Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse

Jan Simson
Department of Statistics
LMU Munich
Munich, Germany
Munich Center for Machine Learning (MCML)
Munich, Germany
jan.simson@lmu.de

Fiona Draxler
University of Mannheim
Mannheim, Germany
fiona.draxler@uni-mannheim.de

Samuel Mehr
School of Psychology
University of Auckland
Auckland, New Zealand
Child Study Center
Yale University
New Haven, Connecticut, USA
sam@auckland.ac.nz

Christoph Kern
Department of Statistics
LMU Munich
Munich, Germany
Munich Center for Machine Learning (MCML)
Munich, Germany
University of Mannheim
Mannheim, Germany
christoph.kern@stat.uni-muenchen.de

Abstract

In light of inherent trade-offs regarding fairness, privacy, interpretability and performance, as well as normative questions, the machine learning (ML) pipeline needs to be made accessible for public input, critical reflection and engagement of diverse stakeholders.

In this work, we introduce a participatory approach to gather input from the general public on the design of an ML pipeline. We show how people's input can be used to navigate and constrain the multiverse of decisions during both model development and evaluation. We highlight that central design decisions should be democratized rather than "optimized" to acknowledge their critical impact on the system's output downstream. We describe the iterative development of our approach and its exemplary implementation on a citizen science platform. Our results demonstrate how public participation can inform critical design decisions along the model-building pipeline and combat widespread lazy data practices.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Collaborative and social computing**; **Human computer interaction (HCI)**; • **Social and professional topics** → *User characteristics.*

Keywords

Participatory Design, Machine Learning, Algorithmic Fairness, Multiverse Analysis, Citizen Science, Garden of Forking Paths

CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713482

ACM Reference Format:
Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. 2025. Preventing Harmful Data Practices by using Participatory Input to Navigate the Machine Learning Multiverse. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/3706598.3713482

1 Introduction

Algorithmic decision-making (ADM) fueled by machine learning (ML) algorithms is becoming ubiquitous in many domains, affecting the lives of millions of individuals. Examples include jobseekers that are classified into different risk groups by profiling models [74], refugees that are re-allocated within their host country based on matching algorithms [7] and the denial or approval of health care coverage for patients [6, 89, 126]. While such systems are introduced with the aim of improving the effectiveness and efficiency of decision-making, there are also serious concerns that algorithmic decisions can treat the affected individuals unfairly [88, 89, 93]. Fairness implications of ADM ultimately depend on how the underlying models interact with biases and deficits in training data, and thus the design, implementation and evaluation of the ML system is of central concern [17, 117]. Addressing fairness and adverse impacts, therefore, does not only include technical measures but rather needs a broader public discourse where developers, stakeholders and affected individuals meet on an equal footing to design and evaluate algorithmic systems within their respective deployment context.

Despite increasing efforts to open up the ML pipeline and involve stakeholders in participatory designs, current research is lacking tools that enable public input where it matters most: the design of the ML model itself.

The ML pipeline includes a multitude of critical decision points – from data selection and curation to pre-processing, modeling and evaluation decisions. As each decision point allows for multiple alternative choices, design decisions in ML resemble a *garden of forking paths* [50] where each fork corresponds to a decision which in turn leads to a set of further scenarios downstream. The full grid of decision combinations can also be understood as introducing a *multiverse* of (potential) ML models, in which each model is defined by a unique path through the set of design decisions upstream [117, 123]. Even with a handful of decisions, a multiverse can quickly grow extremely vast: For the example application we use in this paper, four design decisions lead to a multiverse with 16,352 endpoints (i.e., unique ML models) as illustrated in Figure 1, which grows to 784,896 combinations when four evaluation decisions are included — a detailed description of the use case is presented in Section 3.1.

In ADM contexts, the decision points in the ML multiverse often involve normative considerations and inherent trade-offs: should the model use sensitive attributes such as race and gender as features? Should a more complex or a more interpretable model be used for the task at hand? Which evaluation criterion (error metric) is most important when choosing the final model? Often, such decisions cannot (and should not) be "optimized" based on training data alone. A central result of the fair ML literature, for example, has shown that key fairness notions are incompatible with each other, and thus, it is essential to reason on substantive grounds which type of model error is viewed as most critical in a given application context [27]. Even for "technical" decisions, e.g., in the data pre-processing context, recent research has shown considerable design effects on model fairness downstream [24, 117]. At the same time, questionable design decisions can commonly be observed in data practice: In a review of 280 experiments in the field of fair ML, Simson et al. [116] identify harmful shortcuts such as filtering out members of ethnic minorities in data processing, which are commonly taken even by practitioners in fairness research. They group these shortcuts under the term *lazy data practices*. These observations jointly call for a democratization of the design process not only to provide critical public feedback but to engage diverse stakeholders and affected individuals (as domain experts) along the ML pipeline. As central decisions concerning data processing, evaluation and metrics need to be considered within the given application context. Active public engagement is critical to evaluate and tailor technical decisions according to the needs and perspectives of the communities in which a system is sought to be placed.

In this paper, we introduce a participatory approach to prune the garden of forking paths and help navigate the multiverse of design decisions in ML. While current work in participatory artificial intelligence (AI) rarely focuses on the collaborative shaping of the system's design, including technical decisions such as the type of model and features used [37], these decisions critically affect the eventual functioning of the system, its predictions and fairness properties. In light of inherent trade-offs (including fairness, privacy, interpretability versus performance considerations) and normative questions (e.g., which error notion – for which group – to prioritize), the ML pipeline needs to be made accessible for public input to foster inclusion and participation of diverse populations.

Using a case study of predicting public health care coverage, we demonstrate how participatory input can be implemented to produce meaningful data. Our results show how participants' choices can be better than common practices employed by practitioners and how this can be used to address harmful and lazy data practices in the field. We further demonstrate how results can be used to navigate the machine learning multiverse more efficiently, pruning pathways which are not deemed acceptable by a wide majority of participants.

Our contributions include a reusable workflow for preparing the ML pipeline for participatory input and a case study where this workflow is implemented. We further collect participatory input using a citizen science setup, successfully gathering a diverse sample of participants from across the world. We provide an empirical evaluation of this participatory data and put the results into context by simulating models in the machine learning multiverse.

2 Related Work

This paper links multiverses of decision points in ML engineering with participatory ML. Accordingly, this section introduces the multiverse concept, including application scenarios, its relation to fairness, and multiverse analyses, and it provides an overview of participatory methods in ML and related fields.

2.1 Multiverse Analysis

2.1.1 The Garden of Forking Paths. Whenever one conducts data analysis, there are many different decisions one has to make along the way [113], both explicit and implicit [117]. This has often been likened to a garden of forking paths [50], where each decision creates a fork in the path, creating a multiverse of pathways and destinations. *Many-analyst studies*, where multiple individuals or teams conduct an analysis using the same dataset and research question, show that such decisions can have large effects on the findings [21]. While awareness around this problem has mainly focused on statistical analyses, in particular on null hypothesis significance testing (NHST), it also applies within the context of machine learning and artificial intelligence [64, 90]. In many cases, practitioners may not even be aware of making decisions as they traverse the ML pipeline, although new solutions are being explored to bring potentially problematic decisions to light via notifications during development [60].

2.1.2 Hyperparameter Optimization. The classic scenario where a garden of forking paths is navigated within ML and AI is during hyperparameter optimization (HPO) [14, 47]. Here, a grid of different parameter configurations is created and traversed using either an efficient search algorithm (e.g., Bayesian Optimization [120]) or a full scan of parameter combinations. Research in this field tends to focus on efficiency and finding better search algorithms. However, it has also led to the development of new methods to better understand variance in the space, which can be adapted to the garden of forking paths or multiverse, such as efficient implementations of the functional analysis of variance [65, 68]. As the name suggests, the HPO literature focuses heavily on hyperparameters. However, it usually ignores other critical decision points in the ML pipeline, such as data selection, preprocessing or evaluation decisions. These decisions govern how the model interacts with (biases in) data and have been shown to affect fairness outcomes [24, 117].

2.1.3 Fairness & Multiplicity. Research on algorithmic fairness aims to address and reduce disparities in algorithmic systems. This

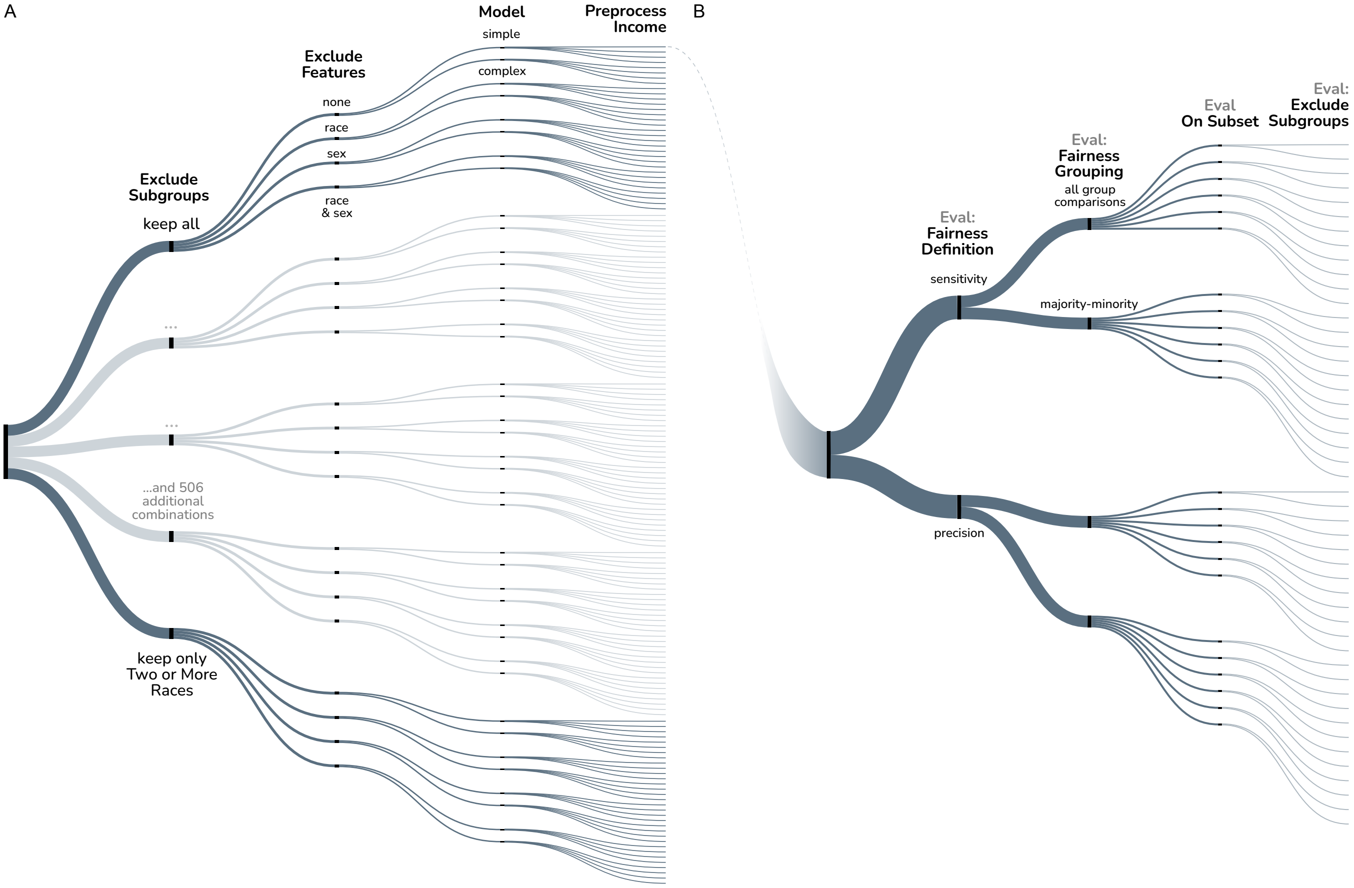

Figure 1: The full multiverse of different decisions is incredibly vast. Illustration of the multiverse of different ML models (A) and evaluation strategies (B). Due to practical limitations, only a small fraction of the multiverse of ML models is shown here, with 506 additional branches hidden, as illustrated by the reduced transparency and only 320 out of 16,352 endpoints visible (<1%). As the two multiverses are not mutually exclusive, their combined total are 784,896 combinations of model and evaluation.

includes formalizing fairness notions in a way that they can be applied to the predictions or scores of an ML system. Group fairness metrics typically compare (different types of) prediction error between groups defined by *protected attributes* [8, 93]. The selection of protected attributes may be based on anti-discrimination law, with *race* and *gender* [116] being common examples. Many specific fairness metrics, workflows, auditing and bias mitigation strategies exist, and the choice between them is neither trivial nor independent of context [8, 85, 109].

A consequence of the machine learning multiverse is that there are often different ML models with equal or comparable performance (examples in [25, 34, 107, 131]). This is termed the *Rashomon Effect* [20] and the set of comparable (best) models is commonly referred to as the *Rashomon set* [110, 131]. One benefit of the Rashomon Effect (among others [108]) is that it allows optimizing for secondary objectives such as fairness, e.g., by selecting the model with the optimal fairness metric from the Rashomon set [17, 69]. It has been argued that there may even be a legal duty to search through the Rashomon set for fairer alternative models [16].

A potentially more troubling result of the Rashomon Effect is that models in the Rashomon set can give very different predictions for a given individual [17, 29]. This phenomenon is commonly referred to as *individual arbitrariness* or *predictive multiplicity* [17, 82], although other names exist [54]. When the choice between different models and, thereby, individual predictions is arbitrary, multiplicity can lead to issues in the justifiability of predictions and decisions [17, 29].

The sources of multiplicity can be diverse, and multiplicity has been demonstrated across different dimensions, such as random seeds [29], target variables [128], different sparse decision trees [131] and the dataset generation process [91] as well as model design decisions [117]. Multiple measures have been proposed to quantify these effects [29, 41, 66]. Related work has examined the influence of different forms of imputation for missing data [24] and hyperparameter selection [42] on algorithmic fairness, albeit not through the lens of multiplicity. A similar concept has also been demonstrated in the context of explainable ML through *fairwashing*, where equivalent models can be generated that show little dependence on sensitive features [2]. Thus, from a fairness angle, decisions along the model-building pipeline are critical: They can introduce individual arbitrariness, exacerbate or hide data deficits, and can result in a variety of models downstream, which differ strongly in their fairness properties.

2.1.4 Multiverse Analysis & Fairness Hacking. While the influence of a variety of decisions on algorithmic fairness has been well-documented in the field, it is usually only examined along a single dimension. As different decisions often interact [117], however, they should also be analyzed together. This can be done through a multiverse analysis, where all possible decision combinations – the forking paths – are evaluated.

Multiverse analyses [123] first emerged in response to the reproducibility crisis [28], where large amounts of research failed to be reproduced. A similar type of analysis, *specification curve analyses* [114], emerged around the same time. Specification curve analyses are characterized by their use of a particular type of visualization to show results from a multiverse analysis. The idea of multiverse analyses is also closely related to that of a *sensitivity analysis* [78], although multiverse analyses are typically bigger in scope, trying to include more decisions and especially more interactions between decisions.

Multiverse analyses have been successfully used in machine learning to study performance [10] and fairness [117] of models. Their most useful application in ML may be, however, as a potential solution for issues of fairness hacking [117]. *Fairness hacking* describes practices of presenting unfair models as fair [87]. This can be achieved by iterating over different definitions and metrics of algorithmic fairness [87] or evaluation strategies [117] and selecting the most favorable one while keeping the actual model fixed. A similar concept has been described under the name *d-hacking* [15].

In the ML context, a multiverse analysis allows us to explicate and structure important data processing, modeling and evaluation decisions and makes them visible and accessible. However, the decision points themselves often involve difficult trade-offs and choosing an option based on metrics alone is commonly not an option for ethical reasons. Even if a decision does not touch upon ethical issues, multiverse analyses can become unfeasible due to computational limitations, especially when costly and complex models are being fit. Participatory input has great potential in these cases to make more informed decisions and constrict the number of pathways in the multiverse.

2.2 Participatory Machine Learning

Participatory machine learning (and often used interchangeably, participatory AI) emerged as a response to power imbalances between system engineers and those affected by or those using a system [76], which can result in biases such as a disproportional impact on marginalized groups [72]. The core idea in participatory ML is to involve stakeholders in the design, development, and deployment of ML models or AI systems [13]. Thus, the aim is to empower stakeholders and to increase fairness, transparency, and accountability [37, 46]. In the context of this work, we specifically address participatory ML for *fairness*.

Application domains of participatory ML range from estimated content quality of Wikipedia edits [57] to various healthcare applications [38]. Contributors shape ML systems as participants in goal-setting or algorithm design workshops [22], model builders or adapters [57], evaluators in result assessments [79], etc. Past work has suggested a variety of tools that support participation: To name just a few examples, model cards help identify and discuss trade-offs in ML model design [112], visualizations and comparisons of model predictions for different inputs encourage reflection [26] and re-designed visualizations of prediction accuracy targeted specifically at non-experts enable them to better assess ML model performance [111].

In contrast to defining the goals of an ML project or eliciting user preferences, participatory approaches for decisions concerning the design and evaluation of ML models are relatively rare. Notably, in a review of 80 participatory AI projects, only 10% included stakeholders in design or specification tasks such as choosing models or features [37]. A related review reports that only a small share of participatory AI projects involved stakeholders beyond a "consultation" stage [31].

However, participatory model and evaluation design can be beneficial for multiple reasons. First, they increase the diversity of perspectives to take into account to avoid pitfalls that professionally blinkered ML engineers may run into. For example, past work highlights that developers often make implicit model-building and evaluation decisions that may impact model fairness [60, 117]. However, adjusting attribute weights in an ML model based on non-experts' fairness perception can actually make a system fairer – although some individual input may also worsen fairness ratings [94]. Second, many problems are only discovered when stakeholders are involved in the discussion [134], as highlighted in Criado-Perez's [33] book *Invisible Women: Exposing Data Bias in a World Designed for Men.* She argues that people must be asked to identify their needs and requirements, e.g., for reliably recognizing heart attack symptoms in women. Therefore, in our work, we look at the relationship between being a member of a minority group and the decisions that participants make.

Participatory ML also entails certain caveats. Notably, balancing power between participants and those asking for their participation is crucial to avoid exploitative practices [30, 31, 118]. There is a danger of co-optation, i.e., superficially acknowledging input and involvement, while the actual influence remains minimal [13]. It is also important to address aspects that non-experts may not be aware of. For example, when non-experts build ML models, they often optimize toward percentage accuracy and may overlook issues such as overfitting [132].

In this paper, we investigate the possibility of using participatory input for actual model decisions (design and evaluation) with a focus on fairness. Thus, we focus on specific steps of a full participatory ML pipeline for which participatory methods are currently underexplored. We rely on iteratively refined descriptions of decision choices to make them accessible for non-expert participants.

3 Methods

For our participatory approach, we chose an exemplary ML use case situated in the ADM context. We derive a set of relevant decisions on model design and model evaluation and set up an online experiment on a citizen science platform.

3.1 Decisions

We use a case study of predicting whether an individual is covered by public health care in the U.S. based on socio-demographic information with data from the American Community Survey Public Use Microdata Sample (ACS PUMS) [23]. In particular, we use the *ACSPublicCoverage* problem, one of a set of problems commonly referred to as "folktables" [40]. We opted for this problem, as it is prototypical for ADM scenarios, due to being a binary classification task which can be framed as a risk prediction problem with race defined as the protected attribute.[1] We also chose this particular problem due to its practical relevance, as healthcare is a highly important domain, with commonly reported fairness issues [6, 89, 96, 126]. In an ADM setting where decisions would be made based on the model's predictions, incorrect predictions of individual health care coverage could lead to communities falling under the radar of preventative measures or information campaigns that are allocated based on the predicted risk of non-coverage. This is particularly concerning for minority groups with historically low coverage rates. Specifically, incorrectly predicting that an individual is covered (i.e., a false positive) might exclude them from preventative measures or targeted campaigns, while incorrectly predicting non-coverage (i.e., a false negative) could lead to a misallocation of such measures. While these implications likely unfold differently if individuals are covered by private insurance plans instead, the *ACSPublicCoverage* task focuses on individuals with an income of less than 30,000 U.S. Dollar per year who may depend more strongly on public offers.

When determining the list of decisions to include in the study, we focused in particular on decisions which may be made "ad-hoc" (sometimes even without the awareness of making a decision), but which eventually introduce trade-offs and involve normative considerations. We selected, adapted and extended upon a list of decisions identified in prior work [117]. Implicit decision-making is particularly common for decisions regarding the evaluation of an ML system. Therefore, we ultimately decided to include four design decisions affecting the model itself and four affecting only its evaluation.

For each decision, we crafted a brief introductory text describing the trade-offs inherent to the decision, taking care not to present any one option as more favorable than others. We refined this introductory text as well as the descriptions of each option in the decision across multiple iterations to make them understandable without prior knowledge of machine learning or artificial intelligence concepts.

A brief explanation of each decision and its options can be found below. The actual wording of each decision, its introductory text and options can be found in Appendix B.3. An overview of all decisions and options can be found in Table 2, and the resulting multiverse is illustrated in Figure 1.

3.1.1 Model Design Decisions.

We included four decisions on model design, covering preprocessing, data and model selection. Each decision includes between two and nine distinct options. However, as the decision *Exclude Subgroups* makes use of the combination of its nine options, it allows for a total of 511 unique combinations of these options. Together, the four model design decisions create a multiverse of 16,352 potential ML models. A subset of this multiverse is illustrated in Figure 1A.

Exclude Subgroups. Related work shows that the exclusion of certain subgroups before model training is common. For example, a recent review focusing on the popular *COMPAS* [5] dataset found that 38 out of 59 studies excluded data from subgroups of the protected attribute [116]. While certainly problematic, as it can lead to representation bias [88, 125], this does not have to be out of malicious intent: One may want to exclude data from certain subgroups to protect their privacy or to make analyses less complicated. Indeed, many commonly used fairness metrics are first and foremost designed with the assumption of only two protected groups, requiring adaptations to work with more nuanced protected attributes. However, we also want to clearly note that this is a problematic trend, and our inclusion of the decision in this study stems from the intent to highlight this issue rather than normalize the practice. We advise for careful deliberation and against the exclusion of subgroups in most real-world scenarios. In our experiment, we present all nine racial and ethnic groups available in the data of our modeling task to participants, using the order suggested by the ACS PUMS [23]. Participants have the option of combining the groups as they see fit to construct the list of included subgroups.

Participants were randomly assigned to one of two conditions for this decision: Either they were shown the percentages illustrating the relative size of each group next to the group name or not. We added this differentiation because overestimation of minority group sizes is common [3].

Exclude Features. It is common in fairness-related contexts to not include potentially sensitive attributes as predictive features in models due to privacy reasons or with the intent to produce less biased models [55, 77]. This practice does not necessarily produce fairer models, as fairness through unawareness has been shown not to work [8]. Nonetheless, we included this decision to represent the popular practice. We include four different options for this decision: (1) To exclude the protected attribute *race* as a feature for the model, (2) to exclude the potentially sensitive attribute *sex* as a feature for the model, (3) to exclude both *race* and *sex* as features or (4) to exclude neither of the two and use both as features for the model.

Preprocess Income. We included the preprocessing of the variable *income* as an example of a feature that is often binned into different categories. Income data is highly relevant to the outcome we are trying to predict, and the correct processing of income data is usually an arbitrary choice without a clear consensus. Preprocessing of income data was also shown to be an influential decision among several comparable preprocessing decisions in its effect on algorithmic fairness [117]. Indeed, the arbitrariness of thresholding income data (as a target) has been criticized [40] as one of the issues

[1] At the same time, datasets that feature other (more prominent) examples from the fairness literature (recidivism prediction, credit scoring) have been shown to suffer from considerable quality issues [43].

in *Adult* [9], a popular dataset based on U.S. Census data with an associated machine learning task of predicting whether an individual's income is above $50,000. We included four different options for processing income data: (1) Keeping it as is, (2) binning it into bins of $10,000 and binning it into (3) three or (4) four equally sized groups.

Model. Choosing the model type is a critical and consequential decision point in the ML pipeline. We thus wanted to include at least one decision on the type of ML model that is used. However, during the development of the decision and option descriptions, we quickly realized that explaining the nuances of different ML models is beyond the scope of a single study and may be challenging to understand for non-experts. We therefore opted to only present participants with the choice between (1) a simple and (2) a more complex model, highlighting the classic trade-off between performance and interpretability with increased complexity [51]. We decided to use a logistic regression [32] for the simple and a random forest [63] for the more complex model. Models were fit using default hyperparameters.

3.1.2 *Evaluation Decisions.*

We included four distinct decisions on model evaluation, each with two to six options. Together, these decisions allow for 48 different strategies of how one might choose to evaluate a (fixed) ML model (Figure 1B). As they can be applied to each of the models created in the previous step, they increase the size of the multiverse from 16,352 unique models to 784,896 different evaluations.

Eval Fairness Definition. Deciding on a particular metric in fair ML is one of the most critical decisions, with multiple valid and conflicting options. While one may be able to narrow down the list of metrics using, for example, fairness-specific metric decision trees [85, 109], alternative choices and, thereby, metrics are usually also plausible. As it has been shown that one cannot optimize for all possible notions of fairness at the same time, this means that one will eventually have to prioritize some metrics over others [8]. One complicating factor with this is that a malicious actor can abuse this ambiguity to pick a definition that produces more favorable scores post-hoc, a practice termed *fairness hacking* [87] (cf. Section 2.1.4).

The full list of potential fairness metrics is large, with the nuances and trade-offs between different definitions often quite subtle and hard to judge, even for experienced practitioners. We, therefore, chose to focus on the important trade-offs between competing concepts of fairness in ML, which can be traced back to different ways of conceptualizing errors (and, conversely, prediction performance). The two options we included for this decision are (1) a focus on sensitivity, corresponding to the fairness concept *separation* and (2) a focus on precision, corresponding to the fairness concept *sufficiency* [85].

Eval Fairness Grouping. When identities from subgroups are not excluded in analyses, they are often aggregated away instead. The most common form of aggregation is to aggregate multiple different groups (e.g., several racial subgroups) into a majority (the biggest subgroup) and minority group (all other subgroups). This is often done when there are substantial imbalances in group size or for convenience reasons such as enabling simpler analyses.

The same study mentioned earlier [116] found that of the 21 studies (using *COMPAS*) that did not exclude data from subgroups, all but one aggregated data from different groups together. Across aggregation and exclusion, 53 out of 59 studies reduced the protected attribute to just two groups.

While we believe that the normalization of such practices is harmful, we still want to represent them in this study, with the hope that participatory input may help us to better understand public opinion of such practices. We, therefore, include two options for this decision: (1) To aggregate the protected attributes into two groups (majority, minority) when calculating fairness metrics or (2) to calculate fairness metrics as the maximum difference between all different combinations of subgroups of the protected attribute[2].

Eval On Subset. Machine learning systems are often evaluated using data that may not represent the eventual target population. This can be due to practical constraints such as limited resources or because only a certain smaller subset of the target population is reachable. A system may have also been developed with a certain target population in mind, but then the scope widened during or after development.

While one would generally want to evaluate a system using a sample that resembles the population eventually affected by the tool, we included this decision to also represent these more practical trade-offs between the cost of data collection and representation in the data. We included the following options for this decision: (1) Using data from the most populous area, (2) using data from the area where the most people have public health insurance, (3) using data from the closest major city, (4) using data from as many people as possible, but excluding military veterans, (5) using data of only U.S. citizens and (6) using data from the overall population in the United States. Note that in our case study, data from the overall population is available, but we create subsets for the decision options to represent and model different data collection practices.

Eval Exclude Subgroups. When data from a subgroup is excluded, it is typically excluded during both training and evaluation of a system. This is problematic, as it will hide any resulting issues affecting the excluded groups during evaluation, whether they be related to fairness or the quality of predictions. We decided to include this decision not due to inherent trade-offs but to see whether participants would pick up on this potential issue and whether their input could help address it.

There were two options available for this decision: (1) To exclude the same subgroups as in the training data or (2) to include all subgroups for evaluation. This decision was only shown to participants if it made logical sense based on their response to the decision *Exclude Subgroups*. When participants chose to exclude certain subgroups from the training data earlier, they were therefore shown this decision later on, asking whether they also wished to exclude data from groups when evaluating the system.

3.2 Procedure

The study was created using jsPsych version 7.3.1 [36], with data collected in World-Wide-Lab version 0.4.1 [130]. Our goal with the

[2] This corresponds to the default behavior in the commonly used Python library fairlearn [12]

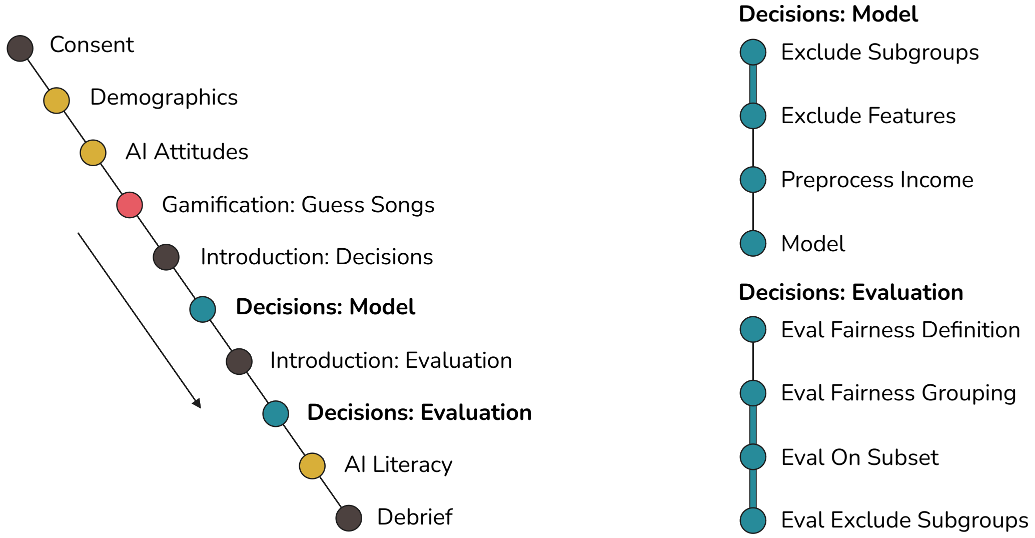

Figure 2: Diagram illustrating the different sections of the study and how they followed each other. The right-hand side of the graphic illustrates the different decisions which were presented in their respective sections. Decisions connected by a thicker line are considered to be within the same logical block during model design, and their order was randomized.

study was to recruit a diverse sample, as different peoples' identities and lived experiences can lend them unique forms of expertise [39] and shape their views on AI ethics [71, 72, 102, 122]. Since gamified citizen science studies have shown promise in recruiting diverse populations [62, 81], we opted for a similar approach in this study. We added a short game as an incentive and embedded the study on a citizen science website. We chose this particular website for its popularity and since its theme was unlikely to strongly bias recruitment in relation to the case study. As this website happened to be music-themed, the gamified section was about identifying whether or not a piece of music was generated by AI. To at least *partially* address the global north bias present in many fields of research (including AI ethics [106]), we chose to also make the study available for non-U.S. participants. Given the diversity of healthcare systems across the world, we believe that non-U.S. participants might indeed have valuable insights from living with different healthcare systems. Due to different notions of race across regions [70] we chose to slightly adapt demographics for U.S. versus non-U.S. participants (see below) and examine the data for differences in this regard later on.

Participants were first presented with a screen where the study was briefly explained, and they could then provide informed consent to participate. This was followed by a list of demographic items, in particular age, country of residence, primary and secondary language, race (U.S. only) and self-assessed membership of a minority (non-U.S. only). Afterwards, participants completed the ATTARI-12 [124], a 12-item questionnaire assessing their attitudes towards AI. Then, participants completed the recruitment game of the study, listening to a random collection of 6 AI-generated and 4 human-made 12-second music samples. After each song, participants were asked to guess whether a song was AI-generated, and they received feedback immediately afterwards. Next, participants were shown an introductory text illustrating the case study and why their input would be valuable. After the introduction, participants were shown the individual decisions, with decisions presented in the order they would be encountered during the design of an ML system. If there was no clear order between decisions, their order was randomized[3]. A second block of decisions related to the *evaluation* of an ML system was preceded by another brief introduction explaining basic concepts of fairness in ML. Each decision was presented with a brief introductory text explaining the decision, followed by a list of options. If there was no inherent order to options, their order was randomized. In addition to the actual decision options, each decision always presented the options of “I don't understand the description”, “I prefer not to answer” and to “Suggest an alternative option”. The evaluation decisions were followed by a short 16-item questionnaire assessing AI literacy [103]. For consistency, both attitudes towards AI [124] and AI literacy [103] were collected using 7-point response scales, labeled *strongly disagree* and *strongly agree* at their ends. Respondents were not required to select any option for the decision trials or AI attitudes / literacy response scales. At the end of the study, participants were shown a debrief with basic information about the study, their final score in the gamified section and a detailed list of the songs they listened to. Before the final screen, participants had the option to provide general feedback about the study and its design.

A high-level overview of the study procedure can be seen in Figure 2. The complete list of decisions and options presented to participants can be found in Appendix B. The study and its analyses were preregistered before any data were collected. The preregistration can be found at https://aspredicted.org/dgyp-bs3b.pdf.

[3] Due to an error in an earlier version of the experiment code, 10.67% of participants saw decisions in a fixed order.

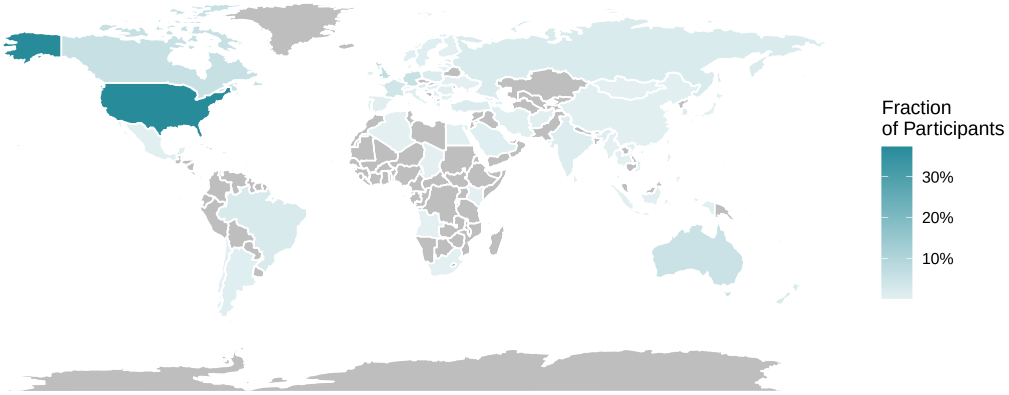

Figure 3: Participants were recruited from across the world with an over-representation of Western countries, especially the United States. Choropleth map of the world, shaded based on self-reported country of residence. A detailed breakdown of sample size per country can be found in Table 1.

3.3 Participants

Participants were mainly recruited through organic internet traffic to the music-focused citizen science website https://themusiclab.org. The study was also shared on a mailing list for auditory experiments and posted on social media platforms (Bluesky, Reddit). After a brief pilot, data for the study itself was collected over 20 days.

A total of 1403 sessions were recorded in the study, with 375 sessions completing the whole study[4]. As a new session is recorded every time someone navigates to the study or refreshes the webpage, this rate of completion (26.73%) is well within the expected range and slightly higher than the overall completion rate of other studies on the website during this time (22.10%).

We restricted the sample to sessions with data available for at least one of the decision trials. This left us with a final sample size of N = 534 individual sessions by n = 517 participants. As only 17 participants had multiple sessions with decision data, we decided to retain their data. Unless stated otherwise, the following analyses will use all the available data and will include every session with data available for a particular decision.

While there is a strong over-representation of both Western and especially English-speaking Western countries, there is also a significant number of non-Western countries present in the data. A graphical overview of participation rates by country can be seen in Figure 3, with detailed counts available in Table 1. Further information on the sample composition is available in Section A of the appendix.

The distribution of AI attitudes [124] ($M = 3.18$, $SD = 1.13$) and AI literacy [103] ($M = 3.20$, $SD = 1.22$) among participants displayed a high degree of variation, with the sample slightly leaning towards more positive AI attitudes and higher AI literacy (Figure 13).

4 Results

Detailed information on the software used for analyses is available in Section C. Code for the multiverse analysis simulation was adapted and extended from prior work [117] and implemented using an early version of the package multiversum [115]. The source code of simulations conducted in this study is available at https://github.com/reliable-ai/participatory-multiverse.

Below, we first address the validity of the survey and responses. Then we proceed to the core parts of the analysis: (1) The decision distributions including relevant differences between groups and (2) the resulting multiverse.

4.1 Quality of Decisions

Decisions were generally perceived as clear, with < 10% of participants checking that they did not understand a description across any decision (Figure 4). A sizable fraction of participants either did not check any answer option at all or indicated that they prefered not to answer a decision, an option we purposefully allowed them to do. This is not surprising, given that there were no monetary incentives and that the study was presented as a secondary objective to the gamified section to participants. While only a very small fraction of participants suggested alternative options during the piloting of the study, this number increased during data collection for the main study. In practice, the option to suggest an alternative was also used as a general feedback outlet by participants. In a small number of instances, participants also used this field to provide detailed and nuanced suggestions. Only $n = 83$ out of $N = 3,315$ responses were excluded for being unreasonably fast, at under two seconds, as the majority of people who did not bother to read descriptions would just not respond at all.

[4] A total of $n = 2$ sessions were excluded from analyses due to corruption in their data: In one case, this seems to have happened due to a particularly unreliable internet connection and in the other due to use of a non-standard in-browser translation feature which interfered with data collection.

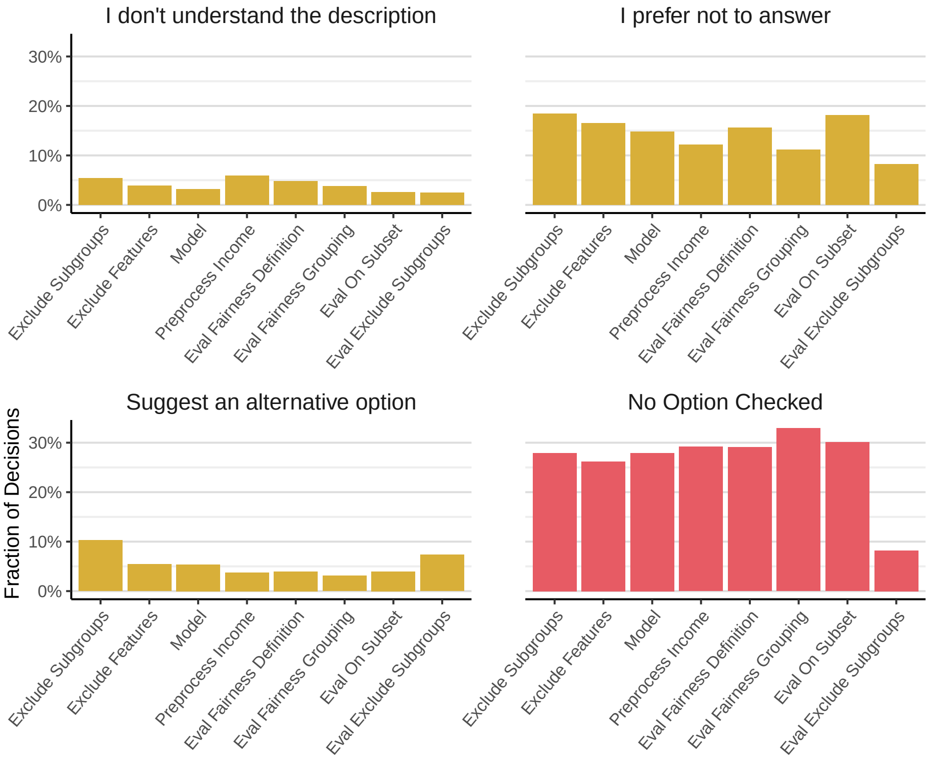

Figure 4: Decisions were generally well understood, and only a few additional options were suggested, but a large fraction of participants clicked through decisions without ticking any option or indicated that they prefer not to answer. Prevalence of different non-option answering trends across decisions in the study.

4.2 Distribution of Ratings

An overview of the overall distribution of ratings in the participatory multiverse can be seen in Figure 5. The figure shows an illustration of the different paths that participants took through the multiverse, weighted by how many participants chose a particular path and split into the multiverse of models (Figure 5A) and evaluations (Figure 5B). It becomes immediately visible that there are a few popular pathways and that if a participant does not respond to a decision or indicates that they prefer not to answer, they tend to do this for all decisions (red path "empty response"). The individual distributions of ratings for all countries with sufficient data are available as an interactive analysis at https://reliable-ai.github.io/participatory-multiverse/.

The degree of agreement between participants is generally high but varies by decision. Figure 6 illustrates this by showing the cumulative prevalence of different *combinations of options*. Naturally, decisions with a high number of different options generally have a lower prevalence per combination. Still, about half of all participants chose the same combination of options out of 511 theoretically possible combinations for *Exclude Subgroups*, indicating strong agreement across participants. The opposite is the case for the choice of *Eval Fairness Definition*: Here, participants had to chose one of two possible metrics (a combination was not possible), and choices are almost exactly equally distributed.

Figure 7 illustrates the different combinations of options observed in the data for the decisions *Exclude Subgroups* and *Eval On Subset*. It has to be noted here that for the decision *Exclude Subgroups*, the combination of choices is indeed what is used in the end, whereas for the decision *Eval On Subset*, each selected option is a separate valid strategy to be explored. Interestingly, there is sometimes little overlap between combinations of options that are of similar popularity. The results here also indicate that participatory results need to be taken with caution, as the second most popular option for the decision *Exclude Subgroups* was to use only data from people who identify as White.

4.2.1 Differences between Groups. We investigated whether participants' answers differed based on different characteristics and conditions. Due to the large number of possible combinations between participant characteristics and decisions, we chose to examine only a small subset of possible decisions of particular interest. We evaluated group differences based on individual votes, including partial data, but excluding responses where the respective grouping information was missing. We calculated comparisons for each option using Bonferroni correction to correct for multiple testing within

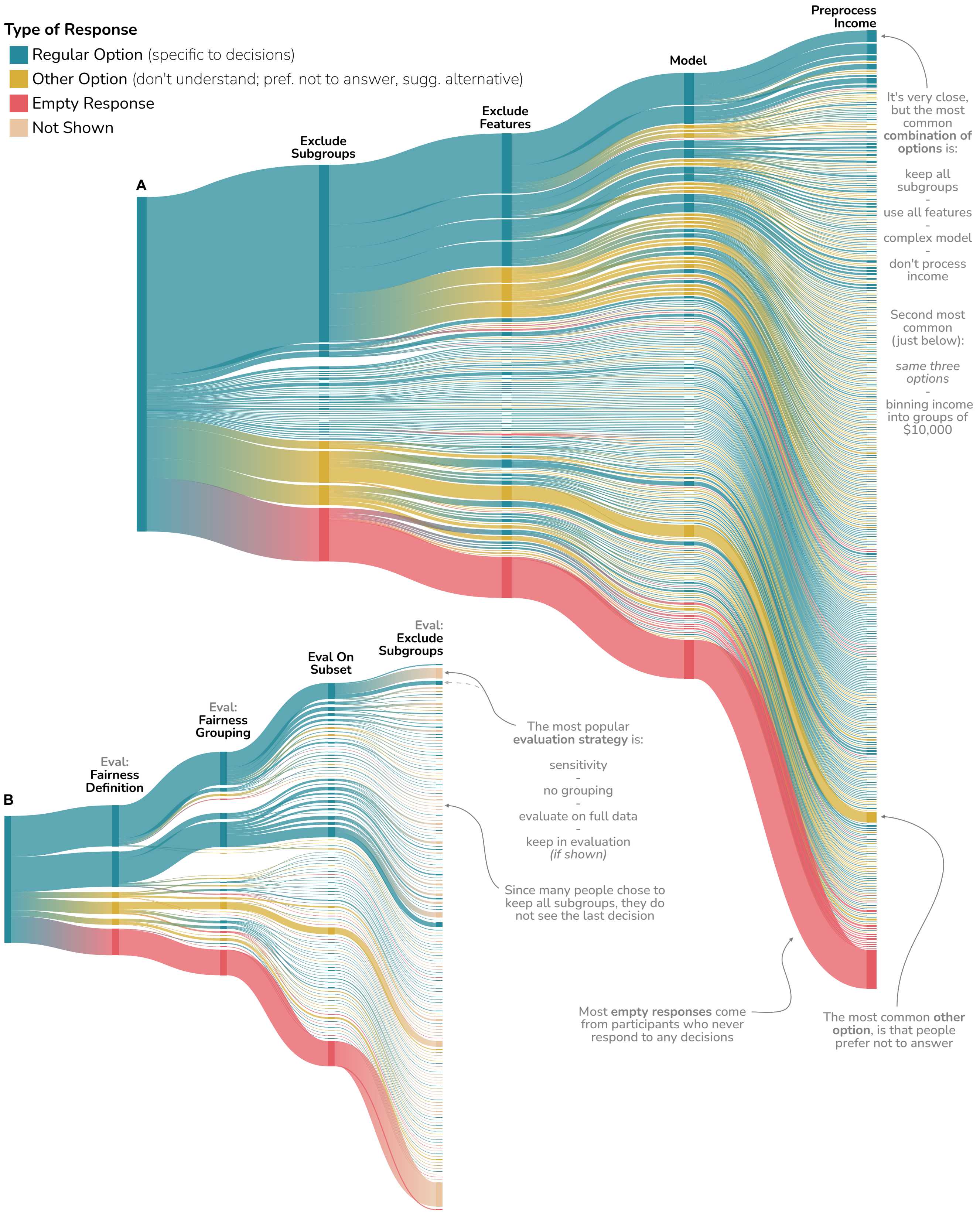

Figure 5: Specific paths in the participatory multiverse are significantly more popular than others, and if participants decide not to respond, they do this consistently. Weighted illustration of the multiverse of model design (A) and evaluation (B) decisions based on participants' votes. Each split corresponds to a decision taken by participants. Only data from participants with data available for all four decisions represented in each diagram were included.

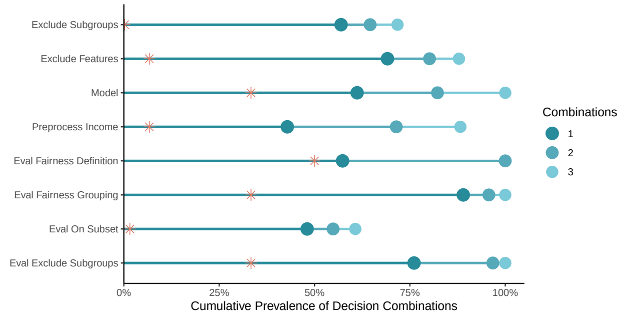

Figure 6: Agreement differs greatly across different decisions. Cumulative frequency of the three most common combinations of options across decisions. The minimum rate of agreement for each decision is highlighted with a star. The decision *Eval Fairness Definition* did not allow the selection of a combination of options.

each decision. For each comparison, frequencies of votes were compared using Fisher's exact frequency test [49] with a significance level $\alpha < 0.05$.

We examined three different comparisons for the decision *Exclude Subgroups*: First, whether participants were more likely to include a subgroup if they were also a member of it, second, whether responses differed based on the country of residency, and third, whether displaying percentages next to the different groups would have an effect on choices. Participants were indeed more likely to include subgroups if they were members of them, although the effect was small, especially in comparison to the overall tendency of participants to include many different subgroups ($p = 0.02$, $OR = 2.43$; Figure 8A). Responses also differed between data from the U.S. and other countries, with higher rates of inclusion in data from the United States (Figure 8B, Table 3). While the effect was only significant for the group "White", this could be due to an interaction of it being the biggest group and the previously described effect of higher rates of including one's own race. Whether or not percentages were shown next to the different groups had a negligible effect on participants' choices, as can be seen in Figure 15 (Table 4).

We examined whether participants who identify with a certain gender (Figure 16A, Table 5) or as being part of a minority (Figure 16B, Table 6) were more or less likely to exclude certain (corresponding) features from the model (*Exclude Features*). While we do see differences between the different groups here, results are hard to interpret due to significant imbalances between group sizes, and differences were not statistically significant.

In order to enable group comparisons between different levels of AI literacy and AI attitudes, we created three equally sized groups on each scale. The distribution of the three groups on either scale can be seen in Figure 14 and compared to the overall distributions in Figure 13.

The decision of which metric to prioritize (*Eval Fairness Metric*) is one of the most technical decisions encountered within the study. We therefore examined whether responses to it would be different based on self-reported AI literacy. Interestingly, we did not observe any significant differences on the regular response options. However, there were significant differences in the number of suggested alternatives and empty responses (Figure 9, Table 7).

We further examined whether there are differences in both explicit non-responses (checking "I prefer not to answer") or empty responses (checking none of the options) based on participants' self-reported AI attitudes. While most of these comparisons did not show significant differences, the observed frequencies show an interesting counter-play with both positive and negative AI attitudes generally showing a higher tendency to explicitly check "I prefer not to answer" (Figure 17, Table 8), whereas the middle group tends to not respond at all more often (Figure 10, Table 9). This is most likely related to overall response tendencies, with participants who are less engaged in the study opting for satisficing response strategies [75], such as responding closer towards the center of a scale and opting not to respond in the later sections of the study. A statistically significant difference in non-response between groups was found for the decisions *Exclude Features* and *Model* (Table 9).

4.3 Multiverse

Besides examining participants' ratings directly, we also conducted a multiverse analysis by traversing the complete multiverse of models, building and evaluating each of the models. In this section, we first present data from the complete multiverse of models before integrating it with participatory input and examining the intersection.

4.3.1 The Full Multiverse. We fit all 16,352 ML models in the multiverse and evaluated each using the 48 different evaluation strategies. This resulted in a total number of $N = 784,896$ different scores.

As our primary metrics of algorithmic fairness, we calculate the difference of either sensitivity (Eq. 1) or precision (Eq. 2) across groups of the protected attribute *race*. Across all combinations of any two racial groups (i, j), the maximum of the differences is recorded as the fairness score[5]. If there are only two groups due to aggregation (via *Eval Fairness Grouping*) or exclusion (via *Exclude Subgroups* and *Eval Exclude Subgroups*), only the difference between those two groups is used. Whether to use sensitivity or precision here corresponds to one of the decisions in the multiverse (*Eval Fairness Metric*). As a reference to these two metrics, we also use Equalized Odds Difference [1, 58], a commonly used fairness metric.

$$\text{Sensitivity} = \text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

$$\Delta\text{Sensitivity}_{\text{max}} = \max_{i,j} \left|\text{Sensitivity}_i - \text{Sensitivity}_j\right| \tag{1}$$

$$\text{Precision} = \text{Positive Predictive Value} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$\Delta\text{Precision}_{\text{max}} = \max_{i,j} \left|\text{Precision}_i - \text{Precision}_j\right| \tag{2}$$

[5] We note that this form of aggregation, which is commonly used in practice [12] and used here for descriptive purposes, can lead to an overestimation of performance differences between groups when the number of groups compared is large [83].

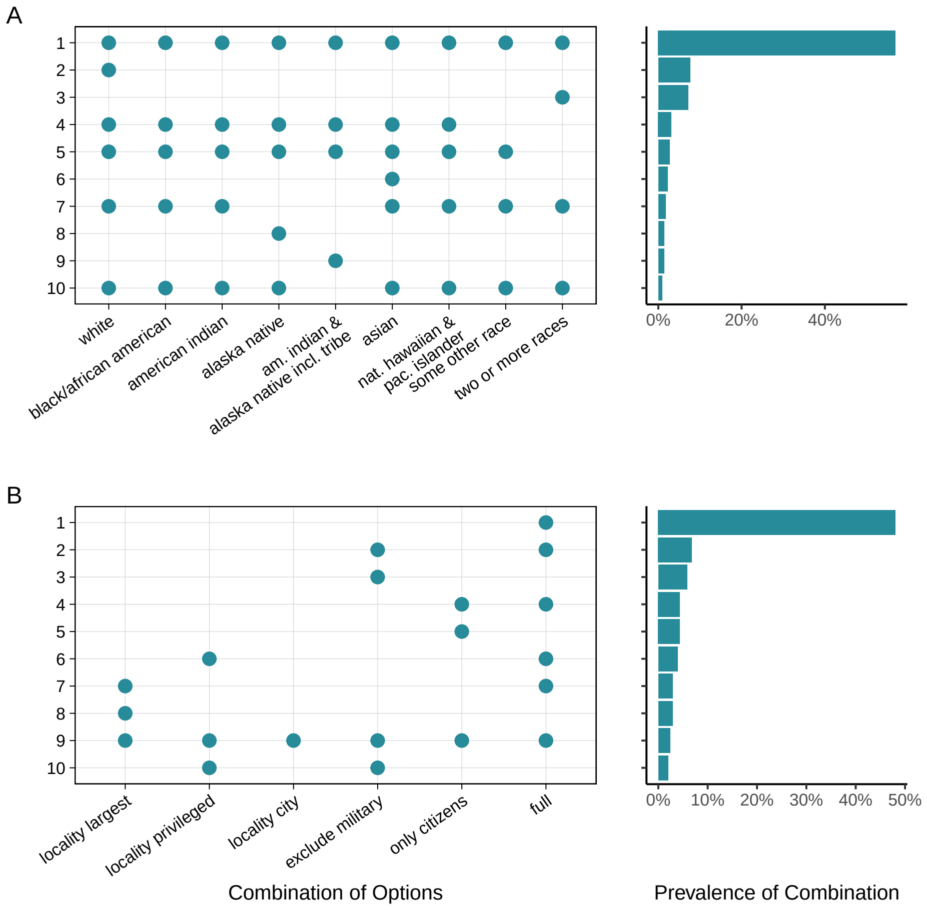

Figure 7: While participants often agree on a single combination of options, the degree of their agreement differs between decisions. The ten most common combinations of chosen options for the decisions *Exclude Subgroups* (A) and *Evaluate on Subset* (B), alongside their respective prevalence.

An overview of the complete multiverse of fairness scores for both metrics ($\Delta\text{Sensitivity}_{\text{max}}$ and $\Delta\text{Precision}_{\text{max}}$) can be seen in Figure 11. Variation within the multiverse is high, with values of both fairness metrics spanning their full possible range from 0 to 1, with a standard deviation of $SD = 0.347$ for $\Delta\text{Sensitivity}_{\text{max}}$ and $SD = 0.353$ for $\Delta\text{Precision}_{\text{max}}$. When examining the figure, the large spread of scores and a clustering towards the two ends of the scale become evident. A large degree of this extreme variation, however, can be attributed to different evaluation strategies.[6] Using a fixed evaluation strategy (see below), a more condensed distribution emerges (Figure 11, in red).

This brings up the question of which evaluation strategy one should use to evaluate the multiverse of models. In the present example, we suggest a strategy of more conservative choices, opting to not group the protected attribute, to use data from all subgroups during evaluation and to evaluate on the full subset of data. While we believe this to be a reasonable choice, it is by no means commonly used in the literature (as discussed in earlier sections, see also Simson et al. [116]). Luckily, this issue can be addressed well using the participatory data: Exactly this combination of evaluation choices is also the most popular in the present data. Therefore, we use this evaluation strategy to fix evaluations for the following analyses.

As there is no clear reason to favor one of the two definitions of fairness metrics over the other and since the decision did not impact scores in a very strong matter, we chose to not fix this decision, but rather continue to examine the two separately going forward.

4.3.2 The Participatory Multiverse. Participants' votes can be combined with data from the multiverse to create a participatory multiverse[7]. Participatory data can be combined with the theoretical full

[6] While some differences in fairness scores across, e.g., different data subsets may be expected, note that the comparisons made by the metrics (performance difference between racial groups) remain the same. The large spread of scores visible here highlights the susceptibility of fairness results of models trained for the same task to changes in the evaluation protocol, opening up opportunities for fairness hacking [15, 87] when evaluation strategies remain unchecked.

[7] As participatory data can contain missing data or non-responses, the combination is actually not a straightforward join. Rather, we assigned equal weights to all participants and spread these weights across all endpoints in the multiverse that match

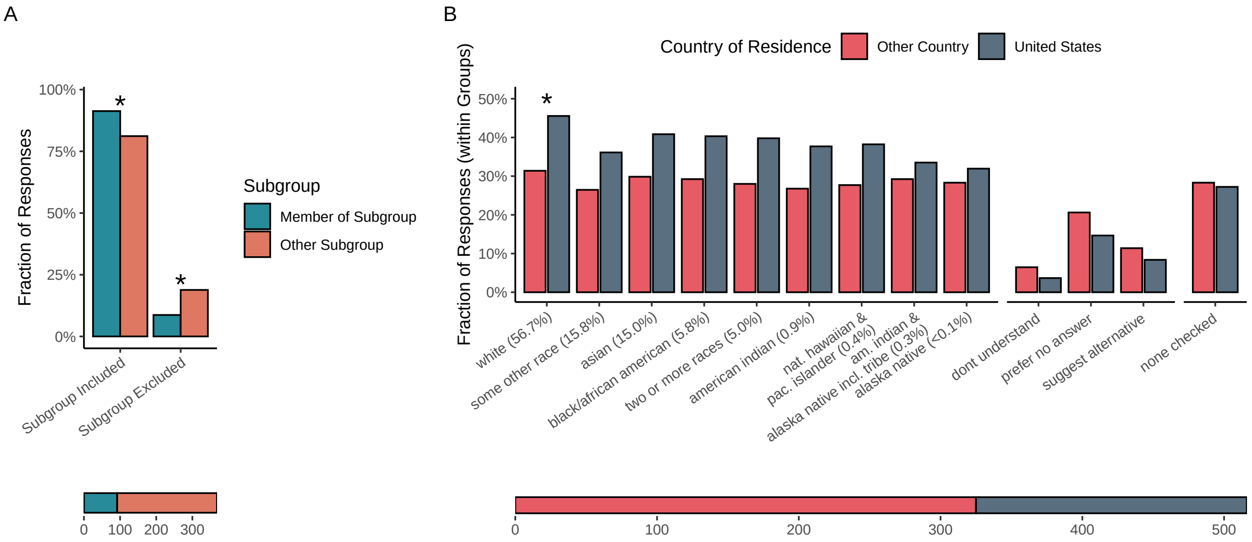

Figure 8: Participants were more likely to include a group that they are a member of (A) and more likely to include subgroups when from the United States (B). Inclusion of subgroups split by membership of participant (A) and country of residence (B) for the decision *Exclude Subgroups*. Bars below plots indicate the raw group distribution and number of votes.

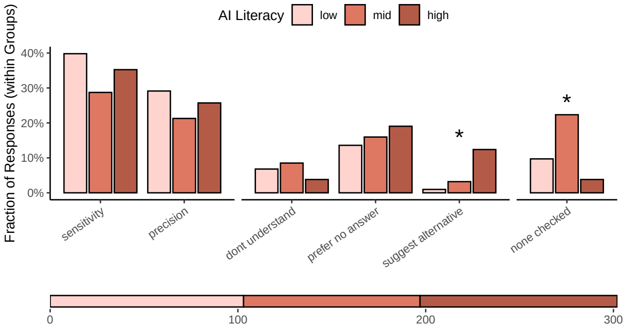

Figure 9: Participants with higher AI literacy opted to suggest alternatives for the fairness metric more often. Response to the decision *Eval Fairness Metric* split by self-reported AI literacy. Bar below plot indicates the raw group distribution and number of votes.

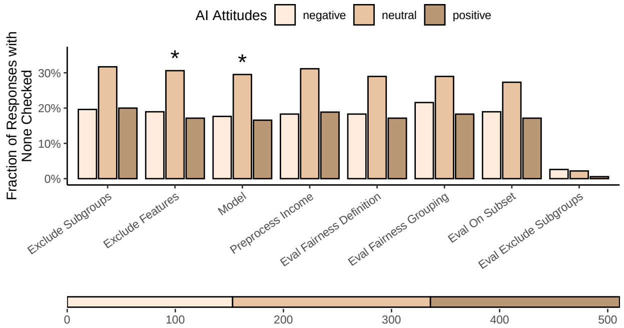

Figure 10: Participants with neutral AI attitudes showed a higher rate of non-response. Fraction of participants choosing *not to check any option* across decisions split by self-reported AI attitudes. Bars below plot indicate the raw group distribution and number of votes.

multiverse (Figure 1) to put variation in the participatory data into context. The resulting data can also be used to prune the full multiverse before its computation. This makes it possible to only examine and compute the most popular branches in the multiverse, greatly reducing computational costs. We discuss several approaches for this in Section 5.2.

Participatory data can also be combined with results from a multiverse analysis. This makes it possible to put the participatory multiverse of models and scores into context using actual metrics. This is exactly what we did here: We combined participants' votes with data from the multiverse analysis to investigate how a more narrow multiverse, weighted by participants' choices, compares to the one created by the complete multiverse analysis.

As can be seen in Figure 12, there are a significant number of models with equal or near-equal performance but significant variation in their fairness scores, illustrating the Rashomon Effect. However, besides areas of equal performance with "free" fairness gains, there also exists a Pareto front of models where any increase in either

their decisions. A very precise combination of responses will, therefore, assign more weight to its resulting endpoints than a very broad one. Multiple other algorithms for combination are conceivable.

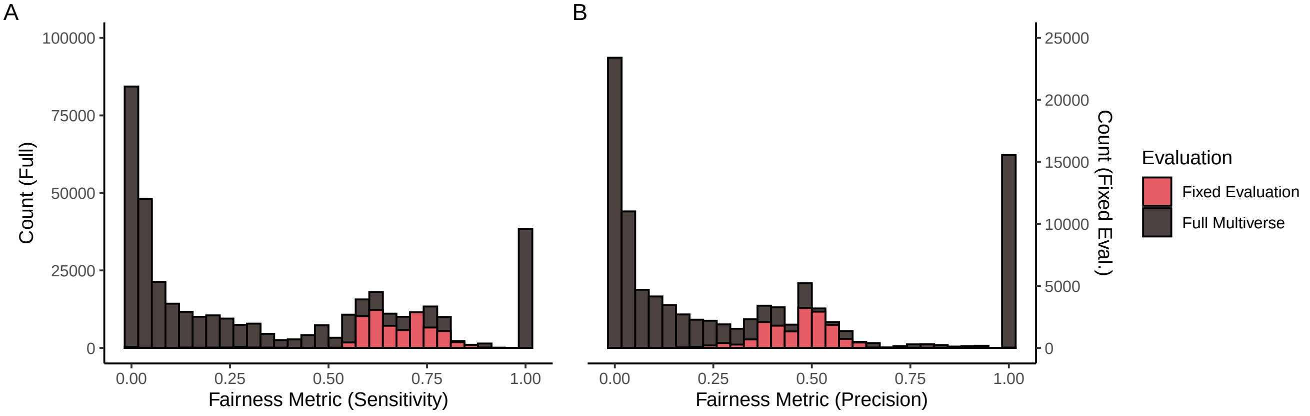

Figure 11: Examining the complete multiverse is unrealistic; rather one will want to use a fixed evaluation strategy to meaningfully compare different models. Histogram of evaluation scores across the complete multiverse (left axis) and with a fixed evaluation strategy (right axis). The multiverse with a fixed evaluation is scaled by a factor of four to be well visible. Fairness metric corresponds to $\Delta \text{Sensitivity}_{\text{max}}$ (A) and $\Delta \text{Precision}_{\text{max}}$ (B), both ranging from 0 to 1 with lower scores being preferable.

performance or fairness would come at a cost to the other. Interestingly, the distribution of models chosen by participants is clustered closer to these Pareto fronts. In particular, the most popular model is situated very closely to the technically "optimal" combination, indicating a competitive model. Repeating the analysis using Equalized Odds Difference yields similar results (Figure 18). Still, given the various trade-offs and nuanced implications of the different decision points, the "performance" of the most popular models cannot be fully reflected by the present metrics. Participants' preferences introduce a new dimension in its own right.

5 Discussion

In this work, we explore a novel workflow that uses participatory input to restrict the decision space during the design of an ML system. Using a case study of predicting public health care coverage on U.S. data, we illustrate this workflow, prototype a selection of decisions, collect citizen science data on which options people deem appropriate and evaluate data in the light of the resulting machine learning multiverse.

5.1 Summary of Findings

Our results indicate that the participatory multiverse approach shows great promise to be applicable in real-world scenarios, although care needs to be taken during design and evaluation. We were able to successfully collect diverse input from across the world, indicating a high willingness of participants to engage with the topic. We collected meaningful input on (complex) modeling and evaluation decisions, exhibiting differing levels of agreement across decisions. It should also be noted that while a large fraction of participants opted not to respond to the decisions, this behavior was to be expected as we solicited input for a hypothetical ML task on a citizen science platform rather than engaging affected individuals of an actual ADM system.

The most popular combinations of people's choices were very reasonable options across the different decisions. Especially in the context of evaluation strategies, the most popular evaluation strategy excluded "lazy" practices, such as evaluating fairness using a simplified setup of majority and minority data. This is a positive result, as these practices commonly occur in the literature, and it illustrates the opportunity for participatory input to combat them.

There were a few significant differences in participants' responses based on their membership to certain groups. However, many more comparisons could be made than the ones we presented here. We therefore argue for the importance of collecting data from a wide and diverse sample of people to represent all identities which will be impacted by a potential system, as anticipating the degree of diversity in views ahead of time will be difficult.

When comparing the complete multiverse of plausible ML models with one weighted by participatory input, we see that the participatory multiverse leans towards favorable models in both performance and fairness. Especially the most popular model, which would follow from a democratic vote, was situated closely to the Pareto front across different definitions of algorithmic fairness. This also aligns with findings from a study of participatory feature weighting, where participants' decisions tended to improve fairness evaluation [94].

As participatory input is a democratic process, one also has to anticipate and embrace disagreement. In the present data, we saw both high agreement and disagreement for different decisions. This should inform how different decisions are handled. For example, since agreement on the definition of a particular fairness metric was quite low, we chose to explore both options of the decision in more detail than other decisions which displayed significantly higher rates of agreement.

5.2 The Participatory Multiverse Workflow

We present the following steps as an outline for our workflow of collecting participatory input when designing an ML system.

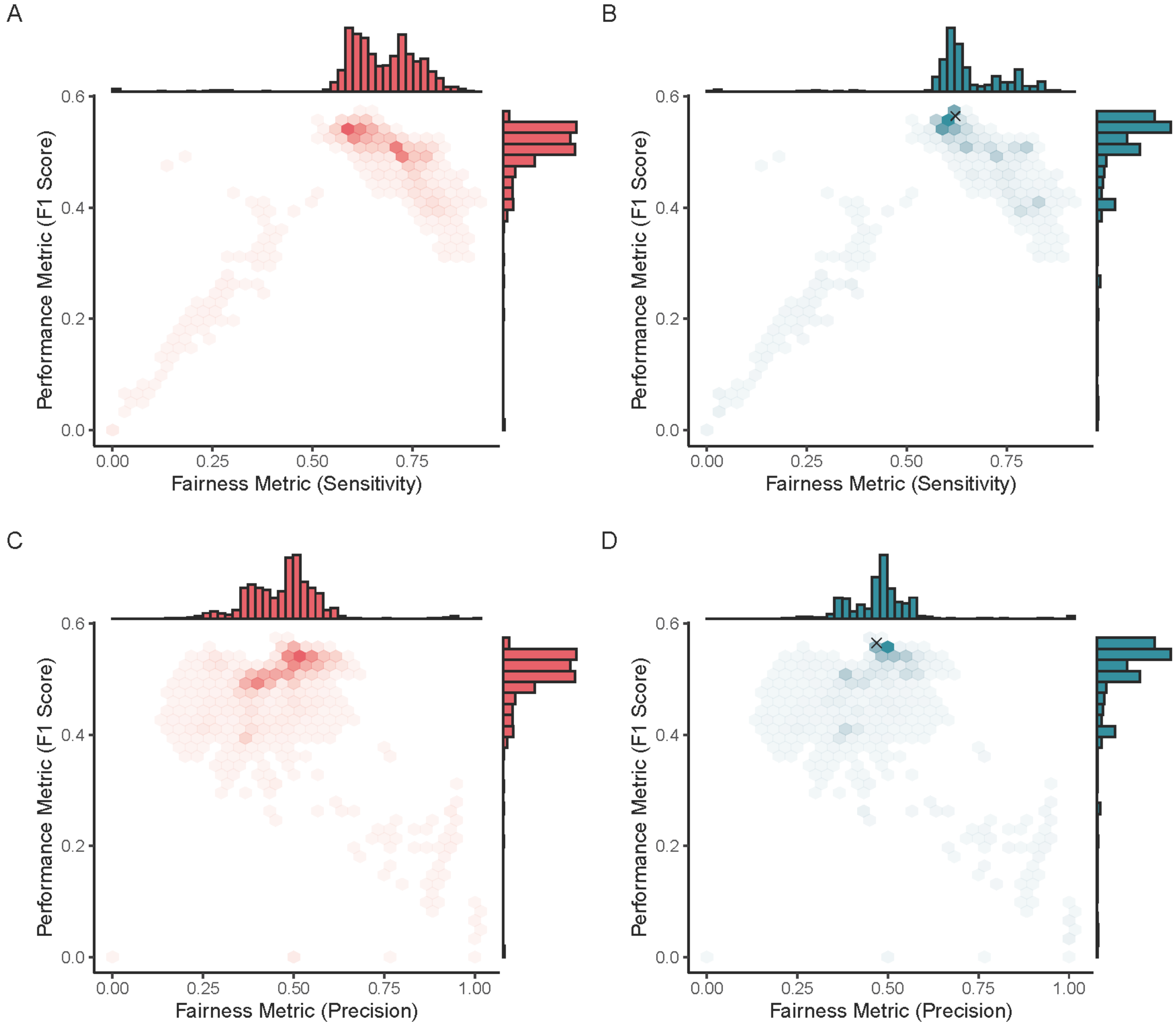

Figure 12: Models from the multiverse weighted by participants' votes are close to the Pareto front. Comparison between a complete multiverse of models (A, C) and one based on participants' votes (B, D) on performance (F1 Score, higher is better) and fairness metric (lower is better). Both multiverses are evaluated using a fixed strategy, split by definition of fairness metric (ΔSensitivity$_{max}$: A, B; ΔPrecision$_{max}$: C, D). Darker areas correspond to a higher clustering of models. Crosses indicate the most popular models among participants.

We also add recommendations for implementing the workflow in practice.

(1) **Decide on your target population and platform.** It is important to be clear on who the target population of both, the participatory input, but also the ML / ADM system is. While this would ideally be the same population in both cases, it may not always be feasible in real-world scenarios. Great care also needs to be taken in the actual sourcing of a chosen target sample. As demonstrated in this work, participants' characteristics may influence their responses to decisions, and it is therefore vital to include all relevant groups in the discussion. For both aspects, measurement and representation, participatory approaches can draw on the rich insights of survey science [11, 56]. Moreover, the platform in which participatory input is collected should fit and inform the intended scope and may range from a large-scale online platform to a town hall for in-depth discourse with stakeholders.

(2) **Source a list of decisions that affect your ML pipeline.** We recommend focusing on the complete pipeline here, including often-neglected steps such as the sourcing and (pre-)processing of data. It can be helpful to involve multiple people into this process to spot potentially overlooked

decision points. Co-designed fairness checklists [84] and documentation aids like model cards [92] can also be a starting point for identifying decisions.

(3) **Identify the different options for each decision.** One should keep an open mind when collecting the list of options, but also only include actually feasible options in each case.

(4) **Prepare the final list of decisions that you will gather participatory input on.** Depending on the application scenario, there may be practical limitations on which kinds of decisions you can present to participants and how many. In this case, it may be useful to prioritize value-related decisions, where stakeholders' attitudes are more likely to differ from those of ML engineers [71, 72, 134]. However, to the extent possible, we advocate for the inclusion of decisions that may seem to have an obvious correct choice and decisions that may be challenging for non-experts (cf. Section 3.1.1), as these may unearth implicit biases. In turn, this may foster reflection and counteract lazy practices.

(5) **Develop the actual wording for each decision and its options.** This should include introductory statements for each decision, outlining the trade-offs between different options. Care should be taken not to bias future participants towards any particular option in a decision [18]. We further recommend allowing participants to help highlight potential issues in the text surrounding a decision, e.g., by including other decision options as presented in this study (e.g., I don't understand the description; I prefer not to answer; suggest an alternative option [80]). The decision-design step should consist of multiple iterations where feedback is gathered from stakeholders in between each iteration, potentially including empirical data from a multiverse analysis. We recommend including additional, information-only sections to explain the context in which the system is used as well as more complicated foundations which may be required for certain decisions. We also recommend being as explicit and practical as possible when explaining decisions and options, opting for applied examples within the context of the ML system over more general statements.

(6) **Launch an initial small-scale pilot** to verify that participants understand the decisions and verify data is collected as expected. Participants should have an option here to provide suggestions for each decision as well as overall study feedback.

(7) **Gather participatory input.** Participants should be encouraged to freely choose which decisions to provide input on and which not. In online settings, we recommend only light input validation to allow participants who are not interested to skip parts of the study. The best course of action will depend on the particular context of participation (see Section 5.3).

(8) **Make decisions based on participatory input.** Participatory input will need to be turned into actual decisions to create the final ML model. There are many ways to achieve this, but committing on an approach ahead of time and explicitly communicating it to participants can serve to empower them. Care should be taken when choosing a strategy, as certain strategies, such as majority votes, risk reinforcing existing power imbalances [45]. Alternative approaches [35, 53] explicitly harnessing disagreement and participant characteristics can be worth exploring as well as approaches to quantify consensus, using it for thresholding and mapping out the opinion space [67, 119]. We suggest considering prohibitory approaches where participatory input has the power to rule out certain options and developers are left with options to choose in a limited space that is deemed acceptable by the public.

We highlight that implementing even parts of this workflow can be beneficial, as it allows for critical reflection of the ML pipeline and may be used to inform a follow-up multiverse analysis across decisions where participatory input may not have been feasible. Additional benefits for developers include surfacing new options they were not yet aware of while potentially saving costs by reducing the size of the multiverse left to explore.

5.3 Practical Implications

In this paper, we presented an illustrative application of the workflow focusing on the design of a real ML model, but for a "hypothetical" ADM system without real-world deployment. Below, we highlight key considerations for applying the workflow in ADM practice.

Special care should be taken when determining the *target population* for participatory input in light of the specific use case. When it is unclear which characteristics, expertise and lived experience are truly relevant, we recommend getting input from diverse sets of populations. Here, one should be open to recognize the different forms of expertise [39] different communities might possess, which can be hard to assess from the outside. In any case, all affected populations should always be considered for their input and thereby given a voice, especially so if they are part of marginalized populations, which can be more sensitive to biases [72].

The *mode, degree and setup of participation* will depend on the target population(s) as well as the decisions one plans to gather input on. While we specifically chose online citizen science to allow for broad participation and as a challenging form of participatory input to test the limits of the approach, we encourage considering the full breadth of participatory methods when applying the workflow in ADM practice. This also includes adequate forms of compensation to participants for the time and work they put into giving their input. This is especially important when working with potentially vulnerable populations.

Last, it is important to *avoid exploitative and extractive forms of participation* such as "participation washing" and manufactured consent [118]. At the same time, one should be aware that participatory design will not be able to solve all issues an AI system might face [13], including issues related to the data a system is built upon.

5.4 The Participatory Multiverse as a Participatory ML Method

With the Participatory Multiverse workflow, we contribute a novel method for gathering participatory input in the model design and evaluation steps of the ML pipeline. Thus, the workflow serves as a building block for a part of the pipeline that currently receives less participatory input than earlier steps such as needs assessment [31]. It puts particular emphasis on fairness by reaching out to a wide audience of stakeholders and capturing their attitudes on questions that also address ethical and value-based decisions. Complementing related qualitative participatory approaches like Value-Sensitive Algorithm Design [134] and UI tools for fairness solicitation [26], we gather structured input that can inform decisions in a more quantitative manner. In practice, these methods can (and should) also be combined, e.g., identifying crucial decisions and choices with qualitative methods, inviting a smaller set of people but gathering in-depth insights, before following up with a participatory multiverse survey.

Mapping the Participatory Multiverse workflow to Delgado et al.'s [37] framework for evaluating participation, the *participation goal* of the workflow is to *include* participants to better align the model and its evaluation with stakeholders' preferences and values. The *scope of the participation* is *collaboration* with participants to query their preferences for *system design* aspects such as model features and *consultation* and *inclusion* for stakeholders' feedback and expertise, encouraging the suggestion of other options and providing space for additional comments. Finally, the *form of participation*, at the minimum, includes *consultation* via questionnaires, which makes it easy to apply the results to the multiverse analysis. In our case study, we ran a single iteration of the workflow, situated within an hypothetical ADM example. However, through repeated queries at different points in time, the approach could easily be extended to empower participants as *collaborators* involved in the ongoing decision-making processes. For example, participant input on design or evaluation decisions could be collected to adapt decision options for a second iteration, making these suggestions available to a broader audience. Repeated participatory input on model and evaluation decisions, even after initial deployment, could help adjust ML models in consideration of observed outcomes and potential biases. The participatory multiverse could also be used as a starting point for collaborating with specific groups of participants, e.g. those who voiced opinions that diverge from practices ML engineers would typically apply. For future work, we suggest further expansions towards higher-level participation (*collaborate* and *own*) that reduce power imbalances [13].

5.5 Limitations

There are several important limitations which apply to the results of the case study as well as the workflow itself.

It is important to note that, while geographically diverse, the sample used within this study is a convenience sample recruited from the internet. As such, it is not an accurate representation of the general public, and we do not claim that reported results represent effects, attitudes or convictions of the general public. Rather, results should be interpreted as a case study, illustrating that this particular workflow can work and produced highly useful results in this particular context. Nonetheless, the results emphasize that even a small case study can already provide a benefit by revealing controversial decisions.

Further, the decisions explored in this study represent a small subset of potential decisions encountered during the design of a real-world ML system. While they may serve as an inspiration for applications, they do not represent a holistic overview of decisions and are specific to this particular context. We also purposefully included potentially harmful decisions (related to lazy practices) to understand how these are addressed by participants. Any real-world usage of this workflow should most likely not include these decisions. Applying the workflow will require a careful evaluation of which decisions may be relevant in a particular context.

While able to provide useful information, the participatory workflow is by no means a replacement for expert input and one should not rely solely on participatory data to design an ML system. When implementing the workflow in practice, special care has to be taken when selecting which decisions to present to people, as despite introductory statements, certain concepts, especially in ML, may be too hard to convey within the context of a short survey or study. More elaborate settings, such as workshops with educational sections (e.g. [52]), may be used to address this limitation.

Due to practical limitations the present case study relies on informative participation, mostly *consulting* and *including* participants, rather than granting *ownership* [37] and *control* [31]. While practical limitations may restrict aspects of open online participatory design, one can borrow ideas from Delgado et al. [37] and Corbett et al. [31] to improve the mode of participation, such as including stakeholders in the design of the participatory system itself and enabling them to take part in the formulation of goals.

As monetary incentives may be unrealistic in many real-world scenarios of participatory input, where a system may be created by a government or NGO entity, this puts practical limitations on the scope, number and complexity of decisions one is able to implement. While we have demonstrated here that collection of participatory inputs without monetary incentives is quite feasible, it did inform our choice of decisions as well as their wording and framing. The lack of monetary incentives also meant that we were unable to include more detailed checks to confirm participants' understanding of the decisions asked to them. We did allow participants to explicitly check that they did not understand a decision, however, and we observed that popular choices largely reflected good decisions. Future research focusing solely on this issue will be required to better understand the degree to which participants may or may not be able to answer more technical ML decisions and how best to describe these.

Introductions and explanations for decisions also carry the risk of priming or influencing respondents to choose a particular option. While we had multiple rounds of iterative development in this study to minimize such influences, it is impossible to rule them out completely. In particular, the surprisingly high prevalence of choosing only "White alone" in the decision *Exclude Subgroups* could be influenced by the introductory text to the decision. Participatory input from the workflow should, therefore, always be evaluated with an eye on the specific wording of each decision. The framing of the study as secondary to the gamified section could have also influenced participants' responses. Although such an effect cannot be entirely ruled out, it is unlikely to have significantly impacted the results of this study. Participants who perceived the study as unimportant due to its framing had the option to refrain from responding. Consequently, it is possible that some of the non-responses can be attributed to this framing.

It is important to note that the workflow presented here represents just a small part of the bigger picture of creating and implementing an ML system. While participatory input can help reduce problematic practices, as demonstrated, it is by no means a replacement for critical reflection by the practitioners implementing the system. Care has to be taken to avoid a "tyranny of the majority". Rather, participatory input and careful reflection should be used jointly, each serving to constrain the garden of forking paths. This is especially the case when an ML system is implemented as part of an ADM pipeline.

6 Conclusion and Outlook

This work presents the *participatory multiverse* as a new workflow of incorporating participatory input into ML design and applies the workflow to a case study of predicting public health care coverage in the United States. Results from this case study demonstrate how participatory input can improve and inform the design of ML systems in an effective manner. In particular, results highlight how participatory input can be used to restrict both the multiverse of different models as well as different evaluation strategies, combating widespread lazy practices and aligning nuanced decisions and trade-offs with public preferences.

While we successfully apply the workflow in this work, future, theory-driven work will be necessary to better understand the interaction of participants' characteristics and their responses. The workflow is also currently limited in which decisions are suitable for participatory input by the complexity of certain concepts. Future work may help address this issue by focusing on particular concepts (e.g., metrics of algorithmic fairness) and how participatory input can be sourced on these.

We hope that adopting the participatory multiverse alongside standard ML workflows will lead to better overall systems, especially so in ADM, and give unheard voices a chance to be heard.

Acknowledgments

We extend our heartfelt thanks to all the participants who generously took part in our study. Their contributions were essential to the success of this research, and we deeply appreciate the time and effort they contributed to this project.

We would also like to express our gratitude to Malte Schierholz for his input and feedback during the writing process and Frauke Kreuter, Frederic Gerdon, Marcus Novotny and Clara Strasser Ceballos for their input and feedback during the review process of this work.

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research and the Munich Center for Machine Learning.

Large language models (LLMs) were used to generate of a small number of helper functions during the technical implementation of the study and analyses. Code completions from GitHub Copilot were used when conducting the multiverse analysis. LLM-generated code presents less than 0.1% of the total code written for this project. LLMs were used to generate first drafts of figure descriptions for accessibility purposes, with all descriptions being reviewed, edited and often completely rewritten by the authors.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. arXiv:1803.02453 [cs]

[2] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*. PMLR, 161–170.

[3] R. Alba, R. G. Rumbaut, and K. Marotz. 2005. A Distorted Nation: Perceptions of Racial/Ethnic Group Sizes and Attitudes Toward Immigrants and Other Minorities. *Social Forces* 84, 2 (Dec. 2005), 901–919. https://doi.org/10.1353/sof.2006.0002

[4] J.J. Allaire, Charles Teague, Carlos Scheidegger, Yihui Xie, and Christophe Dervieux. 2022. Quarto. https://doi.org/10.5281/zenodo.5960048

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (May 2016), 254–264.

[6] Daniel Atherton. 2023. Incident Number 608. *AI Incident Database* (2023). https://incidentdatabase.ai/cite/608

[7] Kirk Bansak, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, and Jeremy Weinstein. 2018. Improving refugee integration through data-driven algorithmic assignment. *Science* 359, 6373 (Jan. 2018), 325–329. https://doi.org/10.1126/science.aao4408

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. Classification - No Fairness through Unawareness. In *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press, Cambridge, Massachusetts.

[9] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[10] Samuel J. Bell, Onno P. Kampman, Jesse Dodge, and Neil D. Lawrence. 2022. Modeling the Machine Learning Multiverse. https://doi.org/10.48550/arXiv.2206.05985 arXiv:2206.05985 [cs, stat]

[11] Paul P Biemer. 2010. Total survey error: Design, implementation, and evaluation. *Public opinion quarterly* 74, 5 (2010), 817–848.

[12] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[13] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Arlington VA USA, 1–8. https://doi.org/10.1145/3551624.3555290

[14] Bernd Bischl, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, Theresa Ullmann, Marc Becker, Anne-Laure Boulesteix, Difan Deng, and Marius Lindauer. 2023. Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges. *WIREs Data Mining and Knowledge Discovery* 13, 2 (March 2023). https://doi.org/10.1002/widm.1484

[15] Emily Black, Talia Gillis, and Zara Yasmine Hall. 2024. D-Hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 602–615. https://doi.org/10.1145/3630106.3658928

[16] Emily Black, Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. 2024. The Legal Duty to Search for Less Discriminatory Algorithms. arXiv:2406.06817 [cs]

[17] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 850–863. https://doi.org/10.1145/3531146.3533149

[18] Kathrin Bogner and Uta Landrock. 2016. Response biases in standardised surveys. GESIS survey guidelines.

[19] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec. 2011), 2301–2309. https://doi.org/10.1109/TVCG.2011.185

[20] Leo Breiman. 2001. Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statistical science* 16, 3 (2001), 199–231. https://doi.org/10.1214/ss/1009213726

[21] Nate Breznau, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K. Andersen, Daniel Auer, Flavio Azevedo, Oke Bahnsen, Dave Balzer, Gerrit Bauer, Paul C. Bauer, Markus Baumann, Sharon Baute, Verena Benoit, Julian Bernauer, Carl Berning, Anna Berthold, Felix S. Bethke, Thomas Biegert, Katharina Blinzler, Johannes N. Blumenberg, Licia Bobzien, Andrea Bohman, Thijs Bol, Amie Bostic, Zuzanna Brzozowska, Katharina Burgdorf, Kaspar Burger, Kathrin B. Busch, Juan Carlos-Castillo, Nathan Chan, Pablo Christmann, Roxanne Connelly, Christian S. Czymara, Elena Damian, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea Forster, Chris Gaasendam, Konstantin Gavras, Vernon Gayle, Theresa Gessler, Timo Gnambs, Amélie Godefroidt,

Max Grömping, Martin Groß, Stefan Gruber, Tobias Gummer, Andreas Hadjar, Jan Paul Heisig, Sebastian Hellmeier, Stefanie Heyne, Magdalena Hirsch, Mikael Hjerm, Oshrat Hochman, Andreas Hövermann, Sophia Hunger, Christian Hunkler, Nora Huth, Zsófia S. Ignácz, Laura Jacobs, Jannes Jacobsen, Bastian Jaeger, Sebastian Jungkunz, Nils Jungmann, Mathias Kauff, Manuel Kleinert, Julia Klinger, Jan-Philipp Kolb, Marta Kołczyńska, John Kuk, Katharina Kunißen, Dafina Kurti Sinatra, Alexander Langenkamp, Philipp M. Lersch, Lea-Maria Löbel, Philipp Lutscher, Matthias Mader, Joan E. Madia, Natalia Malancu, Luis Maldonado, Helge Marahrens, Nicole Martin, Paul Martinez, Jochen Mayerl, Oscar J. Mayorga, Patricia McManus, Kyle McWagner, Cecil Meeusen, Daniel Meierrieks, Jonathan Mellon, Friedolin Merhout, Samuel Merk, Daniel Meyer, Leticia Micheli, Jonathan Mijs, Cristóbal Moya, Marcel Neunhoeffer, Daniel Nüst, Olav Nygård, Fabian Ochsenfeld, Gunnar Otte, Anna O. Pechenkina, Christopher Prosser, Louis Raes, Kevin Ralston, Miguel R. Ramos, Arne Roets, Jonathan Rogers, Guido Ropers, Robin Samuel, Gregor Sand, Ariela Schachter, Merlin Schaeffer, David Schieferdecker, Elmar Schlueter, Regine Schmidt, Katja M. Schmidt, Alexander Schmidt-Catran, Claudia Schmiedeberg, Jürgen Schneider, Martijn Schoonvelde, Julia Schulte-Cloos, Sandy Schumann, Reinhard Schunck, Jürgen Schupp, Julian Seuring, Henning Silber, Willem Sleegers, Nico Sonntag, Alexander Staudt, Nadia Steiber, Nils Steiner, Sebastian Sternberg, Dieter Stiers, Dragana Stojmenovska, Nora Storz, Erich Striessnig, Anne-Kathrin Stroppe, Janna Teltemann, Andrey Tibajev, Brian Tung, Giacomo Vagni, Jasper Van Assche, Meta van der Linden, Jolanda van der Noll, Arno Van Hootegem, Stefan Vogtenhuber, Bogdan Voicu, Fieke Wagemans, Nadja Wehl, Hannah Werner, Brenton M. Wiernik, Fabian Winter, Christof Wolf, Yuki Yamada, Nan Zhang, Conrad Ziller, Stefan Zins, and Tomasz Żółtak. 2022. Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty. *Proceedings of the National Academy of Sciences* 119, 44 (Nov. 2022), e2203150119. https://doi.org/10.1073/pnas.2203150119

[22] 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*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300271

[23] US Census Bureau. 2021. Understanding and using the American Community Survey public use microdata sample files: What data users need to know.

[24] Simon Caton, Saiteja Malisetty, and Christian Haas. 2022. Impact of imputation strategies on fairness in machine learning. *Journal of Artificial Intelligence Research* 74 (2022), 1011–1035. https://doi.org/10.1613/jair.1.13197

[25] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2018. An Interpretable Model with Globally Consistent Explanations for Credit Risk. https://doi.org/10.48550/arXiv.1811.12615 arXiv:1811.12615 [cs, stat]

[26] 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*. ACM, Yokohama Japan, 1–17. https://doi.org/10.1145/3411764.3445308

[27] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[28] OPEN SCIENCE COLLABORATION. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (08 2015), aac4716. https://doi.org/10.1126/science.aac4716 Publisher: American Association for the Advancement of Science.

[29] A. Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. 2024. Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 20 (March 2024), 22004–22012. https://doi.org/10.1609/aaai.v38i20.30203

[30] Ned Cooper and Alexandra Zafiroglu. 2024. From Fitting Participation to Forging Relationships: The Art of Participatory ML. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–9. https://doi.org/10.1145/3613904.3642775

[31] Eric Corbett, Emily Denton, and Sheena Erete. 2023. Power and Public Participation in AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Boston MA USA, 1–13. https://doi.org/10.1145/3617694.3623228

[32] David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958), 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x Publisher: Wiley Online Library.

[33] Caroline Criado-Perez. 2019. *Invisible women: exposing data bias in a world designed for men.* Chatto & Windus, London. OCLC: 1084316434.

[34] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2022. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research* 23, 226 (2022), 1–61. https://www.jmlr.org/papers/v23/20-1335.html

[35] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics* 10 (Jan. 2022), 92–110. https://doi.org/10.1162/tacl_a_00449

[36] Joshua R. de Leeuw. 2015. jsPsych: A JavaScript Library for Creating Behavioral Experiments in a Web Browser. *Behavior Research Methods* 47, 1 (March 2015), 1–12. https://doi.org/10.3758/s13428-014-0458-y

[37] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Boston MA USA, 1–23. https://doi.org/10.1145/3617694.3623261

[38] Kerstin Denecke, Elia Gabarron, Rebecca Grainger, Stathis Th. Konstantinidis, Annie Lau, Octavio Rivera-Romero, Talya Miron-Shatz, and Mark Merolli. 2019. Artificial Intelligence for Participatory Health: Applications, Impact, and Future Implications: Contribution of the IMIA Participatory Health and Social Media Working Group. *Yearbook of Medical Informatics* 28, 01 (Aug. 2019), 165–173. https://doi.org/10.1055/s-0039-1677902

[39] Mark Diaz and Angela D. R. Smith. 2024. What Makes An Expert? Reviewing How ML Researchers Define "Expert". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (Oct. 2024), 358–370. https://doi.org/10.1609/aies.v7i1.31642

[40] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 6478–6490. https://proceedings.neurips.cc/paper_files/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf

[41] Jiayun Dong and Cynthia Rudin. 2020. Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models. https://doi.org/10.48550/arXiv.1901.03209 arXiv:1901.03209 [cs, stat]

[42] Samuel Dooley, Rhea Sukthanker, John Dickerson, Colin White, Frank Hutter, and Micah Goldblum. 2024. Rethinking bias mitigation: Fairer architectures make for fairer face recognition. *Advances in Neural Information Processing Systems* 36 (2024).

[43] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* 36, 6 (Sept. 2022), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z

[44] Mike FC, Trevor L. Davis, and ggplot2 authors. 2024. *ggpattern: 'ggplot2' Pattern Geoms.* https://CRAN.R-project.org/package=ggpattern R package version 1.1.1.

[45] Michael Feffer, Hoda Heidari, and Zachary C. Lipton. 2023. Moral Machine or Tyranny of the Majority? *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 5 (June 2023), 5974–5982. https://doi.org/10.1609/aaai.v37i5.25739

[46] Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. 2023. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Montréal QC Canada, 38–48. https://doi.org/10.1145/3600211.3604661

[47] Matthias Feurer and Frank Hutter. 2019. Hyperparameter Optimization. In *Automated Machine Learning: Methods, Systems, Challenges*, Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren (Eds.). Springer International Publishing, Cham, 3–33. https://doi.org/10.1007/978-3-030-05318-5_1

[48] Sam Firke. 2023. *janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor R package version 2.2.0.

[49] Ronald A Fisher. 1922. On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the royal statistical society* 85, 1 (1922), 87–94. https://doi.org/10.2307/2340521

[50] Andrew Gelman and Eric Loken. 2014. The statistical crisis in science. *American scientist* 102, 6 (2014), 460–465.

[51] Rayid Ghani and Malte Schierholz. 2021. Machine Learning. In *Big data and social science* (second edition ed.). CRC Press.

[52] Global AI Dialogue. 2024. A Workshop Series on the Impact of Artificial Intelligence (AI) on our Everyday Lives. https://perfectfuturedesign.com/global_ai_dialogue_info_english/.

[53] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association

for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3502004

[54] Travis Greene, Galit Shmueli, Jan Fell, Ching-Fu Lin, and Han-Wei Liu. 2022. Forks Over Knives: Predictive Inconsistency in Criminal Justice Algorithmic Risk Assessment Tools. *Journal of the Royal Statistical Society Series A: Statistics in Society* 185, Supplement_2 (Dec. 2022), S692–S723. https://doi.org/10.1111/rssa.12966

[55] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS Symposium on Machine Learning and the Law*, Vol. 1. Barcelona, Spain, 11.

[56] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey methodology.* John Wiley & Sons.

[57] 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 (Oct. 2020), 1–37. https://doi.org/10.1145/3415219

[58] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[59] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[60] Galen Harrison, Kevin Bryson, Ahmad Emmanuel Balla Bamba, Luca Dovichi, Aleksander Herrmann Binion, Arthur Borem, and Blase Ur. 2024. JupyterLab in Retrograde: Contextual Notifications That Highlight Fairness and Bias Issues for Data Scientists. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–19. https://doi.org/10.1145/3613904.3642755

[61] Ewen Henderson. 2024. *ghibli: Studio Ghibli Colour Palettes.* https://CRAN.R-project.org/package=ghibli R package version 0.3.4.

[62] Courtney B. Hilton, Cody J. Moser, Mila Bertolo, Harry Lee-Rubin, Dorsa Amir, Constance M. Bainbridge, Jan Simson, Dean Knox, Luke Glowacki, Elias Alemu, Andrzej Galbarczyk, Grazyna Jasienska, Cody T. Ross, Mary Beth Neff, Alia Martin, Laura K. Cirelli, Sandra E. Trehub, Jinqi Song, Minju Kim, Adena Schachner, Tom A. Vardy, Quentin D. Atkinson, Amanda Salenius, Jannik Andelin, Jan Antfolk, Purnima Madhivanan, Anand Siddaiah, Caitlyn D. Placek, Gul Deniz Salali, Sarai Keestra, Manvir Singh, Scott A. Collins, John Q. Patton, Camila Scaff, Jonathan Stieglitz, Silvia Ccari Cutipa, Cristina Moya, Rohan R. Sagar, Mariamu Anyawire, Audax Mabulla, Brian M. Wood, Max M. Krasnow, and Samuel A. Mehr. 2022. Acoustic Regularities in Infant-Directed Speech and Song across Cultures. *Nature Human Behaviour* (July 2022), 1–12. https://doi.org/10.1038/s41562-022-01410-x

[63] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Vol. 1. 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

[64] Jake M. Hofman, Amit Sharma, and Duncan J. Watts. 2017. Prediction and Explanation in Social Systems. *Science* 355, 6324 (Feb. 2017), 486–488. https://doi.org/10.1126/science.aal3856

[65] Giles Hooker. 2007. Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics* 16, 3 (2007), 709–732. jstor:27594267

[66] Hsiang Hsu and Flavio Calmon. 2022. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 28988–29000.

[67] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1395–1417. https://doi.org/10.1145/3630106.3658979

[68] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. 2014. An Efficient Approach for Assessing Hyperparameter Importance. In *Proceedings of the 31st International Conference on Machine Learning*. PMLR, 754–762. https://proceedings.mlr.press/v32/hutter14.html

[69] Rashidul Islam, Shimei Pan, and James R. Foulds. 2021. Can We Obtain Fairness For Free?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 586–596. https://doi.org/10.1145/3461702.3462614

[70] Sofia Jaime and Christoph Kern. 2024. Ethnic Classifications in Algorithmic Fairness: Concepts, Measures and Implications in Practice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 237–253. https://doi.org/10.1145/3630106.3658902

[71] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 310–323. https://doi.org/10.1145/3531146.3533097

[72] Sara Kingsley, Jiayin Zhi, Wesley Hanwen Deng, Jaimie Lee, Sizhe Zhang, Motahhare Eslami, Kenneth Holstein, Jason I. Hong, Tianshi Li, and Hong Shen. 2024. Investigating What Factors Influence Users' Rating of Harmful Algorithmic Bias and Discrimination. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 12 (Oct. 2024), 75–85. https://doi.org/10.1609/hcomp.v12i1.31602

[73] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter development team. 2016. Jupyter Notebooks - a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, Fernando Loizides and Birgit Scmidt (Eds.). IOS Press, Netherlands, 87–90. https://eprints.soton.ac.uk/403913/

[74] John Körtner and Giuliano Bonoli. 2023. Predictive algorithms in the delivery of public employment services. In *Handbook of Labour Market Policy in Advanced Democracies*. Edward Elgar Publishing, 387–398. https://doi.org/10.4337/9781800880887.00037

[75] Jon A. Krosnick. 1991. Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology* 5, 3 (1991), 213–236. https://doi.org/10.1002/acp.2350050305

[76] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory Approaches to Machine Learning. International Conference on Machine Learning Workshop.

[77] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[78] Edward E. Leamer. 1985. Sensitivity Analyses Would Help. *The American Economic Review* 75, 3 (1985), 308–313. jstor:1814801

[79] 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. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–35. https://doi.org/10.1145/3359283

[80] Timo Lenzner and Natalja Menold. 2016. Question Wording. GESIS survey guidelines.

[81] Bria Long, Jan Simson, Andrés Buxó-Lugo, Duane G. Watson, and Samuel A. Mehr. 2023. How Games Can Make Behavioural Science Better. *Nature* 613, 7944 (Jan. 2023), 433–436. https://doi.org/10.1038/d41586-023-00065-6

[82] Carol Long, Hsiang Hsu, Wael Alghamdi, and Flavio Calmon. 2023. Individual Arbitrariness and Group Fairness. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 68602–68624.

[83] Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-Biasing "Bias" Measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 379–389. https://doi.org/10.1145/3531146.3533105

[84] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376445

[85] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the Applicability of Machine Learning Fairness Notions. *SIGKDD Explor. Newsl.* 23, 1 (may 2021), 14–23. https://doi.org/10.1145/3468507.3468511

[86] Michele Mauri, Tommaso Elli, Giorgio Caviglia, Giorgio Uboldi, and Matteo Azzi. 2017. RAWGraphs: A Visualisation Platform to Create Open Outputs. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter* (Cagliari, Italy) *(CHItaly '17)*. ACM, New York, NY, USA, Article 28, 5 pages. https://doi.org/10.1145/3125571.3125585

[87] Kristof Meding and Thilo Hagendorff. 2024. Fairness Hacking: The Malicious Practice of Shrouding Unfairness in Algorithms. *Philosophy & Technology* 37, 1 (Jan. 2024), 4. https://doi.org/10.1007/s13347-023-00679-8

[88] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (2021). https://doi.org/10.1145/3457607

[89] Michelle M. Mello and Sherri Rose. 2024. Denial—Artificial Intelligence Tools and Health Insurance Coverage Decisions. *JAMA Health Forum* 5, 3 (March 2024), e240622. https://doi.org/10.1001/jamahealthforum.2024.0622

[90] Lisa Messeri and M. J. Crockett. 2024. Artificial Intelligence and Illusions of Understanding in Scientific Research. *Nature* 627, 8002 (March 2024), 49–58. https://doi.org/10.1038/s41586-024-07146-0

[91] Anna P. Meyer, Aws Albarghouthi, and Loris D'Antoni. 2023. The Dataset Multiplicity Problem: How Unreliable Data Impacts Predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 193–204. https://doi.org/10.1145/3593013.3593988

[92] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 220–229. https://doi.org/10.1145/3287560.3287596

[93] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[94] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward Involving End-users in Interactive Human-in-the-loop AI Fairness. *ACM Transactions on Interactive Intelligent Systems* 12, 3 (Sept. 2022), 1–30. https://doi.org/10.1145/3514258

[95] Nteract Contributors. 2017. Papermill: Parameterize and Run Jupyter and Nteract Notebooks.

[96] 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 (Oct. 2019), 447–453. https://doi.org/10.1126/science.aax2342

[97] Observable Team. [n. d.]. Observable: Build Expressive Charts and Dashboards with Code. https://observablehq.com/.

[98] Jeroen Ooms. 2014. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *arXiv:1403.2805 [stat.CO]* (2014). https://arxiv.org/abs/1403.2805

[99] The pandas development team. 2020. *pandas-dev/pandas: Pandas*. Zenodo. https://doi.org/10.5281/zenodo.3509134 DOI: 10.5281/zenodo.3509134.

[100] Thomas Lin Pedersen. 2024. *patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork R package version 1.2.0.

[101] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[102] Emma Pierson. 2018. Demographics and Discussion Influence Views on Algorithmic Fairness. https://doi.org/10.48550/arXiv.1712.09124 arXiv:1712.09124

[103] Marc Pinski and Alexander Benlian. 2023. AI Literacy - Towards Measuring Human Competency in Artificial Intelligence. In *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2023.021

[104] Pipenv Maintainer Team. 2017. Pipenv: Python Development Workflow for Humans.

[105] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[106] Cathy Roche, Dave Lewis, and PJ Wall. 2021. Artificial Intelligence Ethics: An Inclusive Global Discourse? *arXiv preprint arXiv:2108.09959* (2021).

[107] Kit T. Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. 2020. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 142–153. https://doi.org/10.1145/3351095.3372863

[108] Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. 2024. Amazing Things Come from Having Many Good Models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[109] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2019. Aequitas: A Bias and Fairness Audit Toolkit. https://doi.org/10.48550/arXiv.1811.05577 arXiv:1811.05577 [cs]

[110] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2022. On the Existence of Simpler Machine Learning Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 1827–1858. https://doi.org/10.1145/3531146.3533232

[111] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. https://doi.org/10.1145/3415224

[112] Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 440–451. https://doi.org/10.1145/3531146.3533110

[113] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22, 11 (Nov. 2011), 1359–1366. https://doi.org/10.1177/0956797611417632

[114] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. 2020. Specification Curve Analysis. *Nature Human Behaviour* 4, 11 (Nov. 2020), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

[115] Jan Simson. 2024. Multiversum: A Helper Package to Conduct Multiverse Analyses in Python. https://pypi.org/project/multiversum/

[116] Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy Data Practices Harm Fairness Research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 642–659. https://doi.org/10.1145/3630106.3658931

[117] Jan Simson, Florian Pfisterer, and Christoph Kern. 2024. One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1305–1320. https://doi.org/10.1145/3630106.3658974

[118] 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*. ACM, Arlington VA USA, 1–6. https://doi.org/10.1145/3551624.3555285

[119] Christopher Small. 2021. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *RECERCA. Revista de Pensament i Anàlisi* (July 2021). https://doi.org/10.6035/recerca.5516

[120] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* 25 (2012).

[121] Andy South. 2011. rworldmap: A New R package for Mapping Global Data. *The R Journal* 3, 1 (2011), 35–43. https://doi.org/10.32614/RJ-2011-006

[122] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (July 2022), 205395172211151. https://doi.org/10.1177/20539517221115189

[123] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11, 5 (2016), 702–712. https://doi.org/10.1177/1745691616658637 arXiv:https://doi.org/10.1177/1745691616658637 PMID: 27694465.

[124] Jan-Philipp Stein, Tanja Messingschlager, Timo Gnambs, Fabian Hutmacher, and Markus Appel. 2024. Attitudes towards AI: Measurement and Associations with Personality. *Scientific Reports* 14, 1 (Feb. 2024), 2909. https://doi.org/10.1038/s41598-024-53335-2

[125] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3465416.3483305

[126] United States District Court, District of Minnesota. 2023. Lokken v UnitedHealth Group Inc. CASE: 0:23-cv-03514.

[127] Guido Van Rossum and Fred L. Drake Jr. 1995. *Python Tutorial*. Vol. 620. Centrum voor Wiskunde en Informatica Amsterdam.

[128] Jamelle Watson-Daniels, Solon Barocas, Jake M. Hofman, and Alexandra Chouldechova. 2023. Multi-Target Multiplicity: Flexibility and Fairness in Target Specification under Resource Constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 297–311. https://doi.org/10.1145/3593013.3593998

[129] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Pedersen, Evan Miller, Stephan Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the Tidyverse. *Journal of Open Source Software* 4, 43 (Nov. 2019), 1686. https://doi.org/10.21105/joss.01686

[130] World-Wide-Lab Developers. 2024. World-Wide-Lab. https://worldwidelab.org

[131] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. 2022. Exploring the Whole Rashomon Set of Sparse Decision Trees. https://doi.org/10.48550/arXiv.2209.08040 arXiv:2209.08040 [cs]

[132] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, Hong Kong China, 573–584. https://doi.org/10.1145/3196709.3196729

[133] Achim Zeileis, Jason C. Fisher, Kurt Hornik, Ross Ihaka, Claire D. McWhite, Paul Murrell, Reto Stauffer, and Claus O. Wilke. 2020. colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes. *Journal of Statistical Software* 96, 1 (2020), 1–49. https://doi.org/10.18637/jss.v096.i01

[134] 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 (Nov. 2018), 1–23. https://doi.org/10.

Table 1: Distribution of participants across different countries. Due to the high number of countries, different countries with the same sample size are grouped into one row. Sample size and percentages apply to each individual country.

N	Percentage	Countries
199	37.27%	United States
37	6.93%	United Kingdom
31	5.81%	Canada
30	5.62%	Germany
27	5.06%	Australia
21	3.93%	France
10	1.87%	Aotearoa New Zealand; Brazil
9	1.69%	Russia
8	1.50%	Denmark
7	1.31%	Austria; Poland; Turkey
6	1.12%	India; South Korea
5	0.94%	Argentina; Belgium; Italy; Norway; Saudi Arabia; Sweden
4	0.75%	Hong Kong; Romania; Spain; Switzerland
3	0.56%	China; Iran; Ireland; Japan; Portugal; Serbia; Taiwan; The Netherlands
2	0.37%	Afghanistan; Albania; Andorra; Azerbaijan; Egypt; Finland; Greece; Hungary; Indonesia; Mexico; Philippines; Singapore; Vietnam
1	0.19%	Algeria; Angola; Armenia; Artsakh; Barbados; Bulgaria; Chad; Chile; Dominican Republic; Estonia; Georgia; Israel; Kenya; Kuwait; Latvia; Lithuania; Mongolia; Myanmar; Slovenia; South Africa; Sri Lanka; Thailand; Ukraine

1145/3274463

A Participants: Sample Composition

Participants self-reported a total of 69 different countries as their country of origin. The three most frequent countries are the United States (n = 199, 37.26%), the United Kingdom (n = 37, 6.92%) and Canada (n = 31, 5,80%). Participation rates by country are shown in Figure 3, with detailed numbers per country available in Table 1.

The majority of participants identified as male (n = 329, 61.61%), with 31.65% identifying as female (n = 169) and 6.74% identifying with another gender (n = 36). The average reported age is 29.11 years (*SD* = 15.56).

Participants from the United States were asked about their race and identified as predominantly White (n = 109, 54.77%), Asian (n = 38, 19.1%) or with more than one race (n = 18, 9.05%). Only a small number of people identified with other races (Black or African American n = 9, 4.52%; American Indian/Alaska Native n = 4, 2.01%; Native Hawaiian or other Pacific Islander n = 1, 0.5%) and about 10% of participants preferred not to answer the item (n = 20).

As racial identities are a highly complex topic and vary greatly within different geographical and social contexts [70], non-U.S. participants were not asked about their race but rather whether they identify as a minority in their respective country. Wording of the question was adapted from the European Social Survey. The majority of people did not identify as members of a minority (n = 260, 77.61%), with 14.03% of participants identifying with one (n = 47). For analyses using minority status, we also coded U.S. racial data into minority membership with the biggest group (White) coded as majority and all other racial groups as minorities.

B Research Materials

B.1 Introductory Text

In this last section, we have some questions for you about artificial intelligence (AI). We are building an AI system — and we need your help!

Our system is going to try to predict something important in the USA: **whether or not someone has public health insurance**. By "public health insurance", we mean free healthcare provided by the government, like Medicare or Veterans Health Administration coverage.

Here's how it will work. We'll give the AI some information (like a person's age and income) and it will learn how this information can reliably predict — or not — whether or not someone has health insurance. Then we'll see where its predictions went wrong, and help the AI to figure out how to make better predictions.

This could be really useful for helping to figure out who needs more support in getting health insurance. It could help local governments to look out for people in need.

Here's the problem: Sometimes, the people who design AI systems make bad decisions about how to train these models, and those decisions can lead to unfair, biased AI. **We want to know which sorts of decisions you think are good and which sorts you think are bad in designing an AI system.**

B.2 Introductory Text (Evaluation)

How should we evaluate whether an AI system is fair for different groups of people? The fairness of an AI system can be assessed with fairness metrics. These metrics allow quantifying the degree to which an AI system treats different groups of people equally based on sensitive characteristics like race, sex, or age. These metrics evaluate how well a system's predictions or decisions align across these groups to identify and reduce biases.

We want to make our AI system as fair as we can across different **races** (i.e., the AI system's predictions should work equally across different groups of the characteristic "race").

B.3 Decisions

B.3.1 Exclude Subgroups. **Should the AI system analyze all groups of people, or only some groups of people?** When we work with data from different groups, especially when some groups are very small or uncommon, it can be challenging to decide how best to handle their data.

Sometimes, small groups are left out to protect people's privacy, because the data might not be reliable, or excluding them might make the data easier to analyze. But doing so means they are not represented in the data anymore.

Here are a set of race and ethnicity subgroups of the US population. **Which groups do you think should be included in the AI system?** (you can choose as many as you like)

Answering options were not randomized for this question. Order and options based on the ACS PUMS [23].

Variant 1: No Percentages

- □ **White alone:** Use data from everyone identifying mainly as White.
- □ **Black or African American alone:** Use data from everyone identifying mainly as Black or African American.
- □ **American Indian alone:** Use data from everyone identifying mainly as American Indian.
- □ **Alaska Native alone:** Use data from everyone identifying mainly as Alaska Native.
- □ **Anyone who indicated that they are American Indian and/or Alaska Native and specified a tribe:** Use data from everyone identifying as American Indian or Alaska Native and who specified information about their tribe.
- □ **Asian alone:** Use data from everyone identifying as Asian.
- □ **Native Hawaiian and Other Pacific Islander alone:** Use data from everyone identifying mainly as Native Hawaiian or Pacific Islander.
- □ **Some Other Race alone:** Use data from anyone identifying mainly with another race than the ones mentioned here.
- □ **Two or More Races:** Use data from anyone identifying with two or more races (biracial).
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: ______

Variant 2: With Percentages

The percentages correspond to the size of the race/ethnicity group in the dataset.[8]

- □ **White alone (56.7%):** Use data from everyone identifying mainly as White.
- □ **Black or African American alone (5.8%):** Use data from everyone identifying mainly as Black or African American.
- □ **American Indian alone (0.9%):** Use data from everyone identifying mainly as American Indian.
- □ **Alaska Native alone (<0.1%):** Use data from everyone identifying mainly as Alaska Native.
- □ **Anyone who indicated that they are American Indian and/or Alaska Native and specified a tribe (0.3%):** Use data from everyone identifying as American Indian or Alaska Native and who specified information about their tribe.
- □ **Asian alone (15.0%):** Use data from everyone identifying as Asian.
- □ **Native Hawaiian and Other Pacific Islander alone (0.4%):** Use data from everyone identifying mainly as Native Hawaiian or Pacific Islander.
- □ **Some Other Race alone (15.8%):** Use data from anyone identifying mainly with another race than the ones mentioned here.
- □ **Two or More Races (5.0%):** Use data from anyone identifying with two or more races (biracial).
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: ______

B.3.2 Exclude Features. **What kind of information should an AI system use?** Sometimes AI systems take into account potentially sensitive characteristics (like a person's sex or race) and sometimes these characteristics are excluded due to legal or privacy reasons. But excluding this information doesn't always make AI systems fairer, as these characteristics can be related to other information.

Which of these options do you think are acceptable? (you can choose more than one)

Answering options were not randomized for this question.

- □ **Do not exclude any sensitive characteristics:** This means the AI is trained with all available information, including sensitive characteristics like race and sex.
- □ **Exclude race from the system:** This means the AI uses all information available but **not** race.
- □ **Exclude sex from the system:** This means the AI uses all information available but **not** sex.
- □ **Exclude both race and sex from the system:** This means the AI uses all information available but **not** race and sex.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: ______

B.3.3 Preprocess Income. **How should the AI system handle numbers?** When working with data that are numerical (like household income, or people's ages), it is often useful to *bin* these numbers into categories. This can make the data easier to understand and compare, but it also means the system is using less detailed information.

Which of these options do you think are acceptable for income data? (you can choose more than one)

- □ **No binning:** Keep the income data as it is.
- □ **Binning into bins of size $10,000:** Put each income into a group that covers ten thousand dollars, like $0-$9,999; $10,000-$19,999; and so on.

[8] *This text was erroneously displayed with brackets for a brief part of data collection (5.4% of sessions), before being updated. Previous version:* (the percentages correspond to the size of the race/ethnicity group in the dataset).

- □ **Binning into three evenly sized groups:** Divide all incomes into three equal groups: lower income, middle income and higher income.
- □ **Binning into four evenly sized groups:** Divide all incomes into four equal groups: lower income, middle income, upper middle income and higher income.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: ______

B.3.4 Model. **How complicated should the AI system be?** One of the key decisions in designing an AI system is choosing the type of model to use. Below are some possibilities. (it's okay if you don't know how these work!)

Which of these options do you think are acceptable? (you can choose more than one)

- □ **Simple, more understandable model (Logistic Regression):** This type of model is easier to understand and interpret than many alternatives. It may not be able to learn as many relations as other, more powerful models.
- □ **Complex, more flexible model (Random Forest):** This type of model is able to learn intricate relations in the data, but is harder to interpret and understand.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: ______

B.3.5 Eval Fairness Definition. **How should we evaluate our AI system?** Two important metrics are available to evaluate AI systems. Since we can often not be perfect on both of them, we may have to focus on one in particular.

For an AI system that predicts whether a person has public health insurance, which is most important?

This decision allowed selecting only a single answering option.

- □ **Sensitivity:** It is more important that out of all people without public health insurance, the AI system correctly identifies as many as possible. This would minimize the number of people who really do not have public insurance who are incorrectly identified as having public insurance.
- □ **Precision:** It is more important that out of all people where the AI system thinks that they do not have health insurance, many individuals indeed have no insurance. This would minimize the number of people who really have public insurance, but are incorrectly identified as not having public insurance.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: ______

B.3.6 Eval Fairness Grouping. **How should our data be grouped?** When checking whether our AI system is fair or not, we need to choose how to group the data. This could influence how we calculate whether the AI system is fair or not.

Which of these options do you think are acceptable? (you can choose more than one)

- □ **Two Groups:** Create just two groups: the largest group and all other groups combined. We then check whether the AI system works equally well for these two groups. The difference between these two groups is used as the fairness metric.
- □ **All Group Comparisons:** Consider all the different groups separately without combining them. We then check whether the AI system works equally well for each possible pair of groups. The largest difference between any two groups is used as the fairness metric.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: ______

B.3.7 Eval on Subset. **Which people should an AI system be tested on?** It's not practical to test out an AI system on everybody in a whole country, but it's also hard to choose who it should be tested on.

For an AI system that predicts whether a person has public health insurance, Which of these options do you think are acceptable? (you can choose more than one)

- □ **Collecting data from the most populous area:** This means testing the AI system using data from the area with the most people living in it, like a big city.
- □ **Collecting data from the area where the most people have public health insurance:** This means testing the AI system using data from an area where lots of people have public health insurance, but a few people don't have public health insurance.
- □ **Collecting data from the closest major city:** This means evaluating the AI system using data only from a city close-by to the people building the AI system.
- □ **Collecting data from as many people as possible, but excluding military veterans:** Being a veteran can impact healthcare needs, so it might change the AI's predictions to test the model on veterans. This option means testing the AI system only using data from non-veterans living in the area where the AI is being tested.
- □ **Collecting data of only U.S. citizens:** This means testing the AI system using data from U.S. citizens and not other people living in the area where the AI is being tested.
- □ **Collecting data from the overall population:** This means testing the AI in a similar way to how political polls are conducted, by studying a representative sample of people in the US.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: ______

B.3.8 Eval Exclude Subgroups. When creating an AI system, one might exclude certain smaller groups from the data to simplify the process or with the intention to protect their privacy. There already was an earlier question about this regarding the exclusion of data from **certain** groups when creating the AI system.

This decision now is about including or excluding the same small groups when **evaluating** how good the AI system works.

Which of these options do you think are acceptable? (you can choose more than one)

This decision was only displayed if (1) an answer for Exclude Subgroups was provided, (2) that answer was one of the valid options and (3) not all options of Exclude Subgroups were selected.

- □ **Keep all groups for evaluation:** This means evaluating the AI system with data from all groups, also ones that were excluded earlier.
- □ **Exclude the same groups during evaluation:** This means using only data from the groups that were also included when creating the AI, excluding data from the same groups that were excluded earlier.
- □ I don't understand the description
- □ I prefer not to answer
- □ Suggest an alternative option: ______

C Software used in Analyses

Analyses were conducted with R version 4.2.2 [105] using packages from the tidyverse [129] with support of multiple other packages [4, 44, 48, 61, 98, 100, 121, 133].

The complete multiverse of decisions was simulated and explored with Python version 3.8 [127], using pandas [99] for data manipulation and scikit-learn [101] for modeling alongside multiple other packages [12, 40, 59, 73, 95, 104]. Diagrams of the multiverse were generated using RawGraphs [86], d3 [19] and Observable [97].

D Supplementary Tables and Figures

This section contains supplementary tables, figures and information on statistical analyses described in the main body of this work.

Tables with statistical details for group comparisons contain test details for each comparison which was calculated. Odds ratios are provided for comparisons between two groups. The column *Sig. Threshold* refers to Bonferroni corrected significance thresholds for a p-value of $\alpha = 0.05$.

Table 2: Overview of the two decision blocks, the actual decisions examined in the case study and their respective options. For the decision *Exclude Subgroups* the combination of options was used. The decision *Eval Fairness Definition* allowed choosing only one option. For the participatory input, each decision includes three additional *other* options: "I don't understand the description," "I prefer not to answer" and "Suggest an alternative option".

Block	Decision	Options
Model Design Decisions (Section 3.1.1)		
Data Selection	*Exclude Subgroups*	(1) white-alone; (2) black-or-african-american-alone; (3) american-indian-alone; (4) alaska-native-alone; (5) american-indian-and-or-alaska-native-and-tribe; (6) asian-alone; (7) native-hawaiian-and-other-pacific-islander-alone; (8) some-other-race-alone; (9) two-or-more-races
	Exclude Features	(1) none; (2) race; (3) sex; (4) race-sex
Preprocessing	*Preprocess Income*	(1) none; (2) bins-10000; (3) quantiles-3; (4) quantiles-4
Modeling	*Model*	(1) simple; (2) complex
Evaluation Decisions (Section 3.1.2)		
Metric	*Eval Fairness Definition*	(1) sensitivity; (2) precision
Evaluation	*Eval Fairness Grouping*	(1) majority-minority; (2) race-all
	Eval On Subset	(1) locality-largest-only; (2) locality-most-privileged; (3) locality-city; (4) exclude-military; (5) exclude-non-citizens; (6) full
	Eval Exclude Subgroups	(1) keep-in-eval; (2) exclude-in-eval

Table 3: Statistical details of group comparisons for the decision *Exclude Subgroups*, comparing different groups by *Country Of Residence*.

Option	Odds Ratio	p-value	Sig. Threshold
White (56.7%)	0.55	0.0018	0.0038
Some Other Race (15.8%)	0.64	0.0224	0.0038
Asian (15.0%)	0.62	0.0123	0.0038
Black/African American (5.8%)	0.61	0.0119	0.0038
Two Or More Races (5.0%)	0.59	0.0064	0.0038
American Indian (0.9%)	0.60	0.0104	0.0038
Nat. Hawaiian & Pac. Islander (0.4%)	0.62	0.0143	0.0038
Am. Indian & Alaska Native Incl. Tribe (0.3%)	0.82	0.3244	0.0038
Alaska Native (<0.1%)	0.84	0.4247	0.0038
Dont Understand	1.81	0.2278	0.0038
Prefer No Answer	1.51	0.1003	0.0038
Suggest Alternative	1.40	0.2973	0.0038
None Checked	1.06	0.8392	0.0038

Table 4: Statistical details of group comparisons for the decision *Exclude Subgroups*, based on whether percentages of the relative size of each subgroup were visible.

Option	Odds Ratio	p-value	Sig. Threshold
White (56.7%)	1.12	0.5836	0.0038
Some Other Race (15.8%)	1.27	0.2491	0.0038
Asian (15.0%)	1.09	0.7100	0.0038
Black/African American (5.8%)	1.07	0.7794	0.0038
Two Or More Races (5.0%)	1.21	0.3468	0.0038
American Indian (0.9%)	1.09	0.7030	0.0038
Nat. Hawaiian & Pac. Islander (0.4%)	0.98	1.0000	0.0038
Am. Indian & Alaska Native Incl. Tribe (0.3%)	0.95	0.8488	0.0038
Alaska Native (<0.1%)	0.95	0.8472	0.0038
Dont Understand	2.20	0.0786	0.0038
Prefer No Answer	1.03	1.0000	0.0038
Suggest Alternative	0.96	1.0000	0.0038
None Checked	0.79	0.2803	0.0038

Table 5: Statistical details of group comparisons for the decision *Exclude Features*, comparing different groups of the attribute *gender*.

Option	p-value	Sig. Threshold
None	0.0112	0.0063
Race	0.7743	0.0063
Sex	0.2613	0.0063
Race Sex	0.2362	0.0063
Dont Understand	0.7586	0.0063
Prefer No Answer	0.0685	0.0063
Suggest Alternative	0.0152	0.0063
None Checked	0.9346	0.0063

Table 6: Statistical details of group comparisons for the decision *Exclude Features*, comparing different groups of the attribute minority status. Minority status was self-reported for participants outside of the U.S. and computed based on majority-group membership based on self-reported race for U.S. participants.

Option	p-value	Sig. Threshold
None	0.0132	0.0063
Race	0.0929	0.0063
Sex	0.4921	0.0063
Race Sex	0.7785	0.0063
Dont Understand	0.4599	0.0063
Prefer No Answer	0.0680	0.0063
Suggest Alternative	0.7344	0.0063
None Checked	0.6405	0.0063

Table 7: Statistical details of group comparisons for the decision *Eval Fairness Definition*, comparing three equally-sized groups based on self-reported *AI Literacy*.

Option	p-value	Sig. Threshold
Sensitivity	0.2756	0.0083
Precision	0.4475	0.0083
Dont Understand	0.3651	0.0083
Prefer No Answer	0.5722	0.0083
Suggest Alternative	0.0009	0.0083
None Checked	0.0002	0.0083

Table 8: Statistical details of group comparisons for the option *prefer not to answer*, comparing three equally-sized groups based on self-reported *AI Attitudes*.

Decision	p-value	Sig. Threshold
Exclude Subgroups	0.1427	0.0063
Exclude Features	0.0290	0.0063
Model	0.0065	0.0063
Preprocess Income	0.0619	0.0063
Eval Fairness Definition	0.5947	0.0063
Eval Fairness Grouping	0.9266	0.0063
Eval On Subset	0.4232	0.0063
Eval Exclude Subgroups	0.2421	0.0063

Table 9: Statistical details of group comparisons for the option *none checked*, comparing three equally-sized groups based on self-reported *AI Attitudes*.

Decision	p-value	Sig. Threshold
Exclude Subgroups	0.0128	0.0063
Exclude Features	0.0051	0.0063
Model	0.0053	0.0063
Preprocess Income	0.0064	0.0063
Eval Fairness Definition	0.0141	0.0063
Eval Fairness Grouping	0.0514	0.0063
Eval On Subset	0.0476	0.0063
Eval Exclude Subgroups	0.3343	0.0063

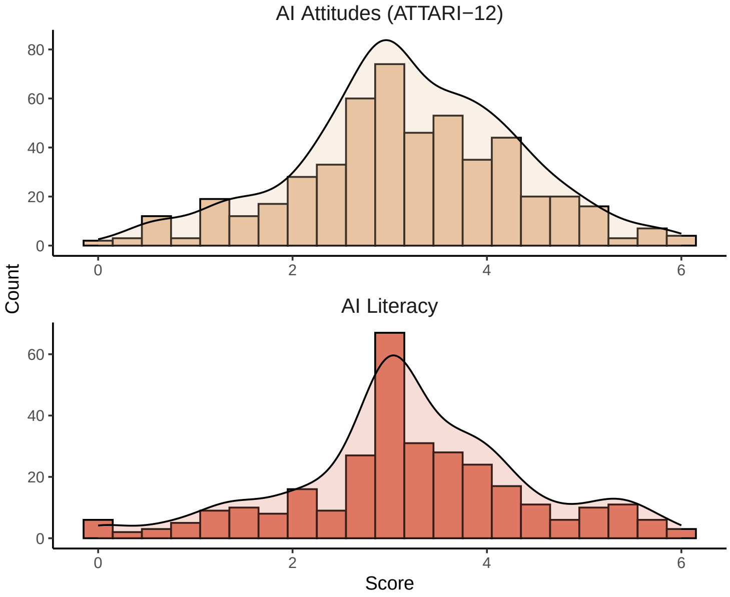

Figure 13: Histograms showing the overall distribution of AI attitudes [124] (above) and AI literacy [103] (below) scores across participants. Both scales emit a high degree of variation in the present sample, with a slight tendency towards more positive AI attitudes and higher AI literacy.

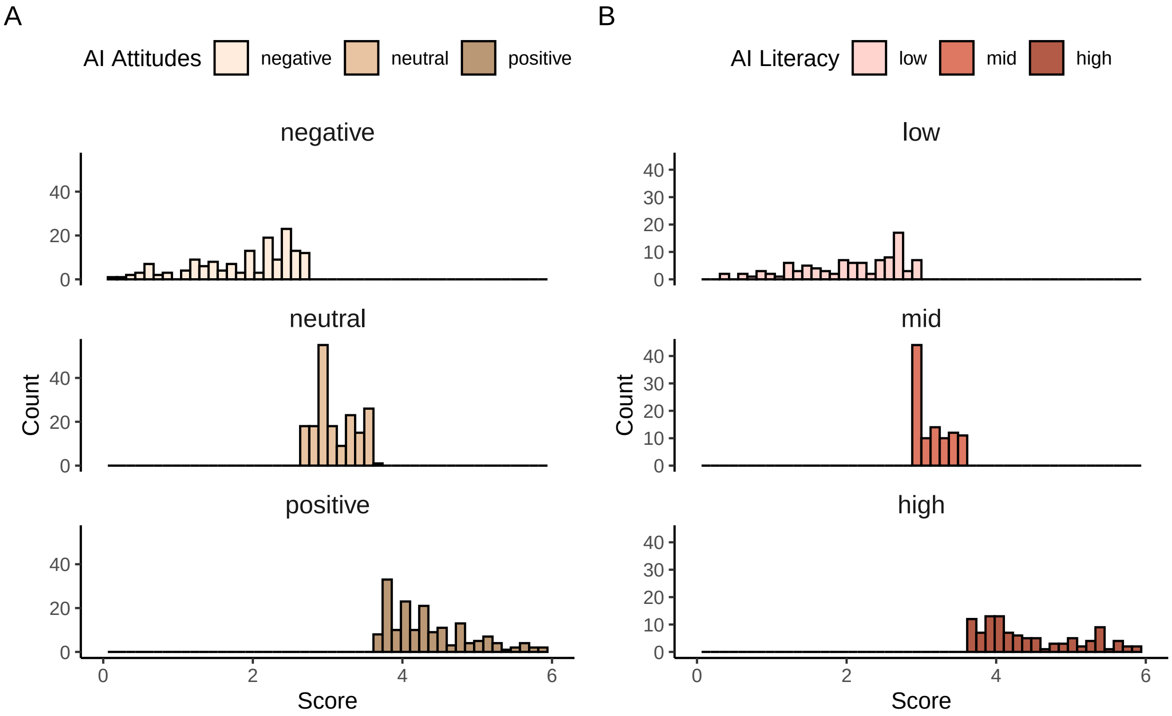

Figure 14: Histograms showing the distribution of AI attitudes (A) and AI literacy (B) scores across the three equally sized groups per metric, which were used for later group comparisons. The overall distribution of both scales is shown in Figure 13.

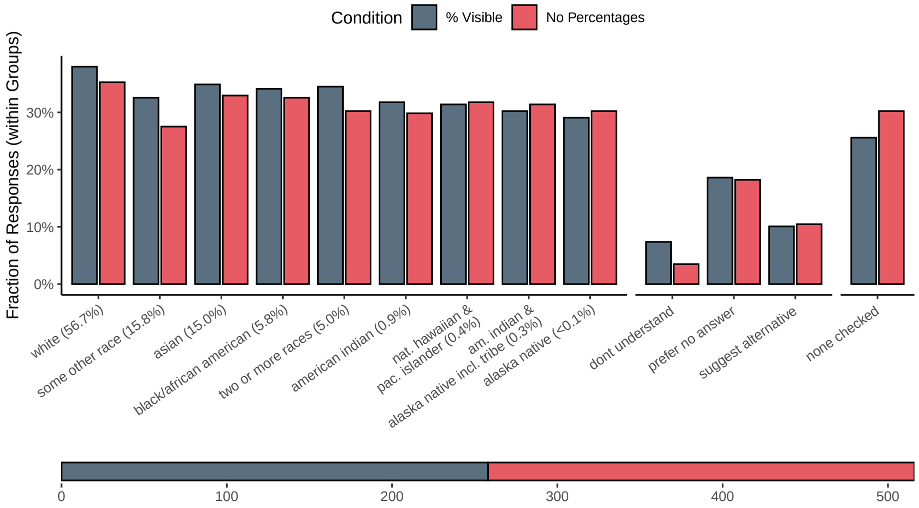

Figure 15: Inclusion of subgroups split by whether or not percentages were displayed next to groups for the decision *Exclude Subgroups*. The bar below the plot indicates the raw group distribution and number of votes.

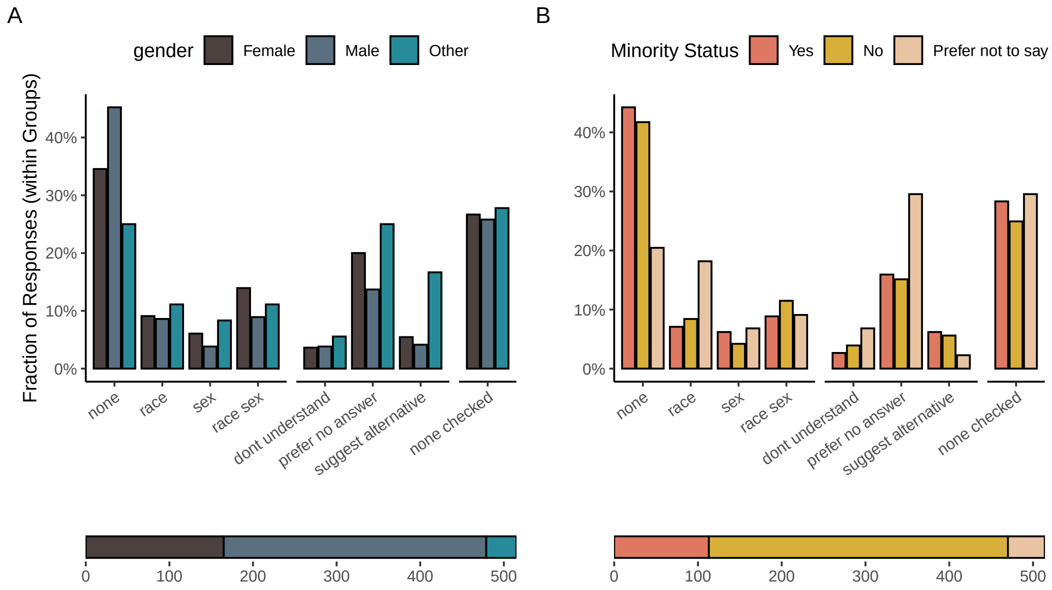

Figure 16: Exclusion of sensitive features split by gender (A) and minority status (B) for the decision *Exclude Features*. Bars below plot indicate the raw group distribution and number of votes.

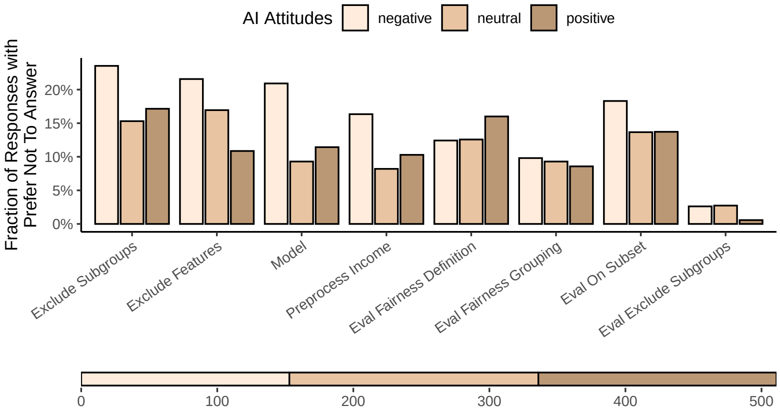

Figure 17: Fraction of participants choosing *"I Prefer Not To Answer"* across decisions split by self-reported AI attitudes. The bar below the plot indicates the raw group distribution and number of votes.

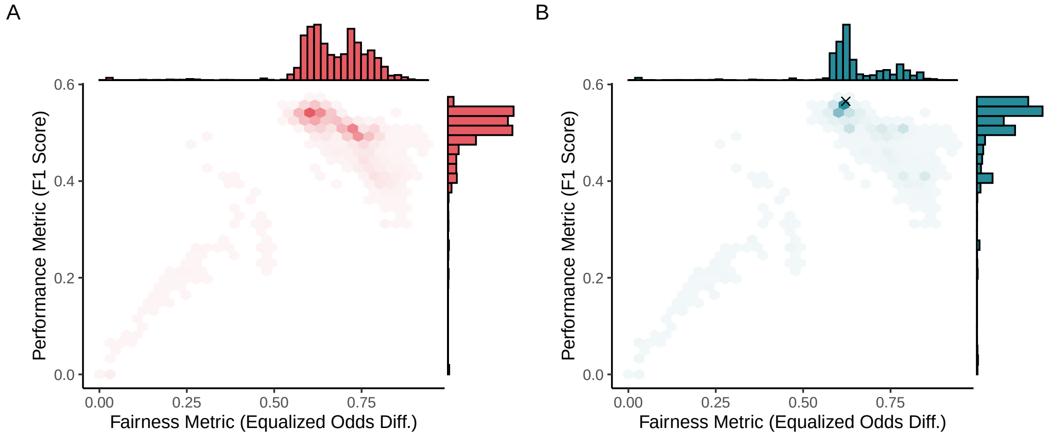

Figure 18: Comparison between complete multiverse of models (A) and one based on participants' votes (B), both evaluated using a fixed strategy with equalized odds difference as fairness metric and F1 score as performance metric. Darker areas correspond to a higher clustering of models. Cross indicates the most popular model among participants.