Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems

Hao-Fei Cheng [1] **Paige Bullock** [2] **Alexandra Chouldechova** [3] **Zhiwei Steven Wu** [1] **Haiyi Zhu** [3]

Abstract

Recent work in fair machine learning has proposed dozens of quantitative notions of algorithmic fairness and methods to enforcing these notions. However, we still lack an understanding of how to develop machine learning systems with fairness criteria that reflect human stakeholders' nuanced viewpoints in the real-world contexts. To address the gap, we propose a framework for eliciting stakeholders' subjective fairness notions. Combining a user interface that allows stakeholder to examine the data and algorithm's predictions, and an interview protocol to probe stakeholders' thoughts while they are interacting with the interface, we can identify stakeholders' fairness beliefs and principles. We propose a user study to evaluate our framework in real world settings of a child maltreatment predictive system.

1. Introduction

Machine learning (ML) algorithms are increasingly being used to support human decision-making in high-stakes contexts such as online information curation, resume screening, mortgage lending, police surveillance, public resource allocation, and pretrial detention. However, concerns have been raised that algorithmic systems might inherit human biases from historical data, and thereby perpetuate discrimination against already vulnerable subgroups. These concerns have given rise to a rapidly growing research area of fair machine learning. Recent work in this area has produced dozens of quantitative notions of algorithmic fairness (Verma & Rubin, 2018; Narayanan, 2018; Hardt et al., 2016; Agarwal et al., 2019; Corbett-Davies & Goel, 2018), and provided methods for enforcing these notions (Dwork et al., 2012; Kamiran & Calders, 2012; Zafar et al., 2017; Agarwal et al., 2018; 2019; Kearns et al., 2018).

[1]University of Minnesota, Minneapolis, Minnesota, USA [2]Kenyon College, Gambier, Ohio, USA [3]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence to: Hao-Fei Cheng <cheng635@umn.edu>.

Existing research on fair machine learning has primarily focused on fairness at the level of pre-defined groups. This *group fairness* approach first fixes a small collection of high-level groups defined by protected attributes (e.g., race or gender) and then asks for approximate equality of some statistic of the predictor, such as positive classification rate or false positive rate, across these groups (see, e.g., (Hardt et al., 2016; Agarwal et al., 2018; Kleinberg et al., 2017b)). While notions of group fairness are easy to operationalize, they are aggregate in nature and make no fairness promises on finer subgroups or individuals (Dwork et al., 2012; Kearns et al., 2018; Hébert-Johnson et al., 2018). In contrast, the *individual fairness* approach aims to address this limitation by asking for explicit fairness criteria at an individual level. For example, Dwork et al. (2012) propose an individual fairness notion that requires that similar people are treated similarly. Their formulation of fairness crucially relies on a task-specific metric that captures whether two individuals are similar for the purpose of the task at-hand. Due to the challenges of specifying such a metric in any given real world decision-making problem, it remains difficult to operationalize individual fairness in practice.

Irrespective of the approach one takes to quantifying fairness, it is important to engage relevant stakeholders in the design of a real-world decision making system. As Shah (2018) has argued, achieving legitimacy or "social license" from the broader community is critical to the ability of even the best-conceived technologies to have a positive social impact. One technology that failed to be adopted due to a lack of public support is a school start time scheduling tool proposed in Boston intended to decrease bussing costs while improving racial equity and better accommodating differences in circadian rhythms across students of different ages. The system's design failed to account for the excess burden that the proposed times would place on families with multiple children who attend different schools, particularly for lower-income parents who tend to have inflexible work schedules (Whittaker et al., 2018). Such gaps are very common. A recent study by Veale et al. (2018) interviewed 27 public sector ML practitioners across 5 OECD countries and noted numerous disconnects between current fair ML approaches and the organizational and institutional realities, constraints and needs.

Thus, there is a growing need to understand how to develop ML decision systems that reflect nuanced task-relevant objectives and constraints. To this end, we develop novel methods to elicit context-specific fairness principles that reflect the perspectives of task-relevant stakeholders. In this paper, we propose a framework for eliciting fairness notions from human stakeholders. The framework includes two components: an ***interactive interface*** that allows stakeholders to examine the data and audit an algorithm's predictions, and an ***interview protocol*** that is designed to probe stakeholders' thoughts and beliefs on fairness and biases of the algorithm while they are interacting with the interface. While the primary focus of our study and interface design is on fairness preference elicitation, the interview protocol enables us to learn about stakeholder perspectives along other dimensions as well.

We will evaluate the framework in a high-stake context of the *Allegheny Family Screening Tool* (AFST)—a ML-based risk assessment tool that assists child maltreatment hotline screening (Vaithianathan et al., 2017). We will conduct studies with stakeholders (parents and social workers) and examine whether the framework can effectively enable stakeholders to express their fairness viewpoints. Furthermore, through the open-ended interview protocol, we will aim to better understand stakeholder perceptions and preferences about dimensions of the algorithm, such as the use of particular non-protected features, that are not strictly speaking related to fairness.

Our work contributes to the fair machine learning research by combining innovative machine learning methods with approaches from human-computer interaction (HCI). We propose a new way for human stakeholders to participate in the process of fairness auditing and bias detection. This raises awareness for future development of algorithms to incorporate fairness notions that better align with stakeholder preferences.

2. Related Work and Research Question

2.1. Machine Learning Fairness

There has been significant development in the research of machine learning fairness and accountability in recent years (Verma & Rubin, 2018; Narayanan, 2018; Hardt et al., 2016; Agarwal et al., 2019; Corbett-Davies & Goel, 2018). Prior literature on ML fairness can generally be classified in two categories: group fairness and individual fairness. The more commonly studied notion, group fairness, requires statistical parity of some measure across a fixed number of protected groups, which provides no meaningful guarantees of fairness to individuals. Further, subgroups of these protected groups can suffer discrimination unless group fairness definitions are held on an infinite class of groups. On the other hand, individual fairness requires "treating similar individuals similarly (Dwork et al., 2012)," which would appear to solve these issues. This approach assumes the existence of some consistent similarity metric, which is not always available, but metric-free individual fairness may be achieved with the help of a human auditor who "knows unfairness when they see it". Recent research has also called for human involvement in developing algorithms (Lee et al., 2019; Zhu et al., 2018), making final decisions after algorithm recommendations (Lai & Tan, 2019), and making decisions about fairness trade-offs (Yu et al., 2019) (as satisfying the criteria for all fairness definitions is mathematically impossible (Kleinberg et al., 2017a; Friedler et al., 2016)). However, it remains a major challenge to devise mechanisms for involving stakeholders in algorithm development and auditing that do not require unrealistic levels of technical knowledge among participants.

2.2. HCI Research on Algorithmic Fairness

More recently, HCI researchers have begun to investigate human perspectives on algorithmic fairness. Several recent studies have investigated public (Grgic-Hlaca et al., 2018; Scurich & Monahan, 2016; Wang, 2018) and practitioner (Holstein et al., 2019; Monahan et al., 2018; Veale et al., 2018) perspectives on the use of algorithmic systems for public-sector decisions. This body of work suggests that fairness principles need to be context-specific, and the algorithmic systems should embody the fairness notions derived from the community of stakeholders (Brown et al., 2019; Dodge et al., 2019; Lee, 2018; Lee & Baykal, 2017). There has been encouraging work towards this direction. For example, researchers have conducted workshops and interviews to understand what people think fairness means in the context of resource allocation (Lee et al., 2017) or targeted online ads (Woodruff et al., 2018). Researchers have also conducted surveys to understand what features should or should not be used a fair learning algorithm (Grgic-Hlaca et al., 2018; Scurich & Monahan, 2016). To our knowledge, however, little work sought to formalize subjective concepts of fairness. Furthermore, while these studies provide us with a better understanding of general public and user perceptions of justice and fairness, only a few of them have closed the loop on algorithm developments that respond to those concerns (Bolukbasi et al., 2016; Freedman et al., 2018; Lee et al., 2018).

2.3. Research Question

In the paper, we want to answer the following research question: **How can we effectively elicit fairness notions from a community of stakeholders who are not technical experts, and incorporate this information back into the algorithm training process?**

3. Framework of Eliciting Fairness From Stakeholders

To answer the first part of this question, we propose an elicitation framework. The framework includes two components: an ***interactive interface*** that allows stakeholders to examine the data and audit the algorithm's recommendations; an associated ***interview protocol*** that is designed to further probe stakeholders' thoughts and beliefs on fairness and biases of the algorithm.

3.1. Interface Design

3.1.1. Design goals

The purpose of the interface is to enable the stakeholders to examine the data at multiple levels, and audit the recommendations made by a machine learning algorithm. We identify three specific design goals for the interface:

Goal 1: Examine the data and algorithm at the "macro" level. Corresponding to the "group fairness" notion, the interface should enable users to examine the data and algorithm performance in the groups defined by the users (not limited to groups defined by common protected attributes such as gender and race). The interface should present the various statistical metrics for each of subgroups and visualize them for stakeholders to investigate.

Goal 2: Examine the data and algorithm at the "micro" level. Corresponding to the "individual fairness" notion, the interface should enable users to inspect the data and algorithm recommendations at a case-by-case level.

According to (Dwork et al., 2012; Joseph et al., 2016), one effective approach of inspecting individual fairness is to allow stakeholders to make pairwise comparisons and inspect (1) whether the pair of individuals should be treated similarly or not, and (2) whether one individual should be prioritized over the other one or not.

Goal 3: Examine the data and algorithm at the "meso" level. The goal is to enable stakeholders to explore to what extent any individual case should be prioritized compared to all the other cases in the dataset. Different stakeholders may have different criteria for evaluating the similarity and priority across the cases. The interface should allow users to specify their own metrics when exploring the data.

3.1.2. Interface prototype

Our current interface prototype consists of three primary views: (i) a group view corresponding to Goal 1 (see Figure 1a), (ii) a case-by-case view corresponding to Goal 2 (see Figure 1b), and (iii) a similarity view corresponding to Goal 3 (see Figure 1c).

Group view: This view aims at giving users a holistic view of the algorithm performance by showing how the algorithm's performance according to different metrics varies across groups. Users have the option to select from a list of common classification performance metrics. The dropdown menus allow the user to select attributes with which to separate the data into subgroups. The interface displays a bar chart depicting the algorithm's performance across the specified subgroups. A textual description is also provided below the graph to provide an alternate description of the algorithm's performance.

Case-by-case view: This view allows users to audit the algorithm performance at a granular level, by inspecting each individual algorithm prediction. Each case of algorithm prediction is presented as a card, and the interface shows two cases at a time for pairwise comparisons. On each card, the algorithmic prediction is shown on top, followed by features the algorithm used to make the prediction. Hovering over each feature will show users the detailed description of that feature, and the possible values the feature can take.

Users can browse through the cases back and forth. The tool will randomly select a new case from the dataset, and replace the currently displayed case. Users can explore new cases by either changing the case on the left or right.

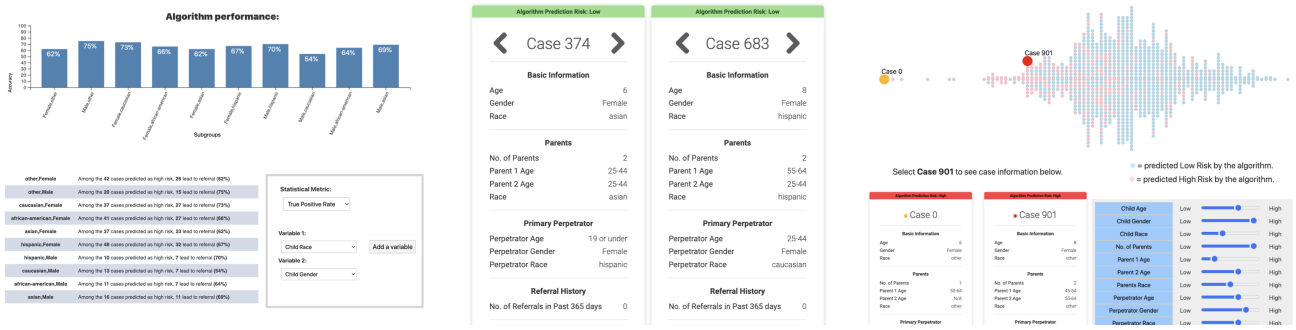
Similarity view: This view shows a one-dimensional scatter plot that compares a selected reference case with all other cases in the dataset. This allows users to explore the dataset at a macro view and narrowing down to individual cases for inspections.

This scatter plot displays all the cases in the dataset, with each case represented by a dot on the plot, color-coded according to the algorithm prediction. The reference case is positioned at the far left of the plot, with other cases ordered by similarity to the reference case along the x-axis. A weighted Euclidean distance metric is used to calculate the similarity of the cases[1]. The y-axis shows the distribution of the cases at that similarity level. A control panel allows users to change the weight associate with each feature. Users can customize the weights to re-rank the cases in an order that aligns with their viewpoints. Users can select a case from the plot to compare with the reference case, or set a new case as the reference case.

3.2. Interview Protocol

To complement the interface, we develop interview protocols to probe stakeholders' fairness viewpoints and principles. The protocol is based on the think-aloud approach, which one of the most valuable usability engineering methods in HCI (Nielsen, 1994). We will ask stakeholders to

[1] weighted Euclidean distance between case p and q is calculated by: $\sqrt{\sum_{i=1}^{n} w_i(q_i - p_i)^2}$, w_i denotes the user assigned weight for feature i.



(a) Group view (b) Case-by-case view (c) Similarity view

Figure 1. Our interface prototype contains three different views, which allows stakeholders examine the algorithm at different levels.

use the interface we described above "while continuously thinking out loud - that is, verbalizing their thoughts as they move through the user interface"(Nielsen, 1994). Think aloud serves as "a window on the soul", letting us discover what participants really think about the fairness and bias of the algorithm.

First, we show stakeholders the group view of the interface. Stakeholders can define the groups they want to inspect, and see the algorithm's performance on the groups. We ask stakeholders whether the algorithm is biased according to their belief. If participants believe any particular groups (and subgroups) are being treated unfairly, and we will ask follow up questions to probe the reasons they believe so.

Second, we ask stakeholders to compare pairs of cases in the data *without showing* the algorithmic prediction, in the case-by-case view. We ask stakeholders if both cases should receive be treated equally (or receive the same decision in the context), and if not, what alternative outcomes should the two cases receive to align with their fairness principles. In this stage, we only show users the features for the cases, as we aim to collect stakeholders' fairness notions regardless of the predictions of those outcomes, and the factors they would consider when evaluating the cases in the context.

Third, we ask stakeholders to make pairwise comparison again, with cases *showing* the algorithmic prediction. We ask users to identify and explain (pairs of) cases that are being treated unfairly. We also asked them to evaluate if the algorithm predictions are in general biased according to their fairness notions. The goal of this stage is to have stakeholders audit the individual fairness of the algorithm.

Last, we ask stakeholders to use the similarity view to compare reference cases with all the other cases in the data. We ask stakeholders to define their own similarity metrics, then identify cases that should be prioritized by the algorithm. We ask stakeholders to explain the reasons behind selecting those cases, and the information they rely on to identify them.

Throughout the interview, participants are encouraged to share their views on the cases before them even if those views do not reflect perceptions of fairness per se. Participants may indicate, for instance, that they are uncomfortable with the use of algorithms in certain cases, that particular case characteristics are of paramount important to the decision-making process, or that having model explanations would improve their understanding of the tool. This is all valuable, actionable feedback that may be incorporated into the algorithm re-training process.

4. User Study: Eliciting Subjective Fairness Notion in Child Maltreatment Prediction

We will evaluate our framework in a real-world high-stake context—child maltreatment prediction.

4.1. Context and Data

In December 2016, Allegheny County Child Welfare Office's intake office started to use a screening tool (Allegheny Family Screening Tool or AFST) to aid call screeners in making recommendations about screening decisions regarding further investigation (Vaithianathan et al., 2017). For each referral case, call screeners are presented with an AFST score. Generation of the score is based on data related to the victim child(ren), parents, legal guardians, perpetrators, prior child welfare history, criminal history, and use of public assistance. Due to the sensitive nature of the data, in the study, we will use a synthetic dataset based on the real dataset provided by the Allegheny County Department of Human Services.

4.2. Experimental Design

We plan to recruit two groups of stakeholders for the user study: social workers with experience of investigating allegations of child abuse, and parents. We will invite each participant to use the prototype tool to audit the child maltreatment prediction data. We will follow the interview

protocol introduced in 3.2 to elicit participants fairness notions in the algorithm decisions.

5. Discussion

We provide a participatory framework that enables stakeholders to perform fairness auditing on a high-stake predictive system and to express their subjective fairness viewpoints. Our ultimate goal is to leverage the elicited feedback to improve the a tool modeled on the call screening system in Allegheny County that includes both the predictive algorithm and the human decision makers. In particular, this will require a further study on how to redesign the system so that it can balance and reflect different stakeholders' values on the algorithm-assisted decision-making in the child welfare context.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 60–69, 2018. URL http://proceedings.mlr.press/v80/agarwal18a.html.

Agarwal, A., Dudík, M., and Wu, Z. S. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 120–129, 2019. URL http://proceedings.mlr.press/v97/agarwal19d.html.

Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4349–4357, 2016. URL http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., and Vaithianathan, R. 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, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, pp. 41, 2019. doi: 10.1145/3290605.3300271. URL https://doi.org/10.1145/3290605.3300271.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018. URL http://arxiv.org/abs/1808.00023.

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., and Dugan, C. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pp. 275–285, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6272-6. doi: 10.1145/3301275.3302310. URL http://doi.acm.org/10.1145/3301275.3302310.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pp. 214–226, 2012. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Freedman, R., Schaich Borg, J., Sinnott-Armstrong, W., Dickerson, J. P., and Conitzer, V. Adapting a kidney exchange algorithm to align with human values. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 115–115, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278727. URL http://doi.acm.org/10.1145/3278721.3278727.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., and Weller, A. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pp. 903–912. International World Wide Web Conferences Steering Committee, 2018. ISBN 978-1-4503-5639-8. doi: 10.1145/3178876.3186138. URL https://doi.org/10.1145/3178876.3186138.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3315–3323, 2016. URL http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.

Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. N. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 1944–1953, 2018. URL http://proceedings.mlr.press/v80/hebert-johnson18a.html.

Holstein, K., Vaughan, J. W., III, H. D., Dudík, M., and Wallach, H. M. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, pp. 600, 2019. doi: 10.1145/3290605.3300830. URL https://doi.org/10.1145/3290605.3300830.

Joseph, M., Kearns, M. J., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 325–333, 2016. URL http://papers.nips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Kearns, M. J., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 2569–2577, 2018. URL http://proceedings.mlr.press/v80/kearns18a.html.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293, 2017a.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS*, 2017b.

Lai, V. and Tan, C. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 29–38, 2019.

Lee, M. K. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1): 2053951718756684, 2018.

Lee, M. K. and Baykal, S. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 1035–1048. ACM, 2017.

Lee, M. K., Kim, J. T., and Lizarondo, L. A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 3365–3376. ACM, 2017.

Lee, M. K., Kusbit, D., Kahng, A., Tae, J. S., Yuan, X., Chan, A. D. C., Noothigattu, R., See, D., Lee, S., Psomas, C.-A., and Procaccia, A. D. Webuildai : Participatory framework for fair and efficient algorithmic governance. 2018.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., and Kusbit, D. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

Monahan, J., Metz, A., and Garrett, B. L. Judicial appraisals of risk assessment in sentencing. 2018.

Narayanan, A. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, 2018.

Nielsen, J. *Usability engineering*. Morgan Kaufmann, 1994.

Scurich, N. and Monahan, J. Evidence-based sentencing: Public openness and opposition to using gender, age, and race as risk factors for recidivism. *Law and Human Behavior*, 40(1):36, 2016.

Shah, H. Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170362, 2018.

Vaithianathan, R., Jiang, N., Maloney, T., Nand, P., and Putnam-Hornstein, E. Developing predictive risk models to support child maltreatment hotline screening decisions: Allegheny county methodology and implementation, Mar 2017. URL https://www.alleghenycountyanalytics.us/index.php/2019/05/01/developing-predictive-risk-models-support-child-maltreatment-hotline-screening-decisions/.

Veale, M., Van Kleek, M., and Binns, R. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–14, 2018.

Verma, S. and Rubin, J. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7. IEEE, 2018.

Wang, A. Procedural justice and risk-assessment algorithms. 2018.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., and Schwartz, O. *AI now report 2018*. AI Now Institute at New York University, 2018.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., and Warshaw, J. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 656. ACM, 2018.

Yu, B., Yuan, Y., Terveen, L., Wu, Z. S., and Zhu, H. Designing interfaces to help stakeholders comprehend, navigate, and manage algorithmic trade-offs. *arXiv preprint arXiv:1910.03061*, 2019.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW*, pp. 1171–1180. ACM, 2017.

Zhu, H., Yu, B., Halfaker, A., and Terveen, L. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):194, 2018.