Towards Supporting and Documenting Algorithmic Fairness in the Data Science Workflow

Galen Harrison, Julia Hanson, Blase Ur
University of Chicago
{harrisong, jchanson, blase}@uchicago.edu

***Abstract*—Recent pushes to replace human decision-makers with machine learning models have surfaced concerns about algorithmic fairness. These concerns have led to a quickly growing literature on defining fairness and making models transparent. In practice, the data scientist building the model necessarily must make difficult tradeoffs choosing between imperfect models, balancing different definitions of fairness with accuracy and other considerations. Because these choices have ethical dimensions, there is a need to better support these choices and both document and justify them for the public. We outline a research agenda towards better visualizing difficult fairness-related tradeoffs between competing models, empirically quantifying societal norms about such tradeoffs, and documenting these decisions. We outline how the best practices that result could enable a consumer protection framework for accountable fairness.**

I. Introduction

In recent years, the rise of artificial intelligence and big data has led to increased deployment of automated decision-making systems. Companies and governments have begun to rely on predictive models to make automated decisions about who gets things like jobs, loans, or bail. Automated decision systems seem attractive because they have the potential to make better decisions than a human decision-maker. At the same time, because these sorts of decisions can seriously impact someone's life, there are concerns about whether these systems can make these decisions ethically.

At the core of these automated decision systems are predictive models created by data scientists to estimate outcomes related to the task. For example, data scientists trying to automate loan decisions may create a model predicting the profit from granting a particular loan. When automating a decision, a data scientist creates many different models and compares them, tinkering with what data to include or exclude, what machine learning algorithm to employ, and what parameter settings to use. We further detail this workflow in Section II.

The widespread impact of automated decision systems has raised questions about the fairness of the data and models they use. Specifically, the question of how to evaluate a model's fairness has generated a large literature of competing proposals for how to define fairness statistically (Section III).

A data scientist in pursuit of a 'fair' model must make a series of key decisions about how to define fairness and how to balance fairness with other considerations. A model might maximize fairness by one definition at the cost of another [1]. Further, a data scientist must weigh the model's fairness against its other properties, like its overall accuracy or computational cost.

A data scientist should be aware of the tradeoffs they are making and also be able to clearly communicate and defend these decisions. Existing tools for "explaining" machine learning models focus on enumerating how a particular model makes decisions (Section IV), rather than highlighting nuanced comparisons between competing models.

In this paper, we outline our vision for empowering data scientists to understand, communicate, and document the tensions and tradeoffs made in model selection for automated decision-making. In particular, we highlight in Section V why the following steps are crucial for a framework of consumer protection that builds fairness into the data science workflow:

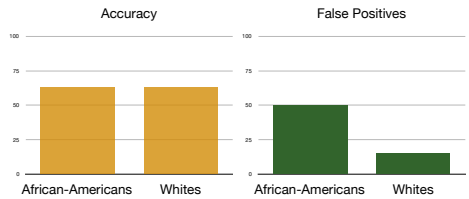
1) Creating information visualizations that compare competing models by fairness and other considerations.
2) Conducting empirical studies collecting opinions about tough tradeoffs made in model selection.
3) Creating protocols and interfaces for documenting why particular decisions were made in model selection.
4) Codifying best practices into a consumer protection framework for fairness in automated decision systems.

II. Fairness in the Data Science Workflow

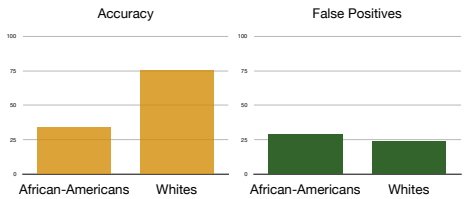
While much of the literature on fair machine learning focuses on quantifying fairness (see Section III), we take a pragmatic view of how fairness will be evaluated and balanced in practice. Fairness, as defined in one particular way, will almost certainly be in tension with other considerations, ranging from model accuracy to other definitions of fairness. In other words, it is highly unlikely that a single model can optimize for every possible definition of fairness, accuracy, computational cost, and all other considerations. The key questions in this paper revolve around how a data scientist would – and should – balance these tradeoffs. Below, we describe the typical data science workflow and then give a simulated example of how fairness may have been considered in one well-studied case.

A. The Structure of the Data Science Workflow

Even when fairness is not a design concern, model building is an iterative process. A typical data science workflow consists of four main phases: *preparation* of data, *analysis* and *reflection* of outputs (of data analysis scripts), and *dissemination* of results, including the model [2]. When using data science



(a) Balancing accuracy.



(b) Balancing the false positive rate.

Fig. 1: A hypothetical choice Northpointe may have faced when deciding which model to use.

to create a predictive model, a data scientist cycles through analysis and reflection phases in a process of trial and error. They create multiple models from prepared data, compare their outputs, tweak the model or create new models according to their findings, and repeat. This tweaking may involve sampling the data in a different way, tuning parameters, or changing the model architecture or loss function.

It is often impossible to maximize all of a model's desirable qualities at once. Thus, a data scientist must prioritize certain metrics above others. These choices may reflect engineering (computational costs) or statistical (accuracy) considerations. However, when the model is being used for a socially significant function, fairness may also be critical. Despite the importance and nuance of these tradeoffs, descriptions of these choices and how they were made are rarely communicated.

B. COMPAS Decisions

To illustrate the difficulty of communicating these choices, consider COMPAS, a predictive model used by judges nationwide to assess defendants' risk of committing more crimes [3]. The fairness of COMPAS was called into question when, in May 2016, ProPublica published an investigation arguing that the model was biased against African Americans [4]. Specifically, their analysis found that the COMPAS algorithm had a higher false-positive rate for African-American defendants than for white defendants. While COMPAS is an archetypal example of algorithmic bias, to our knowledge it has not been analyzed through the lens of model selection.

Northpointe (now Equivant), the company who designed COMPAS, defended their model in a report they published a month later [5]. Their argument relied on the concept of "accuracy equity." That is, they argued COMPAS was fair because it was accurate in predicting recidivism for black and white defendants at similar rates. Thus, it becomes clear that researchers at ProPublica and Northpointe had different definitions of fairness that could not both be satisfied at once.

This raises key unanswered questions: how and when did Northpointe choose their definition of fairness, and what alternatives did they consider? We know from Northpointe's rebuttal that they decided to make the accuracy of their model the same between racial groups. Presumably, at some point Northpointe chose between an equalized accuracy model and models subject to other constraints. Descriptions of, or comparisons to, alternative models are absent from their report.

Hypothetically, Northpointe may have been confronted with the following two options:

1) Equalize accuracy, and accept the consequence of a higher false-positive rate for African-American defendants than for white defendants (see Figure 1a).
2) Constrain the number of false positives, and accept the difference in accuracy between African-American and white defendants (see Figure 1b).

The alternative option in Figure 1b is entirely hypothetical. The option in Figure 1a is accurate only insofar as it was reported that Northpointe equalized accuracy across groups. The public does not know what other model options Northpointe considered, nor why they decided that equalizing accuracy was the correct way to balance their considerations.

Even this relatively constrained scenario is already somewhat complex. There are eight discrete values to keep track of, some of which are comparable to one another, some of which are not. We have not included other arguably pertinent values like false negative rates. We also have limited the number of groups to two and the pertinent category to race. In reality, a data scientist will likely have more than two discrete options. This is why considering and communicating alternatives poses a significant research challenge.

To reason about choices Northpointe made, we argue that data scientists, regulators, and consumers ought to have a clear understanding of the tradeoffs between alternative models and an explanation of how tough choices were made about them. Further, understanding society's beliefs about how these decisions should be made in different contexts will enable data scientists to make better decisions about models, while establishing best practices for documenting these choices could enable a consumer protection framework for fairness.

III. Quantifying Fairness

Fairness is one dimension a data scientist should consider in model selection. Many recent efforts have aimed to characterize and statistically quantify fairness in automated decision systems. There are three aspects of the data science workflow where fairness considerations may be relevant: in the problem construction, the choice of a fairness definition, and the degree of adherence to that definition. Researchers have proposed a number of competing definitions of fairness that imply different, possibly mutually exclusive, understandings of fairness [6]. Whether the application of one of these definitions achieves fairness in practice depends on not just selecting the right definition, but also on applying it sufficiently and in the right context.

A. Problem Construction

A data scientist evaluating a model for fairness needs to decide on more than a metric. They need to identify the groups to whom they are concerned about being fair. Both ProPublica and NorthPointe agreed that race was a meaningful lens through which to examine COMPAS. A data scientist who evaluates a model for fairness with respect to race, but not income or disability, implicitly makes a determination that race is a salient concern, yet income and disability are not. Regardless of whether such assumptions are warranted, they should be made clear.

While the groups chosen for evaluation constitute a highly visible example of how problem construction can influence fairness decisions, it is by no means the only one. Mitchell et al. discuss other aspects of problem construction that can impact fairness [7]. The learning task and the set of outcomes can also affect the fairness of the model.

B. Definition Choice

Several different metrics can be used to measure the fairness of a model. A data scientist seeking to evaluate these must select some subset of these spproaches to apply.

1) Individual Fairness: Individual fairness defines fairness in terms of individual decisions, requiring that similar individuals be treated in a similar fashion [8]. The disparity in outcomes between similar people should then be measured and limited. What constitutes similarity depends on the specific type and context of the data, and it effectively determines the outcome. Individual fairness thus requires a metric of similarity between individuals. For example, a credit score purports to measure a person's creditworthiness, so one can compare two people's creditworthiness by comparing their credit scores. Dwork et al. note that the similarity metric chosen can be controversial [8].

Other definitions of fairness, detailed below, do not necessarily imply individual fairness. The use of a group fairness criterion neither implies that individual fairness is upheld nor that it is violated. However, if individual fairness is achieved, some group fairness criteria may be impossible to achieve.

2) Group Fairness: Group fairness is concerned with outcomes with respect to groups, not individual decisions. For example, while individual fairness might highlight that a particular white individual and a similar black individual receive very different criminal sentences, group fairness might highlight that black defendants have a higher probability of receiving harsh sentences than white defendants. Several methods, including the three below, capture different moral intuitions in group fairness.

Disparate impact is based on the notion in employment discrimination law that a facially neutral test may be discriminatory if it excludes more of one group than another. Feldman et al. show how disparate impact implies a bound on the statistical predictability of the sensitive attribute from the classification [9]. In disparate impact, it is the distribution of outcomes over groups that matter, which is to say that the accuracy of the classifier is potentially ignored.

Hardt et al. instead propose that fairness should consist of equalizing the probability of accurate prediction within subgroups [10]. Zafar et al. generalize this approach by segregating sensitive data to only be necessary when training the model [11]. Heidari et al. have attempted to put various forms of group fairness into a Rawlsian notion of equality of opportunity [12]. Depending on the nature of the underlying population distributions and which specific odds are considered, equalized odds may or may not be reconcilable with disparate impact or individual fairness.

In addition to encoding different notions of what constitutes fairness, certain of these notions are incompatible. Kleinberg et al. found it is only possible to achieve equally accurate risk scores across groups and equally balanced risk quantiles across groups under very specific conditions [13]. Fairness cannot be achieved by applying as many definitions as one can think of. The data scientist will need to make a decision.

3) Process Fairness: In contrast to group and individual fairness, Grgić-Hlača et al. define fairness based on the process by which decisions are made, rather than the outcomes [14]. This notion of process fairness is achieved if the only predictive features used in a model are those that people believe fair to use, quantified empirically via a survey. The survey asks if the respondent deems a particular feature fair to use, fair to use if it increases accuracy, and fair to use even if it increases the likelihoods of false positives for one group. The authors find that one can sometimes achieve both process fairness and outcome fairness, but at an accuracy cost. Process fairness departs from other definitions by explicitly considering input from the general public. Follow-up work seeks to understand why particular features are considered fair or unfair [15].

C. Definition Adherence

Having selected a definition or definitions, the data scientist also needs to determine how strictly the model must adhere to the definitions. Many, if not all, of the algorithmic tools available to achieve a particular fairness definition fulfill the definitions only approximately. In many cases, the closeness of the result to a definition is configurable. For example, the disparate impact measure used by Feldman et al. has a parameter τ, which controls the acceptable difference between positive selection for the majority group and positive selection for the minority group [9]. It is suggested that τ be set to be 80% to be in line with the Equal Employment Opportunity Commission's guidance on disparate impact claims. However, a data scientist who sets it at that level may still create an unfair outcome. Courts have found instances where differences of 1% were sufficient to create a claim under employment anti-discrimination law [16].

Even in instances where the adherence is not directly configured, the data scientist may choose the degree of adherence inadvertently through their pre-processing choices. Friedler et al. conducted a benchmark of various fairness tools and measures, finding that there are significant differences in accuracy and fairness depending on the pre-procecssing methods and algorithms being applied to the data [17].

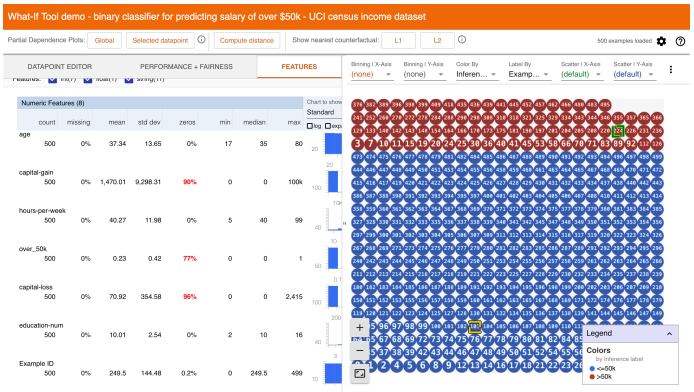


Fig. 2: Google's What-If tool visualizes experiments on machine learning models, aiding in model selection.

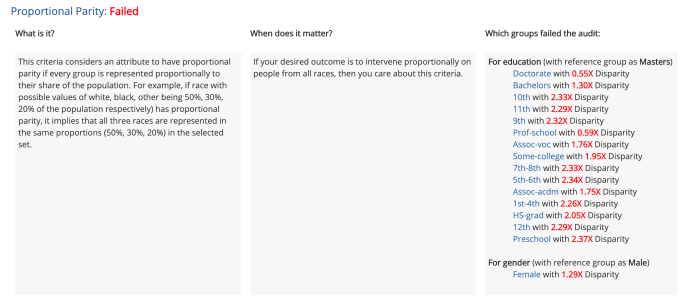


Fig. 3: A portion of a report from the Aequitas tool on the fairness of a given model.

IV. Communicating Fairness

Using any one definition of fairness is itself a decision. Certain definitions may or may not be compatible with others. They may also introduce tradeoffs between equitable outcomes and a model's overall accuracy. In making such decisions, the data scientist may want to consider what the public thinks of these tradeoffs. Once the data scientist has made their decision, they may want to communicate the rationale behind their decision and document the alternatives they considered.

While prior work investigates explaining a given model, to our knowledge no work focuses on how to justify fairness-relevant decisions in model selection. Prior work on model transparency aimed to elucidate how a given model operates, whereas our goal is to communicate why one model was chosen over other options. Understanding what a model is doing is the first step, while we are concerned with the subsequent step. In this section, we summarize current research into model transparency and model explanations, showing why it is insufficient on its own for our ultimate vision.

A. Visualizations of Fairness

An example of a fairness visualization aimed at data scientists is Google's What-If tool [18], shown in Figure 2. The tool facilitates exploratory analysis in model selection, highlighting the impacts of tuning hyperparameters and excluding particular variables on model accuracy. Its visualizations of how these considerations impact different demographic groups can be used to analyze fairness in part. Because it is an exploratory tool, it does not attempt to communicate any sort of set of alternatives, nor is it designed to engage non-expert users.

Fig. 4: IBM's AI Fairness 360 tool at the bias-detection stage.

In contrast, the Aequitas tool focuses on reporting the biases of a model [19]. Given a single model, Aequitas runs tests for a set of fairness definitions like those discussed in Section III-B2. While more straightforward for a non-technical person to understand (see Figure 3), Aequitas reports only on a single model. An Aequitas report does not clarify why the data scientists who built the model made the decisions that led to particular outcomes. Our vision instead requires that future work develop expanded tools for comparing among many models and documenting the difficult decisions made.

Last year, IBM launched AI Fairness 360 (AIF360), an open-source collection of tools that, like Aequitas, run tests for fairness metrics, but also include a suite of possible algorithms for mitigating bias [20]. Along with the toolkit code, the AIF360 website hosts interactive demos of the tools in practice (see figure 4). We observe that the comparisons presented in these demos are limited to pairwise "before and after" views of a model after a single mitigation step. Comparisons between mitigation steps are absent.

B. Transparency and Explanations

Another line of research focuses on making models more transparent. For example, QII quantifies the relative influence of particular variables over the output of a black-box model [21]. For example, QII can show how much race influences the model's decision, and how that compares with the effects of gender. Another transparency tool is Locally Interpretable Model Explanations (LIME) [22]. LIME fits an interpretable model to the area directly around a particular data point. The interpretable model can then be used to explain how the more complicated model made its decisions.

Explaining a single model usually does not provide a mechanism on its own for contrasting different models. Ribeiro et al. conducted a limited user study in which participants used LIME to choose between two models. However, the decision was between a model that generalizes well and one that does not; no ethical choice was involved. We instead propose that when communicating the choice of one imperfect model over other imperfect alternatives, the data scientist should carefully document and justify their choice.

C. Usability of Explanations

Work has started to consider how users understand and interpret model explanations. This literature focuses on effective explanations for individual classifications, whereas any choice a data scientist makes will necessarily be about distributions of classifications. For example, Binns et al. examined whether the type of explanation a person received about a decision affected how they perceived the justness of that decision [23]. They tested explanations based on QII, the amount the input would need to change to affect the output, the most similar case in the training data, and aggregate statistics by category. They found that the style of the explanation only affected perceptions of justice when someone was exposed to multiple different explanations. Unfortunately, an explanation for a single classification on its own communicates little about the choices the data scientist made.

Krause et al. tried to bridge the gap between individual explanations and overall patterns by aggregating explanations provided by LIME [24]. They validated their interface by testing whether participants could identify a deliberately biased model. Aggregating explanations may play a role in comparing two models, but is again insufficient on its own for justifying difficult decisions in model selection.

V. An Agenda for Accountable Model Selection

To remedy the lack of attention paid to supporting, documenting, and communicating data scientists' difficult fairness-related decisions in model selection, we propose four lines of research. First, we propose further research into how to best visualize choices data scientists must make between models. Second, we suggest that data scientists be better supported in these difficult choices by quantifying public opinion and social norms through empirical user research. Third, we propose documentation requirements for explaining and justifying the selection of a particular model based on these visualized tradeoffs. Fourth, we highlight how these techniques can form the basis of a consumer protection framework.

A. Visualizing Model Choices

A first step in our proposed agenda of future work is to develop visualizations for model comparisons that efficiently convey key tradeoffs to a data scientist. A related requirement is determining how to convey differences in the effects of different fairness criteria. For example, a data scientist may want to compare the effect of applying process fairness with the effect of applying equalized odds. What aspects of the model does this data scientist care about? Do we need to compare process fairness in terms of equalized odds, and vice versa? Do the answers to these questions change depending on where one is in the data science workflow? What aspects of the model are most relevant to the public, and how can we emphasize those?

Furthermore, visualizations that are useful for a data scientist may be less helpful for a non-technical audience. Data scientists exploring alternative models may want to have things like counterfactuals or information about the training data available to them. These aspects may be confusing or irrelevant to a non-technical audience.

There are two possible approaches to identifying and eliciting these sorts of considerations. One could approach it as a scientific question, by seeking to develop generalizable rules about how specific kinds of tasks are considered. Another approach is through a design process. In contrast to a scientific approach, a design process elicits considerations for a particular context through iteration and dialogue between the designers and the user base.

Because we expect fairness considerations to be fairly context dependent, a design process is in some ways ideal. However, from a regulatory perspective design processes are harder to assess. How would a regulator be able to identify when a design process properly balanced competing interests, for example? One possible synthesis of these approaches would be to use scientific methods in making a prima facie case of a hypothetical violation. For example, a data scientist documenting the decision in a manner that contravened scientifically produced guidance would be subject to a presumption of illegality. This presumption could be rebutted by producing evidence of a design process.

Pursuant to this goal, we suggest that it should be a research priority to identify what aspects of a machine learning model matter to the general public, and when. There is a strong expectation that the salient fairness qualities in things like bail decisions are different from the salient fairness qualities in an employment context. While the diversity of contexts may seem to make this task difficult, it need not. Employment, the criminal justice system, and credit issuance are all examples of contexts that likely differ in dynamics. While the particular learning task may differ from instance to instance, a broad investigation of the qualities that are significant in fairness assessments in those contexts could guide the investigation. An experimental result on the sorts of salient qualities could then be used by the enforcing agency to identify instances where the documentation is insufficient.

B. Empirically Supporting Difficult Choices

Grgić-Hlača et al.'s idea of process fairness encodes the intuition that a diverse group of voices may be able to make better determinations about fairness than a small group of data scientists. While they focused on usings surveys to document whether laypeople considered different predictor variables fair in a particular decision, we propose extending their idea to empirically documenting laypeople's attitudes concerning the difficult decisions the data scientist will necessarily face. For example, given the choice between 5% greater accuracy overall (with higher inter-group disparity) or lower inter-group disparity (with lower accuracy), which does the public find more fair? The results of such empirical studies could be built into our proposed comparison interfaces to better ground data scientists' difficult decisions in data.

Supporting a decision with data is not the same as determining a decision with data. The public writ large may have biases and prejudices that indicate an ethically impermissible

course of action. Choice documentation is meant to facilitate a dialogue between the data scientist and the affected population. Building an understanding of how the public feels, and why, will enable the data scientist to make and justify decisions that go against general public sentiment.

C. Explaining Decisions

Being able to effectively document and rationalize difficult choices in model selection will facilitate better communication with laypeople potentially affected by the model's decisions. With visualization tools to explain differences between models, data scientists can solicit and use findings on public opinion about model fairness tradeoffs to justify their choices.

Model selection explanations could also be incorporated to improve classification explanations. For example, the plausible explanation "you were classified as a high insurance risk because you are a man, are between the ages of 18-22, and were born in April" communicates that gender, age, and birth month were the criteria used to determine his insurance risk, but fails to explain why those criteria were used. Including descriptions of the decisions made between model alternatives could make this explanation more satisfying. A better explanation could say, for example, that a data scientist chose to include birth month as a variable in their model because its inclusion equalized accuracy between men and women compared to other models that were otherwise similar.

While explanations can empower users, they can also mislead them. Lying with statistics, and in particular with statistical visualizations, is a well-known phenomenon. Data scientists fearing backlash may seek to mislead the public to obscure or reframe the decisions they made in a positive light. Notably, researchers have found that most Facebook ad-targeting explanations are incomplete and vague [25].

A related line of investigation should thus focus on identifying techniques that could be misused to present a decision in misleading ways. While heavily implicated in the visualization of the model qualities, this investigation needs to also look beyond the way the information is presented to the choice of information being presented. For example, one could imagine specifically choosing bad models to present next to the one ultimately chosen to take advantage of anchoring.

VI. Consumer Protection for Fairness in ML

Research into tradeoff visualization, empirical support of difficult decisions, and decision documentation can form the basis of a series of best practices for accountably fair model selection. Best practices crafted along these dimensions could then be used by regulators, legislators, and courts as part of a consumer protection framework for fairness in automated decision-making. In the presence of a rule requiring a good-faith process of discrimination mitigation, insights into communicating about fairness considerations could be used as a template. Furthermore, research into best practices will give regulators and civil society groups a point of comparison when bringing complaints. Even absent regulatory intervention, companies seeking to voluntarily engage in best practices could benefit from inquiry into these questions. A company may seek to dispel concern about fairness in their products by demonstrating the steps and considerations they have used. These benefits can be distilled into two related, but distinct, objectives: effective process documentation and understanding consumer fairness expectations.

A. Process Documentation

Documenting the process by which data scientists selected a particular model for automated decision-making is crucial for transparency. Even if affected individuals receive an explanation of why the model made a given decision the way it did, they may very well ask why they were subject to a decision by that model, and not an alternative model, in the first place. Without an explanation of why this specific model was used, the picture provided to the public will necessarily be incomplete. In our research agenda, best practices for model comparison and choice documentation would form the basis for regulatory intervention. The communication practices we propose will facilitate regulatory scrutiny into whether data scientists' judgments erred, or whether the data scientist picked a reasonable model from among a suite of imperfect alternatives. The documentation can also form the basis of a preemptive defense for the data scientists.

In order to fulfill this role, the documentation must be understandable by a non-technical public. If the fairness decisions the data scientist made are obscured, the only way to convince a data subject that the model is actually correct is by appealing to the data scientist's wisdom and knowledge. This is likely to be unconvincing. Alternatively, the data scientist may have misjudged the correct course of action. Clear and publicly available process documentation will assist in discovering and correcting these misjudgments.

B. Consumer Fairness Expectations

Effective communication about model decisions will not only help data scientists explain their decisions, but also help them make these decisions. In a situation where the data scientist is faced with multiple imperfect options, it may be useful for them to be shown empirical documentation of public opinion on a similar choice in order to inform their decision. Understanding consumer fairness expectations will help them avoid situations where the behavior they encode does not match public beliefs about how the model should operate. Because the decision the data scientist is faced with balances many factors, effective communication about the decision is necessary to gather meaningful feedback. Soliciting opinions can play an additional role in process documentation, increasing the accountability of the process to public opinion.

VII. Conclusion

As data scientists build automated decision systems for widespread deployment, they make a number of decisions in model selection that are opaque to the public and may encode only the data scientist's expectation of fairness, rather than empirically documented societal norms. We have proposed a

research agenda for supporting, documenting, and regulating fairness-related decisions as part of the complex process of model selection in the data science workflow. The best practices we expect to be identified can then support a consumer protection framework for algorithmic fairness.

REFERENCES

[1] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017, http://arxiv.org/abs/1610.07524.

[2] P. Guo, "Data science workflow: Overview and challenges," CACM blog, 2013, https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext.

[3] R. Courtland, "Bias detectives: The researchers striving to make algorithms fair," *Nature*, vol. 558, p. 357, Jun. 2018, http://www.nature.com/articles/d41586-018-05469-3.

[4] J. Angwin and J. Larson, "Machine bias," ProPublica, May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[5] W. Dieterich, C. Mendoza, and T. Brennan, "COMPAS risk scales: Demonstrating accuracy equity and predictive parity," Northpointe Inc., 2016, http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

[6] A. Narayanan, "Tutorial: 21 fairness definitions and their politics," 2018, https://www.youtube.com/watch?v=jIXIuYdnyyk.

[7] S. Mitchell, E. Potash, and S. Barocas, "Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions," Nov. 2018, https://arxiv.org/abs/1811.07867v1.

[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. ITCS*, Jan. 2012, http://arxiv.org/abs/1104.3913.

[9] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," 2014, http://arxiv.org/abs/1412.3756.

[10] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. NIPS*, 2016.

[11] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proc. WWW*, 2017.

[12] H. Heidari, M. Loi, K. P. Gummadi, and A. Krause, "A moral framework for understanding of fair ML through economic models of equality of opportunity," 2018, http://arxiv.org/abs/1809.03400.

[13] J. M. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," 2016, http://arxiv.org/abs/1609.05807.

[14] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller, "Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning," in *Proc. AAAI*, 2018.

[15] N. Grgić-Hlača, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *Proc. WWW*, 2018.

[16] "Jones v. City of Boston," Court of Appeals, 1st Circuit, F. 3d, No. 15-2015, 2016.

[17] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proc. FAT**, 2019.

[18] "What-If Tool," https://pair-code.github.io/what-if-tool/.

[19] "Aequitas - The bias report," http://aequitas.dssg.io/.

[20] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," 2018, https://arxiv.org/abs/1810.01943.

[21] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *Proc. IEEE S&P*, 2016.

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. KDD*, 2016.

[23] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'It's reducing a human being to a percentage'; perceptions of justice in algorithmic decisions," in *Proc. CHI*, 2018.

[24] J. Krause, A. Perer, and E. Bertini, "A user study on the effect of aggregating explanations for interpreting machine learning models: [Work-in-Progress]," in *Proc. IDEA @ KDD*, 2018.

[25] A. Andreou, G. Venkatadri, O. Goga, K. P. Gummadi, P. Loiseau, and A. Mislove, "Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations," in *Proc. NDSS*, 2018.