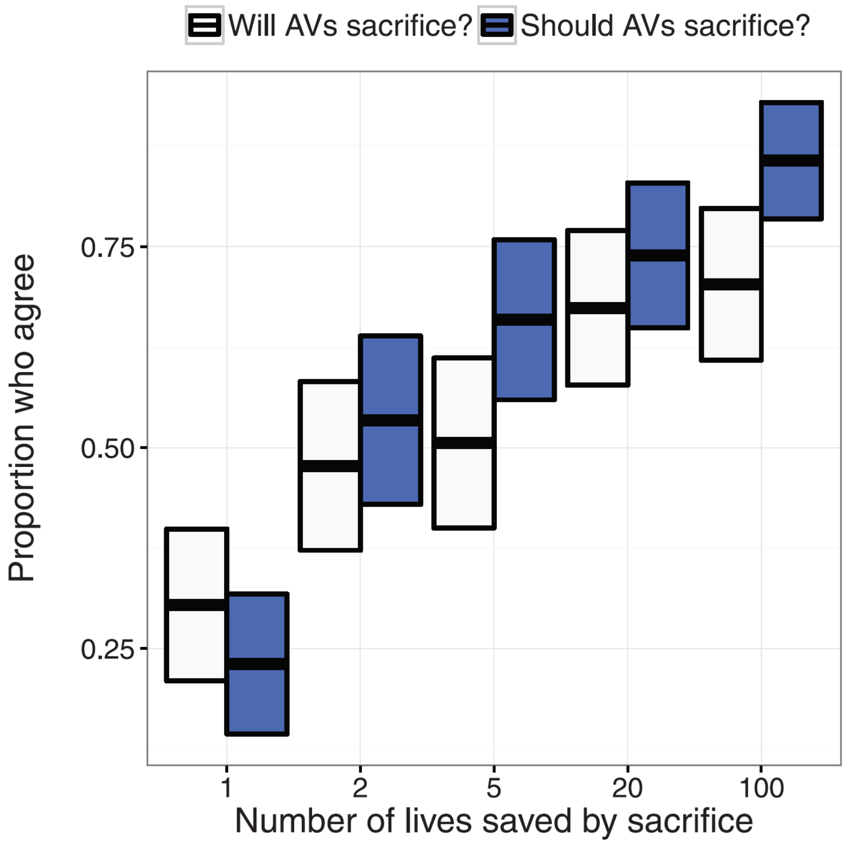
# THE SOCIAL DILEMMA OF AVS

People think an autonomous vehicle should be programmed to minimize the number of casualties, but were less certain that AVs would be programmed that way [Bonnefon et al. 2016]



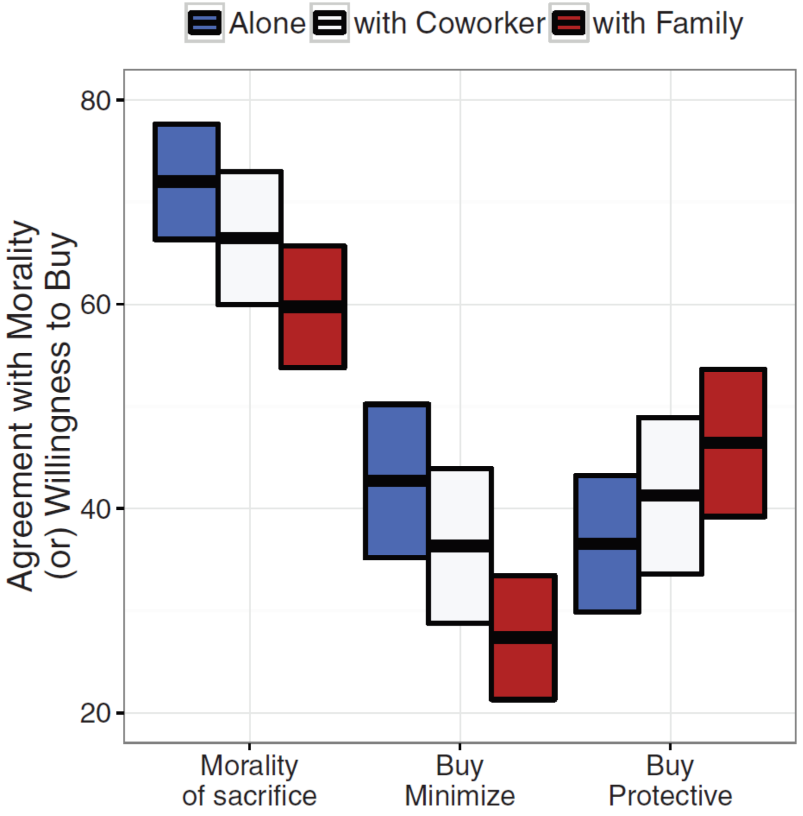
# THE SOCIAL DILEMMA OF AVS

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



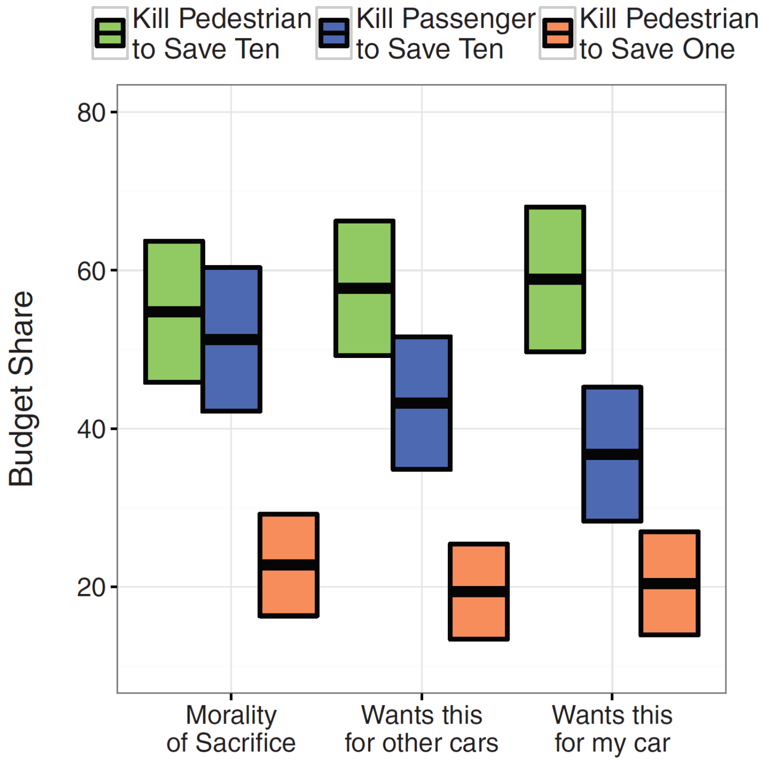
# THE SOCIAL DILEMMA OF AVS

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]

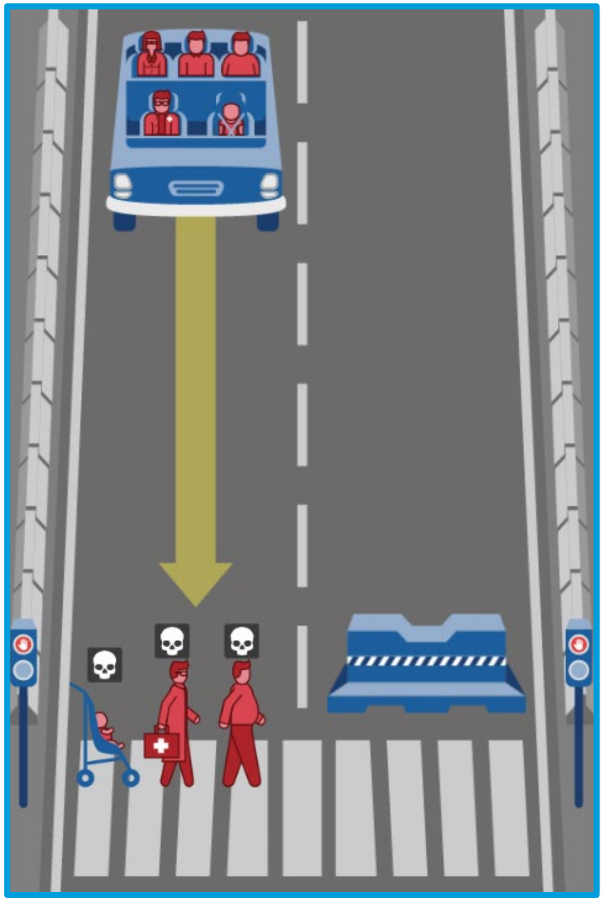


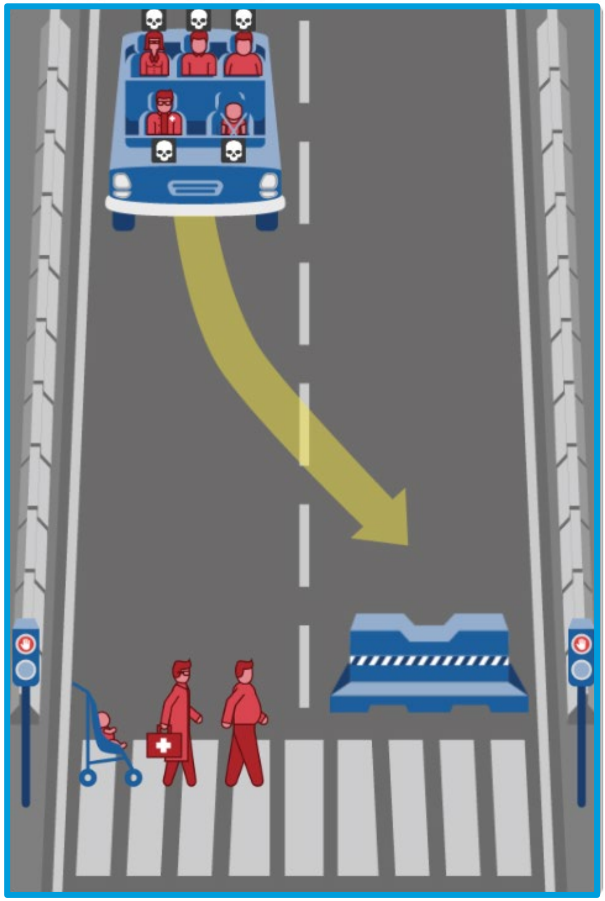
# THE SOCIAL DILEMMA OF AVS

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]



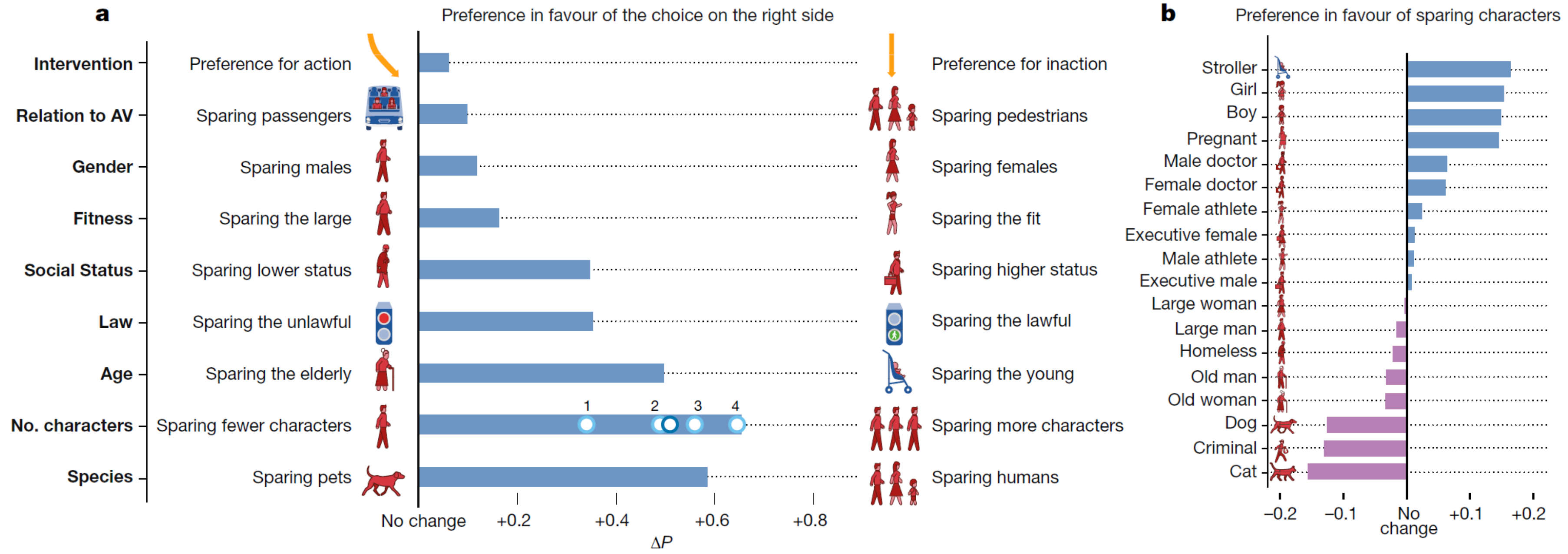
# MORAL MACHINE





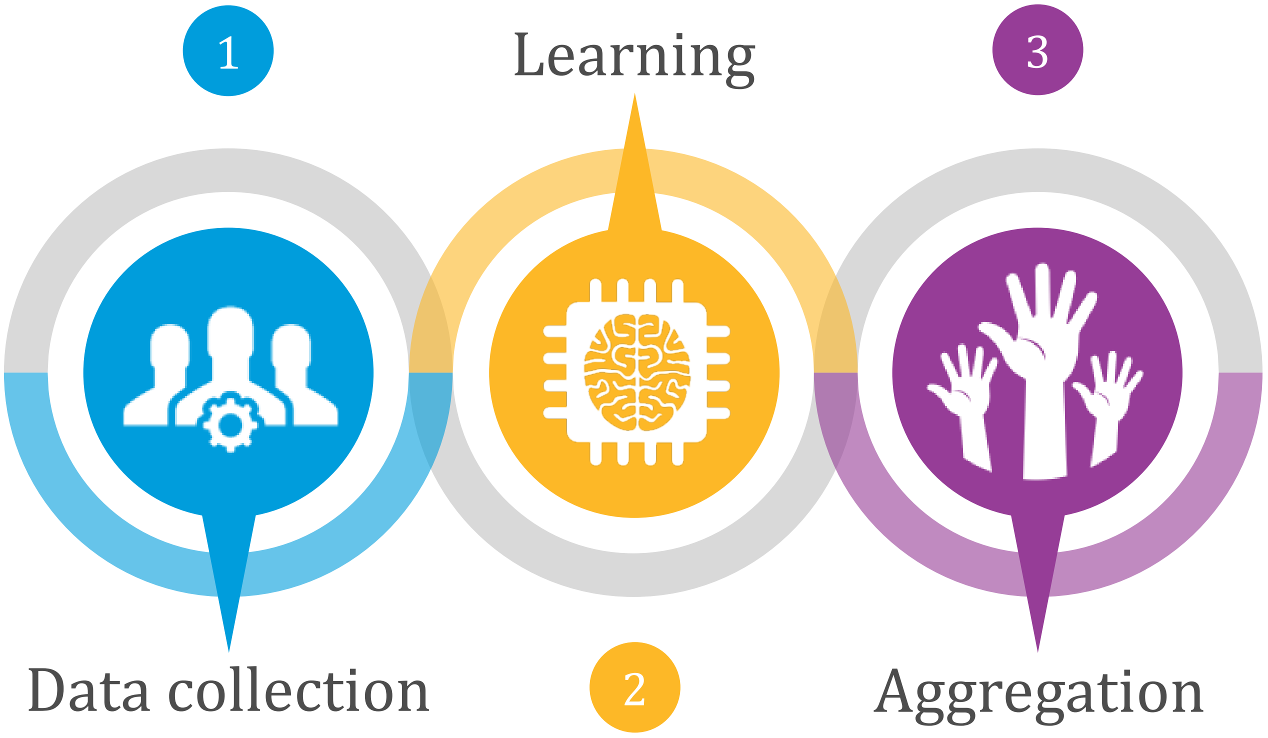
What should the self-driving car do?

# MORAL MACHINE



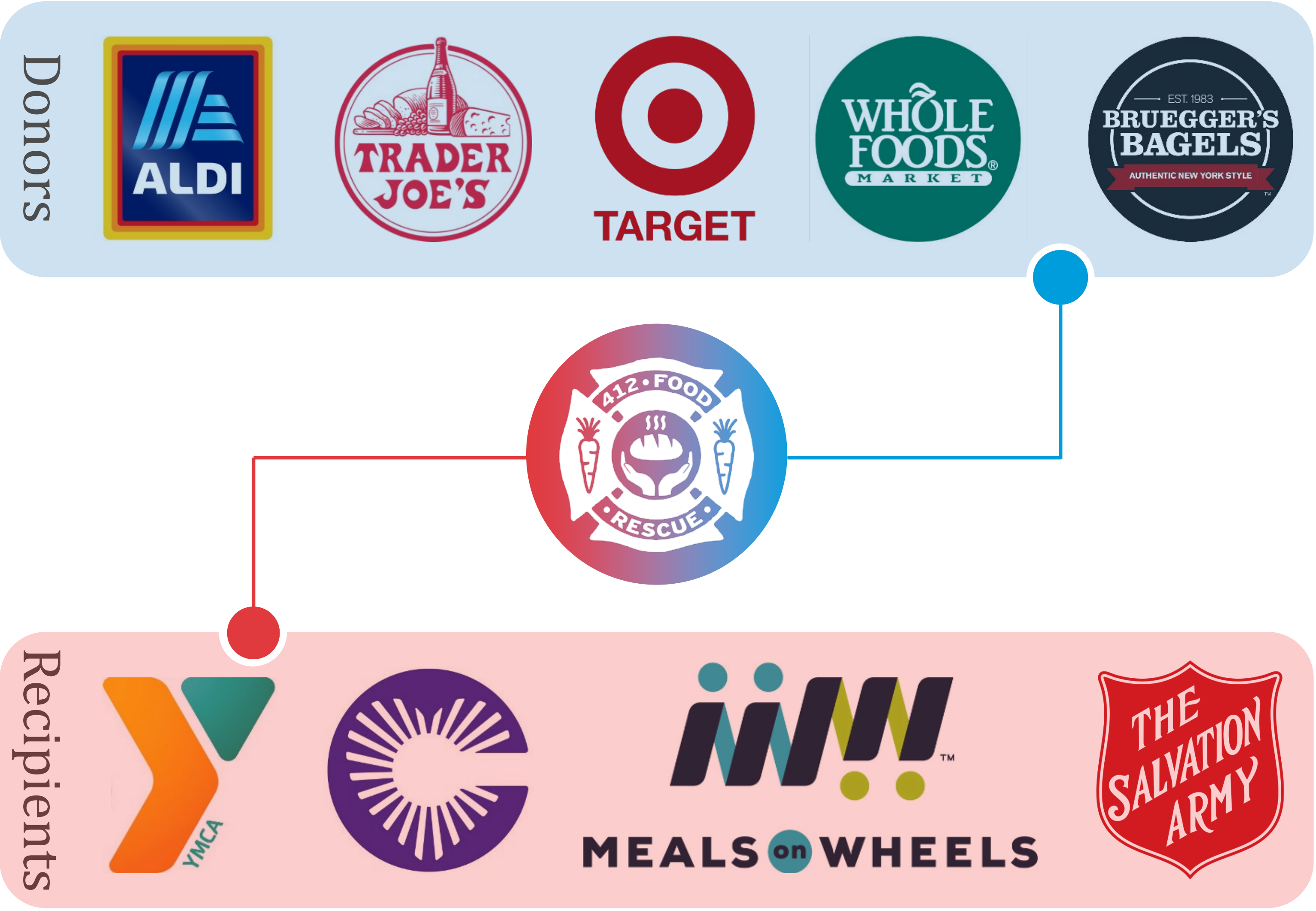
[Awad et al. 2018]

# DECISION MAKING FRAMEWORK



The rest of the lecture based on:
Noothigattu et al. 2018, Kahng et al. 2019, Lee et al. 2019

# FOOD RESCUE



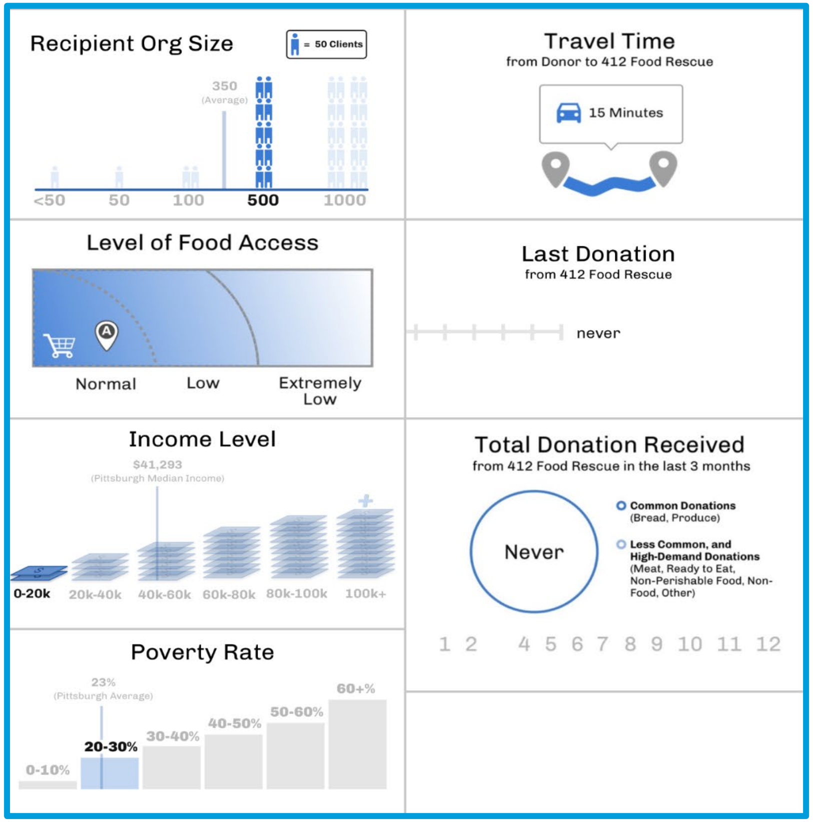
Donors

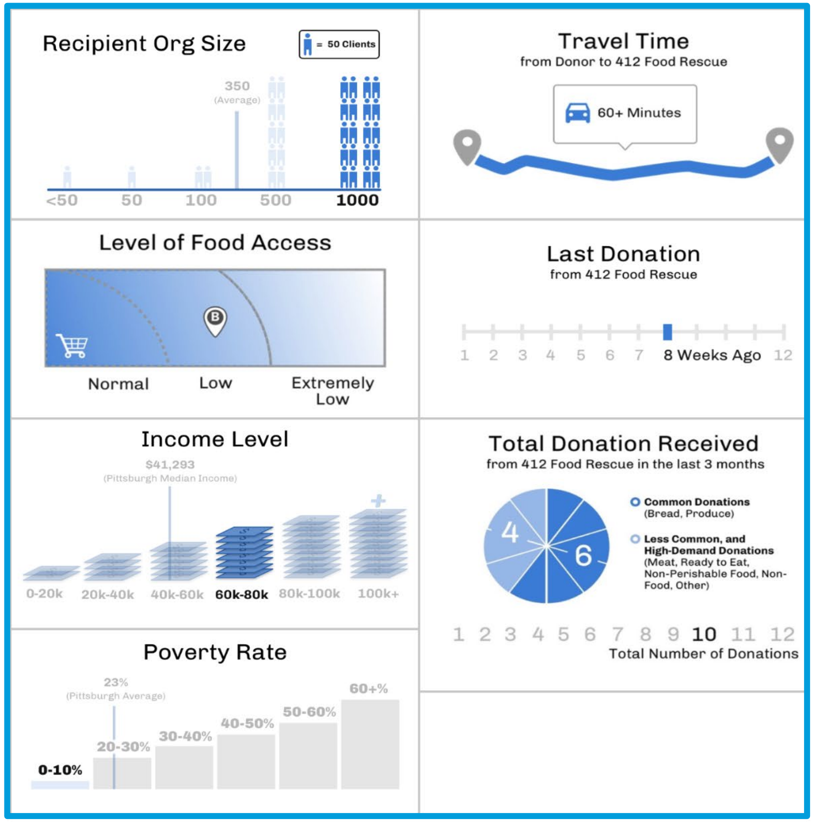
Recipients

# STEP 1: DATA COLLECTION

| Employees | Donors | Recipients | Volunteers |
|---|---|---|---|
| 3 | 6 | 8 | 6 |

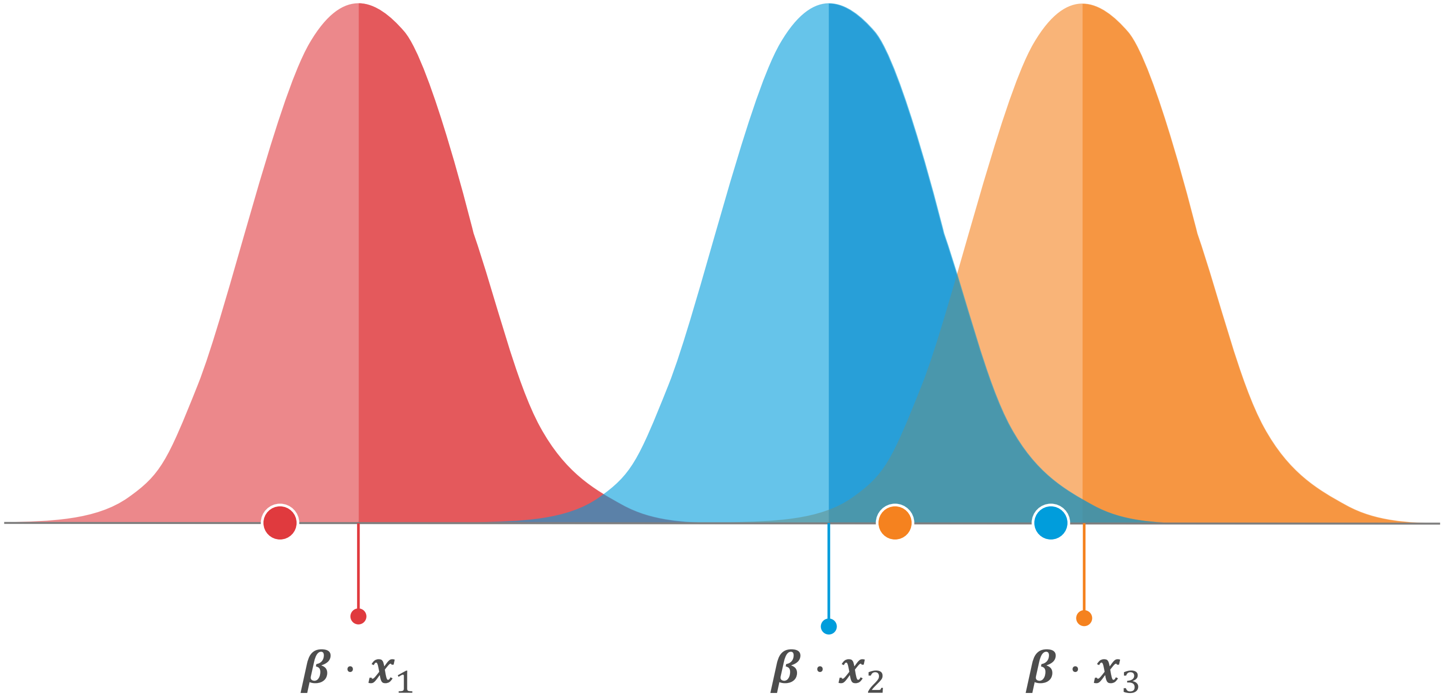
# STEP 1: DATA COLLECTION





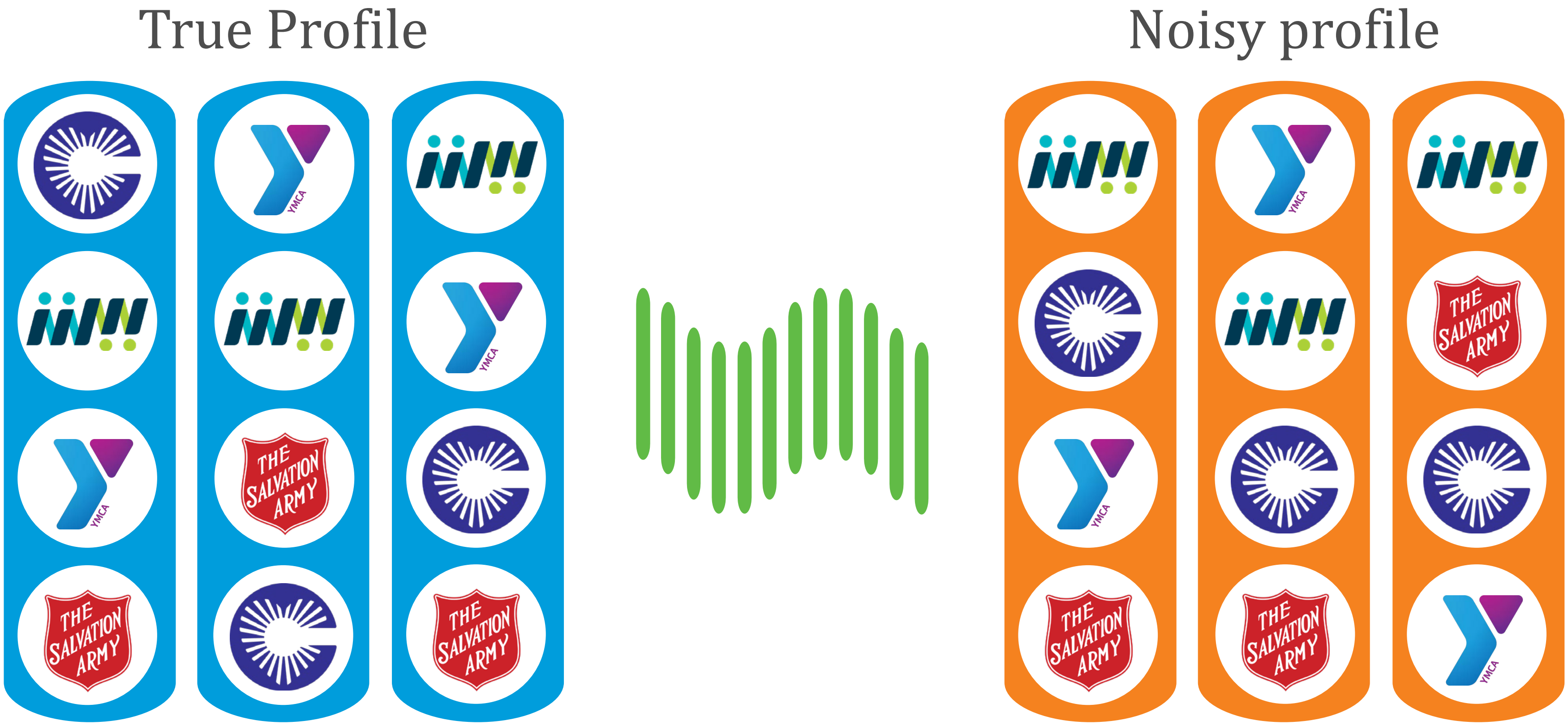
What should 412 Food Rescue do?

# STEP 2: LEARNING



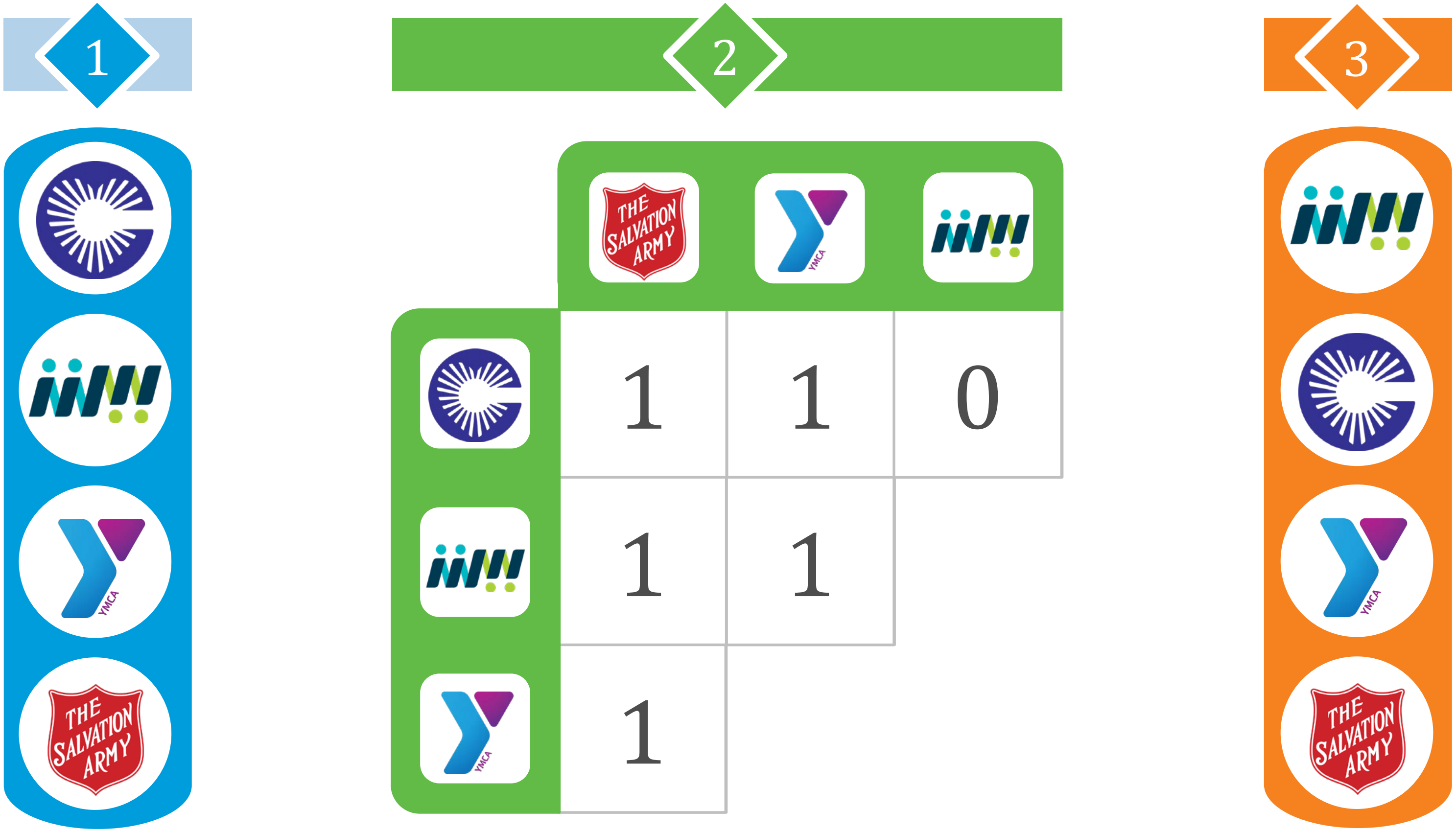
The Thurstone-Mosteller Model

# STEP 3: AGGREGATION



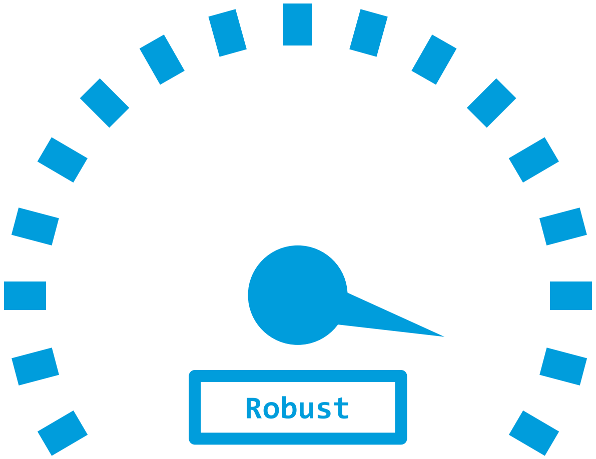
Voting rule should be robust to noise:
Its output ranking from the true profile should coincide with the output ranking from the noisy profile

# STEP 3: AGGREGATION



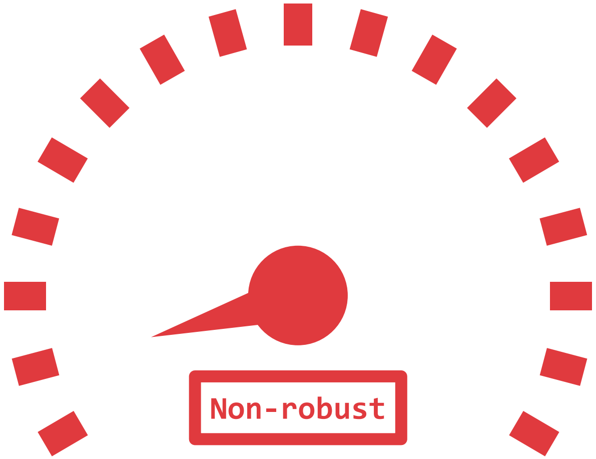
The Mallows Model is an unusually good fit with our setting!

# STEP 3: AGGREGATION



## Borda count

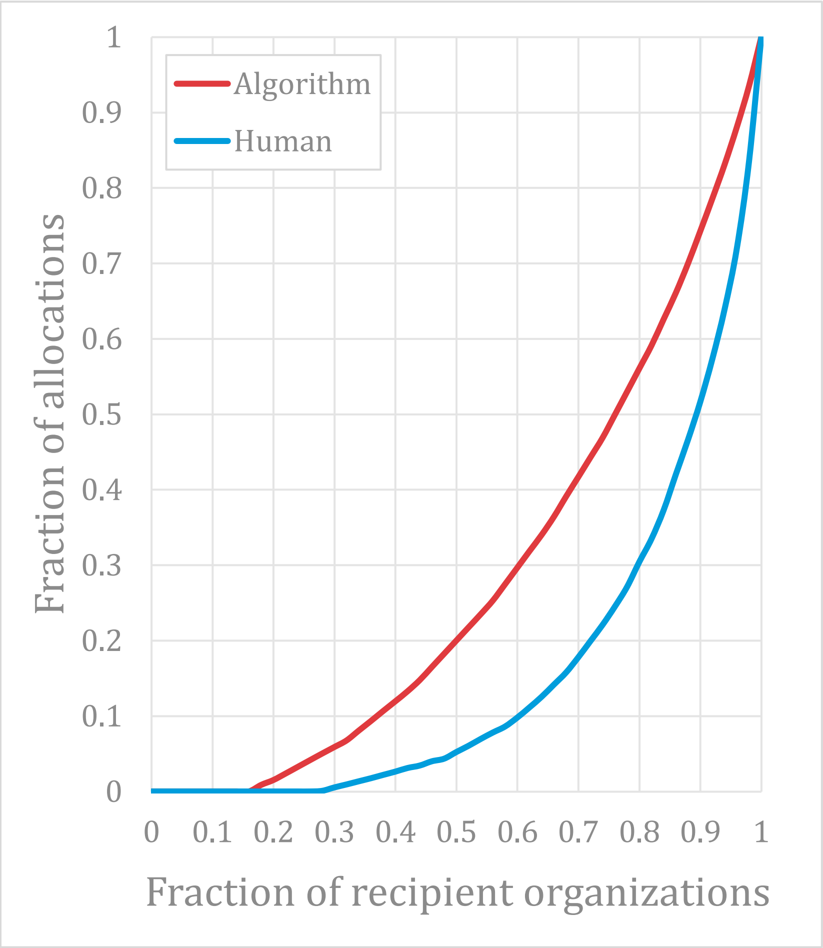
For any true profile, it is unlikely that two alternatives would be ranked differently when Borda count is applied to the true profile and the noisy profile



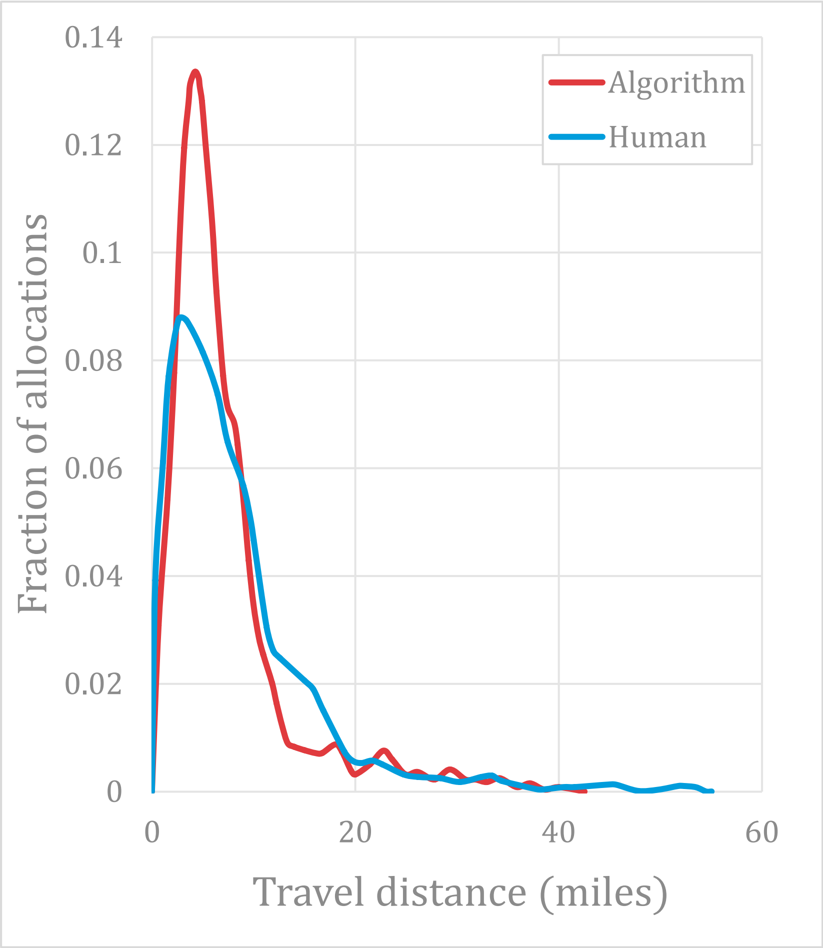
## PMC Rules

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



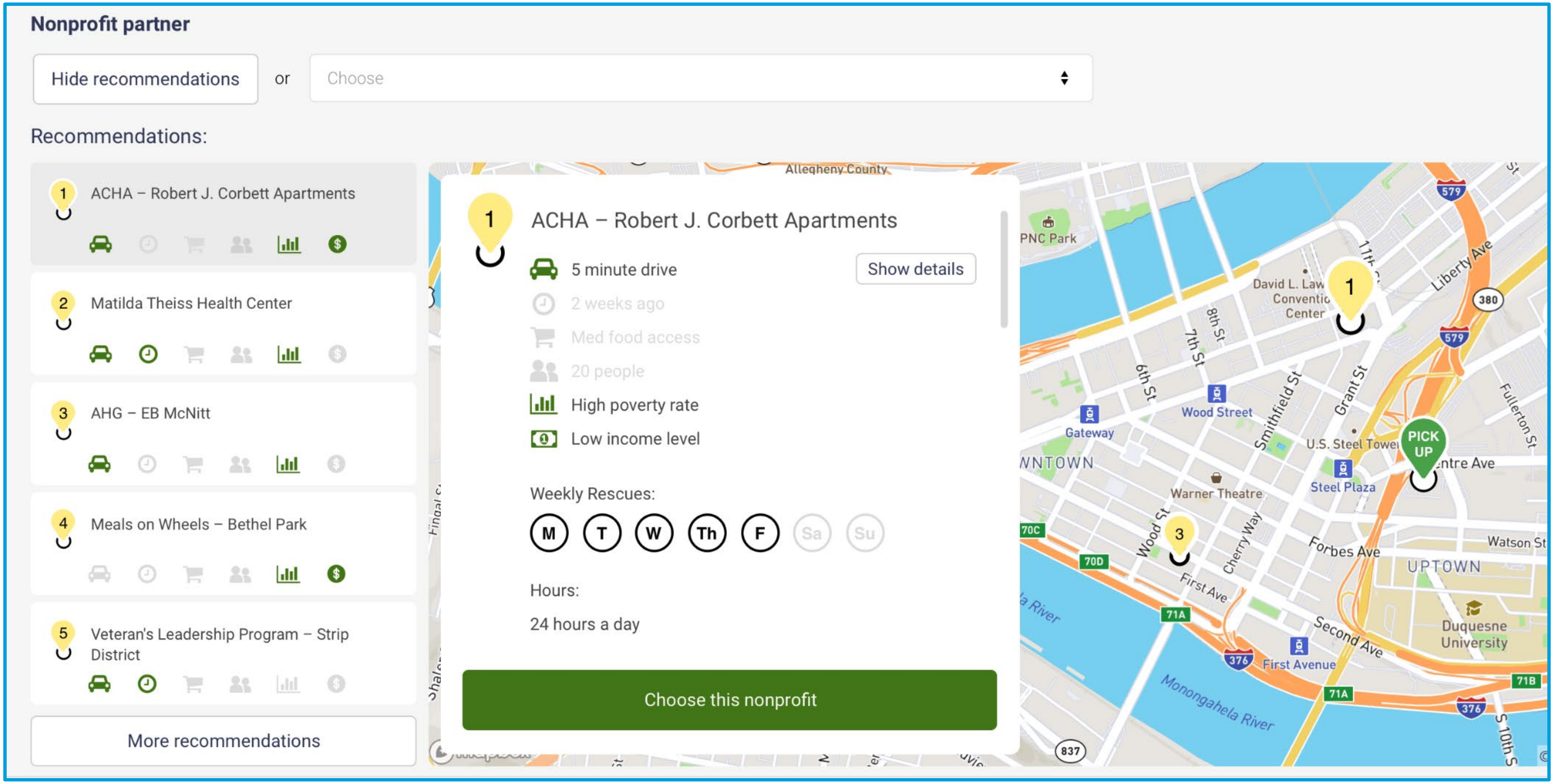
Diversity of allocations



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

Show details

- 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

# PARTICIPANT FEEDBACK

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

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

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

“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.”