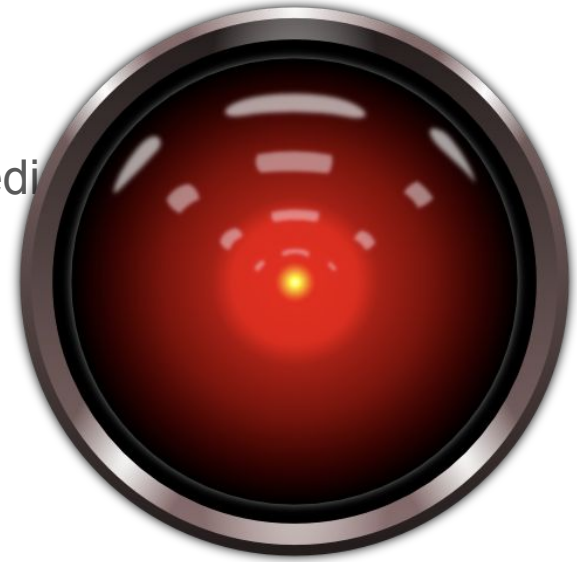
Ethical Decision Making

By Keith Kamons and Will Parkinson

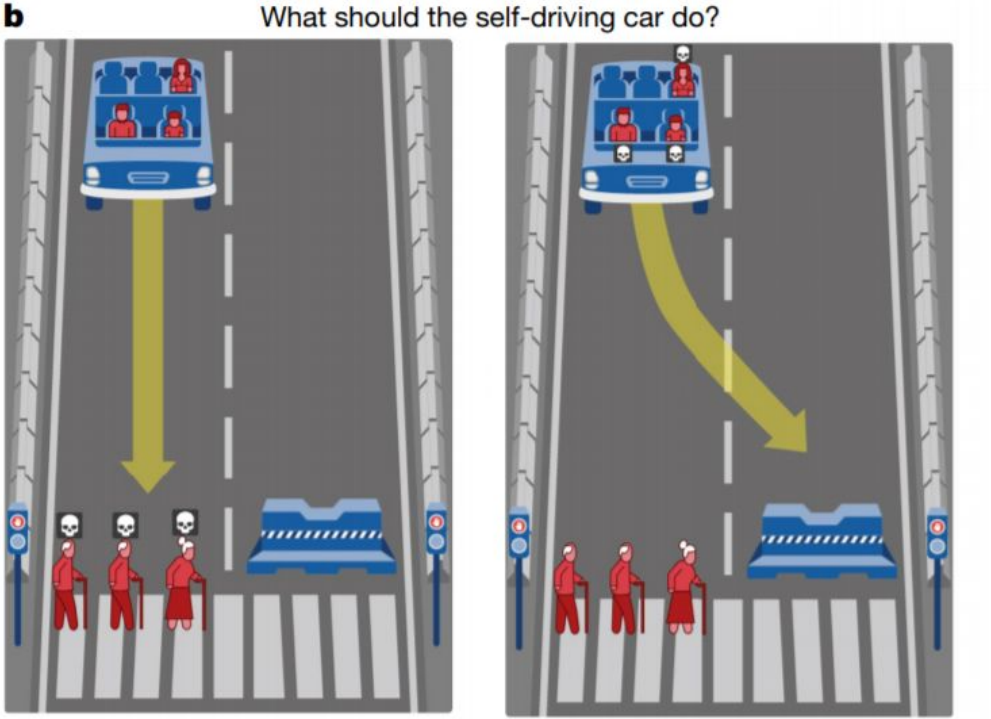
Motivation

- Algorithms are becoming an increasingly large part of daily lives
- With emerging technologies such as autonomous vehicles it is important to think about the how to implement ethical decision making
- Algorithms are already beginning to govern what medi we consume and what we buy



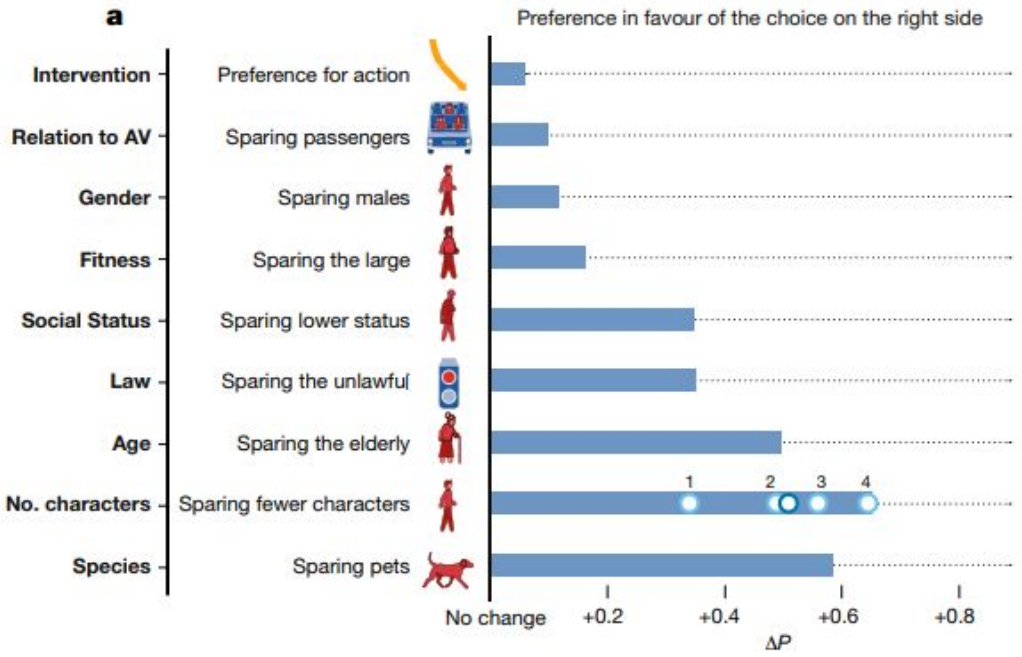
Recap of Reading

- Moral machine is online game for collecting data on how citizens would want autonomous vehicles to solve moral dilemmas in the context of unavoidable accidents



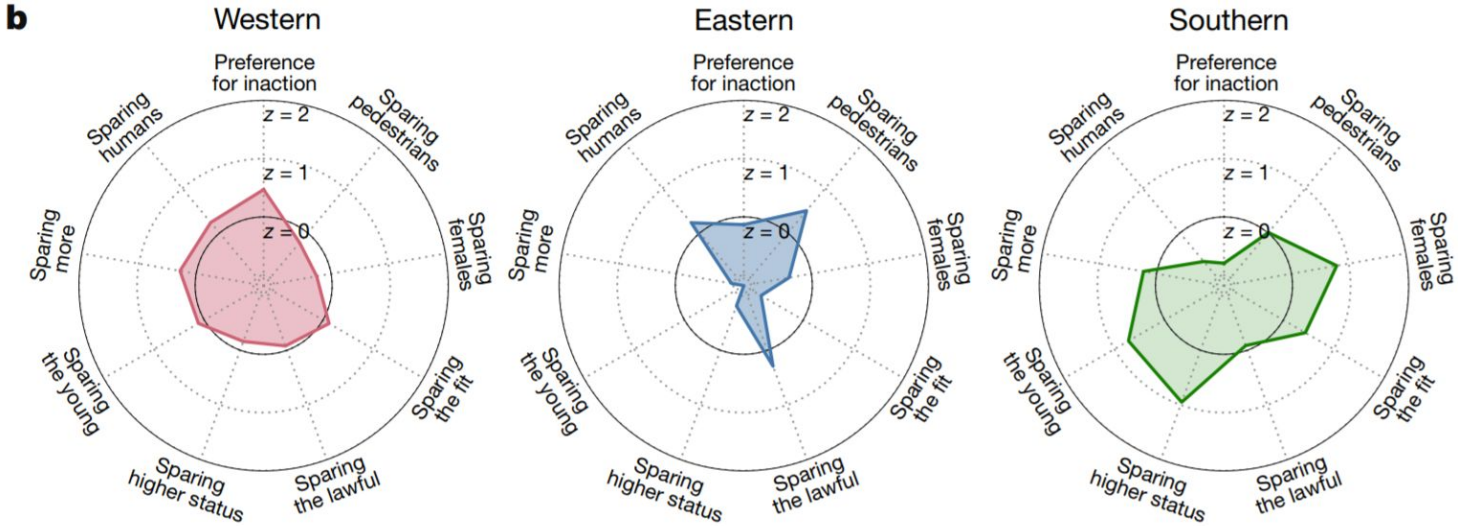
Recap of Reading

- Collected 40M decisions from 233 countries and territories
- Decisions were based on scenarios that ranged from sparing lives of passengers vs pedestrians, then broke down decisions into age, gender, social status, etc.



Recap of Reading

- Reported global, individual and cultural preferences revealed from responses and demographic surveys of respondents



Homework Discussions Questions

- Should ethics vary by region? How could regional ethics be implemented?
- Should ethical decisions be made by a majority vote? Can you think of other ways to form ethical guidelines for autonomous vehicles?
- What are some pitfalls of using popularity voting to determine who lives and who dies?

Automating Moral Decisions

Game theoretic approaches

- Problem: humans are not infinitely rational
- Behavioral game theory can help account for human behavior
- Nuances such as *intent*, distinguishing between *doing and allowing* could enhance automated decisions

Supervised learning approaches

- What are the features of our model?
- No black-box models, encourage interpretability

Moral Decision Making Frameworks for Artificial Intelligence. Conitzer et al. AAAI 2017. Blue Sky Track.

Breaking Down Moral Decisions

“Can we say that using one gallon of gasoline is just as bad for society as creating x bags of landfill trash? How would we arrive at a reasonable value of x?“

- Identify the relevant attributes of the activities/scenarios then we can determine how much each activity contributes to the attributes
  - Attributes can be both objective and subjective (can use crowdsourcing to get societies opinion)
- Examine the tradeoffs among attributes, are some worse than others?
- In the Moral Machine, this is expressed in terms of number of lives saved, age and gender of victims

Crowdsourcing Societal Tradeoffs. Conitzer et al. AAMAS'15. Blue Sky Ideas track.

How can we use the data collected?

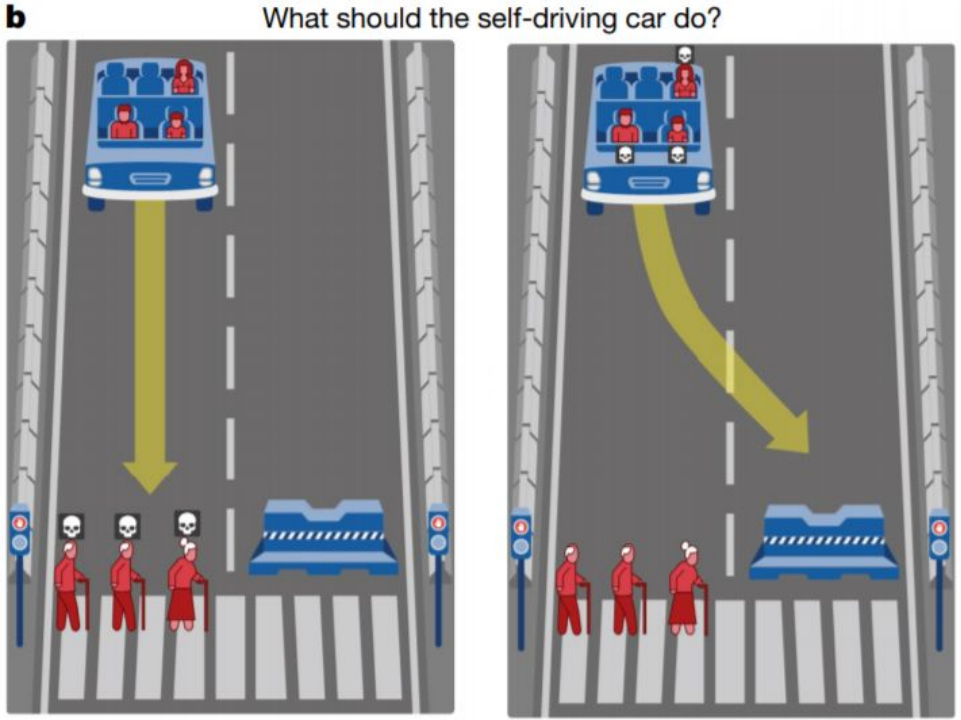
What are some of the challenges we might face if we automate these decisions?

- Lack of ground truth ethical principles
- Approximating ethical principles for a society
- How can we decide what features are important?

A Voting-Based System for Ethical Decision Making. Noothigattu et al. AAAI 2018.

A Voting-Based System

1. Data Collection
    a. Define important features and attributes to a decisions
    b. Present pairs of alternatives to voters to collect feedback



[A Voting-Based System for Ethical Decision Making](). Noothigattu et al. AAAI 2018.

A Voting-Based System

2. Learning preferences
    a. Model for preferences over voters (e.g. rankings over a finite set of alternatives)
    b. Each voter's preferences should have their own model
    c. For an uncountable number of outcomes we can use a permutation process to model the preferences

A Voting-Based System for Ethical Decision Making. Noothigattu et al. AAAI 2018.

A Voting-Based System

3. Summarization
    a. From the previous step, we have N different models (one for each voter)
    b. Combine the models into a single model so that we minimize the total KL divergence between the individual models and the combined model
    c. This step is done offline (computationally expensive) since the decisions need to be made in a split second

A Voting-Based System for Ethical Decision Making. Noothigattu et al. AAAI 2018.

A Voting-Based System

4. Aggregation
    a. Given set of alternatives we need to make a decision
    b. We can aggregate preferences based on the summary preference profile
    c. Aggregation is similar to applying a Borda count
    d. Borda count - single winner election method in which voters rank options in order of preference

Ranking	Candidate	Formula	Points	Relative points
1st	Andrew	n	**5**	**1.00**
2nd	Brian	$n-1$	**4**	**0.80**
3rd	Catherine	$n-2$	**3**	**0.60**
4th	David	$n-3$	**2**	**0.40**
5th	Elizabeth	$n-4$	**1**	**0.20**

A Voting-Based System for Ethical Decision Making. Noothigattu et al. AAAI 2018.
Borda Count

A Voting-Based System applied to the Moral Machine

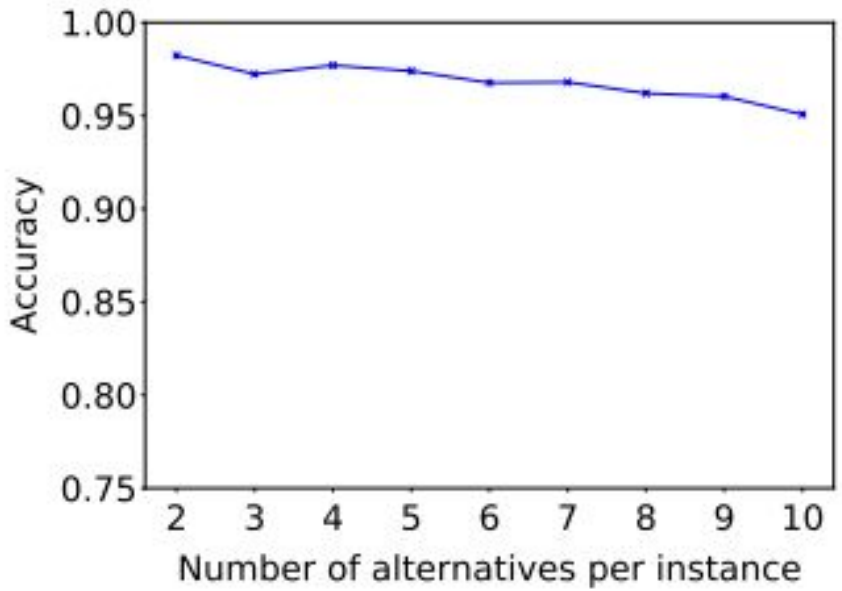
1. Data Collection
    a. Moral Machine defined 22 features (legality, passenger, pedestrian, ...) and 20 types of characters (male, female, pregnant woman, ….)
2. Learning preferences
    a. Learn preferences of 1.3 million votes using TM process
3. Summarization
    a. Sample the voters and calculate a mean of the sample
4. Aggregation
    a. Select an alternative that maximizes the utility of our summarization (done in real time)

A Voting-Based System for Ethical Decision Making. Noothigattu et al. AAAI 2018.

A Voting-Based System applied to the Moral Machine

Using the sampling procedure, they compared the sampled preference summary with the exhaustive preference summary. For two alternatives accuracy was 98% and 95% for 10 alternatives when tested on 3000 instances

Accuracy defined by proportion of agreement between individual model and summarized model



A Voting-Based System for Ethical Decision Making. Noothigattu et al. AAAI 2018.

WeBuildAI

- A Framework for building more “fair” governing algorithms using participatory Design
- Participatory Design: Researchers and users of a technology share power and control in determining its technological future.
- Governing Algorithms: Algorithms that nudge, bias, guide, provoke, control, manipulate, and constrain human behavior
- Emerging work seeks to understand participants’ values with regard to the fairness of actual AI products

Background

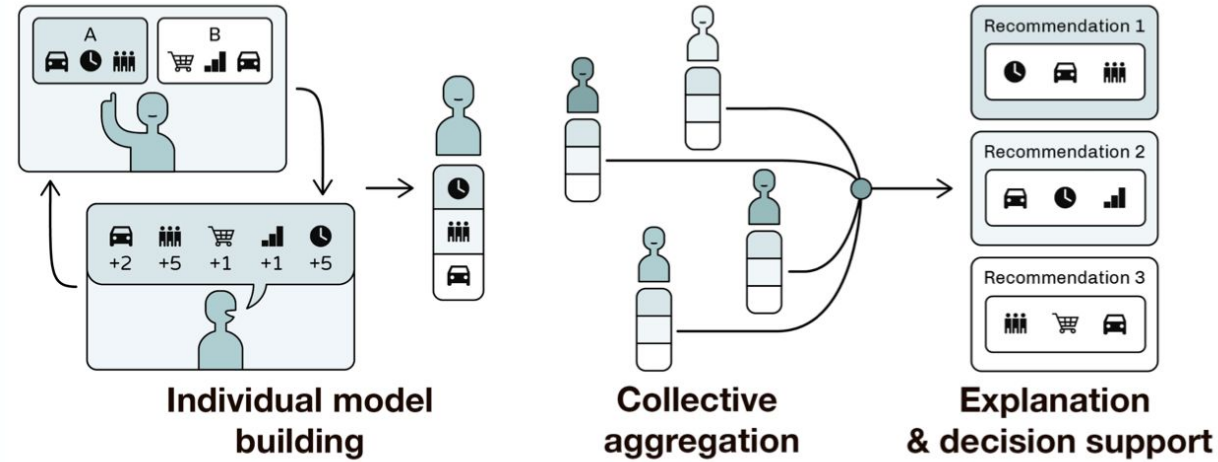
- Designed a matching algorithm that operated on-demand transportation service of a non-profit called 412 Food Rescue
- Food Rescue matches donations of food to non-profit recipient organizations, then a post is made on the app so volunteers can sign up to transport donations to organizations
- Allocation was originally done by humans with a skewed distribution of 20% of organizations receiving 70% of donations.
- Stakeholders: donors, volunteers, recipient organizations, and 412 Food Rescue’s staff

Factors of Matching Algorithm

Factor	Explanation
Travel Time	The expected travel time between a donor and a recipient organization. Indicates time that volunteers would need to spend to complete a rescue. (0-60+ minutes)
Recipient Size	The number of clients that a recipient organization serves every month. (0-1000 people; AVG: 350)
Food Access	USDA-defined food access level in the client neighborhood that a recipient organization serves. Indicates clients' access to fresh and healthy food. (Normal (0), Low (1), Extremely low(2)) [78]
Income Level	The median household income of the client neighborhood that a recipient organization serves (0-100K+, Median=$41,283) [77]. Indicates access to social and institutional resources [69].
Poverty Rate	Percentage of people living under the US Federal poverty threshold in the client neighborhood that a recipient organization serves. (0-60 %; AVG=23% [77])
Last Donation	The number of weeks since the organization last received a donation from 412 Food Rescue. (1 week–12 weeks, never)
Total Donations	The number of donations that an organization has received from 412 Food Rescue in the last three months. (0-12 donations) A unit of donation is a carload of food (60 meals).
Donation Type	Donation types were common or uncommon. Common donations are bread or produce and account for 70% of donations. Uncommon donations include meat, dairy, prepared foods, etc.

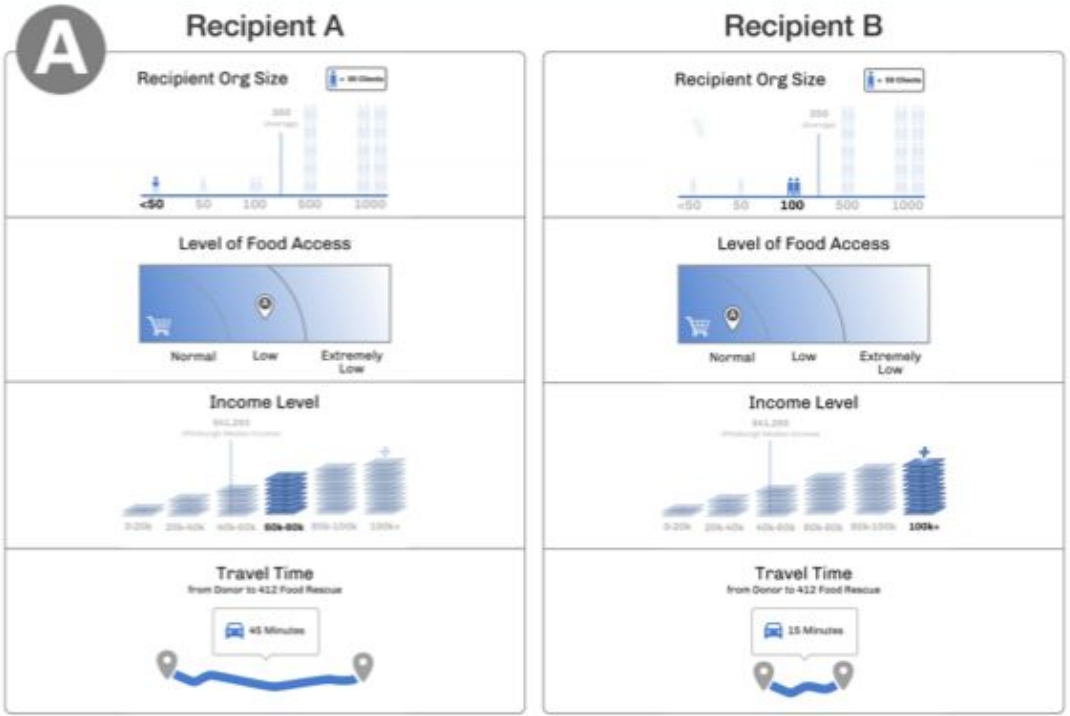
Belief Model Building

- Three sessions were conducted to develop a model to represent each individual in the final algorithm
  - Pairwise Comparisons: Train algorithm using machine learning
  - Explicit Rule Specification: Elaborate on models
  - Choose one of the two models that represented their beliefs most accurately



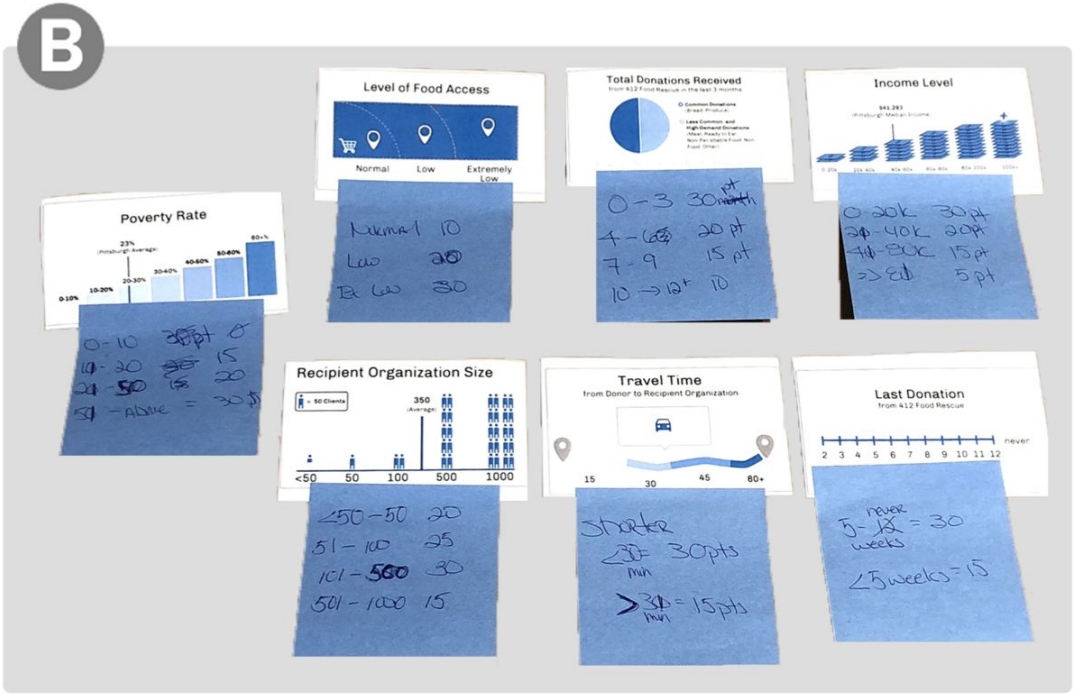
Belief Model Building: Machine Learning Model

- Participants were presented two recipients and then asked to choose which should receive the donation



Belief Model Building: Explicit Rule Model

- Participants specified rules by assigning a score to each factor involved in the algorithm
- If their belief changed after this session they retrained the pairwise generated algorithm



WeBuildAi Case Study: Model Comparison

- Participants are shown a comparison between their Explicit Rule and Machine Learning Model. The identities of the methods were kept anonymous
- Researcher walked the participant through the model graphs

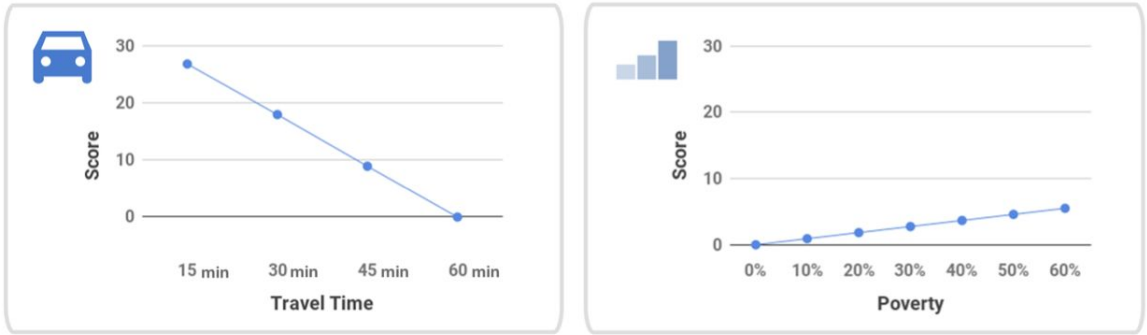


Fig. 3. Model explanations. Both machine learning and explicit-rule models were represented by graphs that assigned scores according to the varying levels of input features.

WeBuildAi Case Study: Collection Aggregation

- Uses a voting method to aggregate individuals' beliefs
- Each individual's model generates a complete ranking of all possible recipients.
- The Borda rule aggregates the rankings to derive a consensus ranking and suggests recommendations

	D2	D4	F2	F3	R1	R2	R3	R5	R7	V1	V3	V4	V5	V6
ML	**0.86**	0.78	**0.92**	**0.92**	**0.90**	0.90	**0.78**	**0.94**	0.74	**0.90**	**0.92**	**0.78**	0.56	0.68
ER	0.68	**0.68**	0.68	0.86	0.80	**0.76**	0.70	0.92	**0.74**	0.76	0.82	0.82	**0.80**	**0.88**

Varying Stakeholders' Voting Influence

- Should all stakeholders votes matter equally?
- The majority of participants believed that voting power should depend on their role.

Participant	Average Percentage of Voting Power
412 Food Rescue	46%
Recipient Organizations	24%
Volunteers	19%
Donors	11%

WeBuildAi Case Study: Results

Matching Algorithm suggested more donations to higher poverty areas with lower income and low access to food, without increasing the transportation distance

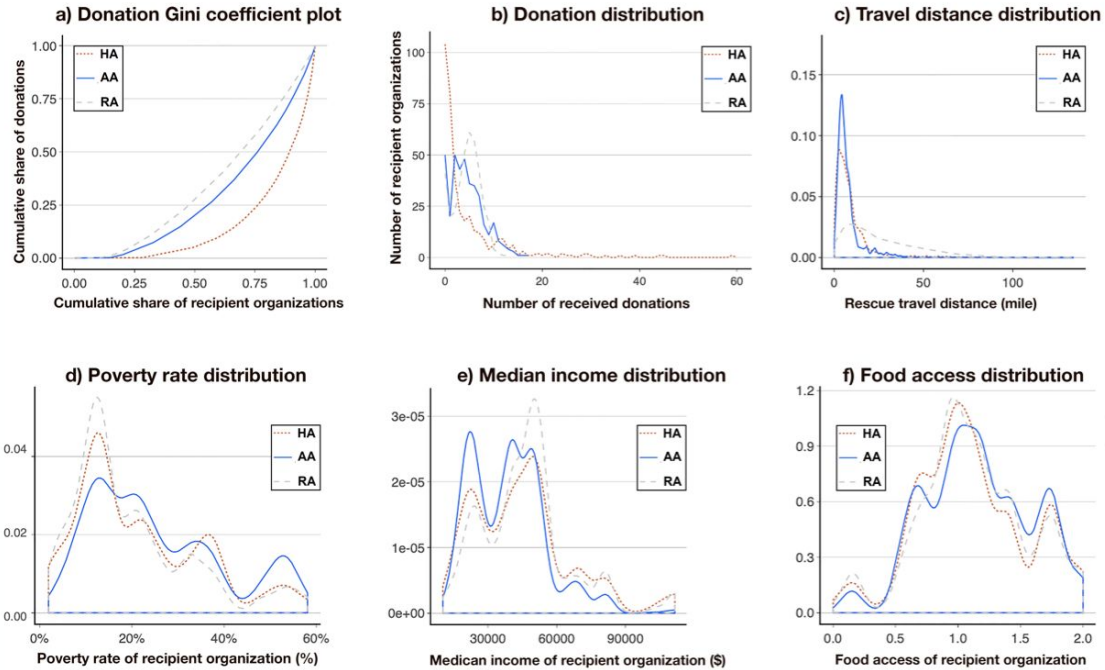


Fig. 5. The performance of our algorithm (AA) versus the human allocation (HA) and a uniformly random allocation (RA), on various metrics.

Discussion

- What are some examples of technologies (governing algorithms) where participatory design should be applied? What would the features/factor for this technology be?
- What types of situations could have nuanced definitions of what is fair for an AI?
- Can you think of some issues/drawbacks with this method of algorithm design? When would participatory design be bad to use?

Closing Remarks and Take-Aways

- Moral Decision making is an important topic to be researching as Algorithms begin to govern more of our lives
- Voting based methods can offer models that reflect the preferences and beliefs of more people.
- Limitations:
  - This sort of design process may be difficult to implement in practice
  - There are not always ground truths when it comes to morality
  - Choosing representative participation can be tricky and not always have ground truths

References

E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, “The Moral Machine experiment,” Nature, vol. 563, no. 7729, pp. 59–64, 2018.

R. Noothigattu, S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, A Procaccia, “A Voting-Based System for Ethical Decision Making,” Carnegie Mellon University, 2018.

M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, D. See, R. Noothigattu, S. Lee, A. Psomas, and A. D. Procaccia, “WeBuildAI,” Proceedings of the ACM on Human-Computer Interaction, vol. 3, no. CSCW, pp. 1–35, 2019.

V. Conitzer, M. Brill, R. Freeman, “Crowdsourcing Societal Tradeoffs,” Duke University, 2015.

V. Conitzer, W Sinnott-Armstrong, J. Borg, Y. Deng, M. Kramer, “Moral Decision Making Frameworks for Artificial Intelligence” Duke University, 2017.