From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research

Michael Feffer
mfeffer@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Michael Skirpan
mskirpan@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Zachary C. Lipton*
zlipton@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Hoda Heidari*
hheidari@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

ABSTRACT

The AI Ethics community faces an imperative to empower stakeholders and impacted community members so that they can scrutinize and influence the design, development, and use of AI systems in high-stakes domains. While a growing chorus of recent papers has kindled interest in so-called "participatory ML" methods, precisely what form participation ought to take and how to operationalize these ambitions are seldom addressed. Our survey of the relevant literature shows that in many papers, participation is reduced to highly structured, computational mechanisms designed to elicit mathematically tractable approximations of narrowly-defined moral values. Of papers that actually engage with real people, these engagements typically consist of one-time interactions with individuals that are often unrepresentative of the relevant stakeholders. Motivated by these clear limitations, we introduce a consolidated set of axes to evaluate and improve participatory approaches. We use these axes to analyze contemporary work in this space and outline future AI research directions that could meaningfully contribute to operationalizing the ideal of participation.

KEYWORDS

Participation, elicitation, value-alignment

ACM Reference Format:
Michael Feffer, Michael Skirpan, Zachary C. Lipton, and Hoda Heidari. 2023. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3600211.3604661

1 INTRODUCTION

With the proliferation of data-driven algorithms automating or assisting high-stakes decisions in diverse societal domains [3, 60], the project of ensuring that these algorithms align with stakeholder values has taken on new urgency. As scholarly work on Fairness, Accountability, Transparency, and Ethics (FATE) has matured, a growing chorus of voices within the research community has called for centering issues of power, agency, equity, and participation [8, 9, 50, 71, 79]. For example, in addressing the goal of achieving *fairness*, scholars have highlighted the importance of determining precisely whose judgments about what constitutes fairness should be prioritized and how those values should be operationalized [67]. Offering appropriate responses to these critical questions requires the research community to design effective processes and mechanisms to involve stakeholders in the ideation, design, development, and use of ML systems in order to make sure these systems reflect their values and make deliberate, morally acceptable trade-offs when those values conflict with one another. Beyond a mechanism for value alignment, *participation* has been hailed as an end on its own and an essential ingredient of broader justice-related ideals, such as procedural fairness and democratic governance [77].

To heed these calls and include non-expert stakeholders in the process of designing, evaluating, and deploying ML-based decision-making systems, a recent line of work in AI/ML has supplied computationally feasible mechanisms to elicit stakeholders' moral preferences and values. For example, one early and influential study of this kind was the Moral Machine study, which went viral and attracted millions of internet users [6]. As described in Section 4, participants were posed pairwise comparison questions in the form of "trolley problems" [73]. In each scenario, they were asked to choose whose lives to prioritize, e.g., passengers' or pedestrians', in the face of an unavoidable accident. In follow-up work, Noothigattu et al. [59] used the data from the study to propose an algorithm to model and aggregate participants' *moral preferences* by estimating and averaging linear utility models in the hopes of reflecting all participants' preferences. Structured, computationally efficient mechanisms of this type are frequently designed to elicit mathematically tractable approximations of narrowly-defined values, and they have sometimes been referred to as *"Participatory ML"*. The rationale behind these methods is to provide the precision, formality, and scalability needed to model and capture moral values in ways that enable ML experts to translate them directly into measures and objective functions for developing and evaluating ML systems.

*These authors contributed equally.

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604661

While the intention behind the above line of work is noble and prior work in the area has provided several intriguing observations [44, 64, 71], we contend that there are fundamental limitations to these so-called "participatory ML" approaches. In particular, we critically examine the leap from structured preference elicitation to participatory design for value alignment. Through an extensive literature review and comparative analysis of several existing methods, we outline ten axes along which participation (by non-technical stakeholders) should be evaluated:

(1) Is the target stakeholder group **represented** appropriately?
(2) At what **stage** of the ML lifecycle is their participation sought?
(3) Is the appropriate **setting** for effective participation provided?
(4) Are adequate **resources** available to facilitate participation?
(5) Are there **communication** channels between participants and researchers to discuss the participatory task and the significance of its outcomes?
(6) Are the affordances and limitations of the **elicitation** mechanism adequately scrutinized, understood, and addressed?
(7) What are the mechanisms for **conflict resolution**?
(8) Do participants get to review and provide **feedback** about the process and outcomes of their participation?
(9) Does the participation benefit and **empower** the target stakeholder group?
(10) And finally, have the researchers properly **evaluated** their proposed approach?

To illustrate the utility of our proposed guidelines, we selected five influential participatory ML articles published in recent years and critically evaluated their contributions through the lens of our ten criteria. While the majority of these contributions required little in the way of resources, they all lacked adequate representation of stakeholder groups. Additionally, four of the five had mixed or unsatisfactory results along half of our axes, notably empowerment and communication channels. These findings suggest that while preference elicitation may pose an interesting computational problem, the corresponding methods are insufficient for addressing stakeholders' needs around participation.

In conclusion, as issues of empowerment, control, and agency take center stage in the AI ethics discourse, the research community must strive to provide real avenues of participation to marginalized stakeholders and impacted community members. We hope that the critique put forward here motivates future research toward closing major gaps and shortcomings of existing approaches and identifies new directions for impactful contributions, including considering participatory methods beyond traditional preference elicitation, increasing the representation of members of target communities in ML research and development teams, and acknowledging the fundamental limitations of ML as a tool to address complex sociotechnical challenges on its own.

2 RELATED WORK

In this section, we provide a brief overview of participatory design, its general critiques, and similar approaches in the context of AI/ML. We also highlight notable recent surveys of participatory design for AI and data practices that, while different in scope and scale from the current contribution, are recommended to the interested reader. We end this section with a brief overview of preference elicitation mechanisms proposed recently in CS venues.

2.1 An Overview of Participatory Design

Participatory design (PD) can be described as an approach to design that centers users in the design process [18, 45, 70]. While it originated in Scandinavian workplaces as a way to empower workers in light of technological changes [70], it has also been deployed in areas of governance and sustainable development as a way to empower citizens, particularly in the Global South [39, 45]. More recently, some have proposed that machine learning and automation can help increase participation in governmental processes, for example, by using natural language processing (NLP) to help citizens audit their government [65] or aid stakeholders in negotiating proposals and peace talks [4, 5].

However, others have argued that due to the large impacts that algorithms can have on people's lives (e.g., [3, 60]), participation in design of *the algorithms themselves* should be prioritized [2]. In line with this argument, researchers have employed qualitative participatory approaches to build algorithmic systems across a range of domains, including but not limited to Wikipedia content moderation [30, 69, 81, 83], teacher assistance tools [37], feminicide news and data collection [23, 72], and legal document review [19]. While our survey centers works that utilize quantitative/computational approaches to preference elicitation to build value-aligned models, it is important to highlight the qualitative work around PD as they share similar goals and are, at least for some design choices, more appropriate for achieving the goals of participation.

2.2 Critiques of Participatory Design

In spite of the general excitement around PD, it is not without its share of critics. In particular, Cooke and Kothari [17] and Mohan [55] criticize participatory approaches to government and development as they have been applied in the Global South. They argue that local practices may not always stem from culture but rather from necessity (i.e., due to scarcity of resources). They also critique homogenizing participants as a single group (resulting in participation benefiting certain subgroups more than others), and assuming norms and communication are similar (enough) to Western counterparts. In response, Kesby [46] concedes that *"[these] are important criticisms"* but nevertheless counters [17]. Using the author's prior work studying gender relations and HIV in Zimbabwe, Kesby highlights that blind resistance to participation on account of it involving power dynamics and domination is dangerous, and argues that well-utilized participation can actually lead to beneficial societal changes.

Focusing on participatory approaches to AI/ML design, Bratteteig and Verne [12] argue that various aspects of AI, including lack of transparency, possibility of biased data, and the need to adapt to new situations through constant training, can make it difficult for AI and PD to work together. Additionally, Delgado et al. [20] note that even though various practitioners are in favor of greater stakeholder participation in algorithm design, what constitutes *participation* in this sense is actually not clearly defined. Robertson and Salehi [63] and Sloane et al. [68] take their criticism a step further,

positing that participation can actually prevent progress or promote exploitation—depending on the choices available to participants.

There are various answers to the above critiques. For instance, recent work has proposed participatory frameworks for handling difficulties posed by PD and AI alike. Martin Jr et al. [53] put forward *Community-Based System Dynamics* (CBSD), which involves engaging stakeholders via causal loop diagrams and simulations to learn and include their viewpoints. Hossain and Ahmed [38] directly respond to Bratteteig and Verne [12] with a different approach they denote as *agile PD*. Drawing parallels with both agile software development and political activism, agile PD centers marginalized voices by leveraging stakeholders' spokespeople, alliances between practitioners and stakeholders, and stakeholder involvement in engineering processes. Hossain and Ahmed note that it is not a panacea to all of the issues raised in Bratteteig and Verne [12]. Nevertheless, they believe that *"agile PD is a first major step towards having a design method used with marginalized people that may be transferable to the design of AI technologies, but also revamped so that it does not encounter and contain the issues that exist with present-day PD."* Bondi et al. [10] respond to Sloane et al. [68]'s concerns of *participation-washing* with another framework called *Participatory Approach to enable Capabilities in communiTies* (PACT). PACT centers stakeholder participation in building AI for social good by inviting stakeholders to evaluate how resulting AI systems distribute and expand human capabilities. We draw from these critiques and responses in formulating our guidelines for participatory ML.

Addressing qualitative participatory approaches to AI algorithms, Birhane et al. [8] perform three case studies. The first involves participatory building of NLP-based translation tools for low-resourced languages in Africa. The second discusses an indigenous community's participatory shaping of data usage agreements. The third details a framework for participatory approaches to dataset documentation. Each of these case studies employs analysis in terms of benefits and shortcomings via priorities and related work similar to ones used here (e.g., [45]). What distinguishes our work is our focus on *quantitative* approaches proposed by AI/ML researchers and our critique structured around a set of axes along which such participatory approaches can be assessed.

2.3 A Survey of Preference Elicitation for ML

To gather papers for our literature review, we employed several approaches. We primarily consulted the proceedings of top conferences and journals that were likely to publish contributions that fit our definition of participatory ML. These venues included but were not limited to AAAI, FAccT, CHI, CSCW, AIES, and EAMMO. We also followed citation trails to and from widely cited papers in our repository, and used searches across Google Scholar and Arxiv to find additional related work.

The results of this review are summarized below. We grouped the articles retrieved by this search into one of two categories based on the goals the participation aimed to achieve: **use cases of moral preference elicitation** or **performance metric elicitation**.

Use cases of moral preference elicitation. Numerous participatory ML approaches have sought to create value-aligned algorithms for certain use cases through *moral preference* elicitation. Awad et al. [6] introduce the Moral Machine experiment, a study in which participants from various countries were posed questions that probed their beliefs about autonomous vehicles. Based on the results of that study, Noothigattu et al. [59] propose a method to construct a utility model to reflect the collective preferences of the participants, which in turn could quickly navigate ethical quandaries in a deployed system. Lee et al. [52] use similar techniques as [59] in conjunction with interview and workshop sessions. They do so to build a donor-recipient matching prioritization algorithm for a food delivery nonprofit based on participatory input from relevant volunteers and stakeholders. Kahng et al. [43] generalize the framework proposed in [52] to motivate algorithms that model participants' beliefs in order to facilitate democratic voting processes via automation. Kahng et al. [43] indirectly builds on earlier work by Lee et al. [51] to motivate participatory democracy via voting rules such as Borda count and Condorcet winner voting. Outside of the algorithmic governance space, Freedman et al. [27] demonstrate how to use participatory input to build kidney exchange algorithms that reflect stakeholders' beliefs. Johnston et al. [41] utilize preference elicitation in the medical resource allocation space, but their application domain is COVID-19 resource triage.

Performance metric elicitation. Another branch of participatory ML work involves building metrics to assess algorithms based on what is most important to participants. Ilvento [40], Jung et al. [42], Mukherjee et al. [58], and Bechavod et al. [7] propose methods to estimate individual-level definitions of fairness (as described in [24]) based on participant queries. Yaghini et al. [78] use Equality of Opportunity, as opposed to individual definitions of fairness, for metric elicitation. Hiranandani et al. [33, 34, 35, 36] apply metric elicitation to derive metrics that pertain to performance or group-level fairness. While these works have sought to build new metrics based on stakeholder input, others have explored which existing notions of fairness and feature selection most align with stakeholders' values. Saha et al. [64], Saxena et al. [66], Srivastava et al. [71], and Harrison et al. [31] assess participants' understanding of different fairness metrics and observe conditions under which they prefer some metrics over others. They do so based on crowdsourced responses to online surveys. In comparison, Cheng et al. [14] propose an interview protocol and user interface to help stakeholders weigh tradeoffs between metrics and gauge responses. Instead of gathering participants' thoughts on metrics, Grgic-Hlaca et al. [29] and Van Berkel et al. [75] explore what feature usage participants consider "fair" to use. Kasinidou et al. [44] further researches subtleties regarding *what participants consider agreeable* versus *what they consider fair* in decisions made by automated systems.

In Section 4, we offer further details and critical evaluations for a small selection of the above quantitative elicitation methods [27, 40, 52, 59, 71]. As we will argue shortly, our focus on these contributions is motivated by the attention they have garnered and is meant to illustrate evaluation via our axes (proposed in Section 3) through several concrete case studies.

3 TEN AXES FOR EFFECTIVE PARTICIPATION

Drawing on our extensive literature review and our sustained direct experience working with impacted community members, we provide a necessary set of axes for a quantitative approach to contribute

to the meaningful involvement of non-technical stakeholders in the design and use of ML systems. Table 1 summarizes our guidelines, and the rest of this section elaborates on each in more detail.

Representation. Our first axis concerns the representation of stakeholders in the participatory activity. We argue that *stakeholder groups should be represented commensurate to their need for/claim to empowerment* and that participants should be *generally representative of their respective stakeholder population.* These considerations serve to center marginalized voices in the activity. As noted by Cooke and Kothari [17] and Mohan [56], failing to do so may result in benefits of participation being enjoyed solely by those with prior privilege(s) and/or good social standing. Ideally, a representative individual or committee should also be placed in the research and development team.

Stage. The next axis pertains to how participants are involved in the ML lifecycle; namely, it concerns which part(s) of the ML pipeline (e.g., ideation, design, development, deployment, or maintenance) participants can affect. Our guidelines stipulate that participation generally requires *engagement as early as possible and at multiple stages of the ML lifecycle* as opposed to a one-time engagement after the system is already built and deployed. For instance, issues could arise if participants were involved in how the model performance was assessed but excluded from the data selection phase. Additionally, given that cultural norms and values evolve and that knowledge of a system's shortcomings accumulates over time, a one-time interaction may not suffice to decide how (or whether) the ML system should be maintained or discontinued.

Setting. The setting in which participation takes place is the next crucial component of our guidelines. Specifically, we contend that participation should be *conducted in an environment that is comfortable, familiar, and beneficial to participants.* If participation takes place in an unfamiliar or uncomfortable setting (e.g., research lab or company headquarters as opposed to one's own neighborhood), processes may not elicit true, underlying views of the participant (e.g., due to pressure or coercion, etc.). Moreover, a beneficial setting guarantees *fair compensation for participants (relative to that earned by the system's researchers and developers)* regardless of outcomes of the participation itself. Sloane et al. [68] argue for *"recogniz[ing] participation as work,"* and one way of doing so is to provide proportionate compensation, especially in cases where downstream deployed systems can yield nontrivial financial benefits for its designers and practitioners.

Resources. We argue that the participatory activity should be designed to be compatible with realistic resource constraints. This axis promotes forms of participation that *require minimal participant resources for effective engagement.* For example, the meaningful participation should not assume background knowledge that the stakeholder group does not possess.

Communication. As argued by Kelty [45], *"The experience of participation must include the sense not only of having spoken, but of* having been heard.*"*[1] To take this into consideration, we recommend that *there should be open-ended communication channels between practitioners and participants to discuss the activity.* In addition, we suggest that practitioners should *provide adequate background information to participants* and should *communicate outcomes of the activity to participants in a way they can understand and probe.* If participants are not provided requisite information on the problem the ML system is trying to solve (e.g., use cases and auditing metrics are obfuscated by technical jargon) and there are no ways to clarify misunderstandings, their participation may not reflect their true beliefs. Additionally, if results are not disclosed or understood, participants may rightfully feel exploited.

Elicitation. The specific mechanism and interface used for eliciting participants' judgements and values can have a significant impact on the the outcomes and perceptions of the activity. We suggest that elicitation should be done in a *multifaceted manner* that requires *reasonable effort* from participants while accounting for *realistic human conditions* (e.g., psychological tendencies and biases). This is in contrast to approaches that may utilize one form of elicitation (such as structured elicitation through pairwise comparisons), saddle participants with cognitive burdens, or assume rational agent models. As Vaughan [76] and Koppol et al. [48] indicate, humans are not oracles and can get tired, make mistakes, or even lie under certain circumstances. Therefore, participatory approaches that assume away these possibilities may fail in practice.

Conflict resolution. Aside from channels of communication between researchers/practitioners and participants, participants *should be empowered to communicate with one another, especially to deliberate and resolve disagreements.* Handling differences solely by crude enforcing mechanisms (such as majority rule) may quash a key aspect of participation and lead to unacceptable outcomes [49, 61, 63].

Feedback. Participatory approaches should allow participants to provide continual feedback to researchers and practitioners about every aspect of the activity, not just the specific question(s) of interest to researchers. For example, participants should be able to voice their concerns about the project generally or about the nature of their participation in particular. Our guidelines support participation that offers *multiple channels for continual feedback.* Failure to provide these channels may result in participants feeling undervalued, unheard, and exploited.

Empowerment. One of the most important axes considered by our guidelines is how participation actually affects the target population and their relevant outcomes. We emphasize that the target stakeholder group *should benefit from participation* beyond adequate compensation for their effort. Participants *should gain better control over the design process and outcomes as well as the benefits the activity produces.* The former is in line with existing human participant research practices (e.g., the Belmont Report, as summarized in [54]). For example, if the participants' involvement leads to significant research insights or accuracy gains, they should be acknowledged as co-authors and co-creators of the resulting artifact.

Evaluation. Lastly, our guidelines encourage researchers and developers of participatory mechanisms to critically evaluate their proposal. We emphasize the need to *verify the efficacy and validate* with people (as opposed to relying on simulations or mathematical proofs). Testing with actual human participants could uncover

[1] Emphasis added by Kelty.

Axis	Sample Prompts	Satisfactory Examples	Unsatisfactory Examples
Representation	*Are all stakeholder groups represented commensurate to their need for/claim to empowerment? Are participants representative of their respective stakeholder population?*	Stakeholder groups are adequately represented; marginalized voices are centered; a representative stakeholder has a long-standing voice in the broader research/development project.	Inadequate representation of key stakeholder groups; marginalized voices remain marginalized and disempowered.
Stage	*At what stage(s) of the ML lifecycle are participants engaged? What is/are the specific choice(s) for which participants can provide input?*	Engagement at multiple stages; providing input on impactful choices in each stage	One-time engagement; focus on unimpactful choices
Setting	*What are the conditions under which participation takes place? Is the setting familiar, comfortable, and beneficial to participants?*	Face-to-face human interactions; familiar location; adequate time and compensation	Virtual interactions; unfamiliar location; insufficient time and compensation
Resources	*What participant resources (e.g., time, money, background knowledge and expertise) are required for meaningful participation?*	Minimal resources needed for practical usage	Nontrivial resources and time investment required
Communication	*Can participants and practitioners communicate about the task? What background information is provided to participants about the ML system and the participation activity? How is the outcome of the activity communicated with them?*	Open-ended communication channels exist; participants are provided enough information accounting for their prior knowledge; results disclosed to participants in an understandable manner	No communication channels exist; participants lack the context required to understand the task; results not disclosed; disclosure is too high-level or technical.
Elicitation	*How are values elicited? What kinds of assumptions are made to capture those values?*	Multiple methods and user interfaces to elicit the same value; accounting for psychological effects	(Only) structured elicitation mechanism; Unrealistic agent models
Conflict resolution	*How are the conflicts of opinion among participants brought forward and handled?*	Channels to handle disagreement and deliberation among participants	(Only) crude voting mechanisms used to handle disagreements
Feedback	*Do participants have effective channels to provide continual feedback and voice concerns both about the participation outcome and the process?*	Feedback channels outside of elicitation	No feedback channels exist
Empowerment	*Does the participation empower/benefit the target population? How much control does it afford to participants? Can participants prevent technologies from being built or suggest entirely new routes?*	Participation provides great control over development and future benefits	Participants have little control over the process/product and little or no access to benefits
Evaluation	*How are participatory mechanisms/frameworks evaluated/validated? Have the findings been reproduced under various conditions?*	Co-design and testing with real human participants	(Only) verification with simulations and mathematical proofs

Table 1: Ten axes to evaluate participatory ML proposals.

practical challenges and limitations that proofs or simulations can never identify.

4 CASE STUDIES

While interest in this area has exploded in recent years, some works in particular have been highly influential. Specifically, they have been widely cited (e.g., on the order of 100 to 1000 times) and have inspired numerous follow-ups. As such, many of the other works in this field utilize similar ideas and fare similarly with respect to our axes. The review and critique in this section consist of in-depth case analyses of a handful of influential works in participatory ML, as a concrete illustration of the utility of our axes. Our assessments of these works are summarized in Table 2.

4.1 Moral Machine Voting

Description. Awad et al. [6] study the Moral Machine experiment in which people from around the world were queried about their personal ethics regarding autonomous vehicles. Specifically, users were posed questions similar to the Trolley Problem [73] in that in the face of an unavoidable accident between a self-driving car and pedestrians, they were asked to answer *whose lives should be prioritized: those of the pedestrians or those of the vehicle's passengers?* under varying conditions (e.g., young passengers, elderly pedestrians, etc.). Noothigattu et al. [59] subsequently use data collected from this experiment as an example and proposes a method to learn linear utility models based on Thurston-Mosteller (TM) processes [57, 74] that approximate individuals' beliefs, after which these models can be averaged together to obtain an overall model that

	Cases				
Axis	Noothigattu et al. [59]	Ilvento [40]	Srivastava et al. [71]	Lee et al. [52]	Freedman et al. [27]
Representation	Unclear target population, possible selection bias, mismatch of norms	Unexplored, no human involvement, left to implementers	Slight deviations from US Census	Selection bias through self-selection and volunteer-based sampling	Utilization of crowdworkers with awareness of divergence from target population
Stage	One-time engagement; at data collection stage	One-time engagement during model evaluation	One-time engagement during model evaluation stage	Interactions at multiple points in process, including model-building, aggregation, and result interpretation	Model training stage
Setting	Clear problem context, minimal language requirements, comfortable environment	Unexplored, left to implementers	Clear problem context, basic language requirements, and comfortable environment	Clear problem context and comfortable environment through in-person meetings	Clear problem context, basic language requirements, and comfortable environment
Resources	Access to internet; social media access	Unspecified, assumes knowledge of problem and access to querying system	Internet access; access to MTurk	Nontrivial time and effort required from participants	Internet access; access to MTurk
Communication	Information communicated through structured UI; No unstructured communication with researchers	Unspecified, left to implementers	Information communicated through structured UI	Frequent, free-form communication through interviews and workshops	Information communicated through structured UI; No unstructured communication with researchers
Elicitation	Structured elicitation via pairwise comparisons	Structured elicitation via pairwise comparisons	Structured elicitation via pairwise comparisons	Structured and unstructured elicitation via pairwise comparisons and interviews	Structured elicitation via pairwise comparisons
Conflict resolution	Aggregation of preferences via voting	None; assumes individual agent or body capable of coming to consensus	None; assumes individual agent or body capable of coming to consensus	Channels for deliberation during discussions and workshops	Aggregation of preferences via BT models
Feedback	None	None	Comment submission form which seemed to have no results on downstream experiments	Channels for feedback during discussions and workshops	None
Empowerment	Unclear	Unclear	Unclear	Control over parts of development, and understanding of results	Unclear
Evaluation	Evaluation via simulation on both synthetic and real-world pair-wise comparison data	Theoretical vetting via proofs of convergence	Human evaluation via MTurk workers	Human evaluation by participants and researchers	Human evaluation via MTurk workers of approach, not outcomes

Table 2: Details of case study assessments across each axis of participation. Green indicates relative satisfaction, orange indicates relative unsatisfaction, and yellow indicates mixed results.

ideally reflects the preferences of all participants. The authors conclude that the resulting model could be deployed at runtime and quickly decide the best alternative (in terms of utility maximization) that should be in line with the population's norms.

Evaluation. Based on Awad et al. [6]'s setup, given the scope and scale of the experiment, the target population is unclear. However, selection bias through requiring interest and internet access to participate may affect stakeholder representation. In particular, a *"[w]orld map highlighting locations of Moral Machine visitors"* in [6] illustrates noticeable sparsity in areas of the Global South, including but not limited to large swaths of Africa, South America, and east and central Asia. Given that existing literature also suggests that Western norms may not cleanly map to non-Western societies [9, 32], the experiment may not have truly elicited global values. Additionally, as described, Noothigattu et al. [59]'s model only allows for a one-time engagement of stakeholders. Pairwise comparison queries were posed via image comparisons to eliminate language barriers (see [6]) and participants could participate in their environment of choice, so the setting is reasonable. Provided participants have access to Internet and social media, resource requirements would also be minimal, as these are the only resources required by this framework. However, the protocol in [59] does not involve communication between participants and practitioners, and there is also only one structured elicitation mechanism through voting that

does not allow for conflict resolution between participants. Moreover, the protocol does not provide feedback channels or empower participants in the process, and simulations and [6]'s dataset were used to evaluate the approach. Beyond our guidelines, Chan et al. [13] highlights that stakeholder identity in terms of gender and perspective (i.e., passenger versus pedestrian) may affect elicited preferences, and El-Mhamdi et al. [25] and Feffer et al. [26] prove that the averaging approach employed here is not robust in the case that participants vote strategically.

4.2 Metric Learning for Fairness

Description. In light of existing work highlighting how various definitions of fairness may be mutually unsatisfiable under certain conditions (e.g., Chouldechova [15], Kleinberg et al. [47]), Ilvento [40] describes mathematically how to elicit metrics of fairness from people. The author does so by introducing an algorithm to obtain a metric grounded in an individual-based definition of fairness (such as the one described at length in [24]) from an agent by posing questions about the distance metric to use for the definition. Specifically, there are two types of queries posed to the agent:

(1) *real-valued distance queries*: questions inquiring about the distance between two individuals (e.g., $\mathcal{D}(u, v)$ for individuals u, v), and
(2) *triplet queries*: questions inquiring about whether one individual in a set of three is closer to one versus the remaining individual in the set (e.g., $\mathcal{D}(a, b) < \mathcal{D}(a, c)$ versus $\mathcal{D}(a, c) \leq \mathcal{D}(a, b)$ for individuals (a, b, c)).

Given these types of queries, the rest of the paper describes how to learn an individual-level fairness metric from these queries based on a finite set of N individuals and proves that their methods of doing so, specifically by choosing a set R of representative individuals and comparing them to other members of the set while using properties of a distance metric to order everyone, converge with total numbers of queries polynomial in $O(|R|N)$.[2] The work assumes that the agent is a single person or body of people "free from explicit biases or arbitrary preferences" but does not perform any analyses with actual human participants.

Evaluation. Participant representation is unconstrained and therefore determined by the researcher(s) or practitioner(s). Given that this method elicits preferences regarding performance metrics, any participants are only involved at a single point of the ML pipeline (the evaluation phase) and only have power over determining the output metric. Setting, communication, and participant resources are also unconstrained beyond assumptions of ample problem context and access to the querying system. The approach uses structured elicitation and utilizes a rational agent model. It also assumes that if the agent is actually a body of people, they should be able to come to consensus and resolve any disagreements amongst themselves. Therefore, there are no methods to handle disagreements between participants.[3] As described, there are also no feedback channels. Evaluations were performed via proofs of convergence and not with live participants.

4.3 Eliciting Perceptions of Fairness

Description. An alternative approach to eliciting value-aligned models or metrics is to determine which type(s) of performance people generally prefer or may want to prioritize for a given situation. To that end, Srivastava et al. [71] conduct three experiments with Amazon Mechanical Turk (MTurk) workers to explore conditions under which stakeholders may want to prioritize a certain definition of fairness as opposed to predictive accuracy or vice-versa. The first two experiments posed pairwise comparison queries to participants that asked which one of two algorithms was more discriminatory based on output of the algorithms in terms of predictions, ground-truth values, and demographics of people affected (namely their race and gender). One experiment involved algorithms to predict criminal recidivism while the second discussed algorithms to predict skin cancer likelihood. The researchers used the Equivalence Class Edge-Cutting (EC^2) algorithm [28] to simultaneously limit the number of queries to ask participants and estimate the mathematical definition of fairness that aligned with participants' responses. They found that demographic parity agreed with participants' answers the most often in both of these experiments, contrary to their hypotheses that equality of false positive or negative rates would be prioritized in the criminal recidivism setting while equality of accuracy would be prioritized in the skin cancer setting. In their last experiment, participants were asked to decide which of three algorithms with different fairness-accuracy tradeoffs should be used in a given setting. For instance, one of these algorithms had high accuracy for both men and women but at different rates while another algorithm with overall lower accuracy had equal accuracy across demographic subgroups. There were four settings in question that varied both the domain and stakes of the decisions being made:

(1) Skin cancer prediction (medical domain, high-stakes),
(2) Flu virus prediction (medical domain, low-stakes),
(3) Jail time prediction (criminal justice domain, high-stakes),
(4) Bail amount prediction (criminal justice domain, low-stakes).

In the end, they found that participants preferred accurate algorithms when the setting involved high-stakes decisions and fair algorithms when the setting involved low-stakes decisions, regardless of the setting's domain. The authors conclude by specifying limitations of their work (such as that they only considered algorithms with similar levels of accuracy, but larger differences in accuracy may have yielded other results) and suggesting future directions, arguing in particular that *"Algorithmic decisions will ultimately impact human subjects' lives, and it is, therefore, critical to involve them in the process of choosing the right notion of fairness,"* and that their work is *"an initial step"* in this direction.

Evaluation. In terms of representation, the sample of participants contained slight deviations from the US Census. While this work involves more than one interaction with participants, all experiments

[2] The bounds reported by Ilvento [40] take the form $O(|R|N)$ multiplied by a logarithmic factor and range from $O(|R|N \log N)$ to $O\left(|R|N \log \frac{1}{\alpha_L}\right)$ depending on the assumptions and querying algorithm provided, where α_L is "the minimum precision with which the arbiter can distinguish elements or distances."

[3] Ilvento [40] explicitly states "When human fairness arbiters strongly disagree, we consider this to be a situation where discussion between the human fairness arbiters, and perhaps additional external parties, is needed."

are limited to determining the fairness notion that most cleanly maps to their intuition. This in turn only relates to one part of the ML pipeline . Even though ample context about each problem was provided to the participants in their place of choice and resource requirements only involved access to MTurk, this work uses structured elicitation via MTurk as opposed to face-to-face interactions. The approach also lacked communication protocols, strategies for resolving conflicts between participants, and participant empowerment. While there were feedback channels for providing additional comments on the experiments, this input seemed to have no effect on determining the flow of the overall study. This being the case, evaluation of the approach still involved humans, in addition to quantitative and qualitative result analysis.

4.4 WeBuildAI

Description. In [52], Lee et al. summarize their work with a local nonprofit food delivery organization. Their goal was to improve the matching algorithm used to connect establishments with leftover food to recipient groups that could use it. As modifications to this algorithm involve a number of stakeholders with different preferences, the researchers believed a participatory approach would work well. The resulting format consisted of three phases. The first involved individual belief elicitation by creating models through pairwise questions (based on TM processes, similar to Noothigattu et al. [59]) or optionally via manual specification of scoring rules. The next used Borda count voting aggregate recommendations from these individual models and involved asking stakeholders *who, if any of you, should be prioritized in this voting process?* (in this case, they almost unanimously chose to prioritize the food delivery organization over donors and recipients). Lastly, the research team built and presented an interface to stakeholders in order to communicate effects of their participation on future decisions, namely in terms of explanations, preference rankings, and vote counts of various outputs. Each part of this process was conducted via in-person workshop and study sessions, and participants were compensated for their time and effort. However, participants were mostly a homogeneous group based on demographics (primarily white female), which the authors attribute to volunteer-based sampling.

Evaluation. While selection bias yielded a participant group that was not very diverse, the resulting participants were involved at multiple points in the algorithmic development process during face-to-face meetings. Sessions were conducted at participants' convenience and for which they were paid, and each provided participants with appropriate context. This being said, this method of participation was resource-intensive for the participants due to the time and effort needed to interact with the researchers. Results of these sessions were communicated to stakeholders during follow-up sessions and the built interface, and even though structured elicitation approaches were used at various points, participants had the ability to modify inputs to these approaches (such as by making individual models through explicit rules or altering Borda count voting power). Stakeholders had channels for feedback and deliberation and were also empowered by having control over several parts of the development process. Researchers' methods were evaluated based on these in-person workshop sessions.

4.5 Value-aligned Kidney Exchange Algorithm

Description. Freedman et al. [27] build on a long line of research on kidney exchanges (e.g., Abraham et al. [1], Dickerson et al. [21, 22]). The authors do so by reasoning about how to incorporate moral preferences into their clearing algorithm (e.g., the belief nonsmoking individuals should be prioritized to receive kidneys over smoking individuals). The work illustrates a proof-of-concept approach in this regard via two experiments in which MTurk workers were used as participants. The first experiment ascertained MTurk workers' thoughts on which patient attributes should be relevant to decisions about prioritization. The next involved asking another set of MTurk participants a number of pairwise comparison questions to learn how they prioritized donating kidneys. Specifically, each comparison involved answering a hypothetical question, namely *based on their patient profiles, which of these two individuals requiring a transplant should receive a kidney?*. Attributes included in these patient profiles were age, general health, and drinking behavior (e.g., young, healthy, rare drinking patient versus old, cancerous, frequently drinking patient) which were in-turn determined based on the results of the first experiment. With the resulting pairwise comparison data, the researchers built Bradley-Terry (BT) models [11] to approximate how the average participant made decisions. They subsequently used the corresponding BT scores for each patient profile to configure tie-breaking behavior of their clearing algorithm such that patients matching profiles with higher scores received kidneys over patients with lower scores if both were otherwise equally prioritized recipients. Finally, they compared this modified algorithm to their original algorithms without participant input through a number of simulation experiments and showed that the new algorithm behaved as expected (i.e., patients with profiles corresponding to higher BT scores received kidneys more often). Overall, they demonstrated that there were no technical barriers to implementing this algorithm.

Evaluation. While this work relies on crowdsourcing via MTurk, the authors clearly state that actually building a model for use in practice would involve working with a number of stakeholders closely related to the problem. There were two engagements with the participants, but both were related to one stage of the model development process through MTurk tasks with no other forms of engagement. However, using MTurk allowed researchers to provide problem context and allow participation in a setting where the participants were comfortable. While their method imposes few resource requirements on participants (primarily access to MTurk), the utilization of (only) structured elicitation means it lacks communication with participants, conflict resolution strategies between participants, channels for feedback, and participant empowerment. Lastly, it allowed them to evaluate their approach through human participation to prove that their method was technically sound, but they were not able to evaluate results in a practical setting.

5 DISCUSSION OF FUTURE DIRECTIONS

Drawing on the limitations of prior work, we conclude this work by outlining several important avenues through which AI/ML researchers and practitioners can effectively contribute to participatory frameworks.

Choosing and justifying the target population. Beginning any participatory project by strongly considering the appropriate target population can facilitate downstream parts of the process. This determination is not necessarily trivial, as questions like *why this group?* and *why this sampling approach?* may not be easy to answer. However, choosing a target population makes it possible to assess whether the sample of stakeholders is conducive to the end goals of the project or not (perhaps due to bias). For instance, Srivastava et al. [71] note that their sample of MTurk workers deviates slightly from the US population. But is this the appropriate target group of stakeholders to include in the process of determining the definition of fairness?

Identifying which choices in the ML lifecycle impact stakeholders' outcomes the most. By being aware of which parts of an ML project have the largest effects on outcomes relevant to stakeholders, AI experts can prioritize engagement with stakeholders on those choices. This prioritization, in turn, can empower participants by improving their control and agency over their subsequent outcomes. Along the same lines, by scrutinizing and potentially relaxing unrealistic assumptions (e.g., participants are rational or oracles of objective truth, preferences are stable and acyclic, etc.), experts can better ensure that proposed participatory approaches can capture the genuine opinions and preferences of the target stakeholder groups.

Acknowledging resource requirements and how they bias the sample. Research protocols that require participant resources, such as time and background knowledge, can hinder or prevent the participation of stakeholders that may otherwise be representative of the target population. This drawback happened to Lee et al. [52] as several participants could only partake in the first sessions due to time and job constraints. Participant barring and dropout can further bias the sample. While such issues may be unavoidable, delineating them as limitations and/or reducing them to the extent possible can promote participation and inform future research.

Meeting participants where they are. Using technical jargon, complicated interfaces, and unfamiliar environments to interact with stakeholders may not produce results in line with what they actually believe or want. In contrast, conducting exercises in a way that makes participants feel at ease can yield more faithful responses. To this end, researchers and practitioners can utilize everyday speech and writing where possible, pilot technical UIs before sharing them with participants, and host their participatory tasks in places stakeholders frequent in their daily lives.

Supplementing elicitation with deliberation. As evidenced by our review, quantitative approaches to preference elicitation have major limitations when used as standalone participatory activities. However, Lee et al. [52] demonstrate the utility of such techniques in conjunction with other forms of engagement and deliberation with stakeholders. Building systems via co-design as exhibited by works in Section 2 (e.g., [30, 37]) and developing new technology and interfaces to handle communications and richer forms of elicitation (e.g., [80, 82]) are among promising paths forward to promote deliberation.

Being receptive to feedback. Many of the works explored here either did not have channels for participants to share their thoughts on the activity with the researchers or did not appear to use input from participants in downstream processes. For instance, Ilvento [40] did not account for feedback in the protocol described, and while the UIs utilized by Srivastava et al. [71] featured open-ended comment boxes, crowd worker input did not appear to influence future experiments. Further work that receives and utilizes feedback can reduce feelings of exploitation, foster goodwill and collaboration, and ameliorate the sense of being heard.

Employing a wider range of frameworks to include non-technical stakeholders. The tacit assumption that experts should lead and execute research and reap its benefits has been challenged in other arenas (e.g., Participatory Design [18, 45, 70]). Research and tech development teams can diversify expertise and include relevant stakeholders as equal team members to incorporate their voices and expertise at various stages of their projects. This level of integration can prevent critical errors, reduce bias, and improve trust between researchers and stakeholder communities.

Conducting contextual, human-centered evaluation with representative participants. Most works referenced here rely on evaluation via simulations, mathematical proofs, or structured interactions with non-representative crowd workers. While these approaches are acceptable for early testing of new proposals, we join Freedman et al. [27] to strongly advocate for further validation studies on these systems (e.g., via usability testing with real stakeholders).[4] Additionally, as argued by Conitzer et al. [16] and Kelty [45], effective evaluation of a participatory activity with actual stakeholders requires both context and locality. As an example of context, garnering effective participation may require establishing long-term relationships with community advocates, representatives, and domain experts. Regarding locality, as we pointed out in Section 4, Western norms may not map well to all societies. For instance, Pugnetti and Schläpfer [62] note that even Swiss citizens (who presumably follow Western norms) have opinions that, on average, differ from those of the average respondent of the Moral Machine study [6].

Understanding the limits of what problems ML expertise can and cannot address. Last but not least, AI experts must avoid using elicitation methods as a way of participation-washing [68]–without empowering or benefiting participants, and to solely make outcomes appear more democratic. AI experts and practitioners must acknowledge that a wide range of skills beyond AI is needed to develop the necessary relationships with community stakeholders, gain their trust, and effectively moderate deliberations and resolve conflicts. ML expertise alone is not the solution to highly complex socio-technical challenges, and "participatory ML" is no exception.

ACKNOWLEDGMENTS

H. Heidari and Z. Lipton acknowledge support from NSF (IIS2040929) and PwC (through the Digital Transformation and Innovation Center at CMU). Z. Lipton additionally acknowledges NSF (FAI 2040929 and IIS2211955), UPMC, Highmark Health, Abridge, Ford Research, Mozilla, the PwC Center, Amazon AI, JP Morgan Chase, the Block

[4] In Freedman et al. [27], the authors note that deployment in the real world would involve medical professionals and other relevant stakeholders but also admit that determining the ideal mixture of medical and non-medical participants is nontrivial.

Center, the Center for Machine Learning and Health, and the CMU Software Engineering Institute (SEI) via Department of Defense contract FA8702-15-D-0002, for their generous support of ACMI Lab's research. M. Feffer acknowledges support from the National GEM Consortium and the ARCS Foundation. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the National Science Foundation and other funding agencies.

REFERENCES

[1] David J Abraham, Avrim Blum, and Tuomas Sandholm. 2007. Clearing algorithms for barter exchange markets: Enabling nationwide kidney exchanges. In *Proceedings of the 8th ACM conference on Electronic commerce*. 295–304.

[2] Evgeni Aizenberg and Jeroen Van Den Hoven. 2020. Designing for human rights in AI. *Big Data & Society* 7, 2 (2020).

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* (May 2016).

[4] Miguel Arana-Catania, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice* 2, 3 (2021), 1–22.

[5] Miguel Arana-Catania, Felix-Anselm van Lier, and Rob Procter. 2022. Supporting peace negotiations in the Yemen war through machine learning. *Data & Policy* 4 (2022), e28.

[6] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[7] Yahav Bechavod, Christopher Jung, and Zhiwei Steven Wu. 2020. Metric-free individual fairness in online learning. *arXiv preprint arXiv:2002.05474* (2020).

[8] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization* (2022), 1–8.

[9] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The forgotten margins of AI ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 948–958.

[10] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A Killian. 2021. Envisioning communities: a participatory approach towards AI for social good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 425–436.

[11] Ralph A. Bradley. 1984. 14 Paired comparisons: Some basic procedures and examples. In *Nonparametric Methods*. Handbook of Statistics, Vol. 4. Elsevier, 299–326. https://doi.org/10.1016/S0169-7161(84)04016-5

[12] Tone Bratteteig and Guri Verne. 2018. Does AI make PD obsolete? exploring challenges from artificial intelligence to participatory design. In *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial-Volume 2*. 1–5.

[13] Robin Chan, Radin Dardashti, Meike Osinski, Matthias Rottmann, Dominik Brüggemann, Cilia Rücker, Peter Schlicht, Fabian Hüger, Nikol Rummel, and Hanno Gottschalk. 2023. What should AI see? Using the public's opinion to determine the perception of an AI. *AI and Ethics* (2023), 1–25.

[14] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[15] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[16] Vincent Conitzer, Markus Brill, and Rupert Freeman. 2015. Crowdsourcing Societal Tradeoffs. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

[17] Bill Cooke and Uma Kothari. 2001. *Participation: The new tyranny?* Zed books.

[18] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

[19] Fernando Delgado, Solon Barocas, and Karen Levy. 2022. An uncommon task: Participatory design in legal AI. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–23.

[20] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2021. Stakeholder Participation in AI: Beyond" Add Diverse Stakeholders and Stir". *arXiv preprint arXiv:2111.01122* (2021).

[21] John P Dickerson, Ariel D Procaccia, and Tuomas Sandholm. 2013. Failure-aware kidney exchange. In *Proceedings of the fourteenth ACM conference on Electronic commerce*. 323–340.

[22] John P Dickerson, Ariel D Procaccia, and Tuomas Sandholm. 2014. *Price of fairness in kidney exchange*. Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

[23] Amelia Lee Dogan. 2022. Participatory Machine Learning Models in Feminicide News Alert Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 13134–13135.

[24] 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*. 214–226.

[25] El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, and Lê-Nguyên Hoang. 2021. On the strategyproofness of the geometric median. *arXiv preprint arXiv:2106.02394* (2021).

[26] Michael Feffer, Hoda Heidari, and Zachary C. Lipton. 2023. Moral Machine or Tyranny of the Majority?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37.

[27] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P Dickerson, and Vincent Conitzer. 2020. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence* 283 (2020), 103261.

[28] Daniel Golovin, Andreas Krause, and Debajyoti Ray. 2010. Near-optimal bayesian active learning with noisy observations. *Advances in Neural Information Processing Systems* 23 (2010).

[29] Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 world wide web conference*. 903–912.

[30] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.

[31] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 392–402.

[32] Joseph Henrich. 2020. *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Penguin UK.

[33] Gaurush Hiranandani, Shant Boodaghians, Ruta Mehta, and Oluwasanmi Koyejo. 2019. Performance Metric Elicitation from Pairwise Classifier Comparisons. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 89)*, Kamalika Chaudhuri and Masashi Sugiyama (Eds.). PMLR, 371–379. https://proceedings.mlr.press/v89/hiranandani19a.html

[34] Gaurush Hiranandani, Shant Boodaghians, Ruta Mehta, and Oluwasanmi O Koyejo. 2019. Multiclass performance metric elicitation. *Advances in Neural Information Processing Systems* 32 (2019), 9356–9365.

[35] Gaurush Hiranandani, Jatin Mathur, Harikrishna Narasimhan, and Oluwasanmi Koyejo. 2020. Quadratic Metric Elicitation with Application to Fairness. *arXiv preprint arXiv:2011.01516* (2020).

[36] Gaurush Hiranandani, Harikrishna Narasimhan, and Oluwasanmi Koyejo. 2020. Fair performance metric elicitation. *arXiv preprint arXiv:2006.12732* (2020).

[37] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019. Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics* 6, 2 (2019).

[38] Soaad Hossain and Syed Ishtiaque Ahmed. 2021. Towards a New Participatory Approach for Designing Artificial Intelligence and Data-Driven Technologies. *arXiv preprint arXiv:2104.04072* (2021).

[39] Sofia Hussain, Elizabeth B-N Sanders, and Martin Steinert. 2012. Participatory design with marginalized people in developing countries: Challenges and opportunities experienced in a field study in Cambodia. *International Journal of Design* 6, 2 (2012).

[40] Christina Ilvento. 2019. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250* (2019).

[41] Caroline M Johnston, Simon Blessenohl, and Phebe Vayanos. [n. d.]. Preference Elicitation and Aggregation to Aid with Patient Triage during the COVID-19 Pandemic. ([n. d.]).

[42] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. 2019. An algorithmic framework for fairness elicitation. *arXiv preprint arXiv:1905.10660* (2019).

[43] Anson Kahng, Min Kyung Lee, Ritesh Noothigattu, Ariel Procaccia, and Christos-Alexandros Psomas. 2019. Statistical foundations of virtual democracy. In *International Conference on Machine Learning*. PMLR, 3173–3182.

[44] Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. 2021. I agree with the decision, but they didn't deserve this: Future Developers' Perception of Fairness in Algorithmic Decisions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 690–700.

[45] Christopher M Kelty. 2020. *The participant: A century of participation in four stories*. University of Chicago Press.

[46] Mike Kesby. 2005. Retheorizing empowerment-through-participation as a performance in space: Beyond tyranny to transformation. *Signs: Journal of women in Culture and Society* 30, 4 (2005), 2037–2065.

[47] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[48] Pallavi Koppol, Henny Admoni, and Reid Simmons. [n. d.]. Iterative Interactive Reward Learning. ([n. d.]).
[49] Hélène Landemore and Scott E Page. 2015. Deliberation and disagreement: Problem solving, prediction, and positive dissensus. *Politics, philosophy & economics* 14, 3 (2015), 229–254.
[50] Benjamin Laufer, Sameer Jain, A Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. Four years of FAccT: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 401–426.
[51] David Lee, Ashish Goel, Tanja Aitamurto, and Helene Landemore. 2014. Crowdsourcing for participatory democracies: Efficient elicitation of social choice functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 2. 133–142.
[52] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.
[53] Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572* (2020).
[54] Vickie A Miracle. 2016. The Belmont Report: The triple crown of research ethics. *Dimensions of critical care nursing* 35, 4 (2016), 223–228.
[55] Giles Mohan. 2006. Beyond participation: strategies for deeper empowerment. In *Participation: The New Tyranny?*, Bill Cooke and Uma Kothari (Eds.). Zed Books, London, 153–167. http://oro.open.ac.uk/4157/
[56] Giles Mohan. 2006. Beyond participation: strategies for deeper empowerment. (2006).
[57] Frederick Mosteller. 2006. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Selected Papers of Frederick Mosteller* (2006), 157–162.
[58] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. 2020. Two Simple Ways to Learn Individual Fairness Metrics from Data. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 7097–7107. https://proceedings.mlr.press/v119/mukherjee20a.html
[59] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. 2018. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[60] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[61] Emma Pierson. 2017. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124* (2017).
[62] Carlo Pugnetti and Remo Schläpfer. 2018. Customer preferences and implicit tradeoffs in accident scenarios for self-driving vehicle algorithms. *Journal of Risk and Financial Management* 11, 2 (2018), 28.
[63] Samantha Robertson and Niloufar Salehi. 2020. What If I Don't Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design. *arXiv preprint arXiv:2007.06718* (2020).
[64] Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. 2020. Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning*. PMLR, 8377–8387.
[65] Paulo Savaget, Tulio Chiarini, and Steve Evans. 2019. Empowering political participation through artificial intelligence. *Science and Public Policy* 46, 3 (2019), 369–380.
[66] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.
[67] Michael Skirpan and Micha Gorelick. 2017. The Authority of "Fair" in Machine Learning. In *2017 ACM Conference on Knowledge Discovery and Data Mining, FATML Workshop*.
[68] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is not a Design Fix for Machine Learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–6.
[69] C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[70] Clay Spinuzzi. 2005. The methodology of participatory design. *Technical communication* 52, 2 (2005), 163–174.
[71] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2459–2468.
[72] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Ángeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 667–678.
[73] Judith Jarvis Thomson. 1985. The trolley problem. *The Yale Law Journal* 94, 6 (1985), 1395–1415.
[74] LL Thurstone. 1927. A law of comparative judgment. 34 (1927), 273–286.
[75] Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.
[76] Jennifer Wortman Vaughan. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.* 18, 1 (2017), 7026–7071.
[77] Brian Wampler. 2012. Participation, representation, and social justice: Using participatory governance to transform representative democracy. *Polity* 44, 4 (2012), 666–682.
[78] Mohammad Yaghini, Hoda Heidari, and Andreas Krause. 2019. A human-in-the-loop framework to construct context-dependent mathematical formulations of fairness. *arXiv preprint arXiv:1911.03020* (2019).
[79] Meg Young, Michael Katell, and PM Krafft. 2022. Confronting Power and Corporate Capture at the FAccT Conference. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1375–1386.
[80] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1245–1257.
[81] Angie Zhang, Alexander Boltz, Chun Wei Wang, and Min Kyung Lee. 2022. Algorithmic management reimagined for workers and by workers: Centering worker well-being in gig work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.
[82] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *arXiv preprint arXiv:2302.11623* (2023).
[83] 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), 1–23.