Toward Algorithmic Accountability in Public Services

A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services

Anna Brown
Massey University
Wellington, New Zealand
a.e.brown@massey.ac.nz

Alexandra Chouldechova
Carnegie Mellon University
Pittsburgh, USA
achould@cmu.edu

Emily Putnam-Hornstein
University of Southern California
Los Angeles, USA
hornste@usc.edu

Andrew Tobin
Massey University
Wellington, New Zealand
a.tobin@massey.ac.nz

Rhema Vaithianathan
Auckland University of Technology
Auckland, New Zealand
rhema.vaithianathan@aut.ac.nz

ABSTRACT

Algorithmic decision-making systems are increasingly being adopted by government public service agencies. Researchers, policy experts, and civil rights groups have all voiced concerns that such systems are being deployed without adequate consideration of potential harms, disparate impacts, and public accountability practices. Yet little is known about the concerns of those most likely to be affected by these systems. We report on workshops conducted to learn about the concerns of affected communities in the context of child welfare services. The workshops involved 83 study participants including families involved in the child welfare system, employees of child welfare agencies, and service providers. Our findings indicate that general distrust in the existing system contributes significantly to low comfort in algorithmic decision-making. We identify strategies for improving comfort through greater transparency and improved communication strategies. We discuss the implications of our study for accountable algorithm design for child welfare applications.

CCS CONCEPTS

• **Information systems → Decision support systems**; • **Applied computing → Computing in government**;

KEYWORDS

Algorithmic Accountability; Algorithmic Bias; Child Welfare Services; Decision-Support; Automated Decision Systems; Participatory Design

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300271

1 INTRODUCTION

Algorithmic decision-making systems are increasingly being adopted by governments in an effort to improve and reform existing public service processes. Decisions about where to police [7], whom to detain[28], which child maltreatment allegations to investigate [20], whom to grant access to permanent supportive housing [41], and what level of unemployment benefit to provide [31] are all being informed—if not performed—by algorithmic systems. These technologies and the opaque manner in which they are deployed have drawn tremendous criticism from researchers [6, 14, 16, 18], policymakers [32], civil rights groups[11], and the media[3]. There is concern that data-driven algorithmic tools may be perpetuating discriminatory practices and having unintended consequences, all while operating outside the scope of traditional oversight and public accountability mechanisms.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300271

Several recent proposals have been put forth in an effort to establish good operating practices for promoting 'algorithmic accountability' [13, 24, 36, 37, 40]. These all offer recommendations for what public agencies can do to enable stakeholders and the public to assess envisaged systems and engage in debate over whether their use is acceptable. Many of the proposals indicate that agencies should specifically seek to engage with affected communities—individuals who are most likely to be subject to, or impacted by, the deployment of algorithmic systems. This helps to ensure that the proposed system will meet the needs of those most directly affected by it, and bestows a kind of 'license to operate'[40].

While public service agencies routinely engage affected communities in deliberative processes, these efforts are generally carried out on an ad hoc basis. The resulting findings remain internal to the given agency, and thus fall short of providing generalizable knowledge from which a broader understanding of effective design strategies could emerge. This paper seeks to take initial steps in bridging this gap.

As part of a broader participatory design effort, we conducted workshops to learn about the concerns of affected communities in the context of child welfare services. These communities include families involved in the child welfare system, and social workers whose professional roles will be impacted by the introduction of algorithmic systems. The work presented here reflects a pilot study conducted with the further intent of designing (with the same participant communities) a blueprint to aid government agencies and data scientists in improving community comfort levels with algorithmic decision-making. Our study was designed to answer three central questions:

(i) How do people who are most likely to be subject to or affected by algorithmic decision-making feel about the deployment of such systems?
(ii) What are the primary sources of community discomfort surrounding the development and deployment of such tools?
(iii) What can researchers and designers do in the development and deployment stages to raise comfort levels among affected communities?

In addition to this research article, we have produced an in-depth report on the study's findings intended for more general consumption. This report contains many more participant quotes and other information that could not be accommodated by the conference proceedings format. The report will be made available for download from the authors' academic website.

2 BACKGROUND AND MOTIVATION

While we are not the first to explore issues of algorithmic accountability in the public sector, a distinguishing feature of our work is that it is directly motivated by ongoing efforts by members of the research team to develop and deploy algorithmic tools for use in child welfare. These tools are intended to inform service delivery and investigation decisions to better focus limited resources on the riskiest and neediest cases. This work has significant buy-in from agency leadership, and is intended to inform the next stages of algorithm design.

The use of algorithmic tools in child welfare is a contentious issue. Several recent attempts at deploying algorithmic systems in child welfare have met with scathing criticism that ultimately resulted in their termination. In December 2017 the Illinois Department of Children and Family Services announced that they were terminating their predictive analytics program for reasons including the perception by DCFS staff that predictions were unreliable, issues with how the predictions were communicated, data quality, and questions surrounding the initial procurement process [21]. Earlier that same year the County of Los Angeles Office of Child Protection released a report examining the use of predictive analytics in assessing child safety and risk in which they cite the black-box nature of proprietary tools and high false positive and false negative error rates as factors driving the county's decision to terminate their project. Studies such as our own can serve to inform more effective and accountable uses of algorithmic systems by helping to identify affected community concerns during the design phase.

Algorithmic systems of the kind we consider here are in many cases already governed by existing data protection regulations such as those articulated in the EU General Data Protection Regulation (GDPR) and Canada's Personal Information Protection and Electronic Documents Act (PIPEDA). Such regulations are primarily intended to govern private sector data use and commonly offer exceptions for public sector use. To the extent that they apply in a given instance, they provide data subject rights, and dictate necessary criteria for data use and storage. Necessary, however, is not sufficient.

Algorithmic systems that are deployed in full compliance with the existing regulation may nevertheless fail to have so-called "license to operate", also known as "social license". In the industrial context, this refers to community and stakeholder acceptance of a company's operating procedures business practices. As Shah [40] argues, without such licence from the public, the promise of algorithmic systems for promoting positive social change may fail to be fully realized. Better understanding affected community concerns is an important step in building toward trust and social license.

Our study is informed by work characterizing the structure of organizational justice perceptions. Colquitt [12] identifies four central components of justice perceptions: distributive (how equitably resources are distributed); procedural (the fairness, consistency and accuracy of the decision process); interpersonal (feelings that one is treated with due respect); and informational (perceived truthfulness and comprehensiveness of justifications provided). Procedural justice is a particularly important concept, as it has been found to mitigate the relationship between outcome favorability and support for decisions [8, 9]. Outcomes in the child welfare system are often unfavorable to families. Understanding how algorithmic systems affect perceptions of procedural justice is therefore central to design and deployment considerations.

We build upon a growing body of work in the HCI community that has explored perceptions of fairness and procedural justice in algorithmic systems [1, 5, 25–27, 35, 44]. Notably, the recent work of [44] reports on a series of workshops that presented participants with scenarios based on internet-related products and services. The scenarios were constructed to reflect different forms of discrimination, stereotyping and exclusion on the basis of race, sex, or geographic location. Study participants were engaged in extensive discussion of their reactions to the instances of "algorithmic bias". The authors specifically recruited participants from traditionally marginalized populations, including racial and ethnic minorities and persons of low socioeconomic status. This focussed the study on those populations most likely to be affected by algorithmic bias in internet services.

Several recent studies have investigated public [17, 39, 43] and practitioner [29, 42] perspectives on the use of algorithmic systems in the context of public sector decision making. Scurich and Monahan [39] and Grgić-Hlača et al. [17] investigate the narrower question how the inclusion of particular features influences public perceptions of justice and fairness. Wang [43] studies public perceptions of procedural justice more broadly in the context of the criminal justice system. These studies of public perception tend to focus on nationally representative samples of the general population, or rely on MTurk workers. In both cases, the participant samples are neither representative of nor targeted towards populations that have direct experience with the public system in question. Participants in our study frequently invoked personal experiences with the child welfare system, which suggests that such first-hand experience is important to informing perceptions. Furthermore, we are not aware of any prior studies that, like our own, involve both affected members of the public and practitioners within the system.

3 METHODOLOGY

The present study reflects just one element of a broader participatory design effort to improve child welfare outcomes through the use of data-driven algorithmic tools. Participatory design methodologies favor a human-centred approach that engages the participation of people most likely to be affected by the services and policies being designed. Within participatory design sits an established set of practices, tools and techniques to explore and understand the 'lived experience' of everyday people [38]. These methods have been successfully applied to address technological design challenges in the public sector. Notably, research [4] commissioned by the Data Futures Partnership formed the basis of national guidelines for New Zealand organizations to obtain social license for data use [34].

Participatory design is characterised by generative, experiential and action-based methods that place emphasis on embodied learning and explorations into future scenarios [22]. To study community perspectives, we adopted a methodology where:

- Meaningful conversations are facilitated with affected communities using scenarios that model a real life situation while exploring data use variables of interest to the study;
- Participants are invited to deliberatively demonstrate their levels of comfort with scenarios using a commonsense trust/benefit matrix;
- Participants are invited to identify what would increase their comfort in the data use and sharing scenarios explored; and
- Participants identify and prioritize the themes and concerns that they consider most important to increasing comfort.

Our study received full IRB approval from all member institutions involved in the human subjects research and primary data analysis. Study participants were provided with an informed consent form at the beginning of the workshop session. In devising the study we consulted with ethics specialists at the local child welfare agencies in order to ensure that our design minimized the risk of unintended harm to study participants, and facilitated an environment in which participants felt comfortable expressing their beliefs and opinions.

Participants

Four participant samples were recruited from a single jurisdiction in a mid-sized US county. The three main groups included in the full analysis are:

- **Families** (n = 18) Two workshops of family participants were recruited through community providers. A condition of recruitment was that peer support people would attend and participate in the discussions. Their responses have been included in the sample and the findings. Four of the eighteen participants in the families workshops were peer support people.
- **Frontline Providers** (n = 38) Three workshops included child welfare worker participants who were in regular contact with families. These workshops were

variably peopled by family support people, peer support people and case workers. These participants were recruited through provider networks by the relevant local state agency.
- **Specialists** ($n = 11$) Specialist providers in the sample above were a subgroup who attended the provider workshop whose roles were identified mostly as not involving regular contact with families. Types of specialists included data workers, supervisors and office managers.

A fourth sample of **Prototype Specialists** included $n = 16$ participants from the city's child protective services head office employed in roles not involving regular contact with families. These participants took part in an earlier prototype of the workshop scenarios, and their perspectives have been included in the findings. Since the prototype scenarios were slightly different from those that all other participants responded to, responses from this group have been filtered to remove any that were attributable to scenario differences.

Family participants were recruited through community providers whose peer support staff approached families with invitations to participate in the workshops. These participants were offered childcare for the full duration of the workshop, a light pizza dinner, and modest compensation for their time. Frontline providers and specialists were recruited through the local human services agency and by community providers. Most of the participants were women, which in the frontline provider and specialist groups is consistent with the over-representation of women in social work and related professions.

Recruitment for one of the workshops specifically focused on participants identifying as persons of colour. While prior studies have found significant differences in perceptions of procedural justice across racial and ethnic groups [19, 33], the responses given in our study were similar overall. The main notable differences arose when participants described their personal experiences as persons of color affected by the child welfare system.

Workshops

We developed a series of progressive scenarios designed to explore variations of interest to the research team concerning the use of algorithmic decision-making by public service agencies. Specifically, the scenarios varied across three distinct dimensions:

- **Decision making framework**: We varied whether the decision was made by a human (human), by a human assisted by an algorithm (algorithm-assisted), or by an automated algorithmic decision-making system (algorithmic).
- **Proactive vs. reactive**: We considered both *reactive* scenarios, where the decision was prompted by a report made to child welfare services, and *proactive* scenarios, where an individual becomes eligible for support or services on the basis of an algorithmic assessment.
- **Data sources**: We varied the breadth of data that was used to inform the decision. Most narrowly, we considered scenarios where only the family's own child welfare data factored into the decision. Later scenarios introduced child welfare data from associates of the family, data from other administrative systems (criminal justice), and local community data not specific to the family (e.g., neighbourhood).

The scenarios were tested and refined via a prototype audience of US child welfare specialists to ensure that they were understandable and relatable. The structure of the final scenarios is summarized below.[1]

- **Scenario 1a**
  Reactive + human decision-making
  Presents a child welfare decision being made by an intake worker in reaction to a call from a member of the public. There is no mention of personal data or a computer tool being involved in making the decision.
- **Scenario 1b**
  Reactive + algorithm-assisted decision-making
  Presents a child welfare action that is reactive, and a decision that is made by a human assisted by an algorithm.
- **Scenario 1c**
  Reactive + algorithmic decision-making + using family's child welfare data
  Presents a child welfare action that is reactive, and a decision that is made by an algorithm including associative data (child welfare investigation).
- **Scenario 1d(i), (ii) & (iii)**
  Proactive + algorithmic decision-making + using family's child welfare data
  Presents a child welfare action that is proactive, and a decision to offer services that is made by an algorithm — with three different ways of communicating this decision to the family.
- **Scenario 1e**
  Reactive + algorithmic decision-making + using administrative data beyond child welfare
  Presents a child welfare action that is reactive, and a decision that is made by an algorithm including associative data (criminal justice data).

[1] Note that the scenarios are all enumerated as "1X". This enumeration reflects our intention to explore other scenarios in later studies. Those scenarios will be coded as "2X", "3X" and so forth.

- **Scenario 1f**
  Proactive + algorithmic decision-making + using administrative and community data
  Presents a child welfare action that is proactive, and a decision that is made by an algorithm including non-associative data (criminal records, neighborhood, age).

Figure 1 displays all six of the scenarios as they were shown to the main workshop participants (those shown to the prototype sample were slightly different). The first scenario outlines a situation that is constructed to be most familiar to the participants' lived experience. This familiarity and engagement is then leveraged to incrementally introduce scenario elements that participants have not experienced, but which they can understand and respond to in the context of the evolving narrative.

The participants at each workshop were split into two groups, each of which was guided by a research facilitator. All participants were provided with a print-out of the scenarios for their reference throughout the discussion. Participants were instructed to not view later scenarios before the current one was finished being discussed. Each scenario was read out by the group's research facilitator. Certain parts of each scenario description were bolded for emphasis. These corresponded to the specific decision and action that was taken in the scenario. However, participant perspectives could be—and were—influenced by any part of the text.

Upon being presented with the given scenario, participants were then asked to place an identifier token on a *comfort board* to indicate their perceived level of 'benefit' and 'trust' in the situation. Here, 'benefit' refers to the amount of benefit they see arising as a result of the decision or action taken; and 'trust' refers to how much trust they have in the decision, action, or system actors. Participants were asked to reflect on the scenarios from the the perspective of an affected parent (for later scenarios, grandparent). Figure 2 shows a photo of the comfort board for one of the groups in response to scenario 1b.

After taking positions on the board, participants were engaged in a facilitated conversation about their perceived levels of comfort with the child welfare decision-making processes and actions in the given scenario. They were then asked to identify what, if anything, would increase their levels of comfort. Upon completing the final scenario the participants were engaged in a broader discussion aimed at synthesizing and prioritizing overarching concerns and recommendations for system design.

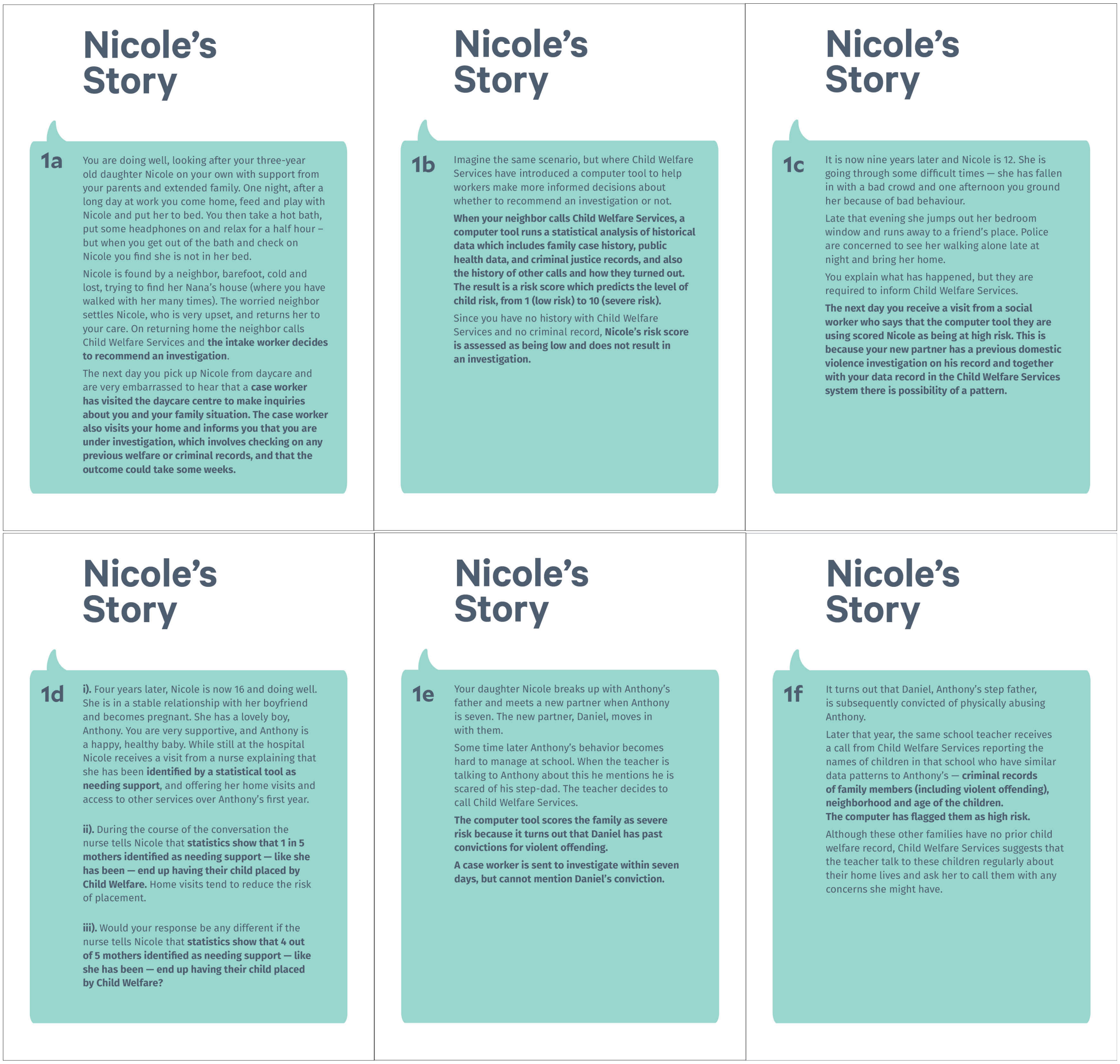
Data collection and analysis

Three sets of data were collected. Data set 1 was composed of photographs of the positions of participants' counters comfort map in response to each scenario. Data set 2 consisted of verbal comments made by workshop participants which were noted verbatim by facilitators and coded according to which scenario they applied to. Data set 3 was composed of summative comments written by participants on sticky notes at the conclusion of each workshop in answer to the question, "Considering all of the scenarios we've discussed, what are the most important things that are needed to increase your comfort with the use of data and computer tools in social welfare decision-making?" Participants were facilitated to sort these comments into thematic groups on a sheet, give each theme a label, and prioritize themes according to their relative importance.

At the conclusion of workshops, Data set 1 (visual comfort maps) was manually translated into derived comfort map matrices (such as shown in Figure 2) that could be examined to identify patterns across scenarios and participants. Data set 2 (verbatim participant comments) was transcribed by each researcher (individually) into a spreadsheet which could be also be filtered by scenario and participant type. Individual researchers used grounded theory to inductively code, conceptualize and categorize data from their workshops, and then came together to inductively assess common and unique findings. At this point, data set 3 (themes self-identified by participants being most important) was inductively analyzed by the research team, and these codes, concepts and categories were compared with those resulting from team analysis of data set 2. Finally, the research team drew theoretical conclusions drawing on all data sets, with the lead researcher reporting these conclusions and evidencing them with verbatim comments and visual comfort maps.

Study limitations

Before proceeding to our main findings, we highlight several limitations of our study that are important to bear in mind while interpreting the results. First, the recruitment strategy was one of convenience sampling. This means that our study sample is almost surely not representative of the full population of individuals affected by the county's child welfare system. Second, the study was conducted in a US county that is already in the process of adopting algorithmic decision-support tools in the child welfare context. The specialist subgroup in particular consists of child welfare workers who may already be familiar with the county's approach. This group is thus likely to be much more informed about the relevant issues compared to workers in other jurisdictions. Lastly, the scenarios we explored are all narrowly focused on algorithmic decision making in the US child welfare system. Our findings cannot directly speak to concerns or experiences of affected communities involved in other systems such criminal justice, homelessness, or unemployment services.



Nicole's Story

1a You are doing well, looking after your three-year old daughter Nicole on your own with support from your parents and extended family. One night, after a long day at work you come home, feed and play with Nicole and put her to bed. You then take a hot bath, put some headphones on and relax for a half hour – but when you get out of the bath and check on Nicole you find she is not in her bed.

Nicole is found by a neighbor, barefoot, cold and lost, trying to find her Nana's house (where you have walked with her many times). The worried neighbor settles Nicole, who is very upset, and returns her to your care. On returning home the neighbor calls Child Welfare Services and **the intake worker decides to recommend an investigation**.

The next day you pick up Nicole from daycare and are very embarrassed to hear that a **case worker has visited the daycare centre to make inquiries about you and your family situation. The case worker also visits your home and informs you that you are under investigation, which involves checking on any previous welfare or criminal records, and that the outcome could take some weeks.**

Nicole's Story

1b Imagine the same scenario, but where Child Welfare Services have introduced a computer tool to help workers make more informed decisions about whether to recommend an investigation or not.

When your neighbor calls Child Welfare Services, a computer tool runs a statistical analysis of historical data which includes family case history, public health data, and criminal justice records, and also the history of other calls and how they turned out. The result is a risk score which predicts the level of child risk, from 1 (low risk) to 10 (severe risk).

Since you have no history with Child Welfare Services and no criminal record, **Nicole's risk score is assessed as being low and does not result in an investigation.**

Nicole's Story

1c It is now nine years later and Nicole is 12. She is going through some difficult times — she has fallen in with a bad crowd and one afternoon you ground her because of bad behaviour.

Late that evening she jumps out her bedroom window and runs away to a friend's place. Police are concerned to see her walking alone late at night and bring her home.

You explain what has happened, but they are required to inform Child Welfare Services.

The next day you receive a visit from a social worker who says that the computer tool they are using scored Nicole as being at high risk. This is because your new partner has a previous domestic violence investigation on his record and together with your data record in the Child Welfare Services system there is possibility of a pattern.

Nicole's Story

1d **i).** Four years later, Nicole is now 16 and doing well. She is in a stable relationship with her boyfriend and becomes pregnant. She has a lovely boy, Anthony. You are very supportive, and Anthony is a happy, healthy baby. While still at the hospital Nicole receives a visit from a nurse explaining that she has been **identified by a statistical tool as needing support**, and offering her home visits and access to other services over Anthony's first year.

ii). During the course of the conversation the nurse tells Nicole that **statistics show that 1 in 5 mothers identified as needing support — like she has been — end up having their child placed by Child Welfare.** Home visits tend to reduce the risk of placement.

iii). Would your response be any different if the nurse tells Nicole that **statistics show that 4 out of 5 mothers identified as needing support — like she has been — end up having their child placed by Child Welfare?**

Nicole's Story

1e Your daughter Nicole breaks up with Anthony's father and meets a new partner when Anthony is seven. The new partner, Daniel, moves in with them.

Some time later Anthony's behavior becomes hard to manage at school. When the teacher is talking to Anthony about this he mentions he is scared of his step-dad. The teacher decides to call Child Welfare Services.

The computer tool scores the family as severe risk because it turns out that Daniel has past convictions for violent offending.

A case worker is sent to investigate within seven days, but cannot mention Daniel's conviction.

Nicole's Story

1f It turns out that Daniel, Anthony's step father, is subsequently convicted of physically abusing Anthony.

Later that year, the same school teacher receives a call from Child Welfare Services reporting the names of children in that school who have similar data patterns to Anthony's — **criminal records of family members (including violent offending), neighborhood and age of the children. The computer has flagged them as high risk.**

Although these other families have no prior child welfare record, Child Welfare Services suggests that the teacher talk to these children regularly about their home lives and ask her to call them with any concerns she might have.

Figure 1: All six scenarios as presented to study participants (see Section 3 for more information). During the workshops each participant was provided with a printout of the scenarios with one scenario appearing per page. Participants were instructed to not advance to later scenarios until the current one had been fully discussed.

We also concede that a full participatory design framework would ideally entail designers, families, frontline providers and specialists all gathering together on an equal basis to learn from each other. For this study, however, we accepted guidance from ethics specialists in the local child welfare agencies that there was a risk of many forms of bias from basing our pilot on such an approach. This was a an especially significant concern with respect to families, whose responses, it was feared, may be unduly influenced or censored in a mixed-group approach. Indeed, we included peer support in the family workshops precisely to further facilitate greater freedom of dialogue among family participants.

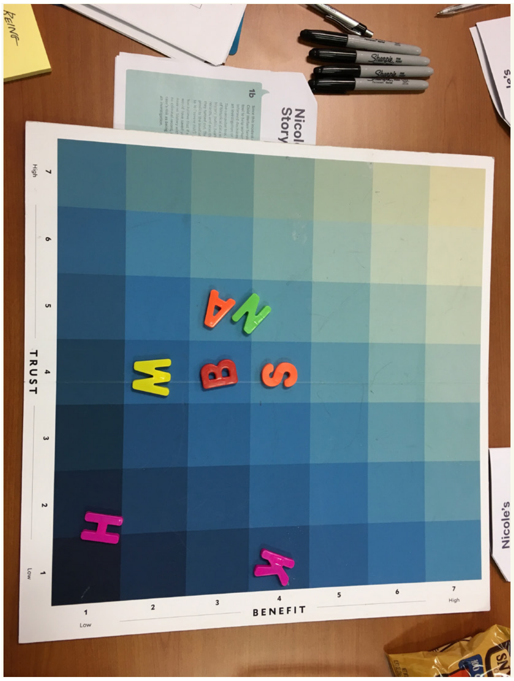


Figure 2: Comfort board for one of the workshop participant groups in response to scenario 1b. Each letter token on the board is a unique participant identifier that is retained across all of the scenarios. The gradient shading on the comfort board is intended to reflect different levels of overall 'comfort' with the scenario. Lighter colors correspond to greater overall comfort.

4 FINDINGS

In this section we summarize several of the major themes that emerged from the workshops. Overall, participants tended towards expressing low levels of comfort across the range of scenarios. Reasons for lower comfort were found to fall under three broad themes: (i) system-level concerns; (ii) scenario-specific concerns about how data and algorithms are used in decision-making; and (iii) concerns about how agencies and service providers relying on algorithmic tools communicate and interact with families. Reasons for higher comfort were more difficult to generalize and tended to be particular to the scenario and participant type. Reasons for higher comfort given in Scenarios 1d(i) and 1e had the most generalizable themes. These included (i) improved access to services; and (ii) the nature and relevance of the data going into the risk score. We elaborate on each of these themes below, highlighting notable differences in how they were expressed by the different participant groups. This section speaks most directly to the first two central questions posed at the outset of the paper: how do affected communities feel, and what are the primary sources of comfort and discomfort?

System-level concerns were the most common reasons given for low comfort in algorithm-assisted and algorithmic decision-making

Participants across all of the represented groups — families, frontline providers and specialists — made frequent reference to 'the system' and 'government' to explain their reasons for positions of low comfort across the different scenarios. In discussion these terms could variously refer, either generally or specifically, to local government or to wider US government agencies (most commonly in child welfare, but also criminal justice, health, education and and social welfare), and to the people, decisions, actions and behaviours of those agencies. These point to low baseline perceptions of procedural and interpersonal justice of the child welfare system as a whole.

Families often referred to negative experiences in their own lives, and reflected on the oppositional nature of the system, saying, for instance, "It's been me versus the system" (Family - 1a). Most family participants had low perceptions of trust in the decisions made, and low expectations of the benefits that the system could provide.

> "They would look at me more because I had previous experience than because they wanted to help me with my daughter." (Family - 1c)

Even in settings where participants saw a clear benefit, system-level factors were cited as primary reasons for low trust, and thus low overall comfort.

> "I see the benefit — I'm the grandma and we need help, my daughter and the young father. I still don't trust the system, though. I'd look for the help elsewhere." (Specialist - 1d)

Frontline providers and specialists expressed concern regarding negative system approaches that placed too much emphasis on the 'deficits' and 'risks' of a case. Providers referred to their professional or personal experiences of 'the system' in describing what they believed to be a 'deficit-based' approach. Specialists often explained their positions of low comfort as a response to the language and terminology used in the scenarios. For instance, the term 'investigation' was commonly perceived as a clear example of negative language, as were technical terms such as 'statistical tool', 'risk score', and 'high risk'.

> "It seems like a deficit model – let's weigh up all the dirty things in your life, nothing good though." (Provider - 1d)

> "'Investigation' says that you've been judged already". (Specialist - 1a)

All groups raised concerns about potential bias on the part of case workers involved in the decision process, as well as bias present in the data or the algorithm

Many participants voiced concerns about bias and discrimination on the basis of race, ethnicity, gender, poverty status, neighborhood, and urban/rural location. Families questioned how and whether the computer tool takes race or ethnicity into account, and whether the data itself might be biased due to selective sourcing. Personal experiences of racial bias were often cited as reasons for low comfort.

> "The system here in America just lets us down, especially if you are Black." (Family-1b)

> "Knowing what 'the system' has done to people of colour, I'm coming from a position of very low trust. As a black mother, I have [an advanced degree] and a career and I still have that fear of a system that is checking on that, on me as being black. That white mother sitting next to me, they won't investigate her like they will me. Knowing that my black friends' babies were drug tested in hospital, it's just a fact that I'm treated a certain way because of who I am." (Specialist - 1b)

> "How honest are we allowed to be? Most of our systems were not made for people of colour, or by people of colour, or have people of colour in them.". (Specialist - 1a)

Where a data-driven algorithm was used in the decision-making process, frontline providers raised concerns about data fidelity and bias, as well as its impact on case worker judgment. Providers questioned the extent to which outdated historical data might be given undue weight in decisions about a new situation, and whether old referrals with no negative findings should be taken into consideration. They also noted that selective sourcing and differential availability of data could result in policies that disproportionately impacted low socioeconomic status populations.

> "My neighbour might be shooting up heroin and their six year old is out in the street, but they have private insurance so their records aren't part of this system. The computer tool is only capturing people who have to use public health so there's a bias to poorer people in the system." (Provider - 1b)

Participants questioned whether a statistical model could adequately account for all relevant decision elements, and emphasized the need for a human in the loop approach

Family participants mostly expressed general concerns about 'human versus computer' capabilities. Some participants were specifically concerned that data-driven algorithmic tools would not be able to produce risk scores that accurately accounted for salient aspects of their circumstances.

> "A computer cannot understand context. My son has autism — how does the data account for this?" (Family - 1b)

Frontline providers and specialists emphasized the need to supplement the algorithmic assessments with a more in-depth look at the family's situation and history. They were vocal in questioning decision-making systems that relied on algorithms alone, stressing that child welfare decisions require an understanding of context that is only possible through human interpretation and contact.

> "The score should be a flag rather than a definitive 'go'. It needs to be approached with curiosity: Where are you at? What are you facing? What are your needs? Would you benefit from home visits, more community? Help put it back together. If child welfare was just a score we wouldn't be sitting here." (Provider - 1b)

> "Use data without removing human decision-making." (Specialist - 1c)

The idea of fully automated algorithmic decision systems was met with low perceptions of procedural justice. While algorithmic risk scores were perceived as potentially helpful as a starting point, they were generally deemed to be an inadequate basis for ultimate decision-making.

Participants wanted more information on how the algorithm weighs different factors, and the ability to dispute the score

Many of the participants expressed discomfort with the black-box manner in which the algorithmic tool was presented. Specifically, they wanted to have more information on how the tool was constructed, and the weights given to different factors included in the model. This indicated low perceptions of informational justice in the algorithm-assisted and algorithmic decision-making scenarios.

> "How do you determine if a person has a record? How did the data get scored? What is the process for deciding the score?" (Provider - 1b)

> "A risk score from 1-10? What is the criteria? I want to know that there is consistency." (Specialist - 1b)

Echoing earlier concerns about a 'deficits-based' approach, participants also questioned whether 'positive' data about families was taken into account, or if the model weighed only 'negative' factors associated with increased risk.

The black-box nature of the algorithmic tools also led to concerns surrounding recourse and contestability. Participants wanted to be able able to dispute input data and risk assessments that they disagree with.

> It's like with credit scores, the Fair Credit [Reporting] Act gives options to contest it if you disagree with data, an official way to fight to get things removed if reported erroneously." (Provider - 1b)

Even potentially beneficial decisions resulted in discomfort due to concerns about how and whether risk information was communicated to families and case workers

All participant groups expressed concern at a perceived lack of supportive communication with families across the range of scenarios.

> "It's all about communication — it starts here and everything else is part of this.' (Family-1d)

> "The computer tool is the 'why' to approach a family NOT the 'way' to approach a family" (Provider - 1c)

This became especially pronounced in the later scenarios such as 1d. Families and providers perceived the explicit mention of a statistical tool as being antithetical to engendering trust with the young mother in the scenario.

> "I see high benefit but low trust. If [the services offered are] voluntary then surely it's good. She's a kid so it's not a bad thing. Often you don't know what's out in your community. But I don't like her being told she was flagged by a statistical tool." (Provider - 1d)

While many saw significant benefit in the proactive offering of services triggered by the tool, this was often outweighed by deep distrust resulting from the communication between the nurse and the mother. Many participants described the explicit mention of risk levels in 1d(ii) and 1d(iii) as coercive and fear-based.

> "It's the way they're doing it—I don't feel you want to do that to anyone. You can provide some of that information without this—almost implying that 'if you don't take this your child will be taken away'." (Family - 1d)

Interestingly, scenario 1d(i) had the greatest divergence in comfort levels across the different participant groups. Figure 3 shows derived plots of the comfort maps for 1d(i) and 1d(iii) for all three main participant groups. Specialists started out expressing uniformly very high comfort. Unlike families and frontline providers, they did not perceive a serious issue with the general mention of a 'statistical tool'. Specialists expressed much lower levels of comfort in scenarios 1d(ii) and 1d(iii), taking issue with the manner in which risk was being communicated.

There was also disagreement on the perceived benefit of the case workers in the scenario having knowledge of the risk score and the risk factors involved. By some, this knowledge was viewed as critical to equipping workers to take informed actions.

> "Social welfare workers need to be equipped to explain, 'this is why we're doing this, these are the factors that made us concerned'." (Specialist - 1c)

> "Does the social worker know about the risk score? If they don't, what if they are walking into a home with domestic violence potential without knowing?" (Specialist - 1c)

However, there was also concern that by knowing the score in advance social workers may be less able to fairly evaluate the situation.

> "But as a social worker you should not know the risk score — you need to be able to give a fair assessment." (Specialist - 1c)

5 DISCUSSION

In this section we turn to the final question raised at the outset of the paper: What can researchers and designers working in partnership with public service agencies do in the development and deployment stages to raise comfort levels among affected communities? These recommendations are drawn from both scenario-specific and general discussions with workshop participants.

Address system-level concerns.

The introduction of an algorithmic tool does not appear to improve trust or perceptions of procedural justice on its own. On the contrary, the very same system-level concerns that pervaded the discussion were readily projected onto the algorithmic decision-making scenarios. If these concerns are not addressed through complementary policy changes, even the best-conceived algorithmic tools and intervention strategies are likely to be met with distrust and low comfort.

Weigh positive and negative data in algorithmic decision-making.

The development of predictive analytics for use in human services follows the tradition of 'risk assessment' wherein an individual's likelihood of an adverse outcome is assessed using a set of predictive 'risk factors'. This framing focuses attention on predicting a negative outcome ('failure') based on negative inputs that capture 'deficits' instead of 'strengths'. There is concern that such approaches risk anchoring workers to a disproportionately negative view of the situation, which may in turn drive negative actions. While capturing 'strengths' in some instances requires the collection of additional data, simpler design changes may also be effective. Modeling success instead of failure may be a matter of recoding the outcome variable of interest, or simply reporting the likelihood of not-failure. Likewise, certain deficit variables

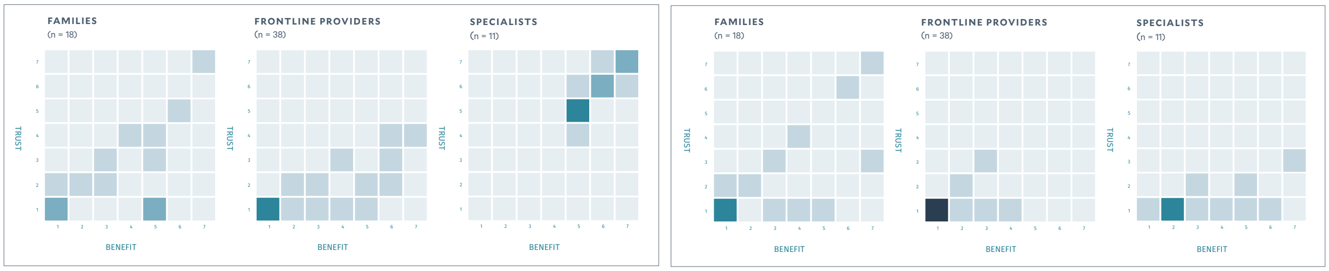


Figure 3: Derived comfort maps for scenarios 1d(i) (left) and 1d(iii) (right).[2] Darker colours indicate more participants in the given cell.

or risk factors can be inverted into more positive variables that capture strengths instead. These simple modifications would produce a model that is mathematically equivalent to the original, but which may be perceived and responded to very differently by users and affected individuals.

Convey how making use of data and algorithms leads to improved family and process outcomes.

Participants generally agreed that algorithmic tools that systematically identify and flag important risk factors such as criminal history provide significant benefit over less systematic approaches. They also viewed the proactive offering of services on the basis of algorithmic assessments to be beneficial to families in need. However, it was unclear to participants whether there were any other benefits to algorithmic decision-making systems. It was also unclear that any such benefits would outweigh the concerns raised by such systems. As one frontline provider put it "How is a 'computer tool' better than 'no computer tool'?" Participants were specifically interested in evidence that algorithmic systems would produce better child welfare outcomes than traditional decision-making approaches. Outcomes of particular interest include increased child safety, decreased disparities, and greater access to strengths-based family support.

Facilitate supportive communication and positive relationships between child welfare workers and families.

Even if an algorithmic back-end is triggering particular actions, those actions are still typically being carried out by child welfare workers through direct interactions with families. Families need to know they can trust frontline providers to act in their best interests. In turn, frontline providers need a clear understanding of how algorithmic decision-making can inform the formation of positive and safe professional relationships with families.

Throughout the workshops, participants raised numerous concerns surrounding the opacity of algorithmic decision-making as described in the scenarios. These concerns may be viewed as obstacles to the ultimate goal of forming supportive relationships. For instance, participants indicated that they would feel more comfortable if they had: (i) knowledge of what data was being used and how; (ii) knowledge of how the score is used, shared, and stored; (iii) the ability to contest inaccurate data; (iv) knowledge of how different factors are weighed in the risk score; and (v) the criteria for acting upon a risk score. However, we caution that simple disclosure of this information is unlikely to have much effect on promoting relationship-building. More research is needed to understand how different elements of algorithmic systems affect perceptions of interpersonal and informational justice.

6 CONCLUSION

The fairness and interpretability of algorithms deployed in consequential decision-making settings has received significant attention in recent years. This has led to a numerous proposals for auditing strategies [15, 45], and the development of models that are constructed to be in some sense 'fair' by design [2, 23, 30]. Existing work has even explored the algorithmic bias properties of tools deployed in the child welfare system [10]. Our study suggests that these technical solutions are in and of themselves insufficient for ensuring that the resulting algorithmic systems are perceived as fair and just. Perceptions of algorithmic fairness are heavily influenced by perceptions of procedural and interpersonal justice of the child welfare system as a whole. Effective design strategies must therefore consider how technological solutions can be designed to work in concert with complementary policy changes to impact community perceptions of system justice.

ACKNOWLEDGMENTS

The authors thank SIGKDD for the seed funding for this project provided through their KDD Impact Program. We also thank the anonymous referees for their many excellent comments and helpful suggestions for improving the manuscript.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 582.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] Toi Aria. 2017. Our data, our way. https://trusteddata.co.nz/massey_our_data_our_way.pdf

[5] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 377.

[6] Robert Brauneis and Ellen P Goodman. 2017. Algorithmic transparency for the smart city. (2017).

[7] Sarah Brayne. 2017. Big data surveillance: The case of policing. *American Sociological Review* 82, 5 (2017), 977–1008.

[8] Joel Brockner and Batia Wiesenfeld. 2005. How, when, and why does outcome favorability interact with procedural fairness? (2005).

[9] Joel Brockner and Batia M Wiesenfeld. 1996. An integrative framework for explaining reactions to decisions: interactive effects of outcomes and procedures. *Psychological bulletin* 120, 2 (1996), 189.

[10] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. 134–148.

[11] Coalition. 2018. The use of pretrial risk assessment instruments: A shared statement of civil rights concerns. http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf

[12] Jason A Colquitt. 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of applied psychology* 86, 3 (2001), 386.

[13] Nicholas Diakopoulos. [n. d.]. Algorithmic-Accountability: the investigation of Black Boxes. ([n. d.]).

[14] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

[15] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[16] Andrew Guthrie Ferguson. 2017. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement.* NYU Press.

[17] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. *arXiv preprint arXiv:1802.09548* (2018).

[18] Bernard E Harcourt. 2014. Risk as a proxy for race: The dangers of risk assessment. *Fed. Sent'g Rep.* 27 (2014), 237.

[19] George E Higgins, Scott E Wolfe, Margaret Mahoney, and Nelseta M Walters. 2009. Race, Ethnicity, and Experience: Modeling the Public's Perceptions of Justice, Satisfaction, and Attitude Toward the Courts. *Journal of Ethnicity in Criminal Justice* 7, 4 (2009), 293–310.

[20] Dan Hurley. 2018. Can an Algorithm Tell When Kids Are in Danger? https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html

[21] David Jackson and Gary Marx. 2017. Data mining program designed to predict child abuse proves unreliable, DCFS says. http://www.chicagotribune.com/news/watchdog/ct-dcfs-eckerd-met-20171206-story.html

[22] Robert Jungk and Norbert Müllert. 1987. *Future Workshops: How to create desirable futures.* Institute for Social Inventions London.

[23] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[24] Christopher Kingsley and Stefania Di Mauro-Nava. 2017. First, do no harm: Ethical Guidelines for Applying Predictive Tools Within Human Services. *MetroLab Network Report* (2017).

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

[26] Min Kyung Lee and Su Baykal. [n. d.]. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division.

[27] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management that Allocates Donations to Non-Profit Organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3365–3376.

[28] Christopher T Lowenkamp. 2009. The development of an actuarial risk assessment instrument for US Pretrial Services. *Fed. Probation* 73 (2009), 33.

[29] John Monahan, Anne Metz, and Brandon L Garrett. 2018. Judicial Appraisals of Risk Assessment in Sentencing. (2018).

[30] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, Vol. 2018. NIH Public Access, 1931.

[31] Jędrzej Niklas, Karolina Sztandar-Sztanderska, and Katarzyna Szymielewicz. 2015. Profiling the unemployed in Poland: social and political implications of algorithmic decision making. *Fundacja Panoptykon, Warsaw Google Scholar* (2015).

[32] Executive Office of the President, Cecilia Munoz, Domestic Policy Council Director, Megan (US Chief Technology Officer Smith (Office of Science, Technology Policy)), DJ (Deputy Chief Technology Officer for Data Policy, Chief Data Scientist Patil (Office of Science, and Technology Policy)). 2016. *Big data: A report on algorithmic systems, opportunity, and civil rights.* Executive Office of the President.

[33] Christopher P Parker, Boris B Baltes, and Neil D Christiansen. 1997. Support for affirmative action, justice perceptions, and work attitudes: A study of gender and racial–ethnic group differences. *Journal of Applied Psychology* 82, 3 (1997), 376.

[34] Data Futures Partnership. 2017. A Path to Social Licence: Guidelines for Trusted Data Use. http://datafutures.co.nz/our-work-2/talking-to-new-zealanders/

[35] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. 2017. Exploring user perceptions of discrimination in online targeted advertising. In *USENIX Security*.

[36] Dillon Reisman, Jason Schultz, K Crawford, and M Whittaker. 2018. Algorithmic impact assessments: A practical framework for public agency accountability.

[37] O'Brien Kirk Roberts, Yvonne H and Peter J Pecora. 2018. Considerations for Implementing Predictive Analytics in Child Welfare. *Casey Family Programs* (2018).

[38] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices.* CRC Press.

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

[40] Hetan Shah. 2018. Algorithmic accountability. *Phil. Trans. R. Soc. A* 376, 2128 (2018), 20170362.

[41] Halil Toros and Daniel Flaming. 2018. Prioritizing Homeless Assistance Using Predictive Algorithms: An Evidence-Based Approach. *Cityscape* 20, 1 (2018), 117–146.

[42] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 440.

[43] AJ Wang. 2018. Procedural Justice and Risk-Assessment Algorithms. (2018).

[44] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 656.

[45] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2016. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *arXiv preprint arXiv:1610.08452* (2016).