WeBuildAI: Participatory Framework for Algorithmic Governance

MIN KYUNG LEE, University of Texas at Austin & Carnegie Mellon University, USA
DANIEL KUSBIT, Ethics, History & Public Policy, Carnegie Mellon University, USA
ANSON KAHNG, School of Computer Science, Carnegie Mellon University, USA
JI TAE KIM, School of Design, Carnegie Mellon University, USA
XINRAN YUAN, Information Systems, Carnegie Mellon University, USA
ALLISSA CHAN, School of Design, Carnegie Mellon University, USA
DANIEL SEE, Decision Science & Art, Carnegie Mellon University, USA
RITESH NOOTHIGATTU, School of Computer Science, Carnegie Mellon University, USA
SIHEON LEE, Information Systems, Carnegie Mellon University, USA
ALEXANDROS PSOMAS, School of Computer Science, Carnegie Mellon University, USA
ARIEL D. PROCACCIA, School of Computer Science, Carnegie Mellon University, USA

Algorithms increasingly govern societal functions, impacting multiple stakeholders and social groups. How can we design these algorithms to balance varying interests in a moral, legitimate way? As one answer to this question, we present WeBuildAI, a collective participatory framework that enables people to build algorithmic policy for their communities. The key idea of the framework is to enable stakeholders to construct a computational model that represents their views and to have those models vote on their behalf to create algorithmic policy. As a case study, we applied this framework to a matching algorithm that operates an on-demand food donation transportation service in order to adjudicate equity and efficiency trade-offs. The service's stakeholders—donors, volunteers, recipient organizations, and nonprofit employees—used the framework to design the algorithm through a series of studies in which we researched their experiences. Our findings suggest that the framework successfully enabled participants to build models that they felt confident represented their own beliefs. Participatory algorithm design also improved both procedural fairness and the distributive outcomes of the algorithm, raised participants' algorithmic awareness, and helped identify inconsistencies in human decision-making in the governing organization. Our work demonstrates the feasibility, potential and challenges of community involvement in algorithm design.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: participatory algorithm design, collective participation, human-centered AI, matching algorithm, algorithmic fairness

Authors' addresses: Min Kyung Lee, University of Texas at Austin & Carnegie Mellon University, USA, minkyung.lee@austin.utexas.edu; Daniel Kusbit, Ethics, History & Public Policy, Carnegie Mellon University, USA; Anson Kahng, School of Computer Science, Carnegie Mellon University, USA; Ji Tae Kim, School of Design, Carnegie Mellon University, USA; Xinran Yuan, Information Systems, Carnegie Mellon University, USA; Allissa Chan, School of Design, Carnegie Mellon University, USA; Daniel See, Decision Science & Art, Carnegie Mellon University, USA; Ritesh Noothigattu, School of Computer Science, Carnegie Mellon University, USA; Siheon Lee, Information Systems, Carnegie Mellon University, USA; Alexandros Psomas, School of Computer Science, Carnegie Mellon University, USA; Ariel D. Procaccia, School of Computer Science, Carnegie Mellon University, USA.

2573-0142/2019/11-ART181 $15.00
https://doi.org/10.1145/3359283

ACM Reference Format:
Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 181 (November 2019), 35 pages. https://doi.org/10.1145/3359283

1 INTRODUCTION

Computational algorithms increasingly take on governance and management roles in administrative and legal aspects of public and private decision-making [26, 27, 47, 79]. 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, advertisers, and the platforms themselves [39]; on-demand work platforms use algorithms to assign tasks, which affects their customers, their workers, and their own profits [41, 51, 72]; and city governments use algorithms to manage police patrols, neighborhood school assignments, and transportation routes [67, 76].

These governing algorithms can have a substantial impact on our society; they can enable efficient, data-driven decisions at massive scale, but they also risk invisibly perpetuating socially undesirable or erroneous decisions. Recent real-world cases suggest that algorithmic governance can lead to compromises in social values and unfairly prioritize a small set of stakeholders' benefits at the cost of others' [4, 21, 82]. For example, the objective of the social media curating algorithms is to maximize the profits of the company and satisfy the advertisement providers, often at the cost of social values such as healthy media consumption and privacy [16]. Algorithms used in public assistance automate decisions for efficiency and risk having disparate impacts on the groups of people affected by the decisions [4]. When algorithms are designed without considering a community's needs, as in the case of Boston's bus scheduling system, they may receive pushback from the community and ultimately not be adopted [82].

Emerging work has called for greater involvement of stakeholders and affected communities in the development of algorithmic systems. These projects have sought to understand the public's expectations of moral behaviors [14, 54, 60] and varying concepts of fairness [48, 50, 84], as well as stakeholders' needs and requirements [2, 87] around Artificial Intelligence (AI) systems; yet translating the results into actual algorithms is difficult, as these studies have often relied on hypothetical moral dilemmas or collected qualitative expectations and opinions that developers and designers need to interpret in order to build the algorithm.

Our vision is to empower people to design algorithmic governance mechanisms for their own communities. We argue that this participatory algorithm design process is a step toward creating algorithmic governance that is effective yet also moral. In traditional participatory governance, stakeholder participation in policy-making improves the legitimacy of a governing institution in a democratic society [36, 38].[1] Participating in service creation has also been shown to increase trust and satisfaction, thereby increasing motivation to use the services [8]. In addition, participation can increase effectiveness. For certain problems, people themselves know the most about their unique needs and problems [36, 56]; participation can help policymakers and platform developers leverage this knowledge pool. Finally, stakeholder participation can help operationalize moral values and their associated trade-offs, such as fairness and efficiency [36]. Even people who

[1]By "legitimacy," we refer to Weber's notion that "persons or systems exercising authority are lent prestige" [81]. A policy or action is legitimate when constituents have good reason to support it [37]. In western democratic societies, the legitimacy of governing systems is often established through the public practice of democracy that seeks to earn the consent of the governed by soliciting their input, often through elections, to influence government and public policy.

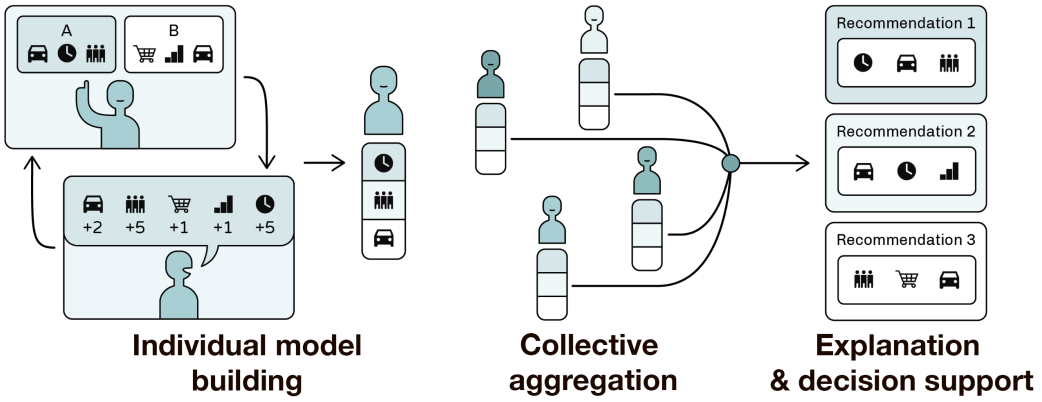

Fig. 1. The WeBuildAI framework allows people to participate in designing algorithmic governance policy. A key aspect of this framework is that individuals create computational models that embody their beliefs on the algorithmic policy in question and vote on the individual's behalf.

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 trade-offs that need to be explicitly codified rather than left up to human judgment.

Enabling stakeholder participation in algorithmic governance raises several fundamental research questions. First, what socio-technical methods will effectively elicit individual and collective beliefs about policies and translate them into computational algorithms? Second, how should the resulting algorithms be explained so that participants understand their roles and administrators can make decisions using the algorithms? How does participation influence participants' perceptions of and interactions with algorithmic governance? Finally, how does the resulting collectively-built algorithm perform?

In order to address these research questions, we propose a framework called WeBuildAI that enables people to collectively design an algorithmic policy for their own community (Figure 1).[2] By "design," we mean having the community members and stakeholder themselves define the optimization goals of the algorithms, the benefits and costs of the algorithmic governance decisions, and the value principles that they believe their community should embody and operate on. The key aspect of this framework is that individuals create computational models that embody their beliefs on the algorithmic policy in question,[3] and then these models vote on their individuals' behalf. This works like a group of people making a decision together: computational models of each individual's decision-making process make a collective choice for each policy decision. The individual models rank possible alternatives, and the individual rankings are then aggregated via the classic Borda rule. The resulting algorithmic recommendations are explained to support administrative decision-makers.

As a case study, we applied this framework to develop a matching algorithm that distributes donations through collaboration with 412 Food Rescue, a nonprofit that provides an on-demand donation transportation service with volunteer support. The algorithm matches donors with

[2]We define "community" according to the Merriam-Webster dictionary as a "unified body of individuals," particularly a group linked by a common interest or policy.
[3]By "belief," we mean a "positional attitude," in other words, "the mental state of having some attitude, stance, take, or opinion about a proposition" [70].

recipient organizations, determining who receives donations and how far volunteers need to drive to deliver donations. We solicited stakeholder participation to adjudicate the tradeoffs involved in the algorithm's design, balancing equity and efficiency in donation distribution and managing the associated disparate impacts on different stakeholders. Over the course of a year, we had the stakeholders—donors, recipient organizations, volunteers, and the 412 Food Rescue staff—use the WeBuildAI framework to design the matching algorithm, and researched their experiences through a series of studies. The findings suggest that our framework successfully enabled participants to build models that they felt confident represented their own beliefs. In line with our original goals, participatory algorithm design also impacted both procedural fairness and distributive outcomes: participants trusted and perceived as fair the collectively-built algorithm, and developed an empathetic stance toward the organization. Compared to human dispatchers, the resulting algorithm improved equity in donation distribution without hurting efficiency when tested with historic data. Finally, we discovered that the individual model-building process raised participants' algorithmic awareness and helped identify inconsistencies in human managers' decision-making in the organization, and that the design of the individual model-building method may influence the elicited beliefs.

Our paper makes three contributions. First, we offer a framework and methods that enable participatory algorithm design, contributing to emerging research on human-centered algorithms and participatory design for technology. Second, through a case study with stakeholders in a real-world nonprofit, we demonstrate the feasibility, potential, and challenges of community involvement in algorithm design. Finally, our work provides insights on the effects of procedurally-fair algorithms that can further understanding of algorithmic fairness.

2 GOVERNING ALGORITHM DESIGN AND PARTICIPATION

Our framework draws from social choice and participatory governance literature to enable participatory algorithm design. In this section, we first lay out normative choices in algorithm design. We then review and identify gaps in participatory design literature and emerging work to introduce stakeholder participation in algorithm design. Finally, we discuss how we leveraged participatory governance literature to inform our framework design.

2.1 Normative Choices in Algorithm Design

In line with Aneesh's definition of "algocracy," when "authority becomes embedded in the technology itself" [3] rather than traditional forms of governance, and Danaher's elaborations, we define "governing algorithms" as algorithms that "nudge, bias, guide, provoke, control, manipulate and constrain human behaviour" [27]. All algorithm design choices cannot be addressed by a purely technical approach [42, 83]; particularly in governing algorithms, some design choices require a normative decision, as they affect multiple stakeholders and need to codify critical social values and associated tradeoffs. We describe three such design choices below.

First, increasingly more research has investigated computational techniques to encode social and moral values in algorithms, yet many still rely on fundamental measures and algorithmic "objective functions" that humans must define. Defining these terms is complex. Fairness, for example, broadly defined as treating everyone equally, has multiple definitions and theoretical roots. In prior work, fairness has been defined as equitable distributive outcomes and just, unbiased, non-discriminatory decision-making processes [11]. Fairness is an important value in governing algorithms as algorithms can perpetuate unfair treatment of different populations or stakeholders [27, 35, 85]. Emerging work develops computationally fair algorithms [17, 34], yet applying these techniques to real-world settings still requires human judgment. For example, individual fairness, or treating similar individuals similarly, requires a definition for "similar individuals" [32].

Second, multiple social values and objectives cannot be satisfied to the same degree, which necessitates making tradeoff decisions. For example, all fairness principles cannot be guaranteed simultaneously [22, 46], so a human decision-maker must determine which fairness definitions an algorithm should use. Similarly, operational efficiency and fairness are often competing values in modern capitalist democracies [61]. Algorithms that aim to achieve both require human judgments about how to balance the two, because there is no fundamental "right" balance and one cannot be determined purely through optimization [9].

Finally, these definitions and values are context-dependent. Recent empirical work on perceptions of "fair" algorithms suggests that different social groups believe in different fairness principles, and even algorithms that embody a fairness principle may not be perceived as fair if the implemented principle is not in accordance with the affected group's beliefs [48]. For example, some groups in the study preferred random allocation that treated everyone equally, and did not consider individual differences to be relevant to task allocation. Other groups desired equity-based allocation, in which the tasks are allocated to satisfy everyone's preferences to a similar degree. Some other groups wanted to consider both preferences and task completion time as fairness factors, so that people work for a similar amount of time and their preferences are satisfied similarly. These findings suggest people believe in epistemically different fairness principles or desire varying ways of operationalizing fairness principles. Real-world examples also suggest that algorithmic software will fail to be adopted if it uses features or objective functions that do not fit the context of the affected community. For example, a "fair" algorithmic school start time scheduling software in Boston received pushback from the community and was ultimately not adopted, because the policymakers' and developers' efforts to decrease racial disparities did not consider important values and constraints of the stakeholders [82]. This body of work suggests that fairness principles must be context-specific, and that algorithmic systems should embody fairness notions derived from the community.

These normative choices in algorithm design are fundamental; how do we understand and formalize context-dependent values? Who should determine these important values and tradeoffs in governing algorithms, and how? Our approach to these questions is inspired by the long line of research on participatory design.

2.2 Gaps in Participatory Design and Human-Centered Research on AI

Participatory design originated in Scandinavia in the 1960s with the intention of involving workers in planning job design and work environments. Participatory design was subsequently adopted in the fields of human-computer interaction and engineering [59, 80], and researchers and designers have included "end-users" in design activities for computing systems in a wide range of domains such as workspaces [13], healthcare [6], and robots [30]. In participatory design, the researchers and users of a technology share power and control in determining its technological future [15, 59, 80], so that the stakeholders or populations that the technology will influence have a say in the resulting design, and the technology can better reflect their needs, values, and concerns. More recently, several scholars have argued that one needs to be more cognizant of the agency and influence of the researchers and designers in "configuring the process participation," and more critical analysis must be done in terms of who initiates participation and who benefits from it [80].

While participatory design has been applied to diverse forms of technology, the research on involving users in the process of designing algorithms or AI is still in its infancy. Rahwan [64] argues for "society-in-the-loop," which stresses the importance of creating infrastructure and tools to involve societal opinions in the creation of AI. Emerging work has also started to explore societal expectations of algorithmic systems such as self-driving cars [14, 60] and robots [54]. This line of work offers an understanding of the public's general moral values around AI through thought

experiments, but it is difficult to translate them into actual AI technology as they have often been done in hypothetical moral dilemma situations.

Emerging work seeks to understand participants' values with regard to the fairness of actual AI products, with the goal of representing these values in the final AI design. For example, Zhu et al. proposed Value Sensitive Algorithm Design [87], a five-step design process that starts with understanding the stakeholders and ends with evaluating algorithms' acceptance, accuracy, and impacts, in the context of Wikipedia bots. In this process, designers interpret stakeholder opinions and make the necessary trade-off decisions. Alvarado and Waern organized a participatory workshop for social media curation algorithms in which people were asked to imagine ideal "algorithmic experiences" [2]. Lee et al. and Woodruff et al. conducted interview and workshop studies on what people think "fair" algorithms are in the contexts of donation allocation [50] and online ads [84]. Other scholars systematically investigated perceived fairness of algorithmic decisions in hiring [47], recidivism [31], child welfare services [18], and resource allocation such as task assignment [48] and goods division [49].

To our knowledge, however, little work has sought to formalize subjective concepts of fairness. Furthermore, while these studies provide us with a better understanding of general public and user perceptions of justice and fairness, they do not close the loop on algorithmic developments that respond to these concerns. Our work proposes a method for directly involving end-users or stakeholders of algorithmic services in determining how the algorithms should make decisions. One aspect that differentiates our work is that we offer a tool through which people without algorithmic knowledge can directly specify or "sketch" [20] how they would like the algorithm to behave; we couple this with a method for aggregating different stakeholders' points of view.

2.3 Participatory Governance

Our framework draws on the literature on participatory governance. A first step in participatory governance is to determine what governance issues participants will consider and how participation will influence final policy outcomes. User groups, or mini-publics [36], can be configured as open forums where people express their opinions on policies; focus groups can be arranged for specific purposes such as providing advice or deriving design requirements. In full participatory democratic governance, citizen voices are directly incorporated into the determination of the policy agenda. Our framework focuses on this last form: direct participation in *designing* algorithmic governance. By "direct participation," we mean that people are able to specify "objective functions" and behaviors in order to create desirable algorithmic policies. This direct approach can minimize potential errors and biases that occur when codifying policy ideas into computational algorithms, which has been highlighted as a risk in algorithmic governance [45].

A key aspect of governance is collective decision-making. Our framework builds on social choice theory. 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 [71]. 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 is typically built on an axiomatic approach, formally defining desirable axiomatic qualities and studying voting rules that satisfy them. Indeed, the Borda voting rule satisfies a number of such properties, including monotonicity (pushing an alternative upwards in the votes should not hurt it) and consistency (if two electorates elect the same alternative, their union does too). We adopted a social choice approach specifically because our ultimate design outcome is an algorithm. While we know "quantification" has limitations in capturing nuances in the real world, quantification is an inevitable step in algorithms as they

need quantitative inputs. Social choice theory provides a framework for formally reasoning about collective decisions at scale.

Implementing participation in algorithmic governance requires addressing the following challenges. First, how can we enable individuals to form beliefs about policies through deliberation and express these beliefs in a format that the algorithm can implement? Second, how do we consolidate individuals' models? Finally, how do we explain the final decisions so that people can understand the influence of their participation on the resulting policy, and administrators can use the collectively-built governing algorithm? In the next section, we describe our framework and how it addresses these challenges.

3 THE WEBUILDAI FRAMEWORK

Here we lay out the basic building blocks of the WeBuildAI framework, which enables participation in building algorithmic governance through a novel combination of individual belief learning, voting, and explanation. Our framework design draws on the field of political theory, which investigates collective decision-making and effective citizen participation in governance.

The key idea of the framework is to build a computational model representing each individual stakeholder, and to have those models vote on their individuals' behalf. This works like a group of people making a decision together: computational models of each individual's decision-making process make a collective choice for each policy decision.

3.1 Individual Belief Model Building

Building a model that embodies an individual's beliefs on policy gives rise to three challenges. First, people need to determine what information, or features, should be used in algorithms. Second, the individual needs to form a stable policy that applies across a broad spectrum of situations. This process requires people to examine their judgments in different contexts until they reach an acceptable coherence among their beliefs, or reflective equilibrium [28, 66]. Third, people without expertise in algorithms need to be able to express their beliefs in terms of an algorithmic model. We address these challenges by deriving a set of features from people's inputs, and then using both bottom-up machine learning training and top-down explicit rule making.

3.1.1 Feature Selection. The first step is to determine features that people believe should be used by the algorithm to make decisions. People's opinions can be solicited through interviews or surveys. The derived set of features will be used to construct pairwise comparisons between alternatives, or allow people to directly specify weights for each of the features.

3.1.2 Model Building. We use both machine learning and explicit rule specification. By allowing people to use both types of models iteratively, we seek to support deliberation. By building a machine learning model via pairwise comparisons, people can develop a policy that works across various contexts; by explicitly specifying a policy that they have been implicitly forming, participants can consolidate and externalize their beliefs; then by answering new pairwise comparisons questions, they can evaluate whether the rules they have in mind work consistently across contexts.

- *Machine Learning Model.* To train an algorithm that reflects people's decision criteria, the machine learning method uses pairwise comparisons between a pair of alternatives that vary along the features derived from the previous step. Pairwise comparisons have been used to encourage moral deliberation and reach a reflective equilibrium in determining fairness principles [66], and have been used as a way to understand people's judgments in social and moral dilemmas in psychology and economics [25]. This method allows people to become familiar with different contexts, and develop and refine their beliefs.

We utilize random utility models, which are commonly used in social choice settings to capture choices between discrete objects [55]. 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. For each participant i, we learn a single vector β_i such that the mode utility of each potential decision 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.

- *Explicit Rule Model.* In this method, participants directly specify their principles and decision criteria as used in expert system design [29]. Human-interpretable algorithmic models [86] 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. Specifically, for each of the features, participants can specify scores to express how much the algorithm should weight different features.

3.1.3 Model Selection. Once people build their models using the two methods, we visualize the models and show example decisions that each model has made so that people can understand each model and select the one that best reflects their beliefs.

3.2 Collective Aggregation

Once participants have built their models, the next challenge is to construct a collective rule that consolidates the individual models. We address this challenge by leveraging social choice, one of the main theories of collective decision-making, which aggregates peoples' opinions according to certain desirable qualities [71]. Voting is one of the most common aggregation methods. In voting, individuals can specify a top choice or rank alternatives, and the alternatives with the most support are selected. In our framework, we use the Borda voting method due to its relative simplicity and robust theoretical guarantees in the face of noisy estimates of true preferences, as shown in a paper by some of the authors [44].

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 $b \succ a \succ c$.

Once stakeholders create their models, the models are embedded in the AI system to represent the stakeholders; for each algorithmic decision task, each individual model ranks all alternatives, and the ranked lists of all participants are aggregated using the Borda rule to generate the final ranked list.

3.3 Algorithm Explanation and Human Decision Support

Finally, the ranked recommendations must be explained to stakeholders to communicate how their participation has influenced the final policy and supported operational decision-making. 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 [36]. Algorithmic governance offers new opportunities in this regard

because the aggregation of individual models and resulting policy operations are documented. A new challenge is how to explain collectively-built algorithmic decisions, an area in which little prior research has been done. We address this challenge by displaying each recommended option's Borda score, its average ranking per stakeholder group, and its "standout" features in order to support the administrators enacting the algorithmic policies.

4 CASE STUDY: MATCHING ALGORITHM FOR DONATION ALLOCATION

We applied the WeBuildAI framework in the context of on-demand donation matching in collaboration with 412 Food Rescue [1].

4.1 Goals of Participation in Matching Algorithm Design

4.1.1 Organizational Context. 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, and then 412 Food Rescue 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 service's success depends on the participation of all stakeholders–a continuous stream of donations, recipient organizations' willingness to accept the donations, volunteers' efforts to transport donations, and 412 Food Rescue's operational support and monitoring. The organization has grown successfully for the past few years. They have rescued over three million pounds of food and are expanding their model into food rescue organizations in four other cities, including San Francisco and Philadelphia. The donation allocation policy is at the core of their service operation; while each individual 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 of the rescues.

412 Food Rescue wanted to introduce an algorithmic donation allocation system for two reasons. First, they currently have a few employees per day, known as dispatchers, manually allocating all donations that come in that day. On a busy day, each dispatcher has to manage over 100 donations, which is too many, so the organization wants to reduce dispatcher workload. Second, 412 Food Rescue wishes to improve equity in their donation distribution. The current donation distribution is quite skewed, with 20% of recipient organizations receiving 70% of donations (Figure 5a), because allocation decisions are often made for convenience.

4.1.2 Equity-Efficiency Tradeoff and Stakeholder Motivation. In designing this matching algorithm, we used participation to determine the tradeoff between equity and efficiency. In this context, we define "equity" as giving donations to recipients with greater need and "efficiency" in terms of the distance each donation travels from donor to recipient. Balancing equity and efficiency is challenging as this design choice has different impacts on different stakeholders. For example, if the matching algorithm prioritizes efficiency and gives donations to recipients closest to donors, volunteers will benefit from shorter driving times, but the donation distribution may be skewed and recipients in wealthier areas may receive more donations, as donors are often located in wealthier areas. On the other hand, if the matching algorithm prioritizes equity, recipients with greater need may receive more donations, but this may increase the distance that volunteers need to drive, as well as the effort 412 Food Rescue must spend in recruiting the volunteers. Finding a collective solution to this problem is critical to the success of the service, because all stakeholders will be more motivated to continue participating in the service if they feel their needs are respected.

4.2 Stakeholder Participants

4.2.1 Volunteer-Based Participation. We used our framework to build the matching algorithm collectively with 412 Food Rescue's stakeholders. One of the important considerations in participatory governance is determining who participates. A widely-used and accepted method is volunteer-based participation [36], which accepts input from people who will be governed by the system and who choose to participate. Many democratic decisions, including elections, participatory forums, and civic engagement, are volunteer-based. In our application, we used a volunteer-based method with stakeholders directly influenced by the governing algorithm. As our first evaluation of the framework, we chose to work with a small focus group of stakeholders who volunteered to participate in order to get in-depth feedback.

4.2.2 Participation Recruiting and Information. Our research took place over a period of one year. We solicited stakeholder participation to determine how the matching algorithm should weight the factors used to recommend recipient organizations. The stakeholders included donor organizations, recipient organizations, volunteers, and 412 Food Rescue staff. We included the governing entity as a stakeholder because they have a holistic viewpoint on logistics: how the donation is collected, handled and delivered to the recipient organization. The mission of the organization is to reduce food waste and serve food-insecure populations, which overlaps with other stakeholders' goals.

The entire staff that oversees donation matching at the organization participated in the study. Recipients, volunteers, and donors were recruited through an email that 412 Food Rescue staff sent out to their contact list.[4] We replied to inquiry emails in the order in which they arrived, and collected information about respondents' experience with 412 Food Rescue and organizational characteristics in order to ensure diversity. We limited the number of participants from each stakeholder group to 5–8 people, which resulted in an initial group of 23 participants (including V4a and V4b, who participated together) with varying organizational involvement (Table 1). Fifteen were female (nine males) and everyone, except one Asian, was white.[5] Sixteen participants answered our optional 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 household income was $65,700 (Median=$62,500 (SD=$39,560); Min-Max:$25,000-$175,000).

4.3 Research Process Overview

Our research goal was threefold: we sought to apply the framework to build a matching algorithm, evaluate the usability and efficacy of the framework, and understand the effects of participation. To this end, we used our framework to allow participants to build their own individual models. We conducted think-alouds throughout the data collection procedure to understand participants' thinking processes. We also showed participants the method and results from each step of our framework—for example, how we aggregate individual models and explain the decisions—and

[4]We did not include recipient organizations' clients for several reasons. First, we asked about service operation in this study. Our previous interviews with clients [50] suggest that recipient organizations do not display where their food comes from at the time of distribution. Thus clients generally have no experience with or knowledge of the food rescue process and lack the hands-on experience required to consider disparate impacts on different stakeholders. Because of this, we represented clients' interests via feedback from the staff of recipient organizations who know and serve client populations. Additionally, 412 Food Rescue did not have recipient client contact information for privacy reasons. In the discussion section, we explain how we will seek out a way to expand participation to include groups, including clients, that are not directly involved in the food rescue process.

[5]Our participants were mostly white, which reflects the population of volunteers and non-profit staff in Pittsburgh. This is the result of a volunteer-based method [36]. In our next step, we will implement targeted recruiting of minority populations.

Role	Studies Involved
412 Food Rescue.*	
F1	Sessions 1-4
F2	Sessions 1-4
F3	Sessions 1-4, w
Recipient organizations. (Clients served monthly, client neighborhood poverty rate)	
R1 Human services program manager (N=150, 13%)	Sessions 1-4
R2 Shelter & food pantry center director (N=50, 20%)	Sessions 1-4
R3 Food pantry employee (N=200, 53%)	Sessions 1-4
R4 Animal shelter staff	Session 1
R5 Food pantry staff (N=500, 5%)	Sessions 1-4
R6 After-school program employee (N=20, 33%)	Session 1, w
R7 Home-delivered meals delivery manager (N=50, 11%)	Sessions 1-4
R8 Food pantry director (N=200, 14%)	Sessions 1-2
Volunteers.	
V1 White male, 60s	Sessions 1-4, w
V2 White female, 30s	Session 1
V3 White female, 70s	Sessions 1-4, w
V4 White female, 70s (V4a), white male, 70s (V4b) †	Sessions 1-4
V5 White female, 60s	Sessions 1-4
V6 White female, 20s	Sessions 1-4
Donor organizations.	
D1 School A dining service manager	Session 1
D2 School B dining service manager	Sessions 1-4
D3 Produce company marketing coordinator	Session 1
D4 Grocery store manager	Sessions 1-4
D5 Manager at dining and catering service contractor	Session 1
D6 School C dining service employee	Session 1, w

Table 1. Participants. Sessions indicate the study sessions that they participated in: *w* represents a workshop study. *Info excluded for anonymity. † A couple participated together.

conducted interviews to study their understanding and responses to the method. Once participants completed all stages of the framework, we conducted interviews to understand participants' attitudes toward the resulting algorithm and the governing organization, 412 Food Rescue.

Overall, our research resulted in 4–5 individual sessions for each participant and a workshop over the course of a year. Because of the extended nature of the community engagement, 15 participants completed all the individual study sessions, while 8 could participate only in the first couple of sessions due to changes in their schedules or jobs (Table 1). 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.

4.4 Researcher Stance

Our research team included people with diverse backgrounds in human-computer interaction, artificial intelligence, theoretical computer science, information systems, decision science, ethics and design, affiliated with Carnegie Mellon University and University of Texas at Austin. We had a constructive design stance and sought to bring about positive change through the creation of

artifacts or systems. Two researchers have conducted research with 412 Food Rescue in the past and one researcher regularly volunteered in homeless shelters and food pantries in Pittsburgh. This relationship and familiarity with public assistance work helped us gain access to the research site.

4.5 Analysis

We report how we analyzed qualitative data from all sessions in this section to avoid repetition. All interviews were audio-recorded and transcribed, and researchers took notes throughout the think-alouds and workshop. The data was analyzed following a qualitative data analysis method [24, 62]. Two researchers read all of the notes and interview transcripts and conducted open coding of the transcripts at the sentence or paragraph level on Dedoose.[6] The rest of the research team met every week to discuss emerging themes and organize them into higher levels. As we progressed in our analysis, we drew from the literature on participatory governance [36] and procedural fairness [49, 52] to see whether the themes that we observed were consistent with or different from previous work. After all sessions were completed, we revisited the themes from each session and further consolidated them into the final themes we present in this paper. In Section 8, we report the number of participants associated with different themes in order to note the relative frequency of different opinions and behaviors in our study. However, as a qualitative study with a small sample size, we note that this should not be taken as an exact weight of whether one opinion is more significant or representative.

5 INDIVIDUAL BELIEF MODEL BUILDING

The first step in building individual belief models is to determine which factors (or features) are relevant and important; we derived these factors from the authors' previous study [50] that examined the 412 Food Rescue stakeholders' concepts of fair donation allocation. A factor that was mentioned most frequently is the distance between donors and recipient organizations. Participants mentioned various other factors that represent the needs of recipient organizations, such as the income level of recipient clients, the food access levels of their neighborhoods, and the size of the recipient organization. Additional factors that were also deemed important were the distributional capabilities of recipient organizations, i.e., how fast they can distribute to their clients, and the temporal regularity in incoming donations. From the factors that participants mentioned, we selected the ones that came up most frequently and had reliable data sources.[7] The selected factors capture transportation efficiency, recipient needs, and temporal allocation patterns (Table 2). For example, poverty rate is an indicator of recipients' needs; distance between recipients and donors is a metric of efficiency; and when each recipient last received a donation is a measure of allocation patterns over time.

We conducted three sessions to develop a model to represent each individual in the final algorithm. Participants first completed pairwise comparisons (Figure 2a, Session 1) to train algorithms using machine learning. Participants who wanted to elaborate on their models participated in the explicit rule specification session (Figure 2b, Session 2). If their belief changed after Session 2, they provided a new set of pairwise comparisons to retrain the algorithm. Participants were later asked to choose one of the two models that represented their beliefs more accurately (Figure 3, Session 3).

[6] https://www.dedoose.com

[7] We did not use organization types (e.g., shelters and food pantries) or addresses because these aspects may communicate the racial, gender, or age characteristics of recipients and elicit biased answers based on inaccurate assumptions or discrimination.

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.

Table 2. Factors of matching algorithm decisions. The ranges of the factors are based on their real-world distributions.

5.1 Machine Learning Model (Session 1)

5.1.1 Pairwise Comparison Scenarios. We developed a web application to generate two potential recipients at random according to the factors (Table 2), and asked people to choose which recipient should receive the donation (Figure 2a).[8] All 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. During the research process, the link to the web application was sent to the participants who wished to update their models on their own. In fact, 13 participants chose to answer an additional 50–100 questions after Session 2 to retrain their machine learning models.

5.1.2 Learning Individual Models. In order to learn individual models, we utilize random utility models, which are commonly used in social choice settings to capture choices between discrete objects [55]. This fits our setting, in which participants evaluate pairwise comparisons between potential recipients. In order to apply random utility models to our setting, we use the Thurstone-Mosteller (TM) model [58, 74], 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. [60], we assume that each participant's mode utility for every potential match is a linear function of the match's feature vector. Therefore, for each participant i, we learn a single vector β_i such that the mode utility of each potential match x is $\mu_i(x) = \beta_i^T x$. We then learn the relevant β_i vectors via standard gradient

[8]Improbable combinations of income and poverty (e.g., very high income coupled with very high poverty) were excluded according to the census data. All factors were explained in a separate page that participants could refer to.

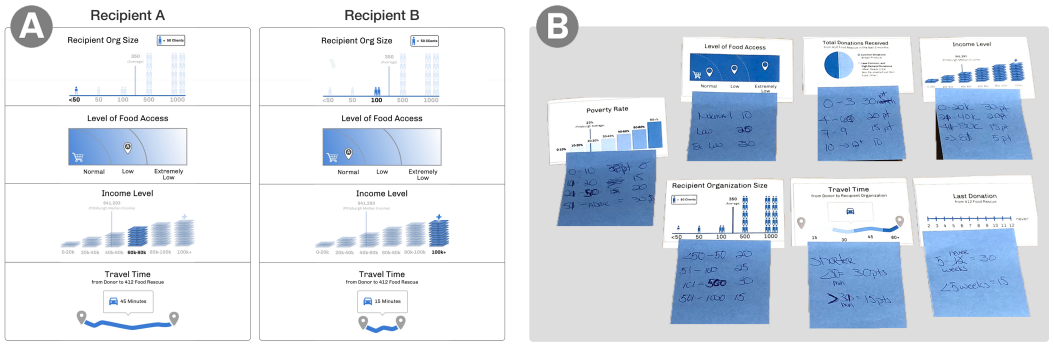

Fig. 2. Two methods of individual model building were used in our study: (a) a machine learning model that participants trained through pairwise comparisons, and (b) an explicit rule model that participants specified by assigning scores to each factor involved in algorithmic decision-making.

descent techniques using Normal loss.[9] 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 Appendix A).

5.2 Explicit Rule Model (Session 2)

To allow participants to explicitly specify matching 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 "balancing" of factors that people identified when answering the pairwise questions.[10] We asked participants to create rules to score potential recipients so that recipients with the highest scores would be recommended. Participants assigned values to different features using printed-out factors and notes (Figure 2b). We did not restrict the range of scores but used 0-30 in the examples in our instruction. Once participants created their models, they tested how their scoring rule worked with 3-5 pairwise comparisons generated from our web application, and adjusted their models in response. At the end of the session, we conducted a semi-structured interview in which we asked participants to explain the reasoning behind their scoring rules, and describe their overall experience. The sessions took about one hour. Two participants wanted to further adjust their models and scheduled 30 minute follow-up sessions to communicate their changes.

5.3 Machine Learning versus Explicit-Rule Models (Session 3)

We asked participants to compare and choose between their machine learning and explicit-rule models, selecting one that best 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[11] 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 with the input range for each feature (Figure 3). In order to prevent any potential bias in favor of a particular method, we anonymized the models ("Model X" or "Model Y"), normalized the two models' parameters (beta values) and scoring rubric using the maximum assigned score in

[9]For participants who consider donation type, we learn two machine learning models, one for common donations and one for uncommon donations.
[10]We also experimented with manually-created decision trees, but the models quickly became prohibitively convoluted.
[11]We used the same set of comparisons for all participants for consistency.

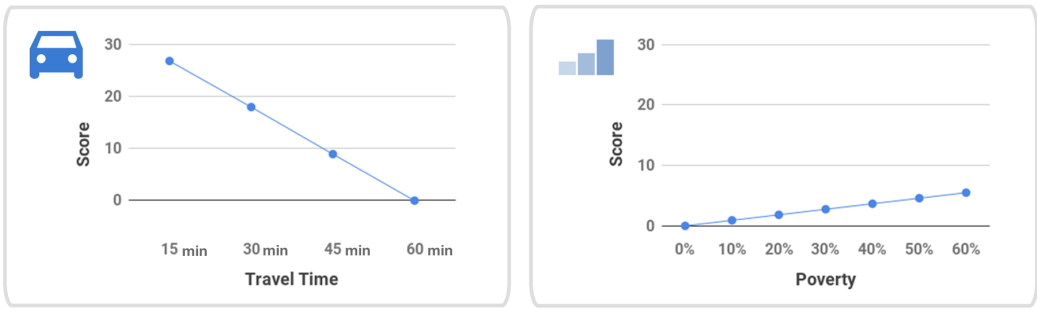

Fig. 3. Model explanations. Both machine learning and explicit-rule models were represented by graphs that assigned scores according to the varying levels of input features.

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 between the two models, 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 about their experience and reasons for their final model choice. We also analyzed individual models 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.

5.4 Final Individual Models

In total, we trained 23 machine learning models[12] and obtained 15 explicit-rule models. Of the 15 participants who completed all studies and were asked to choose models that better represented their belief, 10 of them chose the machine learning models trained on their pairwise comparisons; the other five chose the models that they explicitly specified.

The machine learning models had higher overall agreement with participant's survey answers than the explicit rule models when tested on 50 new pairwise comparisons provided by each participant, as seen in Table 3. However, as our sample size is small, we do not aim to make general claims on which model has better accuracy. In addition, for many, the machine learning model was the one they had built last and therefore reflected their current thinking at the time of comparison; we further elaborate on this in Section 8.1. We also note that we did not observe any differences in participants' perceived accountability in the creation of these models. Both models took an equal amount of participants' time and attention, and participants told us that they felt responsible when making choices and assigning scores.

6 COLLECTIVE AGGREGATION

Our framework uses a voting method to aggregate individuals' beliefs. When presented with a new donation decision, each individual's model generates a complete ranking of all possible

[12]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.

	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**

Table 3. Accuracy of the Machine Learning (ML) model and the Explicit-Rule (ER) model. Bold denotes the model the participant chose as the one that better represented their belief after seeing both models' explanations (Figure 3) and their predictions on the 50 evaluation pairwise comparisons. F1 chose the machine learning model but did not complete additional survey questions to calculate model agreement, so the result is not included in this table.

recipient organizations. The Borda rule aggregates these rankings to derive a consensus ranking and suggest recommendations. We conducted a workshop and interviews to understand participants' perceptions of this method.

6.1 Method (Workshop)

In an early stage of our research, we conducted a workshop in order to gauge participants' perceptions of the Borda aggregation method and determine the method's appropriateness from a social perspective. Five participants (Table 1) who had built their individual models at that time 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 rule worked. The description of the Borda rule given to participants was: "Individuals rank options according to belief. Each option receives a number of points determined by its ranking, with higher-ranked options receiving more points. The points are added up, and the winner is whichever option has the greatest number of points." The words "democratic" or "equal" were not used to avoid potential biases. 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. For participants who joined our research after this workshop, we asked the same questions about the Borda rule and stakeholder opinion weight in the interview in Session 4.

6.2 Varying Stakeholders' Voting Influence

All participants but one believed that the weight given to different stakeholders in the final algorithm should depend 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.[13] Nearly all participants weighted 412 Food Rescue staff as the highest group (n=13 out of 15), 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 out of 15) including the donors themselves, as they are not involved in the process once the food leaves their doors. Recipients and volunteers were weighted similarly because 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, 412 Food Rescue

[13]This is based on the input from participants that participated in the workshop and/or Session 4.

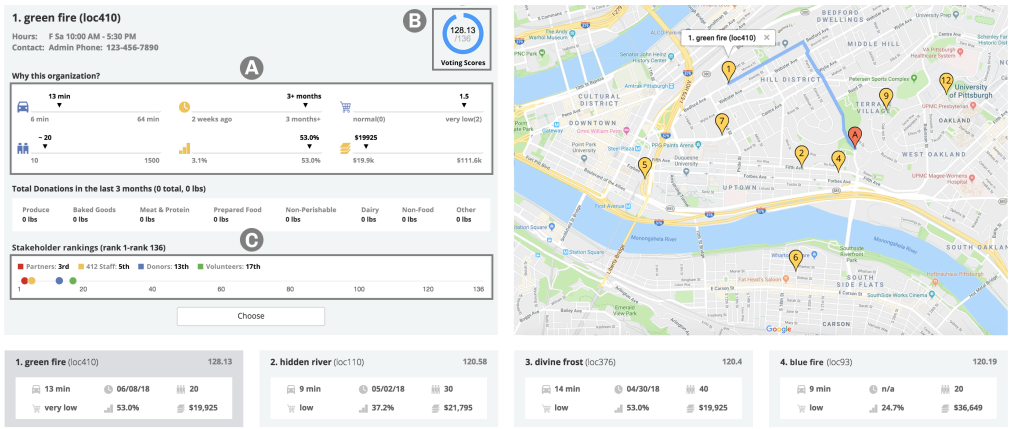

Fig. 4. The decision support tool explains algorithmic recommendations, including the nature of stakeholder participation, stakeholder voting results, and characteristics of each recommendation. The interface highlights the features of the recommended option that led to its selection (marked by A), the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

employees are assigned 46% of the weight; this translates to allocating them 46 votes out of 100 total as a group, where each employee's vote is "replicated" 46/3 times because three 412 Food Rescue employees participated in our study.

7 EXPLANATION AND DECISION SUPPORT

Once recommendations are generated, the decision support interface presents the top twelve organizations, accompanied by explanations, to support the human decision-maker who matches incoming donations to recipients. We used the explanations to demonstrate to participants how their opinions had been incorporated into the algorithm's decision-making. We also explained average stakeholder models to participants so that they could learn about others' models.

7.1 Design of Decision-Support Tool

The interface of our decision support tool is shown in Figure 4. The tool was designed with other considerations, such as choice architecture [73], but they are beyond the scope of this paper. We focus instead on the explanation of decisions made by collectively-built algorithms.

- Decision Outcome Explanation (marked by A in Figure 4): We used an "input influence" style explanation [12]. Features are highlighted in yellow when an organization is in the top 10% of recipient organizations 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 (marked by B in Figure 4): The Borda score for each organization is displayed. It shows this option's score in relation to the maximum possible score that an option could receive (i.e., scores when every individual model picks this option as its first choice). This voting score can indicate the degree of consensus among participants.
- Stakeholder Rankings (marked by C in Figure 4): Stakeholder rankings show how each stakeholder group ranked the given organization on average. It is a visual reminder that

all 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 relations management system currently in use at 412 Food Rescue. Algorithms were coded in Ruby on Rails, the front-end interface used Javascript and Bootstrap, and the database was built with Postgres. The distances and travel times between donors and recipients were pre-computed using the Google Maps API and Python. We used donor and recipient information from the past five months of donation records in the database. On average, the algorithm produced recommendations for each donation in five seconds.

7.2 Method (Session 4)

We conducted a one-hour study with each participant to understand how the decision support and explanation influenced their perceptions of the matching algorithms and their attitude toward 412 Food Rescue. 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.

We first showed participants the graphs of their individual models 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 412 Food Rescue. We also asked participants to reflect on the overall process of giving feedback throughout our studies.

8 FINDINGS: THE IMPACT OF PARTICIPATORY ALGORITHM DESIGN

In the previous sections, we described how stakeholders used the WeBuildAI framework to build the matching algorithm for 412 Food Rescue over multiple sessions and a workshop. We now report the qualitative findings from observations, think-alouds, and interviews to describe the impacts of the WeBuildAI framework and participation.

8.1 Participants' Experience with the WeBuildAI Framework

Overall, the individual belief-elicitation step of the framework—using both machine learning and explicit rule specification methods and visualizing the learned models—successfully enabled participants to build an individual model that represented their beliefs on how the algorithm should make a matching decision. Participants perceived the automatic aggregation method based on the Borda rule as a nuanced, democratic approach; the decision support tool and explanation allowed them to understand how algorithmic recommendations were made.

8.1.1 Effects of Individual Model Building Methods on Elicited Beliefs. Participants told us that performing pairwise comparisons and subsequently specifying explicit rules helped them develop and consolidate their beliefs into a set of principles that they could apply consistently in different decision contexts. Answering pairwise comparison questions helped familiarize participants with the problem setting; however, some participants commented that they felt like they were applying internal rules inconsistently, particularly in their first few questions. Explicitly specifying scores for each feature helped them reconcile their conflicting beliefs. For example, V1 told us that

she originally used organization size inconsistently, sometime favoring smaller organizations or bigger organizations, but when creating a rule, she determined that organization size should not matter. When she answered the new set of pairwise comparison questions to retrain the machine learning model, she further evaluated whether she could consistently apply her belief, i.e., that the organization size does not matter, to different contexts and whether she encountered any new situations in which she would need to further refine her rule.

In choosing the model to use in the final matching algorithm, the most important factor for all participants, except one, was how closely each model represented their beliefs. In Session 3, 10 out of 15 participants chose their machine learning models. For many, this was the model they had built last and therefore that reflected their 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, which were often step-wise functions. Explicit-rule models were chosen by five participants. For four of these participants, their explicit-rule model did a better job of weighing all of the factors that mattered to them and screening off unimportant factors. In other words, machine learning models learned rules that they disagreed with—for example, a machine learning model may give linearly increasing weight to larger organization sizes.

On the contrary, for one participant, the procedural difference in the two methods was why he chose the explicit-rule model. R2 trusted the reflective process of specifying a model and did not trust his pairwise answers nor the machine learning model built from them, even though the accuracy of the machine learning model was 90%, compared to 76% for the model that he created. He believed that determining policy should be based on defining principles, rather than case-by-case decisions; for this reason, he wanted to build a rule and follow the outcomes from the rule.

An unexpected finding was that the methods' procedural differences seemed to influence which aspects participants focused on at the time of decision-making and, in some cases, the rules that participants made. Creating a scoring model from a top-down approach seemed to evoke a higher level of construal [75], eliciting an abstract level of thinking that was absent when answering pairwise comparisons. Many participants stated answering pairwise comparisons felt emotional because it made them think of real-world organizations. For example, V1 said that developing explicit scoring rules felt "robotic"; R3 said that he 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 because she related the questions to real-world decision-making. This contrasted with her explicit-rule model, which favored equity-related factors like income and poverty. When comparing the models in Session 3, she told us that she focused on idealistic matching that prioritized equity when she was specifying scoring rules. In the end, she chose her machine learning model, stating that while her explicit-rule model was appealing as a way of pushing herself beyond her operational thinking, she deemed travel time and last donation date most important in practice.

8.1.2 Responses to the Borda-Based Aggregation. Participants appreciated that the Borda method gave every recipient organization a score (n=5) and that it embodied democratic values (n=4).[14] In the workshop, F1 felt that giving every organization a score captured the subtleties of her thinking better than other methods, such as considering only the top-ranked organization: *"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*

[14]We note that the description of Borda given to participants described a scoring process and did not include words such as "voting" and "democracy" as reported in Section 6.1.

democratic," relating it to a form of human governance. Two other individuals, D2 and D4, also related the method to voting systems in the US. D4 recognized that some US cities in California recently used a similar voting method for their mayoral election. It is worthwhile to note that, when we asked about potential alternatives, participants expressed difficulty thinking of them (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."*

8.1.3 Responses to the Decision Support Interface. Participants were almost universally appreciative of the fact that the system keeps a human dispatcher in the loop to make the final decision rather than automating the decision entirely. While some participants (F1 and R5) acknowledged that full automation could be more efficient than a human-in-the-loop process, most participants expressed that having a human dispatcher overseeing the process was important as they might have knowledge of additional decision factors outside the scope of the algorithm. F3 expressed that the combination of human and computer decision-making elements was "magical" in that it combined the objective data of an algorithm with human elements *"that the computer will never know... like so and so at this place loves peaches and they make peach pies."* Others (e.g., R2) expressed that the algorithm could enable human decision-making in a way that reduces bias or favoritism on the part of the dispatcher, thereby making the decisions of the organization more fair and objective.

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 rankings showed opinions that might differ from those of 412 Food Rescue dispatchers (n=4). V6, 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 in order 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 might use it as a tiebreaker between two similar organizations.

8.2 Participation and Perceptions of Algorithmic Governance

In a manner consistent with theories on procedural justice [49, 52] and participatory policy-making [36], participants believed that having control over the algorithm through participatory algorithm design made the resulting algorithm fair, and this process improved their attitudes toward the organization as a whole.

8.2.1 Procedural Fairness in Participatory Algorithm Design. All participants mentioned that the fact that the organization was putting a priority on fairness, being open to new ideas, and including multiple stakeholder groups improved their perceived fairness and trust of both the matching algorithm and the organization itself. For example, one participant said, *"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."* Some expressed that participation expanded the algorithm's assumptions beyond those of the organization and developers (n=6). V6 noted that it is easy for organizations to remain isolated in their own viewpoints and that building an algorithm based on 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 participation was *"certainly more fair than somebody sitting at a desk trying to figure it out on their own."* 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."*

8.2.2 Empathetic Stance toward the Governing Organization. Participation in algorithm design led many participants to increase the degree to which they viewed 412 Food Rescue positively and develop a more empathetic stance toward the organization (n=8). For some participants, this happened because participation exposed the difficulty of making donation matching decisions and made them realize that there might not be a perfect solution, which in turn made them thankful for the work of the organization (n=4). For example, after experiencing the burden of making the matching decision and seeing how similar the recommended recipients can be in the interface, D2 and V3 both expressed gratitude for 412 Food Rescue. Participants also expressed appreciation for the organization's concern for fairness and the effort needed to continually make such decisions. This shift in perception is particularly important because it can improve people's tolerance for and understanding of tradeoffs in governance decisions.

The participatory algorithm design 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 and V6) or donation acceptance (R2).

8.2.3 Reactions to Other Stakeholders' Models. While sharing other stakeholders' models is not a requirement of our framework, in this work, we showed the models to participants in order to get feedback on the fully transparent implementation of our framework.[15] We report how participants responded to similarities and differences in stakeholder models.

In individual models, all participants considered efficiency and equity factors. 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. Reviewing the models, participants expressed feeling assured that they shared these guiding principles with other participants (n=8). For example, all prioritized higher as opposed to lower poverty, and lower as opposed to higher food access. 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."*

A 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 relatively equal importance. Participants also had divided views on organization size, arguing for larger or smaller organizations, and did not prioritize this factor compared to others. In responses, participants acknowledged these differences and sought to make sense of others' assumptions. For example, R1, referencing how important travel time was to her, mentioned that hers is more of a *"business model"* whereas others were more altruistic, more heavily weighting factors like income and food access. Some participants were even pleased to see differences in the models (n=3). R3 was pleased that other participants were considering unique viewpoints. Likewise V4 and R1 both stated that it was natural to expect differences between stakeholders, as everyone has unique experiences, and that *"this is the point of democracy"* (V4).

However, one participant, V6, was concerned that 412 Food Rescue staff did not weight heavily her most important factors such as food access, income, and poverty. While she said that the

[15]Participants also told us that they were curious about other stakeholders' beliefs.

algorithm was "fair" as it was collectively created, her trust in the organization was lowered as a result, because she inferred that they believe in different principles. She also raised a concern about other participants' input qualities. It took her significant effort to develop a model that accurately represented her views, and she could not judge whether other participants were *"thoughtful enough to really put the effort into their models and capture their own emotions with it."* She concluded that she still trusted the algorithm, but appreciated having human oversight of the final decision.

8.3 Participation and Awareness of Algorithms and Organizational Decision-Making

Our findings suggest that participating in algorithm design improved algorithmic awareness at an individual level, as well as awareness of inconsistencies in decision-making practices at an organizational level.

8.3.1 Increased Algorithmic Awareness. At an individual level, participating in algorithm design changed participants' attitudes toward algorithms.[16] They felt they better understood what an algorithm was and had more appreciation for the kinds of decisions that algorithms could make. 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 Session 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 construction of an 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 in other algorithmic systems would be based on *"how it was developed and who's behind it and programmed [it] and how it's influenced."* D2 felt that the final algorithm for 412 Food Rescue was fair because he came to know and trust the researchers over the course of his participation.

8.3.2 Improved Awareness of Inconsistency in Organizational Decision-Making. The process of eliciting individual models allowed participants from the governing organization to be more aware of internal inconsistencies in decision-making within their organization, and provided an opportunity for them to revisit their own assumptions about other stakeholders. Guided only by the broad goals of the organization's mission, the employees previously made matching decisions according to their own criteria and interpretations of that mission. By externalizing their decision-processes into computational belief models, the employees were able to formalize their own decision-making processes, and see how their models meshed with or differed from other employees' processes, which brought hidden assumptions to the surface. For example, after seeing other employees' models, they discovered that some employees prioritized mid-sized organizations whereas others prioritized larger organizations, and employees differed in the ways they weighted poverty, income, and food access.

Moreover, seeing other stakeholders' models allowed employees to compare their assumptions about other stakeholders with the models actually made by the stakeholders. One common assumption held by the staff was that volunteers would prioritize travel time, but our volunteer stakeholders had diverse models, varying from one that predominantly weighted travel time to one that gave equal weights to travel time and recipient organizations' needs. When F2 saw that

[16]None of our participants had a background in programming.

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."* This reflection opens up the possibility that the organization could seek to appeal to diverse volunteer motivations and tailor recruiting methods accordingly.

9 EVALUATION OF ALGORITHMIC OUTCOMES

Our qualitative findings in the previous section show the procedural effect of participatory algorithm design, but what outcomes do collectively-built algorithms produce? In this section, we evaluate the algorithm's performance on various metrics.

9.1 Evaluation Goal

In the literature on policy-related algorithmic systems, the status quo—current human decision-making practice—is deemed to be an appropriate baseline for comparison to measure the algorithmic tool's efficacy; thus, we compare our algorithm with current human decision-making at 412 Food Rescue. One major reason that the organization wanted to introduce the algorithmic allocation system was to improve equity in donation allocation made by organizational staff and distribute the donations to a larger set of recipients. Indeed, the skewness of their current distribution of donations (i.e., 20% of the organizations receiving 70% of the donations (Figure 5a)) is not the result of conscious strategy, but rather the result of, for example, the memory bias of human decision-makers selecting recipients that they have given donations to recently.

9.2 Dataset

The final matching algorithm included 23 individual models (Section 5) that generated complete rankings of possible recipients for each incoming donation; the rankings were then aggregated using the Borda method with the stakeholder weights provided in Section 6. We ran this collectively-built algorithm on historical allocation data from 412 Food Rescue containing a total of 1,760 donations from 169 donors over the course of five months (March–August 2018).[17] There were 380 eligible recipient organizations in the database, and 277 of those received donations in the timeframe we considered.[18] We compared our algorithm (AA) with two benchmarks: human allocations recorded in historical data (HA), and a random algorithm that selected a recipient uniformly at random (RA). In the simulations for our algorithm and the random algorithm, we applied some of the real-world constraints that influenced human dispatchers' decisions: for any given donation, we filtered out recipients that did not handle the donation type or were not open for at least 2 hours between the incoming donation time and 6 pm.

9.3 Results

The results indicate that our algorithm can make donation allocations more equitable compared to human allocation without hurting efficiency (Figure 5).

9.3.1 Number of Donations Allocated to Recipient Organizations. Our algorithm resulted in a more equal donation distribution compared to human allocation, as illustrated in Figure 5b. As the human donation distribution is skewed, we conducted a Mann-Whitney U test, a nonparametric test that

[17] The original data set had 1,862 donations from 177 donors given to 305 recipient organizations. 412 Food Rescue staff told us that 28 of the recipient organizations were either backup recipient organizations or became inactive at the time of the evaluation, thus we excluded them from the data.

[18] 46 recipients were added during the course of the five months, and for each day, we filtered out organizations based on the date when the recipient organizations were added in algorithm testing.

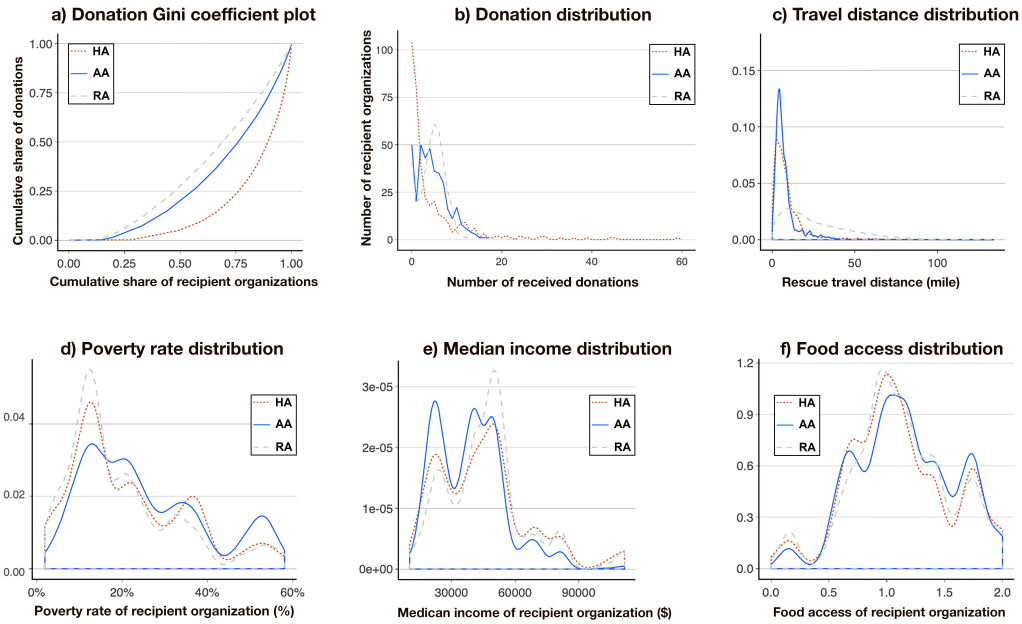

Fig. 5. The performance of our algorithm (AA) versus the human allocation (HA) and a uniformly random allocation (RA), on various metrics.

does not require the data to be normally distributed, to compare the number of donations allocated to recipient organizations.[19] The results show that algorithmic allocation was significantly more equally distributed than human allocation (AA Median = 4 donations (SD = 3.73), Min-Max:0-20; HA Median = 2 donations (SD = 7.26) Min-Max:0-59, U = 57814, p <.00000001).

We also conducted a Gini coefficient analysis, a standard economic inequality measure of income [40] or other kinds of resources [5]. A Gini index of zero means perfect equality, with everyone getting the same number of donations, and an index of 100 means maximum inequality, with one organization receiving all donations. Algorithmic allocation resulted in a Gini index of 42, which was lower than the Gini index of 68 in human allocation; this indicates that the algorithmic allocation was more equal. The random allocation algorithm achieved a Gini index of 32, which intuitively is close to the minimum possible, subject to the constraints. Graphically, as seen in Figure 5a, the closer the allocation line is to the diagonal line $y = x$, the fairer the allocation. Additionally, the x-axis is ordered from lowest to highest, so, for instance, our results show that the lowest 50% of all recipient organizations received about 5% of all donations under a human dispatcher, but received about 20% of all donations under our algorithm.

9.3.2 Poverty, Income, and Food Access of Recipients. When considering poverty, income, and food access levels, random allocation can be seen as uniformly sampling from the poverty, median income, and food access rates of all recipients because these features are completely recipient-specific. As illustrated in Figure 5d, Figure 5e, and Figure 5f, the human dispatcher's decisions closely followed the underlying population distributions, but our algorithm donated to recipients with higher poverty rates, lower median incomes, and worse food access. A Mann-Whitney U test shows that the algorithmic allocation gave donations to areas with higher poverty rates (Median =

[19]The convention is to report medians as the data is not normally distributed.

21.6%, SD = 14.44%) significantly more than human allocation (Median = 18.3%, SD = 13.73%, U = 1303400, p < .00000001). Indeed, Figure 5d shows that the human and the random algorithm gave more donations to areas with 10%-15% poverty rates, whereas our algorithms gave more donations to areas with about 50% poverty rates. Algorithmic allocation also gave more donations to recipients with lower income (Median = $40,275, SD = $16,312) than human allocation did (Median = $42,255, SD = $22,037, U = 1773200, p < .00000001), and the same pattern is observed to a lesser degree in the recipients' access to food levels (AA Median: 1.15 (SD=0.42), HA Median: 1.06 (SD=0.44), U=1414400, p =.0002; 0=Normal access, 2=Extremely low access).

9.3.3 Distance and Efficiency. One of the concerns of the organization was that distributing the donations more equitably could lead to longer and less efficient donation allocation. Our simulation results suggest that algorithmic allocation did not increase rescue distance, as illustrated in Figure 5c. A Mann-Whitney U test shows that the distance of rescues under algorithmic allocation, whose median is 5.5 miles, is significantly shorter than under human allocation, whose median is 6.15 miles (U = 1646900, p = 0.001).

10 DISCUSSION

In this paper, we envision a future in which people are empowered to build algorithmic governance mechanisms for their own communities. Our framework, WeBuildAI, represents one way to realize this goal. We have implemented and evaluated a system of collective algorithmic decision-making, contributing to the emerging research agenda on algorithmic fairness and governance by advancing understanding of the effects of participation.

10.1 Summary of the Research Questions and Results

We summarize our results in response to the research questions raised in the introduction.

10.1.1 What socio-technical methods will effectively elicit individual and collective beliefs about policies and translate them into computational algorithms? How should the resulting algorithms be explained so that participants understand their roles and administrators understand their decisions? (Section 8.1).

- The WeBuildAI framework successfully enabled participants to build models that they felt confident represented their own decision-making patterns. Participants understood graphical representations of individual models (Figure 3) and felt that collective aggregation via the Borda rule was fair. The decision support helped organizational administrators and other stakeholders understand how the final recommendations were made.
- Our findings suggest the elicitation method design could influence elicited beliefs. The top-down explicit-rule method may have promoted idealistic beliefs, while the bottom-up pairwise comparison-based machine learning method may have promoted realistic beliefs that accounted for emotions and constraints associated with tasks.

10.1.2 How does participation influence participants' perceptions of and interactions with algorithmic governance? (Sections 8.2 and 8.3).

- Participation not only resulted in new technology design but also affected participating individuals and organizations [36, 80]. Our participants reported greater trust in and perceived fairness of the matching algorithm, the governing institution, and administrative decisions after participating. Some participants were more motivated to use the services, felt respected and empowered by the governing institution, and reported a greater empathy for difficulties in the organization's decision-making process.

- Our participatory algorithm design, particularly the individual model building method, increased participants' algorithmic awareness and literacy. Through the process of translating their judgments into algorithms, they gained a new understanding and appreciation of algorithms. The method also revealed inconsistencies in employee decision-making in the governing organization, and made employees revisit their assumptions of other stakeholders.

10.1.3 How does the resulting collectively-built algorithm perform? (Section 9).

- The comparisons of the collectively-built matching algorithm and human allocation, using five months of historic data, suggest that the matching algorithm makes donation allocations more even, and gives more donations to recipient organizations in areas with higher poverty, lower income, and low access to food, without increasing the transportation distance.

10.2 Contributions to Research on Human-Centered Algorithmic Systems

10.2.1 Fairness and Moral Behavior in AI. In response to recent scholarly and journalistic work that has pointed out the need for "fair" algorithms, much research has been done to devise computational techniques that guarantee fairness in algorithmic outcomes. Our work offers a method for building procedurally-fair governing algorithms [49]. Our findings also offer empirical evidence of the effects of procedural fairness from the perspectives of both those who are affected by algorithms and those who use algorithms; the framework not only increased perceived fairness and trust of the algorithm but also influenced the organization by making the disparate effects of the algorithm more salient in their daily operation.

Our work also suggests that ongoing research seeking to understand people's moral concepts for algorithms and AI needs to be more cognizant of the design of the stimuli. (Some studies use more illustrative, vivid descriptions, whereas others use abstract textual descriptions.) Previous work in experimental moral psychology suggests that the vividness and realism of stimuli influences participants' answers. Consistent with this literature, our work suggests that the top-down versus bottom-up approach of building an algorithm may elicit different levels of construal, resulting in qualitatively different algorithmic models. It is important to choose an elicitation method and level of abstraction appropriate for the task context, and to take a reflective approach so that people can be aware of those situational effects and build a model in accordance with their beliefs.

10.2.2 Community Engagement in Algorithm Design. Our work contributes to recent research that calls for community engagement in AI design by offering a method to leverage varying stakeholders' participation directly in the design of the algorithm. By working with real-world stakeholders with various educational and economic backgrounds to build an algorithm that operates a service, we demonstrate the feasibility and potential of community involvement in algorithm design. At the outset of our research, we were unsure whether participants would feel confident and comfortable enough to express their beliefs on algorithms, and were concerned they might mistrust AI due to negative representations in popular media. It has been a rewarding experience to see participants not only expressing their beliefs, but also gaining trust in and becoming empowered through algorithmic systems. AI systems should be designed to facilitate these changes.

10.3 Levels of Participation in Algorithmic Governance

In this section, we define levels of participation in algorithmic governance. We discuss the upsides and downsides of different forms of governance and when collective participation is appropriate, reflecting on our research.

10.3.1 Closed, Non-Participatory Governance. Institutions can design a governing algorithm without involving stakeholders by drawing from their existing data and assumptions. This form of

governance is cost-effective compared to participatory governance, which requires effort and resources in soliciting and synthesizing participation. Closed governance is appropriate when there are legitimate metrics for algorithm design. For example, it might be appropriate if the goal is solely to minimize the volunteers' travel time. In our research, the organization was open to stakeholder participation because the staff were unclear on how to balance efficiency and equity in their daily operations. Additionally, closed governance may not inherently earn stakeholders' trust; it works best when the governing institution has already established trust with those being governed. Otherwise the algorithmic decisions may be challenged, mistrusted, or not adopted.

10.3.2 Mediated, Indirect Participatory Governance. Another form of governance is the mediated use of participants' input, resulting in participants' indirect influence on final algorithmic policy. In this form, stakeholders provide input to inform the designers and policymakers, who later design and implement the governing algorithms. The input can be collected through interviews or tools such as individual belief modeling, as in our framework. This form allows the governing organizations to operate on more accurate stakeholder assumptions, and communicating about the stakeholders' involvement can cultivate trust and increase the chances of adoption by those who are governed. This form is most appropriate when the organization seeks to use participatory feedback while retaining full control of the algorithm's design.

10.3.3 Direct Participatory Governance. In fully participatory algorithmic governance, stakeholders' participation is directly implemented in the final algorithm. In this form, participants feel most empowered and responsible, according to both existing literature and our work. However, the governing organization has less control over the final algorithm design. Direct participatory governance is most appropriate in contexts where stakeholders' trust and motivation to participate in the governing organization are critical, when a high level of procedural fairness is required, or in organizations and communities that are already self-governed, such as Reddit.

10.4 Extension of the WeBuildAI Framework and Future Work

Our application of the framework to 412 Food Rescue is a case study that implements participatory governance in one context. Our framework can be used and extended to support both mediated and direct participatory governance, and potentially for other algorithmic governance situations that involve normative design decisions and associated tradeoffs. For example, our framework could be used to create 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 candidate merit with equity issues. Extending our framework to new contexts requires addressing several challenges.

10.4.1 Individual Model Building as a Design Tool. Our findings suggest that the process of building individual models of algorithmic policy has many benefits. Externalized models provide a concrete place for starting a conversation about similarities and differences among the stakeholders or staff members of the organization. Designers and policymakers can use the models to inform algorithm design, or as an auditing or evaluation metric to assess the algorithm's effects from diverse stakeholders' perspectives. However, our research only used about 8-10 features that people could understand. Further research will be needed to apply the individual modeling method to algorithms with hundreds of features or more complex features. New techniques will be needed to explain and combine the features into a set that people can process.

10.4.2 Collectively Aggregated Decisions for Direct Participatory Governance. Our framework can be applied to enable direct participatory governance, particularly in contexts in which trust, motivation,

and perceived fairness matter, and, in its current implementation, contexts that do not require instantaneous decisions (within, say, less than a second).

One challenge, though, is to determine who participates and whether participation needs to be regulated. Opening up an algorithm to participation means that some participants may potentially hold opinions that are not socially acceptable. One way to avoid this is to limit participation so that democratic control of algorithms is subject to the constraints of public reason [10, 65]. This ensures that the behavior of algorithms is justified by a universally agreed-upon subset of principles. Future work would need to investigate how to broaden participation while respecting diversity within public reason, and devise an ethical way to determine the boundaries of participation.

Another challenge is ensuring the quality of participation, particularly when participation occurs at scale. Techniques used in crowdsourcing for quality assurance could be adopted to judge the quality of participation based on the amount of time and number of iterations people use in creating their models. Anecdotally, in our study, we observed that the machine learning model's accuracy was low when participants told us that they were applying rules inconsistently. Further work needs to investigate whether model accuracy can be another metric.

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. To clarify, we do not mean that the potential manipulation of the systems by the disempowered is a risk. We aim to create benefits for all those in need, and we believe the system could be at risk if some individual parties skew the results to maximize their own benefits when all participating individuals have a similar level of need. Indeed, one of the main topics of research in computational social choice [17] 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, this is not likely to be an issue for our framework because each individual does not have direct control over the final algorithm behavior. One may try to manipulate one's pairwise comparisons or specify preferences to obtain a model that might lead to preferred outcomes in very specific situations, but 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 individuals to benefit by behaving strategically. That said, future work would need to evaluate this question in the real world.

10.4.3 Promoting Representative Participation. One of our goals in designing this participatory framework is to empower stakeholders who typically do not have a say in the algorithms that govern their services, communities, or organizations. By empowering, we mean providing a method or tool that allows people to influence and control a system that they themselves use or an institution to which they belong [23, 33]. This shared power between users and developers, or individuals and governing parties, could increase the self-efficacy [7] and motivation [23] of stakeholders. Empowerment is one of the traditional values of HCI research and practice [68].

However, recently scholars have also pointed out that "material empowerment," or the technical tool itself [68] is not enough to enable people to make positive effects on social problems; one needs to devise solutions that also account for legal, social, and economic constraints [57, 68]. Our framework provides a tool that can enable stakeholders to participate in algorithm design, but it in and of itself will not necessarily result in equal empowerment of all stakeholders. Including representation from communities that are underserved or disadvantaged is a critically important challenge to address in future work. While many in these communities may technically have the opportunity to participate, they may face barriers like time or resource constraints that limit their access to participation. For our context with 412 Food Rescue, we acknowledge for example that

volunteers must have access to at least two relatively scarce commodities: access to a private vehicle and free time. Furthermore, recipient organizations often do not have reliable contact information for their clients, who may not have regular access to email or cell phone service. This poses a practical barrier to participant recruitment. In addition to technological design interventions like those we put forward in this paper, social and economic infrastructure will be necessary to ensure equal participation of all stakeholders.

10.5 Limitations

Like any study, our work has limitations that readers should consider. Our study 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. In the deployment, we will also consider additional evaluation measures for the algorithm, such as stakeholder satisfaction. 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, including recipient organizations' clients, by developing an educational component and targeted recruiting methods. We will also explore the possibility of running an open system, where people can join at any time or update their models by providing more data. We also acknowledge that despite our best efforts to base our design choices on participants' input gained through interviews (for example, who the stakeholders are, what factors to use), our views might have influenced our analysis of participants' inputs. Our plan to have an online system where participants can further comment on the selected features, stakeholders, and evaluation measures may mitigate this in the future. 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. On the one hand, it might be the case that a participatory, voting-based approach would be the only way to find a consensus solution. On the other hand, additional techniques—such as public deliberation through an open forum—might be needed to bring together polarized parties to ensure the efficacy of the resulting algorithms. Future work would need to investigate this question further.

11 CONCLUSION

Increasingly, algorithms make decisions influencing multiple stakeholders in government institutions, private organizations, and community services. We envision a future in which people are empowered to build algorithmic governance mechanisms for their own communities. Toward this goal, we proposed the WeBuildAI framework. In this framework, stakeholders build an algorithmic model that represents their beliefs about ideal algorithm operation. For each decision task, each individual's model votes on alternatives, and the votes are aggregated to reach a final decision.

As a case study, we designed a matching algorithm that operates 412 Food Rescue's on-demand transportation service, implementing the framework with their stakeholders: donors, volunteers, recipient organizations, and 412 Food Rescue's staff. We then evaluated the resulting algorithm with historical donation data, which showed that our algorithm leads to a more even donation distribution that prioritizes organizations with lower income, higher poverty rate, and lower food access clients compared to human allocation decisions. Our findings suggest that the framework improved the perceived fairness of the allocation method. It also increased individuals' awareness of algorithmic technology as well as the organization's awareness of the algorithm's impact and employee decision-making inconsistencies.

Our study demonstrates the value and promise of using the WeBuildAI framework as a design tool in order to achieve human-centered algorithmic governance. Future work needs to investigate

mechanisms to expand the application of the framework and its boundary conditions, as well as ways to overcome existing socioeconomic and institutional barriers to enabling wider participation.

ACKNOWLEDGMENTS

This work was partially supported by the Uptake & CMU Machine Learning for Social Good grant; the CMU Block Center for Technology & Society grant; the National Science Foundation CNS-1651566, IIS-1350598, IIS-1714140, CCF-1525932, and CCF-1733556 grants; the Office of Naval Research N00014-16-1-3075 and N00014-17-1-2428 grants; and a Guggenheim Fellowship. We thank Shiqi Chou and Calvin Lui who helped us analyze data; anonymous reviewers, Benjamin Shestakofsky, and Danielle Wenner who provided helpful comments; and our research participants who shared their valuable insights with us.

REFERENCES

[1] 412 Food Rescue Organization Website. 2018. https://412foodrescue.org
[2] Oscar Alvarado and Annika Waern. 2018. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Paper 286.
[3] A. Aneesh. 2006. *Virtual Migration: The Programming of Globalization.* Duke University Press.
[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (2016).
[5] Yukiko Asada. 2005. Assessment of the health of Americans: the average health-related quality of life and its inequality across individuals and groups. *Population Health Metrics* 3, 1 (2005), article 7.
[6] Madeline Balaam, Stefan Rennick Egglestone, Geraldine Fitzpatrick, Tom Rodden, Ann-Marie Hughes, Anna Wilkinson, Thomas Nind, Lesley Axelrod, Eric Harris, Ian Ricketts, et al. 2011. Motivating mobility: Designing for lived motivation in stroke rehabilitation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3073–3082.
[7] Albert Bandura. 2010. Self-efficacy. *The Corsini Encyclopedia of Psychology* (2010), 1–3.
[8] Neeli Bendapudi and Robert P Leone. 2003. Psychological implications of customer participation in co-production. *Journal of Marketing* 67, 1 (2003), 14–28.
[9] Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. 2012. On the efficiency-fairness trade-off. *Management Science* 58, 12 (2012), 2234–2250.
[10] Reuben Binns. 2017. Algorithmic accountability and public reason. *Philosophy & Technology* (2017), 1–14.
[11] Reuben Binns. 2017. Fairness in machine learning: Lessons from political philosophy. *arXiv preprint arXiv:1712.03586* (2017).
[12] 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*. 377.
[13] Keld Bødker, Finn Kensing, and Jesper Simonsen. 2009. *Participatory IT Design: Designing for Business and Workplace Realities.* MIT press.
[14] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.
[15] Alan Borning and Michael Muller. 2012. Next steps for value sensitive design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1125–1134.
[16] Bianca Bosker. 2016. The Binge Breaker. *The Atlantic* (2016).
[17] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. *Handbook of Computational Social Choice.* Cambridge University Press.
[18] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Paper 41.
[19] 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)*. 89–96.
[20] Bill Buxton. 2010. *Sketching User Experiences: Getting the Design Right and the Right Design.* Morgan Kaufmann.
[21] Anupam Chander. 2016. The racist algorithm? *Michigan Law Review* 115 (2016), 1023.
[22] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.
[23] Jay A Conger and Rabindra N Kanungo. 1988. The empowerment process: Integrating theory and practice. *Academy of Management Review* 13, 3 (1988), 471–482.

[24] Juliet Corbin, Anselm Strauss, and Anselm L Strauss. 2014. *Basics of Qualitative Research*. Sage.
[25] 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), 490–503.
[26] John Danaher. 2016. The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology* 29, 3 (2016), 245–268.
[27] John Danaher, Michael J Hogan, Chris Noone, Rónán Kennedy, Anthony Behan, Aisling De Paor, Heike Felzmann, Muki Haklay, Su-Ming Khoo, John Morison, et al. 2017. Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society* 4, 2 (2017).
[28] Norman Daniels. 2016. Reflective Equilibrium. In *Stanford Encyclopedia of Philosophy*.
[29] Robyn M Dawes and Bernard Corrigan. 1974. Linear models in decision making. *Psychological Bulletin* 81, 2 (1974), 95–106.
[30] Carl DiSalvo, Illah Nourbakhsh, David Holstius, Ayça Akin, and Marti Louw. 2008. The Neighborhood Networks project: A case study of critical engagement and creative expression through participatory design. In *Proceedings of the 10th Anniversary Conference on Participatory Design*. 41–50.
[31] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.
[32] 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.
[33] Marie Ertner, Anne Mie Kragelund, and Lone Malmborg. 2010. Five enunciations of empowerment in participatory design. In *Proceedings of the 11th Biennial Participatory Design Conference*. ACM, 191–194.
[34] FATML. 2018. Fairness, Accountability, and Transparency in Machine Learning Workshop.
[35] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.
[36] Archon Fung. 2003. Recipes for public spheres: Eight institutional design choices and their consequences. *Journal of Political Philosophy* 11, 3 (2003), 338–367.
[37] Archon Fung. 2006. Varieties of participation in complex governance. *Public administration review* 66 (2006), 66–75.
[38] 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.
[39] Tarleton Gillespie. 2010. The politics of "platforms". *New Media & Society* 12, 3 (2010), 347–364.
[40] Corrado Gini. 1921. Measurement of inequality of incomes. *The Economic Journal* 31, 121 (1921), 124–126.
[41] Mary L Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books.
[42] Lucas D Introna and Helen Nissenbaum. 2000. Shaping the Web: Why the politics of search engines matters. *The Information Society* 16, 3 (2000), 169–185.
[43] 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.
[44] Anson Kahng, Min Kyung Lee, Ritesh Noothigattu, Ariel D Procaccia, and Christos-Alexandros Psomas. 2019. Statistical Foundations of Virtual Democracy. In *Proceedings of the 36th International Conference on Machine Learning*. 3173–3182.
[45] Rob Kitchin. 2017. Thinking critically about and researching algorithms. *Information, Communication & Society* 20, 1 (2017), 14–29.
[46] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[47] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 1–16.
[48] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1035–1048.
[49] Min Kyung Lee, Anuraag Jain, Hae Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM : Human-Computer Interaction* 3, CSCW (2019), Article 182, 26 pages.
[50] 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.
[51] 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.

[52] E Allan Lind and Tom R Tyler. 1988. *The Social Psychology of Procedural Justice*. Springer.
[53] R Duncan Luce. 2012. *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation.
[54] 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.
[55] Charles F Manski. 1977. The structure of random utility models. *Theory and Decision* 8, 3 (1977), 229–254.
[56] 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, paper 9.
[57] Jessica Morley and Luciano Floridi. 2019. The Limits of Empowerment: How to Reframe the Role of mHealth Tools in the Healthcare Ecosystem. *Science and Engineering Ethics* (2019), 1–25.
[58] 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.
[59] Michael J Muller. 2009. Participatory design: The third space in HCI. In *Human-Computer Interaction*. CRC press, 181–202.
[60] 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)*.
[61] Arthur M Okun. 1975. *Equality and Efficiency: The Big Tradeoff*. Brookings Institution Press.
[62] Michael Q Patton. 1980. *Qualitative Research and Evaluation Methods*. Sage.
[63] Robin L Plackett. 1975. The analysis of permutations. *Applied Statistics* (1975), 193–202.
[64] Iyad Rahwan. 2018. Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.
[65] John Rawls. 1997. The idea of public reason revisited. *The University of Chicago Law Review* 64, 3 (1997), 765–807.
[66] John Rawls. 2009. *A Theory of Justice*. Harvard University Press.
[67] Dillon Reisman, Jason Schultz, K Crawford, and M Whittaker. 2018. Algorithmic impact assessments: A practical framework for public agency accountability. AI Now Institute.
[68] David Roedl, Shaowen Bardzell, and Jeffrey Bardzell. 2015. Sustainable making? Balancing optimism and criticism in HCI discourse. *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, 3 (2015), article 15.
[69] 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.
[70] Eric Schwitzgebel. 2019. Belief. In *The Stanford Encyclopedia of Philosophy* (summer 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.
[71] Amartya Sen. 2017. *Collective Choice and Social Welfare: Expanded edition*. Penguin UK.
[72] 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.
[73] Richard H Thaler, Cass R Sunstein, and John P Balz. 2014. Choice architecture. Manuscript.
[74] Louis L Thurstone. 1959. *The Measurement of Values*. University of Chicago Press.
[75] Yaacov Trope and Nira Liberman. 2010. Construal-level theory of psychological distance. *Psychological Review* 117, 2 (2010), 440–463.
[76] Huey-Ru Debbie Tsai, Yasser Shoukry, Min Kyung Lee, and Vasumathi Raman. 2017. Towards a socially responsible smart city: dynamic resource allocation for smarter community service. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*. ACM, Article 13.
[77] US Census Bureau. 2018. American FactFinder.
[78] USDA. 2017. Food access research atlas.
[79] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 440.
[80] 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.
[81] Max Weber. 2009. *The Theory of Social and Economic Organization*. Simon and Schuster.
[82] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, and Jason Schultz. 2018. AI Now Report 2018.
[83] Langdon Winner. 1980. Do artifacts have politics? *Daedalus* (1980), 121–136.
[84] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Paper 656.
[85] 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.
[86] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722.
[87] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), article 194.

A 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.

A.1 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 [55]. 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, 58, 63, 74]. 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 [60], 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.[20]

A.2 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

[20]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.

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.

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.

A.3 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 [43]. 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 [43], 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 [19]. We note that this is, in essence, equivalent to learning a linear utility model (in particular, a PL model). However, it slightly out-performs the aforementioned linear utility model.

A.4 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.