WeBuildAI: Participatory Framework for Fair and Efficient Algorithmic Governance

Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Ritesh Noothigattu, Daniel See, Siheon Lee, Christos-Alexandros Psomas, Ariel D. Procaccia

School of Computer Science
Carnegie Mellon University

ABSTRACT

Algorithms increasingly govern societal functions, impacting multiple stakeholders and social groups. How can we design these algorithms to balance varying interests and promote social welfare? As one response to this question, we present WeBuildAI, a social-choice based framework that enables people to collectively build algorithmic policy for their communities. The framework consists of three steps: (i) Individual belief elicitation on governing algorithmic policy, (ii) voting-based collective belief aggregation, and (iii) explanation and decision support. We applied this framework to an efficient yet fair matching algorithm that operates an on-demand food donation transportation service. Over the past year, we have worked closely with the service's stakeholders to design and evaluate the framework through a series of studies and a workshop. We offer insights on belief elicitation methods and show how participation influences perceptions of governing institutions and administrative algorithmic decision-making.

Author Keywords

Participation; human-centered algorithm; social choice

INTRODUCTION

Computational algorithms increasingly take governance and management roles in administrative and legal aspects of public and private decision-making [7, 15, 16]. In digital platforms, bureaucratic institutions, and infrastructure, algorithms manage information, labor, and resources, coordinating the welfare of multiple stakeholders. For example, news and social media platforms use algorithms to distribute information, which influences the costs and benefits of their services for their users, news sources and advertisers, and the platforms themselves [25]; on-demand work platforms use algorithms to assign tasks, which affects their customers, their workers, and their own profits [32, 51]; and city governments use algorithms to manage police patrols, neighborhood school assignments, and transportation routes [47], all of which have significant implications for affected communities.

These algorithmic decisions can have a substantial impact on economic and social welfare due to the algorithms' invisible operation and massive scale. In fact, recent cases suggest that algorithmic governance can lead to compromises in social values or hinder certain stakeholder groups in unfair ways [2, 12]. How can we design algorithmic governance that is effective yet also moral, promotes overall social welfare, and balances the varying interests of different stakeholders, including the governing institutions themselves?

Participation is a promising approach to answering this question. Citizen and stakeholder participation in policy making improves the legitimacy of a governing institution in a democratic society [23, 24]. Enabling participation in service creation has also been shown to increase trust and satisfaction, thereby increasing motivation to use the services [5]. In addition, participation increases effectiveness. For certain problems, users themselves know the most about their unique needs and problems [23, 38]; participation can help policymakers and platform developers leverage this knowledge pool. Finally, stakeholder participation can help operationalize moral values and their associated tradeoffs, such as fairness and efficiency [23]. Even people who agree wholeheartedly on certain high-level moral principles tend to disagree on the specific implementations of those values in algorithms—the objectives, metrics, thresholds, and tradeoffs that need to be explicitly codified, rather than left up to human judgment.

Our goal is therefore to enable stakeholder participation in algorithmic governance. This vision raises several fundamental research questions. First, what socio-technical methods and techniques will effectively elicit individual and collective beliefs about policies and translate them into computational algorithms? Second, how should the resulting algorithmic policies be explained so that participants understand their roles and administrators understand their decisions? Finally, how does participation influence participants' perceptions of and interactions with algorithmic governance? In this paper, we explore these research questions in the context of a fair and efficient matching algorithm for an on-demand donation transportation service offered by 412 Food Rescue, a nonprofit organization that coordinates food deliveries from donors (e.g., supermarkets) to recipients (e.g., food pantries). We first describe

design considerations for enabling participation in algorithmic governance, drawing on the political theory literature. We then introduce the WeBuildAI framework (Figure 1). The framework learns individuals' beliefs on algorithmic policy decisions. The learned individual models predict individuals' rankings of decision alternatives, and are aggregated using a voting method. Finally, the framework explains the resulting algorithmic policy, supporting the decision makers using the matching algorithm.

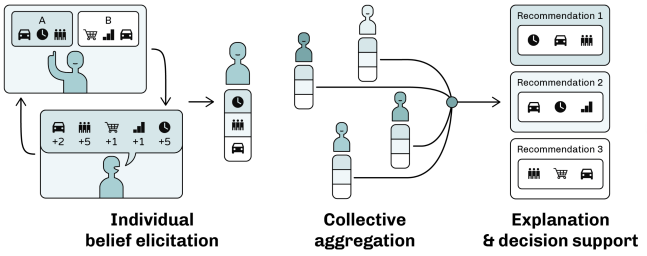

Figure 1. The WeBuildAI framework allows people to participate in building algorithmic governance policy. Individuals deliberate and express their beliefs on algorithmic policy decisions, training machine learning algorithms through pairwise comparisons and explicitly specifying rules and behaviors. Each individual belief model ranks decision alternatives, which are aggregated via the Borda rule to generate recommendations. Recommendations are explained to show how people's input is used in the final algorithmic policy and to support administrative decision-making.

Over the course of a year, we designed and evaluated the framework by closely working with stakeholders through a series of studies and a workshop. We found that our framework was effective in eliciting individuals' belief models. Participants endorsed the voting-based aggregation method and thought the resulting algorithm was trustworthy and fair. Finally, participation led them to trust the organization more and gave them a new perspective on algorithms. Our work contributes to emerging research on understanding and designing human-centered algorithmic systems. We provide a participatory mechanism that directly incorporates individual and collective beliefs into the workings of algorithmic systems, and some early empirical evidence of the impact that participation has on perceptions of algorithms and algorithmic governance.

RELATED WORK

Participation in technology design

In light of the expanding applications of algorithms and artificial intelligence in societal institutions, both industry and academia have begun to emphasize the importance of building and regulating such technology to align with societal and moral values. Rahwan [45] argues for "Society-in-the-loop," which stresses the importance of creating infrastructure and tools to involve societal opinions in the creation of artificial intelligence. Emerging work has also started to explore societal expectations of algorithmic systems such as self-driving cars [8, 41] and robots [36]. Involving stakeholders in the technology design process can be a useful way to encode important social values into new algorithmic systems. Participation in design can be "configured" in a variety of ways, ranging from gathering user insights and requirements to directly involving people in design activities [56]. Value-centered design in particular seeks to understand and design for human values, going beyond utility and usability [39]. A participatory approach to technology has informed many new designs, allowing people to share their knowledge and skills with designers, have control and agency over technologies, and help orchestrate individual and organizational changes [56]. While much research has investigated ways to give users control over algorithmic systems once the systems are already in use, such as supervisory control [50], interactive machine learning [1], and mixed-initiative interfaces [26], artificial intelligence and machine learning systems have rarely leveraged participatory approaches in the design phase. Inspiring emerging work has proposed computational methods and frameworks such as "virtual democracy" [41] and "automated moral decision-making" [22] to incorporate people's moral concepts and judgment in algorithms in the domain of self-driving cars and organ donation matching, but, so far, these frameworks have not been incorporated into real-world algorithmic decision-making systems.

Our work explores a participatory approach to algorithm design, with a particular focus on algorithmic policy—a set of guiding principles for actions, encoded in the governing algorithms. In line with the ethos of work on digital civic technology that helps citizens provide input to governing institutions [38, 4], we directly involve people in the design process of algorithmic technology through a novel combination of individual belief learning, voting, and explanation.

Algorithmic fairness and efficiency

We invited stakeholders to participate in creating an algorithm for resource allocation that involves both the definition of fairness and the adjudication of tradeoffs between fairness and efficiency. Balancing fairness and efficiency is a fundamental theme in modern capitalist democracies [42] as well as in algorithmic governance [16, 58]. Danaher et al. [16] argue that researchers and policymakers should "survey the public about their conception of effective governance to examine the competition between efficiency and fairness". Instead, emerging research suggests that many governing algorithms may in practice focus on efficiency without explicit consideration of fairness or social welfare [2, 12]. Much research has investigated computational ways to make algorithmic decisions more fair. Fair division research has investigated guaranteeing diverse fairness properties [9], and recent work in machine learning has attempted to promote fairness, mostly with a focus on the effects of discrimination against different demographic groups [20]. Other work examines how to balance fairness with metrics such as efficiency [6]. Applying these techniques to real-world applications still requires human judgment and decision-making, as many of these techniques rely on fundamental measures or objective functions that humans must define. For example, all fairness criteria cannot be guaranteed simultaneously, so a human decision-maker must determine which fairness definitions an algorithm should use [13, 30]; individual fairness, or treating similar individuals similarly, requires a definition for "similar individuals" [19]; Likewise, adjudicating the tradeoffs between fairness and efficiency, or

group fairness and individual benefits, requires human judgments about the right objectives for a given problem [6]. Our work leverages participation to make these judgments, thereby increasing the fairness and legitimacy of algorithmic decisions.

CONSIDERATIONS FOR PARTICIPATORY FRAMEWORK

In this section, we draw on the field of political theory, which has investigated collective decision-making and effective citizen participation in governance, and lay out the basic building blocks of the WeBuildAI framework, which enables participation in building algorithmic governance.

Participatory, democratic, algorithmic governance

A first step in participatory governance is to determine what governance issues participants will consider and how directly participation will influence final policy outcomes. User groups, or mini-publics [23], can be configured as open forums where people express their opinions on certain policies; focus groups can be arranged for specific purposes such as providing advice or deriving design requirements. In full participatory democratic governance, citizen voices, whether in open or closed format, can be directly incorporated into the determination of the policy agenda. Our framework focuses on that last form: direct participation in designing algorithmic governance. By direct participation, we mean that people are able to specify objective metrics, functions and behaviors in order to create desirable algorithmic policies. This direct approach can minimize potential errors and biases in codifying policy ideas into computational algorithms, which has been highlighted as a risk in algorithmic governance [29]. In the following sections, we describe the design considerations and associated research questions that motivated our framework.

Individual belief elicitation on algorithmic policy

In order to participate in designing algorithmic governance, individuals should be aware of the policy choices and form their own opinions about those choices. This process requires people to deliberate and examine their judgments across different contexts until they reach a reflective equilibrium, or an acceptable coherence among their beliefs [17, 46]. Our research question is: *How can we enable individuals to form beliefs about policies through deliberation and express these beliefs in a format that the algorithm can implement?*

We explore two ways to promote deliberation and elicit participants' beliefs about algorithmic policy: inferring decision criteria through pairwise comparisons, and asking people to specify their principles and decision criteria. Pairwise comparisons have been used to encourage moral deliberation and determine fairness principles in the form of Rawls' "original position" method [46], and as a way to understand people's judgments in social and moral dilemmas [14], and more recently, moral expectation of AI [41]. Pairwise comparisons can be also be used to capture how people adjudicate varying tradeoffs and conflicting priorities in policy and train governing algorithms, without requiring participants to have an understanding of the specifics of algorithms. The second approach is user creation of rules as used in expert system design [18]. Human-interpretable algorithmic models such as decision trees, rule-based systems, and scoring models have been used to allow people to specify desired algorithmic behaviors. This approach allows people to have full control over the rules and to specify exceptional cases or constraints. However, it can be difficult for people to devise comprehensive rules when making complicated decisions.

Collective decisions

Once individual beliefs are elicited, the next step is to construct a collective concept that consolidates them. Two main theories of collective decision-making can be leveraged: social choice and public deliberation. Social choice theory involves collectively aggregating people's preferences and opinions by creating quantitative definitions of individuals' opinions, utilities, or welfare and then aggregating them according to certain desirable qualities [49]. Voting is one of the most common aggregation methods, in which individuals choose a top choice or rank alternatives, and the alternatives with the most support are selected. Social choice theory offers a scalable approach to produce collective decisions. However, aggregation that satisfies all desirable axiomatic qualities is impossible [49, 3]; a choice needs to be made about what is meant by "desirable" and aggregation guarantees depend on the voting method [34]. The second theory, public deliberation, involves the weighing of competing considerations through a process of public discussion in which participants offer proposals and justifications to support collective decisions [21]. It requires reasoned and well-informed discussion by those involved in or subject to the decisions in question, as well as conditions of equality and respect. In an ideal case, deliberation can result in a final decision through preference assimilation, but it can lead to preference polarization without reaching a final decision.

Both the social choice and the deliberation approaches are feasible for consolidating individual beliefs; a group of individuals can co-construct algorithmic rules through discussion, or individuals' beliefs can be aggregated automatically by voting on alternatives. These methods can also be used in a hybrid manner, such as deliberative polling [21], in which individuals specify their own opinions, deliberate as a group, modify individual opinions, and finally vote. Our framework mainly builds on social choice theory in order to establish baseline models of individual beliefs and to allow new participants to join over time. This framework can be expanded later to incorporate group deliberation as a component in the process as in deliberative polling. Our research question is: *How do people perceive and approve the social choice approach for algorithmic governance?*

Algorithm explanation and human decision support

The final step of the framework is to communicate to participants how their participation has influenced the final policy [23]. Communicating the impact of participation can reward people for their effort and encourage them to further monitor how the policy unfolds over time. While the importance of communication is highlighted in the literature, it has been recognized as one of the components of human governance least likely to be enacted [23]. Algorithmic governance offers new opportunities in this regard because the aggregation of individual models and resulting policy operations are documented. A related challenge in algorithmic governance is how

to support administrators enacting the algorithmic policies. Explaining the logic of any algorithmic decision can be a challenge [33], but this challenge becomes much more complex when algorithms are aggregates of individual decision-making models. Our research question is: *How do we enable people to understand the influence of their participation in resulting policy, and support administrators who use collectively built governing algorithms?*

Who participates

It is important to determine who participates in the creation of algorithmic governance. One widely used and accepted method is volunteer-based participation [23], which accepts input from people who will be governed by the system based on those who choose to participate. Many democratic decisions, including elections, participatory forums, and civic engagement, are volunteer-based. One can also consider expertise or equity issues and focus recruiting efforts on lower-income or minority populations so that the opinions collected are not dominated by the majority, or limit participation to specific groups or stakeholders with certain experiences or expertise. In our application, we used a volunteer-based method with stakeholders directly influenced by the governing algorithm.

WEBUILDAI FRAMEWORK AND APPLICATION

Our framework consists of the three steps identified in the previous section (Figure 1). We applied this framework to the context of on-demand donation matching in collaboration with 412 Food Rescue [57]. 412 Food Rescue is a non-profit that provides a "food rescue" service: Donor organizations such as grocery and retail stores with extra expiring food call 412 Food Rescue, which then matches the incoming donations to non-profit recipient organizations. Once the matching decision is made, they post this "rescue" on their app so that volunteers can sign up to transport the donations to the recipient organizations. The matching policy is at the core of their service operation; while each decision may seem inconsequential, over time, the accumulated decisions impact the welfare of the recipients, the type of work that volunteers can sign up for, and the carbon footprint from the rescues.

Stakeholder participants and research process overview

We worked with a group of 412 Food Rescue stakeholders over a period of a year from September 2017 (Table 1). As our first evaluation of the framework, we chose to work with a small focused group of volunteer-based participants to get in-depth feedback. The entirety of the staff that oversees donation matching at the organization participated. Recipients, volunteers, and donors were recruited through an email that 412 Food Rescue staff sent out to their contact list. We replied to inquiry emails in incoming order, and collected information about their experience with 412 Food Rescue and organizational characteristics to ensure diversity. We limited the number of participants from each stakeholder group to 5–8 people, which resulted in an initial group of 24 participants (including V5a and V5b that participated together) with varying organization involvement.[1] 15 were female and everyone, except one Asian, were white. 16 participants answered our extra demographic survey. Two attended at least some college and 14 had attained at least a bachelor's degree. The average age was 48 (Median=50 (SD=16.4); Min-Max:30-70). The average income household income was $ 65,700 (Median=$62,500 (SD=$39,560); Min-Max:$25,000-$175,000).

We conducted a series of four study sessions with each individual—a combination of survey data collection, participatory model making, think-aloud, and interviews—and a workshop. Because of the extended nature of the community engagement, 15 participants completed all the individual study sessions, and 8 could participate only in the first couple of sessions due to changes in their schedules or jobs. Because participants provided research data through think-alouds and interviews in addition to their input for the matching algorithm, we offered them $10 per hour.

412 Food Rescue†. **F1**$^{@}$ **F2**$^{@}$ **F3**$^{@*}$
Recipient organizations (Client served monthly, client neighborhood poverty rate) **R1**$^{@}$ Human Services Program manager (N=150, 13%) **R2**$^{@}$ Shelter & Food Pantry Center director (N=50, 20%) **R3**$^{@}$ Food pantry employee (N=200, 53%) **R4**^{1} Animal Shelter staff **R5**$^{@}$ Food pantry staff (N=500, 5%) /mkleecheck this number **R6**^{1*} After school program employee (N=20, 33%) **R7**$^{@}$ Home-delivered meals delivery manager (N=50, 11%) **R8** $^{1-2}$ Food pantry director (N=200, 14%)
Volunteers. V1$^{@*}$ White male, 60s **V2**^{1} White female, 30s **V3**$^{@*}$ White female, 60s **V4** dropped out, not counted **V5**$^{@}$ ‡ White female 70s (V5a), white male (V5b) 70s **V6**$^{@}$ White female, 60s **V7**$^{@}$ White female, 20s
Donor organizations. **D1**^{1} School A dining service manager **D2**$^{@}$ School B dining service manager **D3**^{1} Produce company marketing coordinator **D4**$^{@}$ Grocery store manager **D5**^{1} Manager at dining and catering service contractor **D6**^{1*} School C dining service employee

Table 1. Participants. Superscript indicates studies that they participated in: Study 1^{1}, Studies 1-2^{2} , All 1-4 studies$^{@}$, and workshop * †Info excluded for anonymity ‡ a couple participated together

Defining factors for algorithmic policy

We defined inputs for the matching algorithm: factors that have data sources that were deemed to be important by stakeholders in fair allocation [31], and reflect desirable operational behaviors explained by 412 Food Rescue (Table 2).[2] The factors define transportation efficiency, needs of recipients, and temporal allocation patterns. These factors were used to generate the pairwise comparison scenarios and as factors that participants could assign scores to in the user creation model study.

INDIVIDUAL BELIEF ELICITATION

We conducted a series of three studies to develop a model to represent each individual in the final algorithm. Participants first provided pairwise comparisons (Figure 2A, Study 1) to train algorithms using machine learning. Participants who

[1] Our participants were predominantly white female, which reflects the population of volunteers and non-profit administrators in Pittsburgh. This is the result of a volunteer-based method [23]. In our next step, we will do targeted recruiting of minority populations.

[2] We intentionally did not use organization types such as shelters and food pantries, nor location names, because they may communicate racial, gender and age groups of recipients and elicit biased answers based on discrimination or inaccurate assumptions.

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, Low, Extremely low) [55]
Income Level	The median household income of the client neighborhood that a recipient organization serves (0-100K+, Median=$41,283). [10] Indicates access to social and institutional resources [48].
Poverty Rate	Percentage of people living under the US Federal poverty threshold in the clients' neighborhood that a recipient organization serves. (0-60 %; AVG=23% [10])
Last Donation	The number of weeks since the organization last received a donation from 412 Food Rescue. (1 week–12 weeks, never)
Total Donations	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 include meat, dairy, prepared foods and others.

Table 2. Factors of matching algorithm decisions. The range of the factors are based on the real-world distribution

wanted to elaborate on their models participated in the model creation session (Figure 2B, Study 2). If their belief changed after Study 2, they provided a new set of pairwise comparisons to retrain the algorithm. Participants were later asked to choose one of the two models (Study 3).

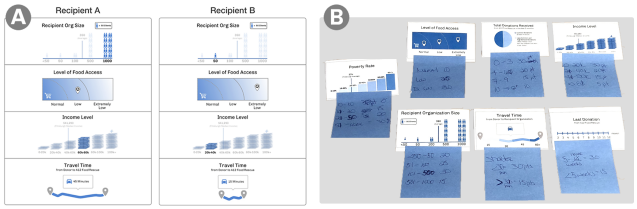

Figure 2. Two methods of belief elicitation were used in our study: (a) algorithm training through answers to pairwise comparison questions, and (b) scoring each factor involved in the algorithmic decision-making.

Model training through pairwise comparisons (Study 1)

Pairwise comparison scenarios

We developed a webapp to generate two possible recipients that randomly vary according to the factors (Table 2), and asked people to choose which recipient should receive the donation (Figure 3a).[3] All the participants completed a one-hour in-person session where they answered 40-50 randomly generated questions. They were asked to think aloud as they made their decisions, and sessions concluded with a short semi-structured interview that asked them for feedback about their thought process and their views of algorithms in general. Throughout the research process, the link to the webapp was sent to the participants who wished to update their models on their own.

[3]Improbable combinations of income and poverty were excluded according to the census data. All the factors were explained in a separate page that participants could refer to.

Learning individual models

We utilize random utility models, which are commonly used in social choice settings to capture choices between discrete objects [37]. This fits our setting, in which participants evaluate pairwise comparisons between potential recipients. In a random utility model, each participant has a true "utility" distribution for each object, and, when asked to compare two potential objects, she samples a value from each distribution. Crucially, in our setting, utility functions do not represent the personal benefit that each voter derives from an allocation. Rather, we assume that when a voter says, "I believe in outcome x over outcome y," this can be interpreted as, "in my opinion, x provides more benefit (e.g., to society) than y." The utility functions therefore quantify societal benefit rather than personal benefit. In order to apply random utility models to our setting, we use the Thurstone-Mosteller (TM) model [53, 40], a canonical random utility model from the literature. In this model, the distribution of each alternative's observed utility is drawn from a Normal distribution centered around a mode utility. Furthermore, as in work by Noothigattu et al. [41], we assume that each participant's mode utility for every potential allocation is a linear function of the allocation's feature vector. Therefore, for each participant i, we learn a single vector β_i such that the mode utility of each potential allocation x is $\mu_i(x) = \beta_i^T x$. We then learn the relevant β_i vectors via standard gradient descent techniques using Normal loss.[4] We also experimented with more complicated techniques for learning utility models, including neural networks, SVMs, and decision trees, but linear regression yielded the best accuracy and is the simplest to explain (see the appendix).

Participant creation of model (Study 2)

To allow participants to explicitly specify allocation rules, we asked them to create a scoring model using the same factors shown in Table 2. We used scoring models because they capture the method of "balancing" factors that people identified when answering the pairwise questions.[5] We asked participants to create rules to score potential recipients so that recipients with the highest scores will be recommended. Participants assigned values to different features using printed-out factors and notes (Figure 3b). We did not restrict the range of score but used 0-30 in our instruction. Once participants created their models, they tested how their scoring rule works with 3-5 pairwise comparisons generated from our webapp, and adjusted their models in response. At the end of the session, we conducted a semi-structured interview in which we asked people to explain their reasons for scoring rules and overall experience. The sessions took one hour. Two participants wanted to further adjust their models and scheduled 30 minute follow-up sessions to communicate their changes.

Machine learning vs user created models (Study 3)

We asked participants to compare and choose between their machine learning and created model, selecting that which best

[4]For participants who consider donation type, we learn two machine learning models, one for common donations and one for uncommon donations.

[5]We also experimented with the use of decision trees, but the models quickly became prohibitively convoluted.

represented their beliefs. To evaluate the performance of the models on fresh data that was not used to train the algorithm, we asked participants to answer a new set of 50 pairwise comparisons[6] before the study session and used them to test how well each model predicted the participants' answers. To explain the models, we represented them both in graph form that showed the assigned scores along the input range for each feature (Figure 3). In order to prevent any potential bias in favor of a model that participants directly specified, we anonymized the models (e.g. Model X), normalized the two models' parameters (beta values) or rubric using the maximum assigned score in each model, and introduced both models as objects of their creation. In a 60-90 minute session, a researcher walked through the model graphs with the participants, showed the prediction agreement scores, and presented all pairwise comparison cases in which the two models disagreed with each other or disagreed with participants' choices. For each case, the researcher illustrated on paper how the two models assigned scores to each alternative. At the completion of these three activities, participants were asked to choose which model they felt best represented their thinking. The models were only identified after their choice was made. A semi-structured interview was conducted at the end asking their experience and reasons for their final model choice.

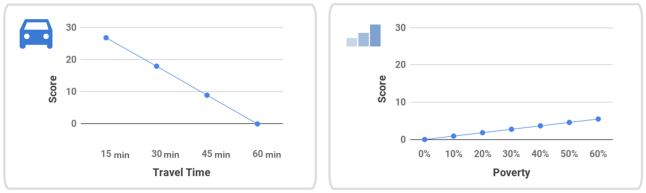

Figure 3. Model explanations. Both machine learning-trained and participant-created models were represented by graphs that assigned scores according to the varying levels of input features.

Analysis

All sessions were audio-recorded and researchers took notes throughout. The qualitative data was analyzed following a qualitative data analysis method [43]. Two researchers read the all notes and documented low-level themes, and the rest of the research team met every week to discuss and organize the themes into higher levels. Individual models were analyzed in terms of the beta values assigned to each factor, or the highest score assigned to each factor. As all the feature inputs were normalized (from 0 to 1), we used the strength of the beta values to rank the importance of factors for each individual.

Results

Final individual models

In total, we trained 18 machine learning models [7] and obtained 15 participant-created scoring models. Of the 15 participants who completed all studies, 10 of them preferred the machine learning model trained on their pairwise comparisons; the other five chose their user created model. In general, the machine learning models resulted in higher overall agreement with participant's survey answers than the user created models when tested on 50 new pairwise comparisons provided by each participant, as seen in Table 3. Either way, we find the prediction accuracy provided by the models to be surprisingly high.

	D2	D4	F2	F3	R1	R2	R3
ML	**0.86**	0.78	**0.92**	**0.92**	**0.90**	0.90	**0.78**
UC	0.68	**0.68**	0.68	0.86	0.80	**0.76**	0.70

	R5	R7	V1	V3	V5	V6	V7
ML	**0.94**	0.74	**0.90**	**0.92**	**0.78**	0.56	0.68
UC	0.92	**0.74**	0.76	0.82	0.82	**0.80**	**0.88**

Table 3. Accuracy of machine learning model (ML) and user created model (UC). Bold denotes the model the participant chose. F1 chose the ML model but did not complete additional survey questions to calculate model agreement.

Effect of elicitation methods

Participants took the process of externalizing their beliefs into the computational models seriously. A few people remarked that creating a model put them under pressure or made them feel that they were "playing God" (V5) by controlling who would receive donations. The sequence of performing pairwise comparisons followed by the model creation session was effective at eliciting and developing their beliefs. Initial pairwise questions helped participants become familiar with the factors and with matching decisions and enabled them to develop initial decision making rules. However, participants said that pairwise feedback could become difficult when presented with alternatives that differed in many features[8] or if they felt they were inconsistent when applying rules to weigh factors.

Building their scoring model helped participants solidify the principles that they began to develop in their first round of pairwise comparisons when they considered all factors at once. Considering one factor at a time in the model creation session enabled participants to identify explicitly which factors mattered and why. For example, nearly all participants began building their scoring models by ordering the factors by importance and identifying some that they did not want to consider. Participants also appreciated that creating explicit rules forced them to reconcile conflicting beliefs that may have been applied inconsistently when judging pairwise comparisons. For example, V1 noted that, in pairwise surveys, he sometimes favored organizations that had not received a donation in a long time because they were receiving less, and sometimes he penalized them, thinking that they were unable or unwilling to accept donations. In the end, his created model favored organizations that had received donations more recently.

Creating a scoring model from a top-down approach evokes a higher level of construal [54] than answering pairwise comparisons. Many participants stated the process of answering

[6] We used the same set of comparisons for all participants for consistency.

[7] We note that there were 8 participants who participated in the first stage of the study but not subsequent stages (Table 1). The average cross-validation accuracy of their linear models was quite high, at 0.819.

[8] E.g., consider a choice between a large recipient organization at a short distance with low poverty rate and two total donations, and a small recipient organization at a further distance with medium poverty rate and one total donation.

pairwise comparisons felt emotional because it made them think of real-world organizations. V1 said that developing scoring rules felt "robotic," but R3 felt that creating the scoring model was easier than the pairwise comparisons because it took the emotion out of the decision making process. For an administrative decision-maker, F3, answering pairwise questions made her focus on day-to-day operational issues like travel time and last donation because she related the questions to real-world decision making. This contrasted with her user-created model, which favored equity related factors like income and poverty. In the end, she chose her machine learning model, stating that while her user-created model appealed as a way of pushing herself beyond her operational thinking, travel time and last donation were just more important. Participants who said they changed their thinking in the user-created model session could answer 50 additional questions to capture their updated beliefs. Some said these extra pairwise comparisons helped them deliberate even further. For example, V7 stated that she tested out whether she actually valued the principles she identified in her user-created model.

In the end, 10 out of 15 participants chose their machine learning models. For many, this was the model they had built last and therefore reflected their current thinking at the time of comparison. Others felt that the machine learning model had more nuance in the way different factors were weighted, and some valued the linearity of the model compared to their manual rules that were often step-wise functions. On the other hand, five participants felt that their user created model better represented their thinking. For four participants, their user created model did a better job of weighing all of the factors that mattered to them and screening off unimportant factors. R2 trusted the reflective process of creating a model and did not trust his pairwise answers nor the machine learning model built from them, given his difficulty balancing all seven factors in his head and fatigue in answering many questions, even though the accuracy of machine learning model was 90% compared to 76% for the model that he created.

COLLECTIVE AGGREGATION

Our framework uses a voting method to aggregate individuals' beliefs. When given a new donation, each individual's model generates a complete ranking of all possible recipient organizations. The Borda rule aggregates these rankings to derive recommendations. We conducted a workshop and interviews to understand participants' approval of this method.

Borda voting

We use the Borda rule to aggregate opinions because it provides robust theoretical guarantees in the face of noisy estimates of true preferences, as shown in a paper by some of the authors [28]. The Borda rule is defined as follows. Given a set of voters and a set of m potential allocations, where each voter provides a complete ranking over all allocations, each voter awards $m-k$ points to the allocation in position k, and the Borda score of each allocation is the sum of the scores awarded to that allocation in the opinions of all voters. Then, in order to obtain the final ranking, allocations are ranked by non-increasing score. For example, consider the setting with two voters and three allocations, a, b, and c. Voter 1 believes that $a \succ b \succ c$ and voter 2 believes that $b \succ c \succ a$, where $x \succ y$ means that x is better than y. The Borda score of allocation a is $2+0=2$, the Borda score of allocation b is $1+2=3$, and the Borda score of allocation c is $0+1=1$. Therefore, the final Borda ranking is $b \succ a \succ c$.

Method

We first conducted a workshop in order to gauge participants' reactions to the Borda aggregation method and get initial approval for its further use. Five participants (Table 1) attended the one-hour workshop. All stakeholder groups were represented. We prepared a handout that showed individuals' and stakeholders' average models at the time, and a diagram that explained how the Borda method worked. We facilitated a discussion of how individuals reacted to the similarities and differences between their model and other groups' models, and had individuals discuss whether all the stakeholders' opinions should be weighted equally or differently. The workshop was audio-recorded and a researcher took notes during the workshop. Following [43], two researchers read the notes and created low-level codes, and the whole research team met later to discuss and create thematic groups. We received approval of, and positive feedback on Borda aggregation from the workshop. In a later study session (Study 4) we conducted individual interviews with all remaining participants in order to solicit their opinions on other stakeholder models and the Borda method.

Results

Responses to the Borda method of aggregation

Participants appreciated that the Borda method gave every recipient organization a score (n=5) and that it embodied democratic values (n=4). In Study 4, F1 felt that giving every organization a score captured the subtleties of her thinking better than other voting methods: "I appreciate the adding up [of] scores. Recognize the subtleties." V3 also stated that being able to rank all recipients is "more true to...[being] able to express your beliefs.' R1 approved of the method saying, "It's very democratic," relating it to forms of human governance. Two other individuals, D2 and D4, approved of the method and related it to voting systems in the US. D4 recognized that some US cities in California recently used a similar voting method for their mayoral election. We acknowledge that their approval is given in isolation and is limited by the lack of comparison to other methods. Participants expressed difficulty thinking of alternatives (n=3), for example, R2 said, "I guess I don't know what the alternative way to do it would be, so I'm okay with it."

Varying stakeholders' voting influence

All but one participant believed that the degree to which different stakeholders should weigh in the final algorithm depended on their roles. On average, participants assigned 46% of the voting power to 412 Food Rescue, 24% to recipient organizations, 19% to volunteers, and 11% to donors. Nearly all participants weighted 412 Food Rescue staff as the highest group (n=13), as people recognized that they manage the operation and have the most knowledge of the whole system. Donors were weighted the least (or tied for least) by nearly all participants (n=14) including donors themselves, as they are

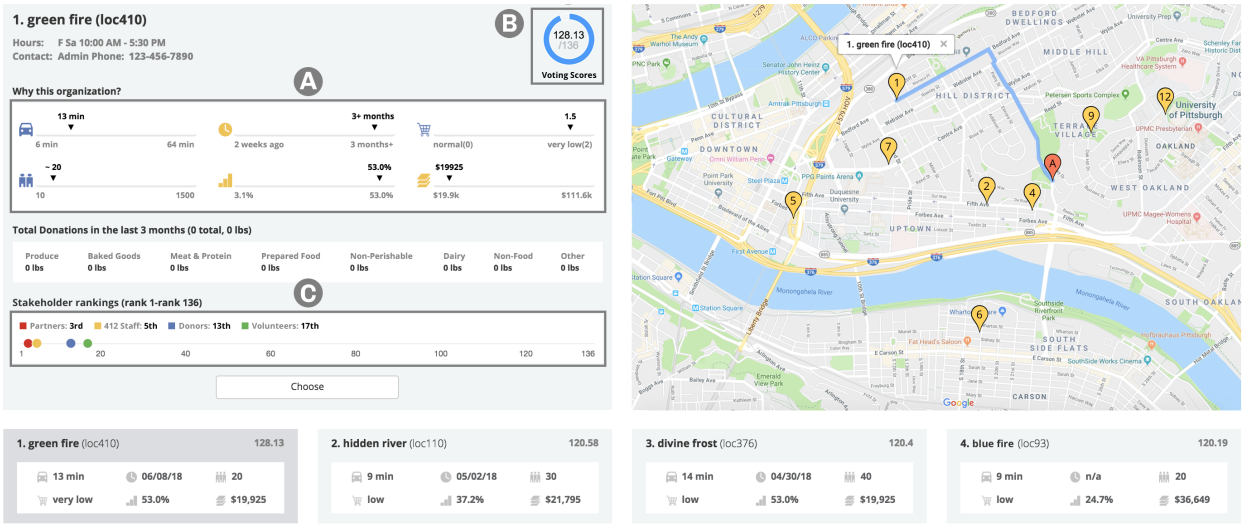

Figure 4. The Decision support tool explains algorithmic recommendations, including the nature of stakeholder participation, stakeholder voting results, and characteristics of each recommendation. The interface highlights (a) the features of the recommended option that led to its selection, (b) the Borda scores given to the recommended options in relation to the maximum possible score, and (c) how each option was ranked by stakeholder groups. (All recipient information and locations are fabricated for the purpose of anonymization.)

not involved in the process once the food leaves their doors. Recipients and volunteers were weighted similarly as participants recognized that recipient opinions are important to the acceptance of donations and volunteer drivers have valuable experience interacting with both donors and recipients. In order to translate these weights to Borda aggregation, we allocated each stakeholder group a total number of votes that was commensurate with their weight, and divided up the votes evenly within each group. For example, if the 412 Food Rescue employees had been assigned 45% of the weight, this translates to allocating them 45 votes out of 100 total as a group, where each employee's vote is replicated 15 times.

EXPLANATION AND DECISION SUPPORT

Once recommendations are generated, the decision support interface presents the top twelve organizations and explains them to support the human decision-maker who matches incoming donations to recipients. We used this explanation to demonstrate to participants how their participation had been incorporated into algorithmic decision-making. We also explained average stakeholder models to participants so that they could learn about others' models.

Design of decision-support tool

We designed the decision-support tool so that a human administrator can use algorithmic decision-making (Figure 4). While the tool was designed with many different considerations, such as choice architecture [52], they are beyond the scope of this paper. We focus on the explanation of decisions from collectively-built algorithms.

- Decision outcome explanation (Figure 4a): We used "input influence" style explanation [7]. Features are highlighted in yellow when an organization is in the top 10% of recipients ranked by that factor. For example, poverty rate is highlighted because the selected organization is in the top 10% of recipients when ranked from highest to lowest poverty rate.

- Voting score (Figure 4b): The Borda score for each option is displayed. It shows this option's scores in relation to the maximum possible score that an option can receive (scores when every individual model picks this option as its first choice). This can indicate the degree of consensus among participants.

- Stakeholder rankings (Figure 4c): Stakeholder rankings show how each stakeholder group ranked the given recipient on average. It is a visual reminder that all 412 Food Rescue stakeholder groups are represented in the final algorithm and gives the decision-maker additional information about the average opinion of each stakeholder group.

We implemented the interface by integrating it into a Customer Relation Management system currently in use at 412 Food Rescue. We used the Ruby on Rails framework. Algorithms were coded in Ruby on Rails, the front-end interface used Javascript and Bootstrap, and the database was built with Postgres. The distance and travel time between donors and recipients were pre-populated using the Google Maps API and Python, and we used the donor and recipient information in the past five months of donation rescue records in the database. On average, the algorithm produced recommendations for each donation in 10 seconds.

Method (Study 4)

We conducted a one-hour study with each participant to understand how this explanation interface influences their perception of governing algorithms and their attitude toward 412 Food Rescue. We first showed participants the graphs of their individual model and graphs of the averaged models for each stakeholder group, and asked participants to examine similarities and differences among these models. We next had

participants interact with the decision support tool run on a researcher's laptop. The researcher walked participants through the interface, explaining the information and recommendations, and asked them to review the recommendations and pick one to receive the donation. After each donation, participants were asked their opinions of the recommendations, the extent to which they could see their models reflected in the results, and their general experience. We concluded with a 30 minute semi-structured interview in which we asked how participation influenced their attitude toward algorithms and their view of the 412 Food Rescue organization. We also asked participants to reflect upon the overall process of giving feedback throughout our studies.

Analysis

The entire interview session was audio-recorded and transcribed. We used a qualitative data analysis method [43]. One researcher read all the transcripts and added initial coding on Dedoose. Low-level codes were organized into emerging thematic groups through discussion with other researchers in research team meetings. Many themes arose, including appreciation of human-in-the-loop governance, but we focus on themes related to participation. In order to generate summary beta vectors for each stakeholder group, we normalized the beta vectors for all stakeholders in the group and took the pointwise average. This yields a summary beta vector where the value of each feature roughly reflects the average weight that stakeholders in the same group give to that feature.

Findings

Reviewing stakeholder models

There were many ideological similarities across stakeholder models. All participants considered efficiency and fairness concerns. For example, all stakeholder group models valued distance as one of the top three factors and favored organizations that were deemed to be in greater need (e.g., higher as opposed to lower poverty, lower as opposed to higher food access). Organization size was the only factor with divided views arguing for larger or smaller organizations. The main source of disagreement among models was how the factors were balanced. 412 Food Rescue Staff tended to weight travel time and last donation significantly more than the other factors. Donors, recipients and volunteers tended to give all factors other than organization size more equal relative importance.

The ideological similarity across models gave participants assurance that they share guiding philosophical principles with other participants (n=8). For example, R7 was pleased to note that all participants were "on the same page" and concluded that "no matter what group or individuals we're feeding, [we] have the same regard for the food and the individuals that we're serving." Participants still acknowledged differences in balancing trade-offs between operational considerations and fairness factors. R1, referencing how important travel time was to her, mentioned that hers is more of a "business model" whereas others were more altruistic by weighting more heavily factors like income and food access. Others reacted to differences by questioning the algorithm (n=2) or their own thinking (n=5). V7 was concerned and upset that 412 Food Rescue staff did not weight heavily her most important factors (food access, income, and poverty) and her trust in the algorithm was lowered as a result. When F2 saw that volunteers did not weight travel time as highly as she had thought, she questioned her evaluation of travel time: "Maybe [volunteers] don't care as much. I think you end up hearing from the people who care... It's like that saying with customer service: Only complain when something's happened." In light of these differences, participants appreciated that the algorithm aggregates multiple models. Finally, others were undisturbed or even pleased to see differences in the models (n=3). R3 was pleased that other participants were considering unique viewpoints. Likewise V5 and R1 both stated that it is natural to expect differences between stakeholders given that everyone has unique experiences and that "this is the point of democracy" (V5).

Reactions to decision support interface

Participants were interested in the stakeholder rankings and asked to see more information. Given that the top twelve results often did not show the first choice for any stakeholder group, several participants wanted to see the first choice for each stakeholder group in addition to the voting aggregation scale (n=7). Participants appreciated that the stakeholder ranking showed opinions that may be different from those of 412 Food Rescue dispatchers (n=4). V7, who was concerned that 412 Food Rescue staff did not heavily weight factors that were important to her, was pleased that the voter preference scale illustrated the difference between her stakeholder group's average model and 412 Food Rescue's average model. She hoped that the staff would see that their thinking differed from other stakeholders and perhaps reconsider their decisions to be more inclusive of other groups' opinions. 412 Food Rescue staff were interested in the information as well and F3 mentioned that, while she would not solely base her decisions on stakeholder ranking information, she may use it as a tiebreaker between two similar organizations.

Participation and perceptions of algorithmic governance

Collective participation strengthened the moral legitimacy of the algorithm for participants (n=12). Some expressed that collective participation expands the algorithm's assumptions beyond those of the organization and developers (n=6). V7 noted that it is easy for organizations to remain isolated in their own viewpoints and that building a base of collective knowledge was more trustworthy to her than "412 [Food Rescue] in a closed bubble coming up with the algorithm for themselves." V3 echoed this sentiment, stating that collective participation was "certainly more fair than somebody sitting at a desk trying to figure it out on their own. These are everybody's brain power who were deemed to be important in this decision... it should be the most fair that you could get." At 412 Food Rescue, F2 stated that "getting input from everyone involved is important" to challenge organizational assumptions and increase the effectiveness of their work. Other participants noted that all stakeholders have limited viewpoints that can be overcome with collective participation (n=3). R1 felt the algorithm would be fair only "if you took the average of everybody. ...[My model] is only my experience. And I view my experience differently than the next place down the road. And my experience is subjective."

However, two individuals had concerns not about the algorithm itself but over the quality of participant input. V7 and F1 questioned the quality of of participant input, doubting both limitations of participant knowledge and limitations in participant ability to construct an accurate model.

Collective participation in the algorithmic building process led many participants to increase the degree to which they viewed 412 Food Rescue positively (n=8). For some participants, this happened because participation exposed the difficulty of making donation allocation decisions which in turn made them thankful for the work of the organization (n=4). For example, after seeing how similar the recommended recipients can be in the interface, D2 and V3 both expressed thankfulness for 412 Food Rescue after experiencing the difficulty and weight of making the final decision. Participants also expressed appreciation for the organization's concern for fairness and the effort needed to continually make such decisions.

The algorithmic building process also increased some participants' motivation to engage with the organization (n=4). Many participants appreciated that their opinions were valued by the organization enough to be considered in the algorithm building process and expressed that they may increase their involvement with the organization in the future either through increased volunteer work (V3&7) or donation acceptance (R2).

Participation and perceptions of general algorithms

For some participants, seeing how the two models predicted their answers in our study session made them rethink their initial skepticism and begin to trust the algorithm. V1, who in earlier studies expressed doubt that an algorithm could be of any use in such a complex decision space, stated at the end of Study 3 that he now "wholeheartedly" trusted the algorithm, a change brought about by seeing the work that went into developing his models and how they performed. F3 expressed that before participating, "the process of building an algorithm seemed horrible" given the complexities of allocation decisions. Seeing how the process of building the algorithm was broken down "into steps ... and just taking each one at a time" made the algorithm seem much more attainable. For D2, interacting with the researchers who were building the algorithm gave him an awareness of the role human developers play in determining algorithms. He said that, after this process, his judgment of an algorithm's fairness would be based on "how it was developed and who's behind it and programmed and how it's influenced." D2 expressed that the final algorithm was fair because he came to know and trust the researchers over the course of his participation.

DISCUSSION

In this paper, we envision a future in which people can collectively build ideal algorithmic governance mechanisms for their own communities. Our framework, WeBuildAI, represents the first implementation and evaluation of a system that enables people to collectively provide inputs to and design real-world algorithmic governing decisions. In doing so, we contribute to the emerging research agenda on perceptions of algorithmic fairness, by advancing the understanding of the effects of participation.

Our findings suggest that participation in algorithmic governance can result in the same positive effects created by participation in human governance and service. Our participants reported greater trust in and perceived fairness of the governing institution and administrative decisions after participating. In addition, they were more motivated to use the services, felt respected and empowered by the governing institution, and felt a collective understanding of its decision-making process. Previous work on participation in technology suggests that participation not only results in new technology design, but also affects participating individuals and organizations [56]. We observed this in our study as well, as participation increased participants' algorithmic literacy. Through the process of translating their judgments into algorithms, they gained new understanding of and appreciation for algorithms.

These findings demonstrate the participatory framework's potential for implementing morally legitimate, fair, and motivating algorithmic governance, whether in bureaucratic government decision-making, digital platforms, or new algorithmic systems such as self-driving cars or robots. There is emerging evidence that algorithms used in public policy and digital platforms have undesirable biases; the "techno-logic" of digital platforms and their neutrality is increasingly called into question, and cannot be fairly determined by a group of engineers alone. Applying a participatory design approach can promote a culture of awareness around the disparate effects that algorithms can have on different stakeholders, as well as distribute accountability for decisions among stakeholders rather than placing the onus of decision-making on the developers alone.

In its current implementation, our framework can be applied to contexts in which instant runtime is not required. For example, our framework can be used in governing algorithms that allocate public resources or contribute to smart planning services, placement algorithms in school districts or online education forums, or hiring recommendation algorithms that balance candidates' merits with equity issues. While our implementation involved all affected stakeholders and made both individual and collective models transparent, both participation and transparency can be tailored depending on organizations' goals and constraints. Additionally, the individual belief elicitation tools in our framework can be used on their own even in settings where direct participatory governance is not feasible. They can be used to understand stakeholders' values with respect to governance issues, or as an evaluative tool to examine how existing algorithms operate and whether they are in line with particular stakeholder groups' beliefs.

When people participate in building systems, those systems become more transparent to them and they gain a deeper understanding of how the systems work. While this is one of the main sources of trust, one potential concern is that people will use this knowledge to game and strategically manipulate the system. Indeed, one of the main topics of research in computational social choice [9] is the design of voting rules that discourage strategic behavior—situations where voters report false preferences in order to sway the election towards an outcome that is more favorable according to their true preferences. However, we view this as a nonissue for our framework,

because voters do not vote directly. Although in theory it is possible to manipulate one's pairwise comparisons or specify preferences to obtain a model that might lead to preferred outcomes in very specific situations, the same model would play a role in multiple, unpredictable decisions. The relation between their models and future outcomes is so indirect that it is virtually impossible for voters to benefit by behaving strategically.

LIMITATIONS AND FUTURE WORK

Our work has limitations that readers must keep in mind when applying the framework. Our studies evaluated people's experiences with participation, as well as their attitudes toward and perceptions of the resulting algorithmic systems. As our next step, we will deploy the system in the field in order to understand long-term effects and behavioral responses. Additionally, in developing our framework, we intentionally used a focused group of participants to get in-depth insights and feedback on our tools and framework. As we implement our next version, we will examine participation with a larger group of people and explore the possibility of running an open system, where people can continuously add their inputs. Finally, our framework needs to be tested with other contexts and tasks that involve different cultures and group dynamics. We are particularly interested in the effects of participation when collective opinions are polarized. In addition, many of the benefits of participation and the resulting algorithms are procedural effects. One interesting future research direction is to empirically evaluate the effectiveness of collectively built algorithms based solely on their outcomes, and see whether the theory of the "wisdom of crowds" applies to algorithms built through participation.

We hope that our work will serve as a building block for a larger process that would ultimately enable us, as a society, to collectively shape the use of emerging algorithmic technology in socially responsible and meaningful ways.

REFERENCES

1. Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
2. Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, and Propublica. 2016. Machine Bias. (2016).
3. Kenneth J Arrow. 2012. *Social choice and individual values*. Vol. 12. Yale university press.
4. Mara Balestrini, Yvonne Rogers, Carolyn Hassan, Javi Creus, Martha King, and Paul Marshall. 2017. A city in common: A framework to orchestrate large-scale citizen engagement around urban issues. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2282–2294.
5. Neeli Bendapudi and Robert P Leone. 2003. Psychological implications of customer participation in co-production. *Journal of Marketing* 67, 1 (2003), 14–28.
6. Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. 2012. On the efficiency-fairness trade-off. *Management Science* 58, 12 (2012), 2234–2250.
7. Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 377.
8. Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.
9. Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. *Handbook of Computational Social Choice*. Cambridge University Press.
10. US Census Bureau. 2018. American FactFinder. (2018).
11. Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*. ACM, 89–96.
12. Anupam Chander. 2016. The racist algorithm? *Michigan Law Review* 115 (2016), 1023.
13. Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.
14. Stéphane Côté, Paul K Piff, and Robb Willer. 2013. For whom do the ends justify the means? Social class and utilitarian moral judgment. *Journal of Personality and Social Psychology* 104, 3 (2013), 490.
15. John Danaher. 2016. The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology* 29, 3 (2016), 245–268.
16. John Danaher, Michael J Hogan, Chris Noone, Rónán Kennedy, Anthony Behan, Aisling De Paor, Heike Felzmann, Muki Haklay, Su-Ming Khoo, John Morison, and others. 2017. Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society* 4, 2 (2017).
17. Norman Daniels. 2016. Reflective Equilibrium. In *Stanford Encyclopedia of Philosophy*.
18. Robyn M Dawes and Bernard Corrigan. 1974. Linear models in decision making. *Psychological Bulletin* 81, 2 (1974), 95.
19. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*. ACM, 214–226.
20. FATML. 2018. Fairness, Accountability, and Transparency in Machine Learning Workshop. (2018).

21. James S Fishkin, Robert C Luskin, and Roger Jowell. 2000. Deliberative polling and public consultation. *Parliamentary Affairs* 53, 4 (2000), 657–666.
22. Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P Dickerson, and Vincent Conitzer. 2018. Adapting a kidney exchange algorithm to align with human values. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*.
23. Archon Fung. 2003. Recipes for public spheres: Eight institutional design choices and their consequences. *Journal of Political Philosophy* 11, 3 (2003), 338–367.
24. Archon Fung. 2015. Putting the public back into governance: The challenges of citizen participation and its future. *Public Administration Review* 75, 4 (2015), 513–522.
25. Tarleton Gillespie. 2010. The politics of "platforms". *New Media & Society* 12, 3 (2010), 347–364.
26. Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the 1999 CHI Conference on Human Factors in Computing Systems*. ACM, 159–166.
27. Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD)*. 133–142.
28. Anson Kahng, Min Kyung Lee, Ritesh Noothigattu, Ariel D Procaccia, and Christos-Alexandros Psomas. 2018. Statistical Foundations of Virtual Democracy. Manuscript. (2018).
29. Rob Kitchin. 2017. Thinking critically about and researching algorithms. *Information, Communication & Society* 20, 1 (2017), 14–29.
30. Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
31. Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3365–3376.
32. Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*. ACM, 1603–1612.
33. Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).
34. Christian List. 2017. Democratic deliberation and social choice: A review. In *The Oxford Handbook of Deliberative Democracy*, Andre Bächtiger, John S Dryzek, Jane Mansbridge, and Mark Warren (Eds.). Oxford University Press.
35. R Duncan Luce. 2012. *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation.
36. Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the 10th annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 117–124.
37. Charles F Manski. 1977. The structure of random utility models. *Theory and Decision* 8, 3 (1977), 229–254.
38. J Nathan Matias and Merry Mou. 2018. CivilServant: Community-led experiments in platform governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 9.
39. Jessica K Miller, Batya Friedman, Gavin Jancke, and Brian Gill. 2007. Value tensions in design: The value sensitive design, development, and appropriation of a corporation's groupware system. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work (GROUP)*. ACM, 281–290.
40. Frederick Mosteller. 2006. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. In *Selected Papers of Frederick Mosteller*. Springer, 157–162.
41. Ritesh Noothigattu, Neil S Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D Procaccia. 2018. A Voting-Based System for Ethical Decision Making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*.
42. Arthur M Okun. 1975. *Equality and Efficiency: The Big Tradeoff*. Brookings Institution Press.
43. Michael Q Patton. 1980. *Qualitative Research and Evaluation Methods*. Sage Publications, Inc.
44. Robin L Plackett. 1975. The analysis of permutations. *Applied Statistics* (1975), 193–202.
45. Iyad Rahwan. 2018. Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.
46. John Rawls. 2009. *A Theory of Justice*. Harvard University Press.
47. Dillon Reisman, Jason Schultz, K Crawford, and M Whittaker. 2018. Algorithmic impact assessments: A practical framework for public agency accountability. (2018).
48. Robert J Sampson, Jeffrey D Morenoff, and Thomas Gannon-Rowley. 2002. Assessing "neighborhood effects": Social processes and new directions in research. *Annual Review of Sociology* 28, 1 (2002), 443–478.
49. Amartya Sen. 2017. *Collective Choice and Social Welfare: Expanded edition*. Penguin UK.

50. Thomas B Sheridan. 2002. *Humans and Automation: System Design and Research Issues.* Human Factors and Ergonomics Society.

51. Will Sutherland and Mohammad H Jarrahi. 2017. The gig economy and information infrastructure: The case of the digital nomad community. *Proceedings of the ACM Conference on Human-Supported Cooperative Work (CSCW)* 1 (2017), 97.

52. Richard H Thaler, Cass R Sunstein, and John P Balz. 2014. Choice architecture. (2014).

53. Louis L Thurstone. 1959. *The Measurement of Values*. University of Chicago Press.

54. Yaacov Trope and Nira Liberman. 2010. Construal-level theory of psychological distance. *Psychological Review* 117, 2 (2010), 440.

55. USDA. 2017. Food access research atlas. (2017).

56. John Vines, Rachel Clarke, Peter Wright, John McCarthy, and Patrick Olivier. 2013. Configuring participation: On how we involve people in design. In *Proceedings of the 2013 CHI Conference on Human Factors in Computing Systems*. ACM, 429–438.

57. 412 Organization Website. 2018. (2018). `https://412foodrescue.org`

58. Tal Zarsky. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values* 41, 1 (2016), 118–132.

APPENDIX

LEARNING MODELS OF VOTERS

Throughout this entire process, we evaluate each model by withholding 14% of the data and using that as a test set. Once we train the models on 86% of the data, we evaluate their performance on the test set and report the average accuracy of the model.

Random Utility Models with Linear Utilities

Random utility models are commonly used in social choice settings to capture settings in which participants make choices between discrete objects [37]. As such, they are eminently applicable to our setting, in which participants evaluate pairwise comparisons between potential recipients.

In a random utility model, each participant has a true "utility" distribution for each potential allocation, and, when asked to compare two potential allocations, she samples a value from each distribution and reports the allocation corresponding to the higher value she sees. Crucially, in our setting, utility functions do not represent the personal benefit that each voter derives, as is standard in other settings that use utility models. Rather, we assume that when a voter says, "I prefer outcome x to outcome y," this can be interpreted as, "in my opinion, x provides more benefit (e.g., to society) than y." The utility functions therefore quantify societal benefit rather than personal benefit.

In order to apply random utility models to our setting, we must exactly characterize, for each participant, the distribution of utility for each potential allocation. We consider two canonical random utility models from the literature: Thurstone-Mosteller (TM) and Plackett-Luce (PL) models [53, 40, 44, 35]. Both of these models assume that the distribution of each alternative's observed utility is centered around a mode utility: the TM model assumes that the distribution of each alternative's observed utility is drawn from a Normal distribution around the mode utility, and the PL model assumes that the distribution of each alternative's observed utility is drawn from a Gumbel distribution around the mode utility.

As in work by [41], we assume that each participant's mode utility for every potential allocation is a linear function of the feature vector corresponding to the allocation; that is, the mode utility is some weighted linear combination of the features. For each participant i, we learn a single vector β_i such that the mode utility of each potential allocation x is $\mu_i(x) = \beta_i^T x$. We then learn the relevant β_i vectors via standard gradient descent techniques using Normal loss for the TM utility model and logistic loss for the PL utility model.[9]

Specific Design Decisions

Separate Models for Different Donation Types

Certain participants consider donation type when allocating donations, whereas most do not. In light of this, we train two separate machine learning models for participants who consider donation type (one for common donations and one for less common donations), and we train one machine learning model for participants who did not consider donation type. Although training two separate models for participants who did consider donation type resulted in roughly half the training data for each model, the models were more accurate overall.

Quadratic Utilities

Many participants had non-monotonic scoring functions for various features. One common example was organization size: multiple participants awarded higher weight to medium-size organizations and lower weight to both small and large organizations. In order to capture non-monotonic preferences, we tested a quadratic transformation of features, where we learned linear weights on quadratic combinations of features. Concretely, given a feature vector $\vec{x} = (x_1, x_2, x_3)$, we transform $\vec{x}$ into a quadratic feature vector $\vec{x_2} = (x_1, x_1^2, x_2, x_2^2, x_3, x_3^2)$ and learn a vector β_i for each participant i. Although this allowed us to more accurately capture the shapes of participants' value functions, it resulted in slightly lower accuracy overall. This is most likely due to the increased size of the β_i vectors we learned—in general, learning parameters for more complex models with the same amount of data decreases performance.

[9] Logistic loss captures the PL model because the logistic function can be interpreted as the probability of one alternative beating the other (implicitly captured by the structure of the PL model), and logistic loss is the negative log of this probability.

TM vs. PL

Overall, learning Thurstone-Mosteller models performed better than learning Plackett-Luce models.

Cardinal vs. Ordinal Feature Values

We also experimented with cardinal vs. ordinal feature values, where cardinal features use the values themselves and ordinal features only take the rank of the feature value among all possible values for the feature.This was only relevant for recipient size, which was the only feature with nonlinear jumps in possible value. Overall, training on cardinal feature values led to slightly higher accuracy than training on ordinal feature values.

Polynomial Transformations of Features

In order to capture nonlinear mode utilities, we tested a polynomial feature transformation where we learned linear weights on polynomial combinations of features up to degree 4. For instance, given a feature vector $\vec{x} = (x_1, x_2, x_3)$, a polynomial combination of these features of degree 2 transforms each feature vector $\vec{x}$ into an expanded feature vector $\vec{x_2} = (x_1, x_2, x_3, x_1^2, x_1x_2, x_1x_3, x_2^2, x_2x_3, x_3^2)$. We again learn a single β_i vector for each participant i on these transformed features; note that the length of the β_i vectors increases, which stretches our already sparse data even further. We observed that accuracy monotonically fell with increasing degree of the transformed feature values; linear features performed the best.

Pair-Based Approaches

We also learned models for straightforward comparisons; i.e., without random utility models. For all of these models, we transformed comparison data of the form (x_i^1, x_i^2, y_i), where x_i^1 and x_i^2 are the feature vectors for the two recipients and y_i is the recipient that is chosen, into $(x_i^1 - x_i^2, y_i)$, as in the work of Joachims [27]. This allowed us to train models with fewer parameters and ameliorate the effects of overfitting on our small dataset.

Rank SVM

We implement Ranking SVM, as presented by Joachims [27], which resembles standard SVM except we transform the data into pairs, as discussed above. We use hinge loss as the loss function, as is standard with SVMs.

Decision Tree

After again transforming the data into pairwise comparison data, we implement a CART decision tree with the standard scikit-learn DecisionTreeClassifier. However, we both limit the depth of the tree and prune the tree in a post-processing step because it overfit tremendously to our data.

Neural Network (RankNet)

Lastly, we implement a single-layer neural network with the pairwise feature transform, identity activation function, and logistic loss. This was based on the RankNet algorithm of [11]. We note that this is, in essence, equivalent to learning a linear utility model (in particular, a PL model). However, as seen below, it slightly out-performs the aforementioned linear utility model.

Final Model

In general, we found that approaches that learned (linear) utilities for random utility models strongly outperformed pair-based approaches.

Therefore, due to both its simplicity and good performance, our final model is the TM utility model with linear mode utility. Crucially, it is quite easy to summarize and explain to constituents, as utilities are linear with respect to features.