How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions

Hao-Fei Cheng*
Carnegie Mellon University
Pittsburgh, USA
haofeic@andrew.cmu.edu

Logan Stapleton*
University of Minnesota
Minneapolis, USA
stapl158@umn.edu

Anna Kawakami
Carnegie Mellon University
Pittsburgh, USA
akawakam@andrew.cmu.edu

Venkatesh Sivaraman
Carnegie Mellon University
Pittsburgh, USA
vsivaram@andrew.cmu.edu

Yanghuidi Cheng
Carnegie Mellon University
Pittsburgh, USA
yangcheng@cmu.edu

Diana Qing
University of California, Berkeley
Berkeley, USA
dianaqing@berkeley.edu

Adam Perer
Carnegie Mellon University
Pittsburgh, USA
adamperer@cmu.edu

Kenneth Holstein
Carnegie Mellon University
Pittsburgh, USA
kjholste@cs.cmu.edu

Zhiwei Steven Wu
Carnegie Mellon University
Pittsburgh, USA
zstevenwu@cmu.edu

Haiyi Zhu
Carnegie Mellon University
Pittsburgh, USA
haiyiz@cs.cmu.edu

ABSTRACT

Machine learning tools have been deployed in various contexts to support human decision-making, in the hope that human-algorithm collaboration can improve decision quality. However, the question of whether such collaborations reduce or exacerbate biases in decision-making remains underexplored. In this work, we conducted a mixed-methods study, analyzing child welfare call screen workers' decision-making over a span of four years, and interviewing them on how they incorporate algorithmic predictions into their decision-making process. Our data analysis shows that, compared to the algorithm alone, workers reduced the disparity in screen-in rate between Black and white children from 20% to 9%. Our qualitative data show that workers achieved this by making holistic risk assessments and adjusting for the algorithm's limitations. Our analyses also show more nuanced results about how human-algorithm collaboration affects prediction accuracy, and how to measure these effects. These results shed light on potential mechanisms for improving human-algorithm collaboration in high-risk decision-making contexts.

*Both authors contributed equally to this research.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3501831

KEYWORDS

human-centered AI; machine learning; algorithmic biases; algorithm-assisted decision-making; child welfare

ACM Reference Format:
Hao-Fei Cheng, Logan Stapleton*, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3491102.3501831

1 INTRODUCTION

Recent years have seen the deployment of AI-based tools either to augment or replace human judgments across a growing range of high-impact decision-making contexts, such as social work, criminal justice, hiring, healthcare, and education [11, 15, 38, 43, 44, 58, 72, 95, 98]. These technologies have often been adopted under the logic that they are more accurate and equitable than human decision makers [15, 49, 58]. Prior work suggests that on various predictive tasks, AI systems are more accurate than human decision makers [28, 30, 55, 89]. However, in many social decision-making contexts, such as recidivism risk assessment, AI systems have been shown to inherit human biases from historical data, and perpetuate discrimination against already vulnerable populations, e.g. [9, 10, 28, 75, 88]. Prior attempts to make these algorithms less discriminatory have largely focused on the technical design of the algorithms—a central focus of the area of *algorithmic fairness*, e.g. [29, 41]. Another possible approach to improving fairness in decision-making may be through *human-AI collaborations*, aimed at combining strengths and mitigating limitations in both AI-based

and human decisions [23, 42, 60, 86, 97]. In some contexts, human-AI collaboration has demonstrated potential to improve fairness and effectiveness of decision-making, compared with human or AI decision-making alone, e.g. [6, 23, 43, 52, 71]. However, empirical results have been mixed. For example, in a real-world pretrial criminal context, human-AI collaboration was shown to exacerbate discriminatory decision-making [3, 84].

In this paper, we examine how call screen workers in Allegheny County, Pennsylvania use the AI-based Allegheny Family Screening Tool (AFST) to make decisions about which reports of child abuse or neglect (henceforth *referrals*) to investigate. Similar to De-Arteaga et al. [23], we compare *automating* versus *augmenting* decision-making with the AFST. Note that the AFST was designed to *"augment the human decision whether to investigate a call"* [14] and Allegheny County Office of Children, Youth and Families leadership assures that the tool is never used to completely automate decisions [5]. However, critics such as Eubanks [31] and the National Coalition for Child Protection Reform [66] worry that the AFST, and other tools like it, may someday be used to automate decisions, for example, as an austerity measure. In this paper, we investigate the effect that such a hypothetical automated decision-making policy would have on racial disparities in child welfare call screening. We then compare this policy to the current standard decision-making process, where workers make decisions with the assistance of AFST. Since child welfare workers themselves are known to make racially disparate decisions [24, 25, 27, 40, 54], it is unclear whether adding them back "in the loop" will do any good in this regard. The central question of this paper is thus: **when people work with algorithms in a child welfare context that is known to have racial disparities, will they serve to mitigate or exacerbate disparities?** The answer to this question can inform the responsible design and use of AI tools in the child welfare context, as well as other high-stakes social decision-making contexts.

The primary racial disparity measure we use in this paper is the difference in the screen-in rates across Black and white children. Where prior work has emphasized the impact that the AFST has had on caseworkers' workloads [35], we think it is also important to focus on the impact that the AFST has on families in Allegheny County. Call screening is most often the first point of contact with the child welfare system, where an agency decides whether to intervene into the life of a family by investigating them. Being screened in may lead to more child welfare involvement into a family's life. AFST documentation states that "screening in and a child protection investigation has some potentially deleterious effects on families. If screening in, however, is a prerequisite to being offered higher quality services or being prioritized for a slot in a desired program, one can argue the benefits of an investigation" [92]. Higher screen-in rates indicate higher levels of state intervention into families' lives, starting with investigation. Racial disparities in screening rates indicate uneven application of interventions or investigations of Black and white families and potentially uneven distribution of the potential harms and benefits of them. Disparities in screening rates may be unjustified if they occur because of unwarranted intervention or lack of intervention, e.g. investigating a family when their child is not at imminent risk of abuse or neglect. As Dorothy Roberts [26] suggests, "[t]he disproportionate number of Black children under state supervision results from discriminatory decision-making within the system as well as racist institutions in the broader society." However, as we will discuss below, disparities may be justified: higher need or higher risk of abuse or neglect among children in one group of people could warrant a higher screen-in rate.

In summary:

- Through quantitative analysis based on the two years of data immediately following the introduction of the AFST—from August 2016 to July 2018—we evaluated racial disparities in AFST-only and worker-AFST screening decisions. Our results show that worker-AFST decision-making served to reduce the disparity in screen-in rate between Black and white children compared to algorithm-only decisions.
- We conducted a contextual inquiry by observing how call screen workers use the AFST to help them make decisions, and we interviewed workers about their experiences working with the algorithm. Through qualitative analysis, we find that by assessing referrals holistically using all of the information available to them and by adjusting for the algorithm's limitations, call screen workers disagreed with the AFST in ways that serve, in aggregate, to reduced the impact of racial disparities in the algorithm. Our findings suggest that the AFST did not supplant call screening discretion and decision processes and that workers were not blindly following the AFST, consistent with the statement from Allegheny County Department of Human Services [5] responding to Eubanks [31].
- We also analyzed the accuracy of AFST-only and worker-AFST decisions. Although the AFST is better than workers at predicting the outcomes that it is trained to predict, our qualitative findings indicate that workers make screening decisions to optimize for fundamentally different outcomes than the AFST.
- Finally, we provide design implications for potential ways to improve the collaboration between call screen workers and the AFST in improving the decision-making process.

The AFST is just one algorithmic system used in child welfare; there are many similar systems used in child welfare agencies across the U.S. [78, 81]. We anticipate that several of our findings may generalize to other public sector, algorithm-assisted decision-making contexts. However, agencies are often reticent about their internal policies, decision-making, and even public information [2, 79], making it challenging to conduct similar analyses across other contexts. We thank the Allegheny County Office of Children, Youth and Families (CYF) for their continued interest in working with external researchers, and for their transparency in providing us data and access to their facilities. We also thank the workers in the Intake Department for taking the time to speak with us, and for sharing their insights. We hope that more agencies will follow Allegheny County's lead in opening opportunities for public and research accountability.

Finally, the data used throughout this paper contained information on 39,429 children who were referred to CYF. We acknowledge all 39,429 of these children and their families, on whom this data was

collected and for whom this data reflects potentially consequential interactions with CYF.

2 RELATED WORK

2.1 Algorithm-in-the-loop decision-making

"Algorithm-in-the-loop" decision-making is commonly characterized by first having an algorithm-produced prediction or classification, with a human making the final decision after considering an algorithmic recommendation [38, 85]. Algorithm-in-the-loop decision-making has been observed in multiple high-stakes scenarios, including pretrial bail decisions [3], recidivism predictions [28], predictive policing [73], and diagnosing patients in clinical settings [62].

As algorithm-in-the-loop decision-making becomes increasingly common in practice, recent research has started to look at how humans work with algorithms when making decisions and at the relative contributions of humans versus algorithms to overall performance. Many studies have focused on prediction accuracy, finding that on many tasks, algorithms can outperform humans in terms of prediction accuracy [8, 37, 39, 56, 57, 99]. However, recognizing that human experts and algorithms may have complementary strengths and limitations, a line of research has sought to understand how to combine the capabilities of each [23, 42, 60, 86, 97]. Some studies have demonstrated that combinations of human and algorithmic judgment can improve prediction and/or decision-making (e.g., [6, 23, 43, 52, 57, 71]). Yet empirical results in this space have been varied so far. In other studies, human-algorithm decision-making has failed to improve or has even *harmed* performance, compared with either human or algorithmic decisions alone (e.g., [3, 37, 74, 84, 99]).

Beyond accuracy, other metrics have been used to evaluate between decisions made by humans, AI, and human-AI combined [3, 33, 38, 51, 59, 74]. Much of this work focuses on minimizing the error of the decisions compared to the ground truth. Most relevant to our work, Albright [3] compares racial disparities in human-only and human-AI decisions in the context of pretrial bail hearings. Here, judges are presented with a risk score and recommendation from a risk assessment algorithm (similar to the AFST) and then must decide whether to give a person who is charged with a crime bail (and keep them in jail until they can pay) or let them go free without paying bail until their trial date. Albright [3] suggested that judges disagreed with the algorithmic recommendations for certain types of defendants such that the judge-algorithm bail decisions were more racially disparate—giving Black people bail rather than letting them free without bail at a higher rate than white people—than both the past judge-only decisions (before the algorithm was implemented) and the algorithm-only recommendations in the same time period. In our work, we present empirical results in the opposite direction: in the context of child welfare call screening, we found that human-algorithm decisions were less racially disparate than algorithm-only decisions would have been, and somewhat less racially disparate than past human-only decisions.

2.2 Risk Assessment Tools in Child Welfare

For an overview of predictive algorithms used in child welfare, see [81] and [78]. We provide a brief history here. For decades, child welfare agencies across the U.S. and abroad have been using risk assessment instruments (RAIs) to assist child social workers in making decisions, such as whether or not to investigate a family or whether to remove a child from their family. Most RAIs have been checklists that workers fill out in order to estimate the risk of child maltreatment. For example, see the Structured Decision Making (SDM) tools used in the California Child Protective Services system and a number of other locales [70]. For a case study of other kinds of RAIs and algorithms used in child welfare, see [80]. However, newer RAIs include automated tools, commonly called *predictive analytics* or *data-driven predictive tools*, which use statistical modeling and machine learning to estimate risk based on historical administrative data. Earlier iterations of these were developed by private companies, such as Eckerd Connects [16], MindShare Technology [87], or SAS [46]. Due to high error rates and their "black box" nature, the Los Angeles County and Illinois child welfare systems dropped private algorithms after brief trials [48, 63, 64]. As such, these private tools have fallen out of favor, though some were still in use in other locations at the time of publication. Other data-driven predictive tools have been or are being developed through public-academic partnerships—such as the Allegheny Family Screening Tool (AFST) in Allegheny County, Pennsylvania [15, 92] or other tools being implemented by the same designers in Douglas County, Colorado [91] and Los Angeles County [76].[1] These publicly-developed algorithms have proved more resilient, with the AFST being the longest-lasting and most prominent predictive tool in use today. Proponents of these newer automated risk assessment tools, such as the AFST, argue that they make more accurate decisions than both child social workers and standard checklist-based RAIs; and that they make more consistent, objective, and equitable decisions [15, 20, 45, 82]. Some critics argue that these automated tools are still too inaccurate, that they do not predict true child abuse or neglect, and that they still make biased decisions because they are trained on biased data [15, 31, 65].

In order to evaluate the AFST based on these opposing concerns, the Allegheny County CYF commissioned Goldhaber-Fiebert and Prince [35] to conduct an Impact Evaluation of the AFST. Among the results, the most relevant to this paper were the following: 1) the AFST *"increased accuracy for children screened-in for investigation and may have slightly decreased accuracy for children screened-out;"* 2) the AFST did not decrease the screen-in rate overall; and 3) the AFST reduced racial disparities in terms of the screen-in rate, but possibly worsened them in terms of accuracy [35]. These findings called into question whether the AFST improves either accuracy or equity. These results were based on an analysis of two years of data immediately following the introduction of the AFST—from August 2016 to July 2018. Our quantitative findings are based on the same data as Goldhaber-Fiebert and Prince [35], which we preprocessed and analyzed to match their work, as well. Rather than reiterate their findings, we use this data to evaluate the effects that automating screening decisions would have on racial disparities. We also expand upon Goldhaber-Fiebert and Prince [35]'s findings with a new mixed-methods approach.

[1] It is notable that Los Angeles County, the largest child welfare department in the U.S., dropped a private model earlier, but is now implementing a new model with the same team who developed the AFST.

Figure 1: The current AFST-assisted call screening process. Call screen caseworkers make screening recommendations and supervisors make the final screening decisions, both with the AFST's risk score and recommendation.

It should also be noted that some critics oppose not only automated predictive tools in child welfare, but also the discourse around accuracy, fairness, accountability and transparency that our paper contributes to, which "does not address the core structural issues at work" with these tools [2, 77]. Abdurahman [2] explicitly names and critiques the central question of our paper—namely, "Does the Allegheny Family Screening Tool (AFST) produce fair outcomes?" We still see some merit in "adjudicating [the] downstream impact" of the AFST [2]. However, we recognize that our paper is limited in that it does consider the larger political economic or social contexts in which these tools are deployed, as in [2, 77].

3 METHODS

3.1 Study Context

In this paper, we studied a high-stakes scenario of child maltreatment referral screening decisions. The Allegheny County Office of Children, Youth, and Families (CYF) has been using an algorithmic tool, the Allegheny Family Screening Tool (AFST), to assist with child maltreatment call screening process since August 2016. The AFST is a machine learning-based predictive risk modeling tool that analyzes county data to predict outcomes related to child abuse or neglect. The AFST Version 1 used demographic data related to the alleged victims, caretakers, alleged perpetrators, prior child welfare history, criminal history, public behavioral health history, and use of public assistance [92].[2] Prior to the introduction of the AFST, call screen workers made all referral decisions without any algorithmic aids. Since its deployment, workers have been presented with an AFST risk score to assist with their call screening decisions for all referrals which were not automatically screened in or out (henceforth *referrals* or *discretionary referrals*).[3] Call screen workers still make the final decisions—they have the option to either agree with the algorithm recommendation, or to disagree and go with their own decisions.

We adopted a mixed-methods approach to investigate how CYF call screen workers work with the AFST, and how the resulting human-AI decisions affect disparities in decision outcomes. We analyzed historical data on call screening decisions prior to and after the deployment of the AFST algorithm. We also conducted contextual interviews with caseworkers and supervisors to support interpretation of findings from these quantitative analyses.

3.1.1 Use of the AFST in call screening. The AFST Version 1 (used from August 2016 to November 2018) was made up of two models—the *re-referral* model was trained to predict whether a child would be reported again within two years of being screened out; the *placement* model was trained to predict whether a child would be removed from their home and placed in foster care within two years of being screened in. Each model produces a risk score ranging from 1 (lowest risk) to 20 (highest risk) associated with the likelihood of the corresponding predicted outcome (re-referral or placement) for each child in the referral. The score is categorized visually into 3 bins: Low risk (score 1-9), Medium risk (score 10-14), and High risk (score 15-20). The caseworker and supervisor sees a single risk score (see Figure 2) associated with the referral. The presented score is the higher of the scores across the two models. If the referral involves multiple children, the score of the child with the highest score in the referral would be shown.[4]

Workers used the AFST as follows: each caseworker first gathered information about the referral, made assessments of risk and safety,[5] made a screening recommendation for the referral, then ran the AFST to generate a risk score and passed the report to a supervisor. The supervisor then reviewed the report on the referral, which included its AFST score, then made the final decision to screen in and investigate the family or not. For all discretionary referrals, the AFST score served as a recommendation; workers had the authority to either agree or disagree with that recommendation when making the final screening decision. In Appendix D, we suggest that CYF workers consider High risk referrals (15-20) to be recommended screen-in, Low risk (1-9) recommended screen-out, and Medium risk (10-14) sans recommendation. However, 29.3% of children in discretionary referrals from August 2016 to May 2018 had a placement model score of 18 or above and were flagged as *mandatory screen-in* [92]. For these referrals, workers were shown the AFST interface on the right side of Figure 2. These referrals were "required to be screened in," but supervisors were able to "override this requirement at their discretion" provided that they "documented and reviewed" their reasons for overriding this requirement [92]. Supervisors overrode these decisions and screened out 21.0% of children labeled *mandatory screen-in*.

3.2 Data analysis

3.2.1 Data. We acquired data from Allegheny County CYF about all children who were referred to CYF from January 2015 to July 2018. The data contains referrals from both before the deployment of the AFST Version 1 (January 2015 - July 2016) and after (August 2016 - July 2018). We excluded all referrals which were automatically screened in or out, since these non-discretionary referrals would

[2]This was true for the first deployed version of the tool. The second version (in use from November 2018 until the time of publication) stopped using public assistance as a predictive feature and started using birth records [93]. For a full list of variables used by the AFST Version 1, see pages 37 to 44 of the documentation [92].
[3]We provide more detail about which referrals are discretionary or not in Section 3.2.1.

[4]For example, if there are two children in the referral: the first child with a re-referral model score of 1 and a placement model score of 15, the second child with a re-referral score of 10 and a placement score of 3; a score of 15 (and a "High risk" message along with Figure 2) would be shown to the call screen workers.
[5]This is the risk of future child maltreatment and immediate safety. A U.S. government source defines a safety assessment as gathering information to "determine the degree to which a child or youth is likely to suffer maltreatment in the immediate future" and a risk assessment as collecting information "to determine the degree to which key factors are present in a family situation that increase the likelihood of future maltreatment to a child or adolescent" [34].

not have been influenced by the AFST. Thus, we excluded all Child Protective Services (CPS) referrals,[6] referrals with active cases, and referrals with completed cases. We excluded referrals which were labeled both CPS and General Protective Services (GPS) in the data. We also excluded all cases which did not include white nor Black children, according to our definitions in Section 4.2. After preprocessing, the data used in our quantitative analyses included GPS referrals without active or completed cases which included white or Black children from January 1, 2015 to May 13, 2018.

We also used AFST scores which were generated retrospectively for the entire time period, which—due to a technical glitch that led the AFST to produce erroneous scores for a subset of referrals during the first year and a half of deployment—means that in some cases the scores we use in our analysis were not the scores that workers were shown from August 2016 until December 2017 [23]. Also, workers were not shown AFST scores from any referral from January 2015 to July 2016, since the AFST was not deployed until August 1, 2016. After December 2017, the AFST scores used in our analysis were the same as those shown to workers. We use the corrected AFST scores instead of the scores shown to workers to more accurately portray the screen-in rate of a hypothetical automated AFST-only policy.

Each entry in the data corresponds to an individual child who was referred at one time. Each child and each referral were associated with unique IDs. If a family with only one child was referred to CYF three times, this would correspond to three different entries in the data with the same child ID but three different referral IDs. If a family with two children was referred to CYF once, this would also correspond to two different entries with two different child IDs but only one referral ID. Each entry included the AFST risk score generated for each child in each referral, and the final call screening decisions made by the call workers for that referral.[7] Throughout our analysis, we report statistics and percentages in terms of entries, where one entry represents a unique child in a unique referral.[8] For shorthand throughout the paper, however, we describe numbers and percentages in terms of children. For example, when we write that *71.0% of Black children were screened-in*, this really means that 71.0% of entries containing Black children in discretionary referrals were not labeled screen out. After preprocessing, our data contains 31,025 entries before the deployment of the AFST Version 1, which include information on 23,230 unique children in 15,179 unique referrals; the preprocessed data from after the deployment of the AFST contained 51,750 entries on 33,613 children and 24,250 referrals.

3.3 Contextual Inquiry and Interviews

To support interpretation of our quantitative findings, we also conducted contextual interviews with call screen caseworkers and supervisors. A group of researchers visited the Intake Department of Allegheny County CYF in July 2021. The visit consisted of two parts: 1) **contextual inquiries**, where the researchers observed how call screen workers worked with the AFST when making screening decisions; and 2) **semi-structured interviews** where researchers were able to ask more in-depth questions. We observed and interviewed 13 participants in total: 9 call screen caseworkers and 4 supervisors over 2 separate visits in a span of two weeks. All participants worked full time as call screen caseworkers or supervisors. We checked with workers before and during our visits to make sure that we did not burden them too much while they were busy with work. To prevent workers from being identified in the workplace, all responses in the paper are anonymous and we report only minimal demographic information. Participants included, but may not have been limited to, white workers, Black workers, women, and men. At the request of the office, we did not provide monetary compensation to the participants. This study was approved by the Institutional Review Board of Carnegie Mellon University.

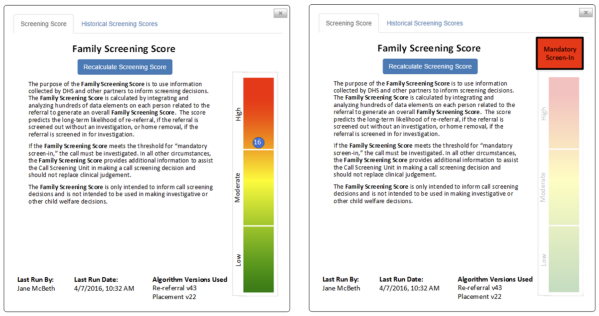


Figure 2: The interface of the AFST which is used by the child welfare workers to make screening decisions. The left figure shows the score the for a normal high risk referral. The right figure shows a referral with an exceptional high risk that triggers the mandatory screen-in policy.

3.3.1 Contextual Inquiry. After obtaining consent from participants, we observed call screen caseworkers and supervisors in their normal workflow, and participants were encouraged to think aloud as they performed their tasks, to make more of their thinking and reasoning visible. See Figure 1 for a visual diagram of the overall workflow. For caseworkers, this included taking phone calls from reporting sources, gathering information for reports, running the AFST to produce a risk score and recommendation, and making the call screening recommendation. For supervisors, this included reviewing the reports made by caseworkers, correcting information in reports (if need be), requesting field screening to gather missing information about a referral (if need be), making the final screening decisions, and overriding the mandatory screen-in referrals that the algorithm enforces.[9] Each contextual inquiry session took about three hours. Researchers took notes on the actions and thought processes of the participants, while asking brief follow-up questions as needed. Due to the sensitive nature of the work, we neither audio recorded the contextual inquiry nor took notes on any personally identifiable information.

[6] According to a Pennsylvania governmental source, "CPS reports are those that allege a child might have been a victim of child abuse" [1]. Pennsylvania law dictated that these referrals were automatically screened in and investigated.
[7] All children in the same referral received the same screening decision.
[8] Within CYF analysis, this is referred to as the *individual* or *child level.*
[9] The AFST automatically flags referrals with the highest risk as "mandatory screen-in." Only supervisors have the authority to override the AFST's decisions for these referrals [68].

3.3.2 Semi-structured Interviews. After the contextual inquiry, we invited each caseworker and supervisor for a semi-structured interview. The goal of the interview was to understand how participants incorporated the AFST in their decision-making process, and in particular, to gain further insight into possible mechanisms underlying our quantitative findings. At the beginning of each interview, we discussed participants' background and experience in child welfare. We also asked any follow-up questions that arose from our observations during the contextual inquiry, including clarifying questions about specific referrals or about their day-to-day workflow. We shared statistics about racial disparities in call screening similar to those in Figures 3, 4, and 5 which compared the AFST recommendations to actual decisions from 2016 to 2018, and asked workers' thoughts about these numbers. We then asked participants how they worked with the AFST to ensure fairness in screening decisions. Lastly, we discussed potential improvements to the design and use of the AFST.

3.3.3 Qualitative Analysis. We first transcribed all interview recordings into text, and used thematic analysis [7] to analyze our data, a constructivist approach inspired by grounded theory [13]. We combined the data collected from both the contextual inquiry and interviews, which contains interview transcripts and field notes. The authors collaboratively conducted open coding on the data, which generated over 1500 open codes. The authors then conducted an iterative affinity mapping process to the open codes, performing constant comparisons and iteratively clustering related codes. In the end, the authors refined the themes that emerged from the affinity mapping.

3.4 Positionality

We authors acknowledge that our positionality shapes our approaches to research, as well as how we interpret and present our findings. Given that the subject of research involves how Black and white families have been treated by the child welfare system in Allegheny County, Pennsylvania, we think it is especially important to acknowledge our racial/ethnic backgrounds, where we live, and our relationships to child welfare. The two lead authors are Asian and white, respectively. The rest of the authors self-described their racial/ethnic backgrounds as Asian, Asian American, Caucasian, Chinese, Filipino and White, in alphabetical order. None of us authors are Black. All but two of us live in Allegheny County; the other two live in Minnesota and California. None of us have been investigated by a child welfare agency, nor were any of us adopted nor involved in the foster care system as children. Throughout this work, we collaborated with Allegheny County CYF in order to gain access to data and to talk with workers, although the analysis and writing were conducted independently.

4 DECISION-MAKING PARADIGMS AND TERMINOLOGY

In this section, we define terminology used throughout the paper.

4.1 Decision-making paradigms

In this paper, we compare hypothetical screening decisions made by the AFST alone versus actual decisions made by child welfare workers when using the AFST. We define these two decision-making paradigms as follows:

(1) **AFST-only decisions:** The hypothetical screening decisions that the AFST algorithm would make if it were the only decision-maker (without workers). For clarity, we suppose that the AFST would screen in any High risk referrals (with a score of 15 to 20) and screen out any Medium and Low risk referrals (scores 1 to 14). We choose a threshold of 15 for our AFST-only policy because it is the threshold between Medium and High risk AFST labels presented to the call screen workers, and because its hypothetical screen-in rate would be close to the actual screen-in rate from 2016 to 2018. This split between High and Low risk referrals follows official AFST documentation, which discusses disparities: "up until the end of 2017, 47% of black children received a 'high'-range score (15–20), compared to 39% of white children. Conversely, 29.6% of white children have received a 'low'-range score (1–9), compared to 10% of black children" [68, p.11]. See Appendix D for evidence that call screen workers see High risk labels as screen-in recommendations and Low risk as screen-out. However, we also conducted sensitivity analyses by replicating our empirical results across alternative thresholds. For example, Figure 6a shows similar screen-in rate disparities across thresholds from 10 to 20.
(2) **AFST-assisted worker decisions:** The actual call screening decisions made by child welfare workers from 2016 to 2018, assisted by the AFST (see Section 3.1.1 for a detailed description of the decision-making process). For brevity, we refer to these decisions as **worker-AFST decisions** throughout the rest of the paper.

4.2 Definitions of Black and white children

In the data, each child had one or more of the CYF race labels: "Black," "'white," "Hispanic," "Asian," "Native American," "other," or "unknown." For our quantitative analyses in this paper, we considered a child Black if they were assigned the CYF race label "Black" alone or "Black" with any other CYF race label. We considered a child white only if they were labeled "white" only —i.e. if the child were labeled "white" plus any other CYF race label, they were not considered white in our quantitative analyses. We considered a child with the CYF labels "Black" and "white" as Black. This follows the same racial classification as the official AFST Impact Evaluation [35].

4.3 Evaluation Metrics

Our primary research questions to evaluate are: 1) how worker-AFST decisions affected racial disparities in call screening, and 2) whether changes in disparities affected the decision accuracy. We adopt the following evaluation metrics:

- **Racial disparity**
  The primary disparity measure used in this paper is the difference in the screen-in rate between Black children versus white children. Differences in the screen-in rates between

different racial groups correspond to one of the simplest and most popular algorithmic fairness notions—statistical parity, e.g. [17, 29, 32]. Specifically, a classifier satisfies statistical parity if the subjects in the protected and unprotected groups have equal probability of being assigned to the positive predicted class. While statistical parity serves as a starting place for our analyses, to assess the robustness of our results we also evaluate the disparities in other metrics, including accuracy, precision, true positive rate and false positive rate.

- **Accuracy**
  We measured accuracy by the percentage of decisions that aligned with a proxy ground truth: for screen-in decisions, we measured the percentage of children that were either removed from their home within 2 years or re-referred again within 2 months of the referral; for screen-out decisions, we measured the percentage of children who were neither removed from their home within 2 years nor re-referred again within 2 months of a referral.[10] Our definition of accuracy differs slightly from prior work on the AFST, in which a screen-in decision was accurate only if the child was later placed in foster care within 2 years and a screen-out decision was accurate only if the child was not re-referred within 2 years [35, 92]. We adopt the former definition of accuracy so that our hypothetical AFST-only accuracy is a decent estimate of what the accuracy of screening decisions would have been had they actually been automated by the AFST-only policy.
  Our measurement of AFST-only accuracy is an imperfect estimate, due to its counterfactual predictions [18]. For any referrals where the AFST-only decision differs from the actual screening decision, ideally we would want want to measure the AFST-only accuracy in terms of counterfactual outcomes —e.g. whether a child would have been placed in foster care or re-referred had they been screened in (when in reality they were screened out). Since we do not know these counterfactual outcomes, we evaluate AFST-only accuracy based on the actual predictive outcomes instead. However, because the screening decisions affect the predictive outcomes, the actual outcomes may have different probabilities than the counterfactual ones. For example, if the AFST-only decision is screen in, its accuracy will be judged based on whether the child is re-referred or placed; but, if they are actually screened out, we assume (but do not know) that the child is less likely to be placed and more likely to be re-referred than if they were actually screened in. This assumption may distort our measurement of accuracy for the hypothetical AFST-only decisions. As Coston et al. [18] note, this is a limitation endemic to risk assessments where the predictions affect the predictive targets.
- **Other metrics**
  We also evaluated disparities and prediction performance for a few additional metrics. We defer these results and the necessary terminology to the Appendix.

[10] We chose to evaluate re-referral within 2 months for consistency with the official AFST Impact Evaluation [35]. However, that the AFST Version 1 predicted re-referral within 2 years [92].

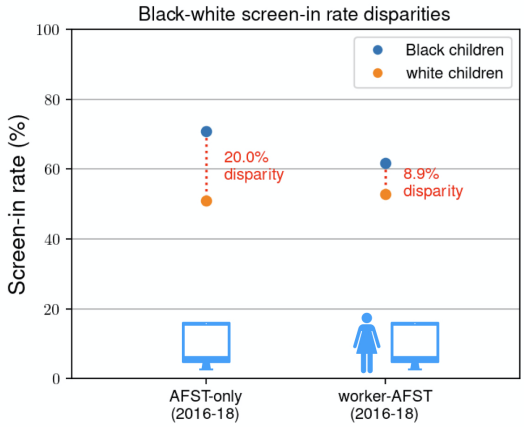


Figure 3: Black-white screen-in rate disparities for AFST-only and worker-AFST decision

	AFST-only		Worker-AFST	
	Black	white	Black	white
Screen-in	18536	11013	16133	11420
Screen-out	7587	10610	9990	10203
Total	26123	21623	26123	21623

Table 1: Total number of children that the AFST-only would have hypothetically screened in and out, compared to the actual decisions made by workers aided by the AFST.

5 HOW AFST-ASSISTED WORKER DECISIONS AFFECTED SCREEN-IN RATE DISPARITIES

5.1 Black-white disparities between AFST-only decisions and workers' final screening decisions

In this section, we look at the disparity in screen-in rates among Black and white children who were referred to CYF for AFST-only and worker-AFST screening policies. In Figure 3, we see the difference in screen-in rates between white and Black children under these two screening policies for all children reported to CYF from August 2016 to July 2018 who were not immediately screened in or out.[11] The screen-in rates in Figure 3 are calculated over the number of discretionary referrals within the race listed, e.g. the 71% AFST-only screen-in rate for reports with Black children means that 71% of reports with Black children would have been screened in from 2016 to 2018 following the AFST-only screening policy. The *Black-white screen-in rate disparity* refers to the difference between the screen-in rate for Black children and the screen-in rate for white children under a given policy for this time period. For example, the AFST-only Black-white screen-in rate disparity from 2016 to 2018 was 20%, because 71% of all referrals with Black

[11] By Pennsylvania law, all CPS referrals are automatically screened in. Some other referrals were marked as automatically screened in or out. All of these referrals were excluded from our analysis. All numbers in this section reflect only *discretionary referrals* for which the AFST could have had some influence on.

children would have been screened in and 51% of all white children would have been screened under the AFST-only screening policy.[12] We calculate that the worker-AFST screen-in rate disparity for the same time period was 9%, since 61.8% of Black children and 52.8% of white children were screened-in. **These results suggest that the Black-white screen-in rate disparity under the worker-AFST screening policy was less than half than that of the AFST-only policy from August 2016 to May 2018, 11% lower to be exact.** Table 1 shows the total number of referrals with Black and white children which would have been screened in and out under the AFST-only policy. An AFST-only policy would have screened in over 7500 more Black children than white children. In actuality, workers only screened in only 4713 more Black children than white. For reference, we calculate that from January 2015 to July 2016, before the AFST was implemented, workers screened in 52.5% of Black children and 41.2% of white children in discretionary referrals. This was a Black-white screen-in rate disparity of 11.3%.

We also examined Black-white screen-in rate disparities using other decision thresholds for the hypothetical AFST-only policy, as well as additional disparity metrics. We found that worker-AFST decisions were less disparate than AFST-only decisions, regardless of which threshold was chosen. It should be noted that the threshold of 15 that we default to throughout the paper has the second-to-highest disparity of any threshold from score 10 to 20 (including "mandatory screen-in"). We default to threshold 15 not to overstate our results, but because we argue that it would be the most reasonable threshold to choose from, given the design of the AFST which splits referrals into High and Medium risk referrals at 15, and given that a score of 15 produces an AFST-only overall screen-in rate comparable to the actual screen-in rate, whereas other thresholds do not. With some exceptions, we observe similar patterns in precision rates, true positive rates, and false positive rates—worker-AFST decisions exhibit less racial disparity in these metrics, as compared to AFST-only decisions defined by all thresholds from score 10 and above. See Appendix C for our complete analysis.

How workers disagreed with the AFST to reduce the Black-white screen-in rate disparity. Figure 4 shows the percentage of children which would have been screened in and out under both the AFST-only and worker-AFST policies, broken down by race. Recall that the Black-white screen-in rate disparity under the AFST-only screening policy was 20%. The simplest way that workers could have reduced this disparity would have been to screen in more white families and screen out more Black families than the AFST-only policy would have. Overall, this is what we observed: however, workers did not disagree with the AFST-only policy *exclusively* in ways that would have lessened this disparity. If workers were heavily guided by the AFST, but intentionally tried to reduce its screen-in rate disparity, we might expect that workers would not have screened in Black families that the AFST-only would have screened out, and that workers would not have screened out white families that the AFST-only would have screened in. However, we see in Figure 4 that this is not the case: workers screened out 15.9% of Black children who would have been screened out under the AFST-only policy and screened out 17.6% of white children who would have been screened in under the AFST-only policy. Because we see disagreement across the board, it is likely that workers are using their best judgment across the board, and they are not simply following the AFST's recommendations. It is also likely that workers are not making screening decisions explicitly in order to reduce racial disparities. These interpretations align with our qualitative findings in Section 5.2.

[12]Note that this 20% difference is in percentage points, i.e. 71% minus 51%.

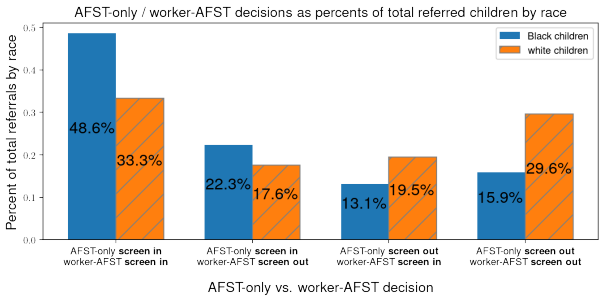


Figure 4: Proportions of Black and white children for which the AFST-only decisions and worker-AFST decisions agreed and differed. Note that the percentages here are over the total number of referrals per race, e.g. the 48.6% in the leftmost bar indicates that 48.6% of all Black children referred to CYF would have been screened in by AFST-only and was screened in by workers aided by the AFST (worker-AFST). Notice that the sum of the two leftmost bars (for a single race) equals the AFST-only screen-in rate —e.g. 48.6%+22.3%=71% for Black children,—whereas the sum of the leftmost and the right middle bars equals the actual worker-AFST screen-in rate—e.g. 48.6%+15.9%=61.7%.

What ultimately led to the Black-white screen-in rate disparity being lower under the worker-AFST policy than under the AFST-only policy is shown in Figure 4. **Among children that the AFST-only policy would have screened in, call screen workers screened out more Black children than white children (22.3% vs 17.6%). Among children that the AFST-only policy would have screened out, workers screened in less Black children than white children (13.1% vs 29.5%).**

5.2 Qualitative: How did workers achieve lower Black-white screen-in rate disparity?

When reviewing a referral, workers can see the races of everyone involved. In theory, one (naive) way workers could have reduced the screen-in rate racial disparity is by simply looking at the race of the children in the referral, screening in more Black children and screening out more white children, regardless of the other factors involved in the referral. Based on both our quantitative results (Section 5.1) and our qualitative findings, it is clear that this was not the case: workers were making decisions in a more sophisticated way. Furthermore, four call screen caseworkers and three supervisors explicitly said they did not make screening decisions based on the race of the family.[13] Two caseworkers and three supervisors said

[13]Note, however, that five caseworkers stated that they did consider the race of the family in order to account for racial biases in reporting. We explain this further below.

Mechanisms	Details
Holistic decisions	Workers considered AFST scores in the context of all the other information in a referral and made holistic, contextual assessments of risk and safety to make screening decisions.
Adjusting for limitations of the AFST	Workers adjusted for what they perceived to be limitations of the AFST and disregarded the AFST's recommendations when they thought it over- or under-scored referrals.
	Workers thought the AFST was over- or under-scoring certain referrals because it did not take the allegation or other current referral information into account properly.
	Workers thought that the AFST over- and under-scored referrals specifically based on system involvement, i.e. welfare, public medical services, criminal history, or CYF history.
	Workers compensated for what they thought were the AFST's racial disparities caused by systemic racial biases in CYF reporting and county data collection.
Collaborative decisions	Workers regularly made decisions collaboratively, both under standard procedures between caseworkers and supervisor, and impromptu between caseworkers.

Table 2: Summary of qualitative findings presented in Section 5.2.

they did not look at race at all. For example, one supervisor said, *"I have no idea what races people are."* Based on contextual inquiries and interviews with CYF call screen workers, we hypothesize that the decrease in Black-white screen-in rate disparity from 2016 to 2018 occurred because of the following reasons (also summarized in Table 2):

(1) Workers made screening decisions based on holistic, contextual assessments of risk and safety.
(2) Workers adjusted for limitations that they perceived in the AFST when making decisions on a case-by-case basis. For some workers, this adjustment was a conscious adjustment to try to reduce racial disparities. For others, it was unintentional. However, the effect of reducing disparity in aggregate was the same.
(3) Workers made collaborative screening decisions about some reports they were uncertain about.

5.2.1 Holistic decision-making. **Workers made holistic, contextual assessments of risk and safety in order to determine whether to screen referrals in or out.**

Interview findings: Four caseworkers and one supervisor said they did not consider the AFST's recommendation as a baseline to guide their final screening decision. Rather, workers said the AFST provided additional information to consider (two supervisors and three caseworkers). For example, one supervisor said, *"[the AFST is] a good tool to have for some extra information, in terms of risk."* Another supervisor said workers *"take in consideration what the computer [i.e. the AFST] is saying, but,... we're not making a decision based on what the computer says. If the computer says, 'This score is a 16,' per the computer we have to assign that family, and we're not doing that, we're using real... information to make the decision and not numbers to make the decision."* This supervisor also said they screen reports based on *"what risk factors are going to impact the children, because it's all about child safety."*

For call screen workers, *safety* refers to present danger or well-being of the children in a report; *risk* refers to the chance that the children will be harmed or neglected in the future. One supervisor said, *"safety is more immediate. And risk is even more long-term."* For example, one caseworker said living in a dirty home is a risk, but not an impending danger. Workers conduct *holistic, contextual assessments* when they consider any piece of information in a report to be relevant only when it's relevant to risk and safety in the context of all other information in the report (cf. [4]).

Contextual Inquiry Observations: For example, we observed a caseworker review a referral where a single mother was reported by a friend for allegedly using drugs and generally neglecting her four children (ages five to seventeen). The report also included pressing concerns about some of the children's dental health. The family was reported a month prior because one of the children was truant. The AFST score was 11 (medium, on the side of screening out), which this caseworker considered to be a low score. However, the caseworker recommended screening in the referral, because of the combination of the drug allegation, the presence of some young children, the past truancy case, and the dental health concerns. The caseworker said they would have screened out the referral if it had included only one or two of these risk factors without the others.

Overall: Because workers considered the AFST risk score in the context of all the other information in a referral, we often observed that they considered other information to be more relevant to risk and safety than the AFST score, and made screening decisions despite the score. We suspect that this holistic decision-making contributed to the baseline level of disagreement across both race and AFST-only screening decisions in Figure 4. One supervisor expressed that they thought this kind of holistic, contextual decision-making led to less racially disparate screening decisions. However, this supervisor said they did so unintentionally and that the 11% reduction in screen-in rate Black-white disparity from 2016 to 2018 was *"not the intention, just the outcome."*

5.2.2 Adjusting for limitations of the AFST. **Next, workers said they adjusted for what they perceived to be limitations of the AFST: they disregarded the AFST's recommendations when they thought it over- or under-scored referrals because it was unable to properly take all referral-relevant information into account.** Instead, workers relied on factors that they believed were more relevant in a given referral to make a decision, as discussed below.

(1) Workers thought the AFST was over- or under-scoring certain referrals because it did not take the allegation or other current referral information into account properly.

Interview findings: For example, one caseworker said that the AFST does not consider the allegation in the report when determining the risk score, *"even if it says, 'Dad killed Mom in front of the kids.' You know what I'm saying? Like, something crazy."* One caseworker gave an example where the AFST over-scored a referral because it did not properly take into account the results of a recent investigation: *"We just investigated and we found that the parents are providing fine... and we just closed it. Then some anonymous person reports the same thing. Then... [the AFST says] it's a high risk again, and we already just previously addressed it."*

Contextual Inquiry Observations: For example, we observed a caseworker review a referral where a mother was reported for allegedly not giving prescribed mental health medicine to her daughter. The caseworker immediately told us that this was a serious allegation. One prior report had also been for withholding medicine. The AFST score was a *Low risk protocol* (i.e. a mandated screen out).[14] The caseworker thought the AFST score was too low. The caseworker said they would override the AFST Low risk protocol and screen the current referral in, because of the seriousness of the current allegation and the one prior referral with a similar allegation.

(2) Workers thought that the AFST over- and under-scored referrals specifically based on system involvement, i.e. welfare, public medical services, criminal history, or CYF history.

Interview findings: Six caseworkers and three supervisors said that the AFST over-scored families with more system involvement and under-scored families with less system involvement. For example, one caseworker said that families who do not use public welfare or medical services get scored lower than families who do: *"if you were poor and you're on welfare, you're gonna score higher than a comparable family who has private insurance. Because those people go to private therapists."*[15] A caseworker said people who do not have a history of county involvement *"could totally be away from Big Brother forever."* Workers' beliefs align with prior work which suggest that the AFST is biased towards poor people and people with system involvement [31, 45]. We observed workers disregard the AFST score because they thought it was relying too heavily on system involvement and not taking relevant information into account.

Contextual Inquiry Observations: For example, we observed one caseworker review a report where they thought the AFST score was high because of system involvement, but they wanted to screen it out. A judge mandated that CYF look into the family after their child came to a juvenile probationary hearing. The child was currently incarcerated. The family had a number of prior referrals and used a lot of public behavioral health services (50+ times). Based on this history, this caseworker said *"I know [the AFST score] is gonna be high."* When the caseworker ran it, the AFST score was a 20. However, the caseworker wanted to screen out the referral, because the child was already incarcerated, so it would be no use to investigate the family. The AFST over-scored this referral based on the family's history and neglected highly-relevant context: that the child was in custody of the state and thus would not need to be investigated.

Overall: Prior work suggests that Black people have higher rates of poverty and system involvement, and that this accounts for racial disparities in the child welfare system [27, 45]. Workers understood that the AFST was over- and under-scoring referrals based on system involvement and were correcting for that by disagreeing with the AFST's recommendations for these referrals.[16] Because system involvement is also correlated with race (Black families having more involvement than white families), this would account for the higher percentage of Black children than white children that the AFST-only decisions would have screened in but workers screened out (22.3% vs. 17.6%) and for the lower percentage of Black children than white children that the AFST-only decisions would have screened out but workers screened in (15.9% vs. 29.6%) as seen in Figure 4. As stated in Section 5, this pattern of disagreement contributed to the 11% reduction in Black-white screen-in rate disparity from the AFST-only to worker-AFST decisions. In sum, because workers disregarded the AFST score more often when they perceived it to be over- or under-scoring based on system involvement (and not considering other relevant information), they were able to reduce racial disparities in call screening.

(3) Workers compensated for what they thought were the AFST's racial disparities caused by systemic racial biases in CYF reporting and county data collection.

Interview findings: Five caseworkers thought AFST-only decisions were racially disparate not only because the AFST over- and under-scored based on poverty and system involvement, but because of systemic racial biases in CYF reporting and in county information collected elsewhere—such as the medical system or the criminal system. These workers did not make screening decisions based on the race of the family, but they did consider the race of the family in order to account for systemic racial biases. These workers also thought that the AFST was biased because of over-reporting on Black families and systemic racism. One caseworker said that *"white people are not reported as much as Black kids"* and that they *"get a lot of reports on African-Americans and a lot of them are bogus.* Another caseworker also agreed with this: *"I also think [the AFST is] very biased, but so is the world."* This caseworker continued, *"the whole system is racially biased. ... It's the people entering the information [i.e. reporting families] that's affecting the [AFST] score."*

Contextual Inquiry Observations: For example, we observed a caseworker review a report which included a fifteen-year-old boy who had not been to the dentist in five years and whose teeth were severely damaged: The caseworker said, *"he needs nine root canals, seven fillings."* This was not reported by the dentist, however. The caseworker said that *"the dentist had all the information of the last five years of them trying to get [the boy] to come to the dentist... And [the dentist] didn't report him, because he's white."*

[14] This *Low risk protocol* label was introduced in the AFST Version 2 [93]. It is analogous to the mandatory screen-in label, but for screening out: so, the default decision is screen out, but it can be overridden by a supervisor.

[15] The official AFST FAQ documentation says that "receiving of public benefits" did not necessarily increase a family's AFST score: "[F]or 45% of families, receiving of public benefits (e.g., SNAP, TANF)... was associated with lower scores than for similar families that did not receive those services" [68]. As one supervisor pointed out, however, workers' perceptions could still be right if there are *"other things that are affecting that score"* that are *"just more associated"* with public welfare records, e.g. public mental health records or criminal records.

[16] Prior work also suggests that CYF workers can reliably correct for limitations in the AFST [23].

Interview findings: Workers said they do not take the race of the family into account when making screening decisions. However, they do consider racial biases in CYF reporting and county data collection. For example, when asked if they take race into account when making screening decisions, one caseworker said *"we don't treat any of the cases differently."* However, this caseworker later gave a hypothetical example of a referral where a Black family might get reported by someone who doesn't *"deal on a daily basis with people of different cultures"* and who might *"automatically assume, like, 'oh my God, holy shit, you can't swear at your kids like that.'"* But, this report would be unjustified: *"you and I may think [swearing at your kids is] tacky, but is that child hurt? Kid's not hurt."* Another caseworker also said that they consider the race of the family in order to account for biases which could affect reporting:

> *"Colorblind assessment also feels like it's ignoring the point. So, I feel like [race is] definitely something that I take into consideration. ... Bias could be affecting the way that the information is being reported. ... I feel like I'm definitely more conscious of it now."*

Overall: Some workers understood that there were systemic racial biases that existed outside of their agency—in reporters and in professionals (e.g., in the medical and criminal systems) who create the data that they then make decisions based on. These workers expressed that they make screening decisions in order to compensate for these systemic racial biases. When it came to the race of the family, most of these workers said they did not consider race qua race, but rather race qua racial biases, which could color the information that they see in a report.[17]

5.2.3 Collaborative decision-making. Screening decisions were not made by siloed, individual workers. Workers regularly made decisions collaboratively. They did so in two ways: First each report went through multiple (at least two) layers of workers in order to make the final screening decision. Second, workers often collaborated in ad hoc, impromptu ways, especially when workers were uncertain about a decision. We suspect that this collaboration contributed to mitigating workers' individual biases, and thus a reduction in the screen-in rate disparity in aggregate. However, this contribution is likely lessened, because final screening decisions are still made primarily by a few supervisors, and because the second kind of collaboration listed above is still informal and not applied consistently across all referrals.

Interview findings: First, as described in Section 3.1.1, the workflow on a standard referral necessitates that one caseworker and one supervisor review a referral. For most referrals, the caseworker receives a report, fleshes it out by searching the KIDS county database, makes a risk and safety assessment, makes a screening recommendation, and runs the AFST to generate a risk score. The report is then sent to a supervisor, who reads it in its entirety. One caseworker said the reports are not written by the caseworker alone—it is often a collaborative effort where the supervisor communicates with the caseworker about any inconsistencies, ambiguities, or mistakes in the report: *"it isn't like, 'I'm doing it. Clickety-click-click-click.'... [The supervisor] knows right away that I screwed something up."* If there are any holes in the report or any children in the report are younger than four, the supervisor may also ask for a *field screen* to send a field caseworker out to conduct a preliminary check on the family or gather more information about the report. Once the report is complete, the supervisor makes the final decision to screen the report in or out.

Supervisors primarily make the screening decisions. That being said, the process of both caseworker and supervisor reviewing a referral, communicating about the referral, making assessments of risk and safety, and making screening recommendations, is an important form of multi-layered, collaborative decision-making. Having the multi-layered process not only adds additional pairs of eyes to review each referral, it also ensures that the referrals are more likely to be reviewed by workers with different (demographic) backgrounds. One caseworker believed that the diversity among caseworkers was helping to reduce biases in the screening decisions: *"It's good to have different backgrounds with supervisors and others who make those decisions after it passes from our hands and it goes on to the next level of folks. And it's good to know there's diversity within those groups."*

Contextual Inquiry Observations: We also observed workers collaborating in ad hoc, impromptu ways that were not built into the standard decision-making process. Workers often discussed reports with other workers beyond the one caseworker and one supervisor assigned to the report. This occurred especially if they were uncertain about it. For example, while observing one caseworker (call them caseworker A), another caseworker (caseworker B) asked caseworker A to double-check that all the family members on a referral were correct, since caseworker B had taken a call about this family previously. Supervisors often talked with each other (or sometimes caseworkers) if they felt uncertain about a screening decision. For example, if they felt uncertain about a report, one supervisor said, *"we'll sit down and discuss it amongst the supervisors. 'Well, what do you think we should do with this?' It's not like we all work in a little bubble."* At one point, we observed a caseworker walk to a supervisor's desk to discuss what to do about a referral that the caseworker had reviewed and the supervisor was making the final decision about.

This process was often informal and ad hoc: Workers worked in tightly-packed cubicles in a single room. So, they would talk to one another over the cubicles, often overhear one another talk about a report and add their two cents, or walk to another worker's cubicle to ask them what they think about a certain referral. One supervisor described the call screening decision-making process as *"very collaborative"* and *"a group effort"*: *"Sometimes I'll be reviewing and I'll be like, 'Hey, [supervisor], what do you think about this?' You know, or we'll just be talking about it in the room. If the reporting source calls back, someone else will hear it and be like, 'Oh, that's my reporting source'."*

Overall: We suspect that this collaboration may have an effect of assuaging workers' biases. Prior work suggests that workers' biases play into the CYF screening process [45]. Some workers acknowledged that they have their own personal biases. For example, a caseworker said, *"I try to be conscious of my biases."* With at least a caseworker and a supervisor reviewing each referral in detail,

[17] As a note, nondiscrimination laws apply to child welfare investigations [67], which may motivate workers not to consider race or to say they do not.

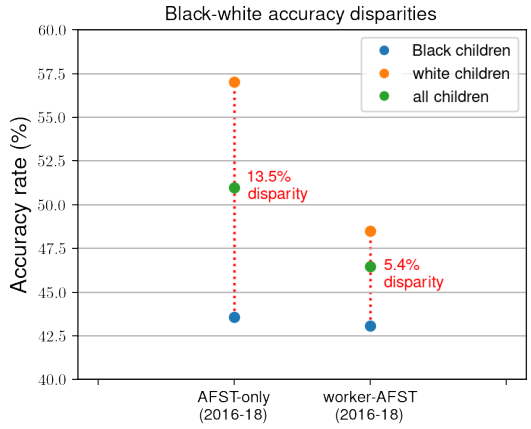


Figure 5: The differences in prediction accuracy between AFST-only and worker-AFST decisions.

workers have more of a chance to correct for each others' biases.[18] For example, one caseworker said the CYF call screening process has *"a lot of checks and balances."* Some workers themselves also expressed this view that multi-layered decision-making curbed biases. For example, when asked how workers make fair decisions, one caseworker said that *"it's good to have... others who make those decisions after it passes from [caseworkers'] hands and it goes on to the next level of folks."* By curbing workers' individual biases, we hypothesize that such collaborative decision-making may have contributed to workers' final screening decisions being less racially disparate in aggregate from 2016 to 2018. That being said, however, any bias-curbing effects of multi-layered, collaborative decision-making may be lessened due to caseworkers having less agency to influence the final screening decisions. For example, one caseworker said, *"I have very limited power... I don't have the power of saying something's not right."* Furthermore, because the second form of collaborative decision-making described above was primarily informal, ad hoc, and not built into the standard decision-making for every referral, it is unclear how many referrals would have been affected by this form of collaboration.

6 HOW WORKER-AFST COLLABORATION AFFECTED ACCURACY AND DISPARITY

6.1 AFST-only screening decisions were more accurate than workers when measured on outcomes that the AFST was trained to predict

So far, we have seen how workers made screening decisions which were less racially disparate than the AFST-only decisions. However, one of the primary arguments for the use of the AFST has been that it is more *accurate* than call screen workers [15, 20, 45, 82].[19] For example, in the official Ethical Analysis of the AFST, Dare and Gambrill [20] claim that the AFST is "more accurate than any alternative" and argue that it "is hard to conceive of an ethical argument against use of the most accurate predictive instrument." After all, an inaccurate screening decision could mean that a family is unjustly investigated—a possibly traumatic experience and a step too far towards child separation. Or it could mean that CYF does not intervene when a child will be harmed. In our setting, one concern is that a screening policy which makes less racially disparate decisions might make less accurate decisions [12, 61, 90]. For example, it may be the case that because workers make less racially disparate decisions than the AFST, they would also make less accurate decisions. In Figure 5, we see that this is ostensibly the case: the AFST-only screening policy would have been more accurate than the actual screening decisions made by workers from August 2016 to July 2018 (51.0% vs 46.5%).[20] In Figure 5, we also see that the worker-AFST decisions were less racially disparate in terms of accuracy than the AFST-only decisions (5.4% vs 13.5%). In sum, although the worker-AFST decisions achieved lower prediction accuracy than the AFST-only decisions, worker-AFST decisions were also less racially disparate than the AFST-only decisions in terms of accuracy.

However, the definition of *prediction accuracy* we adopt is particularly important when interpreting these results. Accuracy is measured against outcomes that the AFST is trained to predict, i.e. re-referral and placement. Thus, a more precise interpretation of the results in Figure 5 is that the AFST-only was 4.5% more accurate than workers at predicting the outcomes that the AFST was trained to predict. However, prior work suggests that these outcomes are biased and contested [18, 23, 50]. Furthermore, as we will discuss in Section 6.2, our qualitative findings indicate that workers disagree with the use of these outcomes, and that they are predicting different outcomes than the AFST day-to-day. Therefore, although the AFST-only screening policy was better at predicting the outcomes that it was trained to predict, it is important to keep in mind that the decisions made by workers aided by the AFST may actually be better at predicting the outcomes that workers find useful or important for preventing child abuse or neglect.[21]

6.2 Qualitative: Workers and the AFST do not agree on prediction outcomes and accuracy measures

6.2.1 Workers believe re-referral and placement are bad proxies for risk of child abuse and neglect. In an official CYF survey of Human Services [68], workers were asked about how confident they are in "the AFST's ability to accurately assess the risk of a future referral or out-of-home placement?" Workers responded with lukewarm confidence because they thought the AFST was unable "to take expected improvement or individual circumstances into account" [68]. Yet, this may be the wrong question to ask entirely: workers may or may not be confident that the AFST can accurately predict the proxy outcomes that it was trained to predict. A more

[18] Here, we mean individual biases, such as implicit biases or overt prejudices. For systemic biases, such as those built into worker training or the bureaucratic process itself, having more eyes on a report is unlikely to curb them.

[19] On the other side, one of the primary arguments against its use has been that it is not accurate enough [31, 65].

[20] To reiterate, this AFST-only policy is defined as in Section 4 with a screening threshold of 15, i.e. screen in all High risk referrals and screen out everything else.

[21] Here, we hearken back to a question Dare and Gambrill [20] bring up in the Ethical Analysis of the AFST: "The question is, how can we make the fewest errors in our efforts to protect children and families?"

fundamental question is: are caseworkers confident that *any* assessments of risk of future referral or out-of-home placement will help them make better decisions to prevent child abuse and neglect?

Interview findings: Based on our observations and interviews with workers, we believe workers lean towards 'No.' Although the AFST was intentionally designed to complement workers judgment by nudging them to consider longer-term risk, in addition to short-term safety concerns [93], workers did not view the AFST's target outcomes as relevant to their decision-making.

Re-referral. Five caseworkers and two supervisors thought that re-referral does not necessarily indicate risk of child abuse or neglect. One supervisor said that re-referral may not mean that the first decision to screen out was incorrect, since the second referral could be for an entirely different reason than the first: *"it could have been referred because the mom was outside on the porch, smoking a cigarette with a newborn baby, and then it comes back mom beat the 10-year-old. I mean, it's just a whole different reason."* Three caseworkers and one supervisor said that some reporters misuse the system and report families for unnecessary reasons. For example, one caseworker gave *"retaliation reports"* as an example: *"I call on you. You call on me. I call on you."* In our contextual inquiry, we observed a caseworker review a referral where a divorced couple called on each other three different times and each report was unsubstantiated: the caseworker said these were likely false retaliation reports. The designers of the AFST themselves noted problems with using re-referral as a target outcome and they removed it from the AFST model in 2018 [93].

Placement. Since August 2018, the AFST Version 2 only predicts placement. However, four caseworkers and one supervisor also found issue with using placement in foster care as a proxy for child abuse or neglect. A caseworker said, *"knowing the risk of removal within two years is not feeling like it's super relevant to the decision that is needed. And it is very little to do with immediate safety or anything like that."* Two caseworkers said that children were often placed in foster care without any concerns of child abuse or neglect: there are often other reasons for placing a child. For example, one caseworker said:[22]

> *"The majority of these cases [when children are placed in foster care] are not child abuse in nature, it's parent-child conflict, the kid doesn't want to live with mom or dad or the grandma, the child is saying, 'I don't feel safe, I don't want to go home.' And if they go to the police and then saying that information, some police will take custody of that child and the court has to place that kid. So it's just a lot of other variables going on that decide whether or not this child is going to be placed."*

On the other side, one caseworker and one supervisor said that placement is not the right option for many families who do have concerns of child abuse or neglect. One caseworker said families are often reported for safety concerns, without any possibility of placing a child in foster care:[23] *"just because the report is being made doesn't mean that a kid is going to go into placement."* In prior work, De-Arteaga et al. [22] similarly note that "[n]ot all cases where there is a risk to the child result in out-of-home placement." Lastly, another caseworker said that placement was not a good outcome to measure because it was a kind of self-fulfilling prophecy: *"risk of removal in two years is inherently going to be increased by our [CYF] involvement, because we're the only ones that can remove the children."* De-Arteaga et al. [22] and Coston et al. [18] also suggest this point.

6.2.2 Workers did not make decisions based on risk of re-referral or placement. Not only did workers say that re-referral and placement were not helpful outcomes to predict, we also observed workers making screening decisions based on assessments of entirely different outcomes.

Contextual Inquiry Observations: We observed that call screen workers did not try to predict whether or not children in a referral would be re-referred or placed in foster care within two years. Workers' screening decisions were based on different, shorter-term outcomes related to child safety.

For example, while reviewing referrals, one caseworker said that call screen workers make decisions by looking for *"safety concerns."* This caseworker looked for the following, for example: signs that the child felt unsafe, would be hurt, was left home alone a lot, was in contact with a child molester, that the caregiver was under the influence of drugs, etc. The caseworker also said that caseworkers asked reporting sources about the following, for example: *Is there food in the fridge? Do the kids have sheets on their beds? Is there furniture in the home? Do the parents have drug or alcohol problems? Mental health concerns? Domestic violence concerns?*

Clearly, re-referral or placement were rarely among the factors that workers assessed for. Furthermore, these factors were not indicative of risk of child abuse or neglect over the span of two years: it will not take a child two years to starve if there is currently no food in the fridge. The factors that workers look for when making their screening decisions are either shorter-term safety concerns, or specific details about the referral that could be longer-term sources of risk, regardless of whether they would lead to removal from home.

7 DISCUSSION

7.1 Interpretation of the findings

In this section, we discuss two implications of our work as it pertains to the AFST specifically. First, our work suggests that Allegheny County CYF's choice to use the AFST to aid call screen workers, rather than to replace them yielded less racially disparate screening decisions from August 2016 to May 2018. Second, our work complicates two of the primary positive arguments that prior work have identified for introducing the AFST in the first place: 1) that it decreases racial disparities; and 2) that it increases accuracy. Although our results do not run entirely contrary to these claims, we suggest that future work is necessary in order to better evaluate how the AFST positively contributes to the decision-making process, if at all.

7.1.1 Automated AFST screening decisions would have yielded larger racial disparities. In this paper, we compared two policies for using the AFST in CYF call screening from August 2016 to May 2018: first,

[22] We leave it to future work to validate this worker's statement, i.e. to see what are the primary causes of removal among discretionary referrals.

[23] This claim reflects this worker's perceptions, yet may not be entirely accurate. We leave it to future work to validate this claim.

the *hypothetical* policy where the AFST entirely automates screening decisions; and second, the *actual* policy where workers use the AFST to inform their decisions. We evaluated these policies in terms of racial disparities in both screening rates and predictive accuracy. Our results in Section 5 suggest that the automated AFST-only policy would have resulted in larger disparities in call screening: the AFST-only Black-white screen-in rate disparity would have been 20%. This is larger than both the 11.3% pre-AFST disparity from January 2015 through July 2016 and the 9% disparity for AFST-assisted worker decisions from August 2016 through May 2018. Had the AFST automated screening decisions at CYF, our results suggest that this would have increased the disparity in the rates at which Black versus white children were screened in for investigation.

Instead, when workers used the AFST, this disparity of 20% decreased to 9%. In order to understand why this occurred, we first looked at how workers disagreed with the AFST. Rather than disagreeing with the AFST in ways that directly reduced disparities, we found that workers disagreed with the AFST across the board. Some of workers' disagreements contributed to greater disparity and some served to reduce disparity. Yet, overall, worker-AFST disagreement led to a reduction in screen-in rate disparity. We then conducted a contextual inquiry and interviews with call screen workers at Allegheny County CYF to further understand this pattern. As shown in Section 5.2, we found that workers were not surprised that the AFST gave racially disparate screening recommendations. Workers made holistic screening decisions based on their knowledge of relevant context beyond just the AFST score, and they believed that this ultimately led them to make less racially disparate screening decisions than the AFST alone would have. Workers made decisions on a case-by-case basis, focused primarily on the risk and safety of the children involved in each individual report. Within these bounds, some workers made a conscious effort to reduce unwarranted racial disparities. Others believed that any reduction in disparities that occurred due to their disagreements with the AFST was incidental and unintentional, viewing this as a side effect of making decisions holistically based on various sources of information available to them. However, the effect of reducing disparity in aggregate was the same.

Our results run counter to much prior work on racial disparities in human-AI systems. As Green [36] writes, the "vast majority of research suggests that people are unable to provide reliable oversight of algorithms" and that human discretion worsens racial disparities. For example, judges have disregarded criminal risk assessment recommendations in ways that worsen racial disparities [3, 19, 47, 83–85, 96]. In Mechanical Turk experiments, people have been shown to use their discretion to make racially biased decisions [37, 38]. We present a case study in the opposite direction: worker discretion served to *decrease* racial disparities in the AFST from August 2016 to May 2018, echoing De-Arteaga et al. [23], who suggest that CYF workers used their discretion to counteract errors in the AFST. As discussed below, further research is needed to understand the impacts of worker-AFST decision-making on predictive accuracy, and to untangle the extent to which worker discretion served to decrease unwarranted versus warranted screening disparities.

In the context of this prior literature, our findings suggest that caution is needed when drawing broad generalizations about the impacts and dynamics of human-AI decision-making (cf. [23, 42, 86]). In practice, the interplay of human and AI judgment may yield very different results depending on the specific domain under study, the abilities and expertise of the people involved, and the organizational contexts in which people and AI are making predictions and decisions.

7.1.2 Was the AFST responsible for the reduction in disparity from pre- to post-AFST? A key concern for any policy change in child welfare is whether it will reduce or amplify existing racial disparities. In their description of the official AFST Impact Evaluation [35], Allegheny County Department of Human Services suggests that the *"AFST led to reductions in disparities of case opening rates between black and white children"* [69]. We also observed a reduction in screen-in rate disparity from before the introduction of the AFST to after (11.3% disparity from January 2015 through July 2016 to 9% disparity from August 2016 through May 2018). However, it is not clear that this reduction in disparity was caused by the introduction of AFST in August 2016. It is difficult to determine whether and to what extent changes in disparities can be attributed to the AFST, since the AFST was implemented across all Allegheny County CYF screening decisions (without A/B testing or randomization) and there were a number of other factors which influenced call screening—including a number of changes in practices and policy internal or external to CYF around the time of or since the deployment of the AFST.[24] It is also possible that the kinds of referrals being referred from August 2016 through May 2018 were different from those in January 2015 through July 2016, regardless of the AFST's recommendations.

Our results show that the AFST gave recommendations which were more racially disparate than workers pre-AFST. So, several interpretations are possible. For instance, it may be that workers reduced the overall disparity on their own, by making decisions largely as they had prior to the introduction of the AFST. It may be that AFST's disparate recommendations could have pressured workers to make more disparate decisions, but workers disregarded it and did the opposite. On the other hand, if the introduction of the AFST did contribute to a reduction in screen-in rate disparities from pre- to post-deployment, our results suggest that workers' discretion was integral to this reduction. One possible explanation is that the introduction of the AFST may have led workers to reflect on their own biases, resulting in decisions that were less racially disparate than their decisions in the year and a half prior to the introduction of the AFST. A second possibility is that workers and the AFST have complementary strengths and biases, and that the interplay of these led to less biased screening decisions overall. These possibilities are neither mutually exclusive nor exhaustive: some combination of these mechanisms may be at play, and additional mechanisms may be possible beyond those mentioned here. We leave it to future work to further explore whether and how the introduction of the AFST may have impacted disparities from pre- to post-deployment. In any case, our results suggest that the observed reduction in screen-in rate disparity can be attributed to human workers, whether or not they used information from the AFST to do so.

[24] Here, we note amendments to the Pennsylvania Child Protective Services Law (CPSL) that went into effect on December 31, 2014 and could have had residual effects on reporting and screening throughout the Allegheny County and the state.

7.1.3 Was the AFST more accurate than workers? Prior work suggests that the AFST makes more accurate predictions than workers [15, 20, 45, 82]. While our results in Section 6.1 show that the accuracy disparity for AFST-only decisions would have been higher than that of worker-AFST decisions, they also show that AFST-only decisions would have been more accurate than worker-only decisions from January 2015 to July 2016 (55.2% vs 49.7%) and worker-AFST decisions from August 2016 to May 2018 (51.0% vs 46.5%).[25] This increased accuracy is one of the primary reasons for the use of the AFST [15, 20, 45, 82]. An official Allegheny County Department of Human Services [5] statement echoed that "not using [the AFST] might be unethical because of its accuracy." Yet, the AFST's accuracy is measured in terms of specific proxies for abuse or neglect (re-referral and placement, for Version 1) over a long time range (two years). In Section 6.2, we found that workers did not agree with these choices of proxies, and that they viewed themselves as predicting immediate safety concerns—in other words, a *different outcome* than the AFST, over a *shorter time span.* This is consistent with suggestions from prior work [21, 22]. Thus, while our results suggest that the AFST is better at predicting the proxy variables it was trained to predict, it remains an open question how the AFST compares at the prediction and decision tasks that workers are actually trying to perform. Vaithianathan et al. [94] suggest that higher AFST Version 2 risk scores identify children with higher rates of injury-related hospitalizations, which may be closer to the kinds of target outcomes that workers actually consider when making decisions.[26] However, our results suggest that workers predict different outcomes depending on the specific referral, many of which may not involve child hospitalization. Overall, our results in Section 6.2 complicate the argument that the AFST alone is more accurate than call screen workers, and point to critical directions for future research. Without first understanding workers' objectives when making predictions and decisions, such accuracy comparisons may essentially be evaluating workers' performance on a game that they are not playing.

7.2 Design Implications for Human-AI Collaboration

Our results suggest that worker-AFST collaboration yielded lower screen-in rate racial disparity than AFST-only decisions, and lend insight into how this reduction occurred. Here, we provide positive design implications for potential ways to improve human-AI collaborative decision-making in child welfare and related contexts.

- **Be cautious about automating decisions in contexts with existing racial disparities and biases.** Our results suggest that workers reduced racial disparities in the AFST through their patterns of disagreement with the AFST score. Based on our findings, we strongly caution against fully automating decisions in high-stakes, real-world social contexts. This recommendation aligns with prior work in the child welfare context [23]. For the same reasons, even partial or 'soft' automation (like the mandatory screen-in policy in the AFST context) should be approached cautiously to ensure that they do not incentivize against potentially *productive* forms of human disagreement with algorithmic recommendations.
- **Explainable AI and interfaces to empower workers in mitigating algorithmic limitations.** Our findings suggest that workers' ability to identify instances where the AFST was over- or under-scoring a referral may have contributed to a reduction in racial disparity (5.2). However, it remains unclear how accurately and reliably workers are able to do so. Thus, one possible design implication is to develop interfaces that assist workers in identifying specific instances where an algorithm may be more or less reliable. For example, in the AFST context, providing local explanations for AFST recommendations may help workers in calibrating their reliance on the tool, case-by-case. However, given recent empirical findings demonstrating ways such local explanations can backfire in ways that harm human-AI decision-making (e.g., [6, 74]), further research is needed to understand how these can be designed and presented effectively.
- **Value sensitive AFST models.** Our results suggest that workers often adjust for what they perceive as limitations of the algorithm. For example, workers believed that the AFST was unable to sufficiently account for certain features (e.g., details of the actual allegation), and the outcomes predicted by the AFST did not align with the outcomes that workers were predicting. Zhu et al. [100] suggest involving stakeholders into the design process to make sure their values are incorporated into the design of an algorithmic system from the beginning. One design implication is for AFST designers to engage with child welfare workers, among other key stakeholders, and to incorporate their insights and feedback in future iterations of AFST. Some value misalignments between workers and the AFST were intentionally designed into the system, with the goal of nudging workers towards different practices. However, in practice we observed that these misalignments meant that workers often needed to work around the algorithm instead of working with it.
- **Promoting more collaborative decision-making.** Child welfare workers often made decisions collaboratively, both formally and informally. We suspect that this may have had the effect of curbing workers' individual biases. However, in formal collaboration, caseworkers felt they had little agency—even though they made screening recommendations. The first design implication is to encourage more regular conversations between caseworkers and supervisors about caseworkers' recommendations, so they do not get overlooked. There may also be opportunities for the AFST interface to promote more informal collaboration. For example, future versions of the AFST or similar tools could include a feature to suggest caseworkers in the office that have dealt with similar referrals in the past. This could enable workers to collaborate with the right person on each referral, which may be particularly helpful when dealing with highly uncertain referrals.
- **Diversity and lived experience among workers.** Child welfare workers were generally aware of their own individual biases. Another way to curb these biases may be to

[25] We measure the AFST's accuracy differently than most prior work, much of which uses AUC (area under the receiver operating characteristic curve).
[26] Note that AFST Version 2 also predicts foster care placement within 2 years, but not re-referral.

increase diversity among the child welfare workers, especially supervisors, who make the final decisions. As one caseworker stated, it is important to ensure that there are screeners and supervisors who have *"lived the experience"* of the families in a report.

7.3 Limitations & Future Work

One limitation of our methods is that the data used in our quantitative analysis is from August 2016 through May 2018, whereas the contextual inquiry and interviews were conducted in July 2021. Thus, if workers changed how they used the AFST significantly or they are unable to remember how they used it previously, then our results from the contextual inquiry and interviews may not reflect the reasons why workers reduced disparities from 2016 to 2018. This is likely not entirely the case, since caseworkers' workflow has remained largely consistent over this time frame (even though there have been some changes in CYF policy around the use of the AFST, including the roll-out of a new V2 model).

Another limitation of the current work is that we focus on the impacts of worker-AFST decision-making on overall disparities. We do not investigate the extent to which worker discretion served to decrease *unwarranted* versus *warranted* disparities in the AFST's screening recommendations. Our results demonstrate that worker-AFST decision-making served to reduce the disparity in child maltreatment screening rate between Black and white children compared to algorithm-only decisions. As discussed, this is an interesting finding in itself, in light of prior empirical research suggesting that human discretion often increases rather than decreases disparities. This finding also has practical implications, as racial disparities in child maltreatment screening have real consequences for children, families, and communities. For example, higher screen-in rates indicate higher levels of state intervention into families' lives, starting with an investigation. Such disparities may indicate uneven application of interventions or investigations. Importantly however, higher disparities in screening, on their own, do not necessarily imply unfairness. Disparities may be warranted if they reflect genuine differences in underlying distributions: a higher screen-in rate for one group may be justified if there is higher need or higher risk of maltreatment among children in that group [24, 28, 32, 53]. For this reason, naively attempting to equalize screening rates across demographic groups without regard for actual children's needs and safety could be harmful and counterproductive. However, it is clear from our interviews and contextual inquiries that this is *not* how workers reduced disparities in algorithmic decisions. Rather, workers appeared to reduce disparities by making holistic assessments of child risk and safety based on all of the information available to them, and by working to mitigate limitations that they perceive in the algorithm. An important direction for future research is to untangle the extent to which worker discretion serves to decrease unwarranted versus warranted screening disparities. To support such investigations, it will be critical to overcome limitations of current approaches for measuring accuracy, discussed in Sections 4.3 and 6.1.

We present two additional avenues for future work, informed by limitations of our current work. First, we leave it to future work to conduct a more comprehensive evaluation of racial disparities between worker-only and worker-AFST screening decisions. We did not focus on these results in the current paper, due to several confounding factors that complicate the analysis. These included changes in external factors—such as a 2015 state-wide policy change in mandated reporting laws and the COVID-19 pandemic since at least March 2020—for affecting the County reporting rates and the CYF screening process.[27] Second, as discussed, measuring prediction accuracy is limited by the use of proxy outcomes (re-referral or placement) that do not align with the outcomes that humans are predicting [8, 23]. It is possible that people are better than algorithms at the prediction task they are actually performing. While beyond the scope of the current paper, this remains a critical direction for future work. For example, future research could involve better understanding what constructs human workers are predicting, operationalizing these as target measures, and then re-running some of the accuracy comparisons presented in this paper using those measures.

ACKNOWLEDGMENTS

We thank Luke Guerdan, our colleagues from GroupLens Research at the University of Minnesota, and our anonymous reviewers for their valuable feedback. This work was supported by the National Science Foundation (NSF) under Award No. 1939606, 2001851, 2000782 and 1952085.

REFERENCES

[1] [n.d.]. Department Of Human Services Releases 2016 Child Protective Services Report. https://www.media.pa.gov/Pages/DHS_details.aspx?newsid=253 Online; accessed 6-January-2022.

[2] J. Khadijah Abdurahman. 2021. Calculating the Souls of Black Folk: Predictive Analytics in the New York City Administration for Children's Services. *Columbia Journal of Race and Law Forum* 11, 4 (2021), 75–110. https://journals.library.columbia.edu/index.php/cjrl/article/view/8741

[3] Alex Albright. 2019. If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions. *Law, Economics, and Business Fellows' Discussion Paper Series 85* (2019).

[4] Ali Alkhatib and Michael Bernstein. 2019. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[5] (DHS) Allegheny County Department of Human Services. [n.d.]. *DHS response to Automated Inequality by Virginia Eubanks.* https://www.alleghenycounty.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=6442461672

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[9] Joy Buolamwini. 2017. *Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers.* Master's thesis. Massachusetts Institute of Technology (MIT).

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research Conference on Fairness, Accountability, and Transparency (Conference on Fairness, Accountability, and Transparency, Vol. 81)*. 1–15.

[11] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners

[27] We visited Allegheny County CYF's Intake Department in July 2021 when many call screen workers were working from home, due to the COVID-19 pandemic. This has had an impact on the CYF call screening workflow since March 2020.

for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.
[12] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *ICDM Workshop Domain Driven Data Mining*. 13–18.
[13] Kathy Charmaz. 2014. *Constructing Grounded Theory*. SAGE.
[14] Marc Cherna. [n.d.]. We will use all resources to keep children safe. *Pittsburgh Post-Gazette* ([n. d.]). https://www.post-gazette.com/opinion/letters/2018/03/23/We-will-use-all-resources-to-keep-children-safe/stories/201803230094 Online; accessed 8-September-2021.
[15] 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*. PMLR, 134–148.
[16] Eckerd Connects. [n.d.]. ECKERD RAPID SAFETY FEEDBACK. https://eckerd.org/family-children-services/ersf/. Online; accessed 8-September-2021.
[17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.
[18] Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 582–593.
[19] Bo Cowgill. 2018. The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities. (2018). http://www.columbia.edu/~bc2656/papers/RecidAlgo.pdf
[20] Tim Dare and Eileen Gambrill. 2016. Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County. (2016). https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf Online; accessed 8-September-2021.
[21] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. 2018. Learning under selective labels in the presence of expert consistency. *arXiv preprint* arXiv:1807.00905 (2018).
[22] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. 2021. Leveraging expert consistency to improve algorithmic decision support. *arXiv preprint* arXiv:2101.09648 (2021).
[23] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[24] Alan J Dettlaff and Reiko Boyd. 2020. Racial disproportionality and disparities in the child welfare system: Why do they exist, and what can be done to address them? *The ANNALS of the American Academy of Political and Social Science* 692, 1 (2020), 253–274.
[25] Alan J Dettlaff, Stephanie L Rivaux, Donald J Baumann, John D Fluke, Joan R Rycraft, and Joyce James. 2011. Disentangling substantiation: The influence of race, income, and risk on the substantiation decision in child welfare. *Children and Youth Services Review* 33, 9 (2011), 1630–1637.
[26] Dorothy Roberts. 2002. *Shattered bonds: The color of child welfare*. Basic Books, New York.
[27] Brett Drake, Jennifer M Jolley, Paul Lanier, John Fluke, Richard P Barth, and Melissa Jonson-Reid. 2011. Racial bias in child protection? A comparison of competing explanations using national data. *Pediatrics* 127, 3 (2011), 471–478.
[28] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.
[29] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
[30] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.
[31] Virginia Eubanks. 2018. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press.
[32] 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*. 259–268.
[33] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.
[34] Child Welfare Information Gateway. [n.d.]. Safety and Risk Assessment. https://www.childwelfare.gov/topics/systemwide/assessment/family-assess/safety/ Online; accessed 25-Dec-2021.
[35] Jeremy D Goldhaber-Fiebert and Lea Prince. 2019. Impact evaluation of a predictive risk modeling tool for Allegheny county's child welfare office. *Pittsburgh: Allegheny County* (2019).
[36] Ben Green. [n.d.]. The Flaws of Policies Requiring Human Oversight of Government Algorithms. ([n. d.]). http://dx.doi.org/10.2139/ssrn.3921216
[37] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of Conference on Fairness, Accountability, and Transparency (FAT*)*. ACM.
[38] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[39] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
[40] Emil J Haller. 1985. Pupil race and elementary school ability grouping: Are teachers biased against Black children? *American Educational Research Journal* 22, 4 (1985), 465–483.
[41] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.
[42] Kenneth Holstein and Vincent Aleven. 2021. Designing for human-AI complementarity in K-12 education. *arXiv preprint arXiv:2104.01266* (2021).
[43] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2018. Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In *International Conference on Artificial Intelligence in Education*. Springer, 154–168.
[44] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. 2020. Shifting Concepts of Value: Designing Algorithmic Decision-Support Systems for Public Services. *NordiCHI* (2020), 1–12. https://doi.org/10.1145/3419249.3420149
[45] Dan Hurley. [n.d.]. Can an Algorithm Tell When Kids Are in Danger? *New York Times Magazine* ([n. d.]). nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html Online; accessed 8-September-2021.
[46] SAS Institute. [n.d.]. Analytics for Child Well-Being. https://www.sas.com/en_us/software/analytics-for-child-well-being.html. Online; accessed 12-December-2021.
[47] Sheriff's Justice Institute. 2016. Central Bond Court Report. https://www.chicagoreader.com/pdf/20161026/Sheriff_s-Justice-Institute-Central-BondCourt-Study-070616.pdf
[48] David Jackson and Gary Marx. [n.d.]. Data mining program designed to predict child abuse proves unreliable, DCFS says. *Chicago Tribune* ([n. d.]). https://www.chicagotribune.com/investigations/ct-dcfs-eckerd-met-20171206-story.html Online; accessed 6-January-2022.
[49] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. 2021. *Noise: a flaw in human judgment*. Little, Brown.
[50] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*. PMLR, 2439–2448.
[51] Emily Keddell. 2019. Algorithmic Justice in Child Protection: Statistical Fairness, Social Justice and the Implications for Practice. *Social Sciences* (2019).
[52] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. 2021. Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration. *Advances in Neural Information Processing Systems* 34 (2021).
[53] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744* (2017).
[54] Hyunil Kim, Christopher Wildeman, Melissa Jonson-Reid, and Brett Drake. 2017. Lifetime Prevalence of Investigating Child Maltreatment Among US Children. *American Journal of Public Health* 107, 2 (2017), 274–280. https://doi.org/10.2105/AJPH.2016.303545
[55] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.
[56] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[57] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.
[58] Karen Levy, Kyla E Chasalow, and Sarah Riley. 2021. Algorithms and Decision-Making in the Public Sector. *Annual Review of Law and Social Science* 17 (2021), 1–38.
[59] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering* 2, 10 (2018), 749–760.
[60] David Madras, Toniann Pitassi, and Richard Zemel. 2017. Predict responsibly: improving fairness and accuracy by learning to defer. *arXiv preprint arXiv:1711.06664* (2017).
[61] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*.

PMLR, 107–118.

[62] Clara Mosquera-Lopez, Sos Agaian, Alejandro Velez-Hoyos, and Ian Thompson. 2014. Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. *IEEE reviews in biomedical engineering* 8 (2014), 98–113.

[63] Judge Michael Nash. 2017. EXAMINATION OF USING STRUCTURED DECISION MAKING AND PREDICTIVE ANALYTICS IN ASSESSING SAFETY AND RISK IN CHILD WELFARE (ITEM NO. 49-A, AGENDA OF SEPTEMBER 20, 2016). http://ocp.lacounty.gov/Portals/OCP/PDF/Reports%20and%20Communication/Safety%20and%20Risk%20Assessment/SDM%20and%20Predictive%20Analytics%20Report%20(Risk%20Assessment%20Tools)%20(May%202017).pdf?ver=2018-10-24-083910-100 Online; accessed 6-January-2022.

[64] NCCPR: National Coalition for Child Protection Reform. 2017. Los Angeles County quietly drops its first child welfare predictive analytics experiment. https://www.nccprblog.org/2017/05/los-angeles-county-quietly-drops-its.html Online; accessed 6-January-2022.

[65] NCCPR: National Coalition for Child Protection Reform. 2018. Predictive analytics in Pittsburgh child welfare: Was the "ethics review" of Allegheny County's "scarlet number" algorithm ethical? https://www.nccprblog.org/2018/03/predictive-analytics-in-pittsburgh.html Online; accessed 8-September-2021.

[66] NCCPR: National Coalition for Child Protection Reform. 2019. No, you can't use predictive analytics to reduce racial bias in child welfare. https://www.nccprblog.org/2019/06/no-you-cant-use-predictive-analytics-to.html Online; accessed 2-December-2021.

[67] U.S. Department of Health & Human Services. [n.d.]. Protection from Discrimination in Child Welfare Activities. https://www.hhs.gov/civil-rights/for-individuals/special-topics/adoption/index.html Online; accessed 6-January-2022.

[68] Allegheny County Department of Human Services. [n.d.]. Allegheny Family Screening Tool, Frequently-Asked Questions | Updated August 2018. https://www.alleghenycountyanalytics.us/wp-content/uploads/2018/10/17-ACDHS-11_AFST_102518.pdf. Online; accessed 8-September-2021.

[69] Allegheny County Department of Human Services. 2019. Impact Evaluation Summary of the Allegheny Family Screening Too. *Pittsburgh: Allegheny County* (2019).

[70] National Council on Crime and Delinquency Children's Research Center. [n.d.]. The Structured Decision Making System for Child Protective Services Policy and Procedures Manual. https://www.cdss.ca.gov/Portals/9/SDM%20Policy%20and%20Procedure%20Manual.pdf. Online; accessed 8-September-2021.

[71] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine* 2, 1 (2019), 1–10.

[72] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 125–134.

[73] Walt L Perry. 2013. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

[74] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[75] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

[76] Naomi Schaefer Riley. 2018. Can Big Data Help Save Abused Kids? https://reason.com/2018/01/22/can-big-data-help-save-abused/ Online; accessed 8-September-2021.

[77] Dorothy Roberts. 2019. Digitizing the Carceral State (Review of [31]). *Harvard Law Review* 132 (2019), 1695—1728.

[78] Anjana Samant, Aaron Horowitz, Kath Xu, and Sophie Beiers. 2021. Family Surveillance by Algorithm: The Rapidly Spreading Tools Few Have Heard Of. *American Civil Liberties Union (ACLU)* (2021). https://www.aclu.org/sites/default/files/field_document/2021.09.28a_family_surveillance_by_algorithm.pdf

[79] Joaquin Sapien. 2016. Foiled by FOIL: How One City Agency Has Dragged Out a Request for Public Records for Nearly a Year. *ProPublica* (April 2016). https://www.propublica.org/article/how-city-agency-dragged-out-request-for-public-records-for-nearly-a-year

[80] Devansh Saxena, Karla Badillo-Urquiola, Pamela Wisniewski, and Shion Guha. 2021. A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. *arXiv* 5 (2021). arXiv:arXiv:2107.03487v2

[81] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms used within the US Child Welfare System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[82] Barbara White Stack. [n.d.]. CYF: An agency that works, helping kids and their families. *Pittsburgh Post-Gazette* ([n. d.]). https://www.post-gazette.com/opinion/Op-Ed/2018/02/11/CYF-An-agency-that-works-helping-kids-and-their-families/stories/201802040037 Online; accessed 8-September-2021.

[83] David Steinhart. 2006. Juvenile Detention Risk Assessment: A Practice Guide to Juvenile Detention Reform. https://assets.aecf.org/m/resourcedoc/aecf-juveniledetentionriskassessment1-2006.pdf

[84] Megan Stevenson. 2018. Assessing Risk Assessment in Action. *Law, Economics, and Business Fellows' Discussion Paper Series 85* 103 (2018), 303–384. https://scholarship.law.umn.edu/mlr/58/

[85] Megan Stevenson and Jennifer Doleac. 2021. Algorithmic Risk Assessment in the Hands of Humans. (2021). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440

[86] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating human + machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123* (2018).

[87] MindShare Technology. [n.d.]. Improving Outcomes using data you already have. https://mindshare-technology.com/analytics/. Online; accessed 8-September-2021.

[88] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. 83–92.

[89] Philipp Tschandl, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer Hofmann-Wellenhof, et al. 2019. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology* 20, 7 (2019), 938–947.

[90] Indrė Žliobaitė. 2015. On the relation between accuracy and fairness in binary classification. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*.

[91] Rhema Vaithianathan, Haley Dinh, Allon Kalisher, Chamari Kithulgoda, Emily Kulick, Megh Mayur, Athena Ning, and Diana Benavides Prado. [n.d.]. Implementing a Child Welfare Decision Aide in Douglas County | Methodology Report. https://csda.aut.ac.nz/__data/assets/pdf_file/0009/347715/Douglas-County-Methodology_Final_3_02_2020.pdf. Online; accessed 8-September-2021.

[92] Rhema Vaithianathan, Nan Jiang, Tim Maloney, Parma Nand, and Emily Putnam-Hornstein. 2017. Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions. https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/04/Developing-Predictive-Risk-Models-package-with-cover-1-to-post-1.pdf

[93] Rhema Vaithianathan, Emily Kulick, Emily Putnam-Hornstein, and Diana Benavides Prado. [n.d.]. Allegheny Family Screening Tool: Methodology, Version 2. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V2-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-7.pdf.

[94] Rhema Vaithianathan, Emily Putnam-Hornstein, Alexandra Chouldechova, Diana Benavides-Prado, and Rachel Berger. 2020. Hospital injury encounters of children identified by a predictive risk model for screening child maltreatment referrals: evidence from the Allegheny Family Screening Tool. *JAMA pediatrics* 174, 11 (2020).

[95] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[96] Human Rights Watch. 2017. "Not in it for Justice": How California's Pretrial Detention and Bail System Unfairly Punishes Poor People. (2017). https://www.hrw.org/report/2017/04/11/not-itjustice/how-californias-pretrial-detention-and-bail-system-unfairly

[97] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).

[98] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. *Conference on Human Factors in Computing Systems - Proceedings* (2019). https://doi.org/10.1145/3290605.3300468 arXiv:1904.09612

[99] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

[100] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

A NUMBERS AND PERCENTAGES OF CHILDREN BY RACE AND RISK LEVEL

See Table 3 for a breakdown of the numbers and percentages of children referred to CYF for discretionary referrals from August 2016 to May 2018 by risk level (*High*, *Medium*, *Low*, and *mandatory screen-in*). Here, we see a full of the AFST's disparate risk scores, as mentioned in Section 5. A higher percentage of Black children than white children were labeled *mandatory screen-in* and *High risk*; a lower percentage of Black children than white children were labeled *Medium risk* or *Low risk*.

B ADDITIONAL TERMINOLOGY

In addition to screen-in rate disparities and decision accuracy, we also evaluate the following additional metrics.

- Precision: the percentage of children that are screened-in and are either removed from home within 2 years, or re-referred again within 2 months of the referral.
- True Positive Rate (TPR): the likelihood that a child who has the positive proxy ground truth label (i.e. either removed from home within 2 years or re-referred again within 2 months of a referral) will be screened-in.
- False Positive Rate (FPR): the likelihood that a child with a negative proxy ground truth label (i.e. neither removed from home within 2 years nor re-referred again within 2 months of a referral) will be screened-in.

C COMPARISONS OF OTHER DISPARITY METRICS AND THRESHOLDS

Figure 6 shows racial disparities in screen-in, accuracy, true positive, false positive, and precision rates across possible AFST-only thresholds from pre- to post-AFST deployment. Figure 7 shows screen-in, accuracy, true positive, false positive, and precision rates by race, possible AFST-only thresholds, and AFST-only versus worker-AFST decisions from August 2016 to May 2018.

D WORKERS SEE RISK LABELS AS SCREENING RECOMMENDATIONS

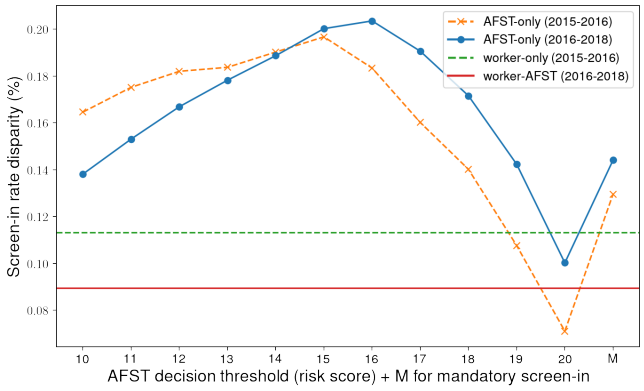
Neither official AFST documentation nor public comments from CYF leadership claim that a Low risk label (scores 1 through 9) means that a referral should should be screened out nor that a High risk label (scores 15 and up) means screen in. However, this is the message that is implied through the interface design of the AFST and is what we heard from workers. For one, the AFST interface shows scores binned into Low, Medium, and High risk levels with clear divisions between them. Low risk is green (calm, low stress) and High risk is red (urgent, severe). Although not explicitly stated in the AFST documentation, the message that Low risk referrals should be screened out and High risk referrals screened in is relayed to workers via these design choices. When we interviewed and observed workers, many implied or echoed this explicitly. One caseworker (erroneously) said that *"with a High risk [referral], you know, we absolutely have to screen them in."*[28] Another said, *"when it is High risk, I've just been going with open investigation,"* i.e. screen in. While observing one caseworker, they got a referral and explained: *"it was a Low risk [referral]... And instead of screening out, I just recommended accept for investigation."* Here, the caseworker implied that Low risk referrals should be screened out. Furthermore, five caseworkers and one supervisor said that they did not know what Medium risk scores meant or did not find them helpful when making screening decisions. In all, workers see an AFST Low risk label as a recommendation or pressure to screen out a referral, High risk label to screen in, and Medium risk confers no recommendation.

[28] Workers are not mandated to screen in High risk referrals. However, this quote exemplifies workers' perceived pressure to screen in High risk referrals.

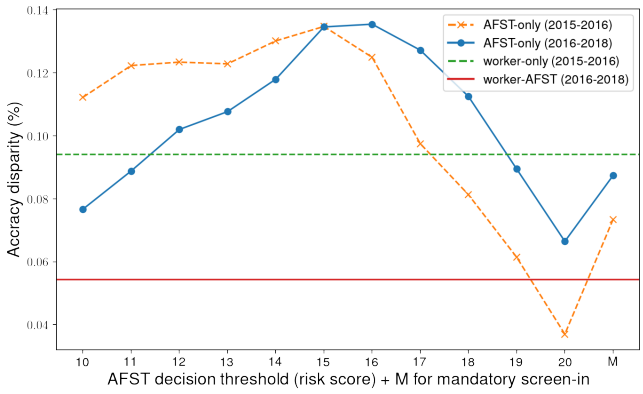
E ACCURACY BREAKDOWN BY SCORE

In this section, we examine the accuracy of AFST-only decisions (with a threshold of 15) versus worker-AFST decisions for referrals binned by AFST score or mandatory screen-in label from August 2016 to May 2018. Accuracy is calculated just as in Section 6.1, except the denominators include referrals with only a single score.

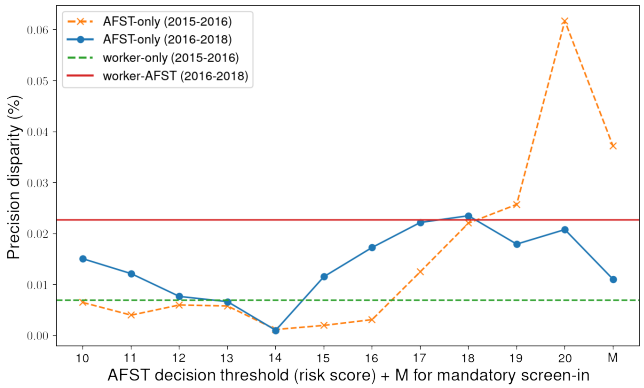
If all 15,182 children labeled as mandatory screen-ins were screened in (as the AFST-only policy would have), then 32.1% of these decisions would have been accurate. The actual screening decisions made by workers were accurate for 37.5% of these 15,182 children. Table 4 shows that worker-AFST decisions were more accurate than AFST-only decisions would have been for every score from 15 and above. Table 5 compares the accuracy of AFST-only decisions versus worker-AFST decisions for Low risk referrals (with scores 1 to 9) from August 2016 to May 2018. Worker-AFST decisions were less accurate than AFST-only decisions for every score from 1 to 9. Overall, these results suggest that worker discretion increased the accuracy for children that the AFST-only would have screened in and decreased it for those the AFST-only would have screened out. In particular, worker discretion increased the accuracy of decisions made in mandatory screen-in referrals, which are recommended to be screened in by default. However, these results are again limited by problems with evaluating counterfactual predictions, as discussed in Section 4.



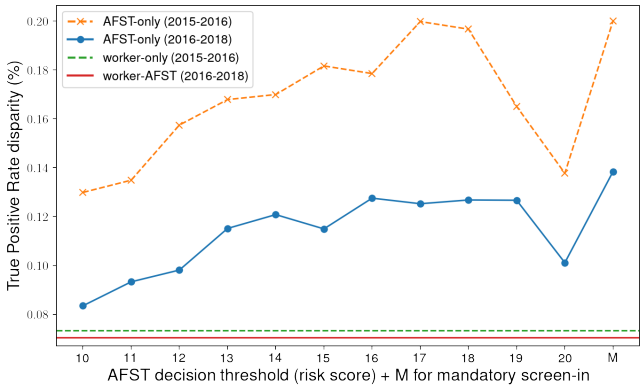
(a) Screen-in rate disparity



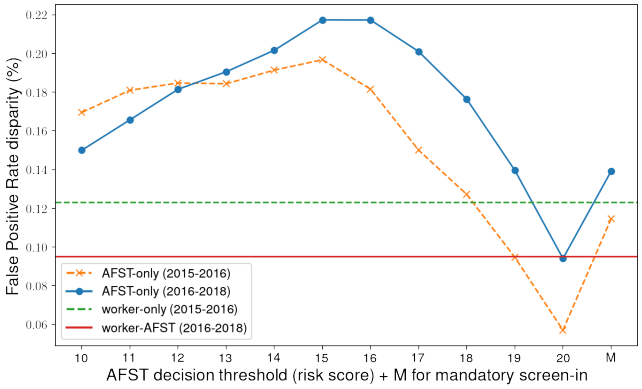
(b) Accuracy disparity



(c) Precision rate disparity



(d) True positive rate disparity



(e) False positive rate disparity

Figure 6: Common group fairness metrics showing disparities between Black and white children. The x-axis is the various decision thresholds for the hypothetical AFST-only decisions (for a given threshold, AFST-only would screen in all referrals with that score or above, and screen out all referrals under). The solid red line represents the worker-AFST final decisions after AFST deployment (2016-2018), the solid blue line with dots represents AFST-only decisions in the same time period (2016-2018). The dashed green line represents the worker-only decisions without before the AFST was deployed (2015-2016), and the dashed orange line with crosses represents the AFST-only decisions generated retrospectively for all referrals in the same time period (2015-2016).

	All children	Black children	white children
All risk levels	51750	26123 (50.5%)	21623 (41.8%)
Mandatory screen-in	15182 (29.3%)	9639 (36.9%)	4863 (22.5%)
High risk	31022 (59.9%)	18536 (71.0%)	11013 (50.9%)
Medium risk	11778 (22.8%)	5208 (19.9%)	5653 (26.1%)
Low risk	8950 (17.3%)	2379 (9.1%)	4957 (22.9%)

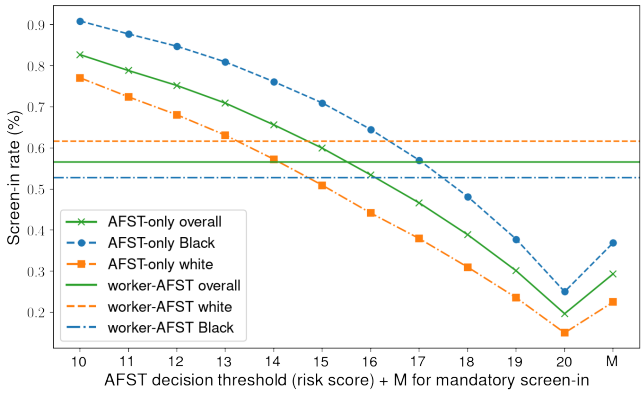
Table 3: Numbers and proportions of children by race and AFST risk level. Percentages are over total children by race, e.g. 2379 Black children labeled Low risk made up 9.1% of all 26123 Black children referred to CYF (2016–2018).

AFST score	Num. of children by score	AFST-only accuracy	Worker-AFST accuracy
15	3374	16.7%	48.1%
16	3520	19.5%	40.7%
17	4003	22.8%	40.3%
18	4543	25.2%	38.7%
19	5406	27.3%	39.3%
20	10176	35.5%	39.2%
M	15182	32.1%	37.5%

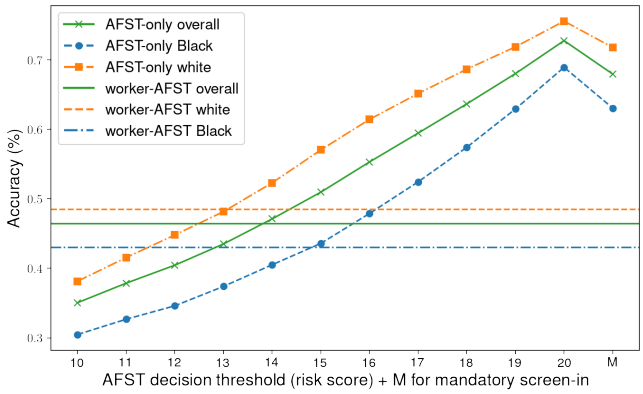
Table 4: Accuracy rates AFST-only and worker-AFST policies binned by AFST score or mandatory screen-in (M) from August 2016 to May 2018 for High risk referrals. The AFST-only policy (with a threshold of 15) would have screened in all referrals in this table, so the accuracy rates in the third column are equivalent to the percent of children who were either re-referred within 2 months or placed within 2 years. For example, of the 15,182 children labeled mandatory screen-in (placement model score 18 or higher), 32.1% were re-referred or placed in foster care. The rightmost column is the accuracy of workers' actual screening decisions.

AFST score	Num. of children by score	AFST-only accuracy	Worker-AFST accuracy
1	81	87.7%	84.0%
2	329	94.2%	79.3%
3	621	93.7%	72.0%
4	805	90.8%	64.2%
5	1063	90.1%	60.6%
6	1243	87.5%	61.5%
7	1366	89.7%	61.9%
8	1589	88.3%	57.7%
9	1853	87.1%	60.4%

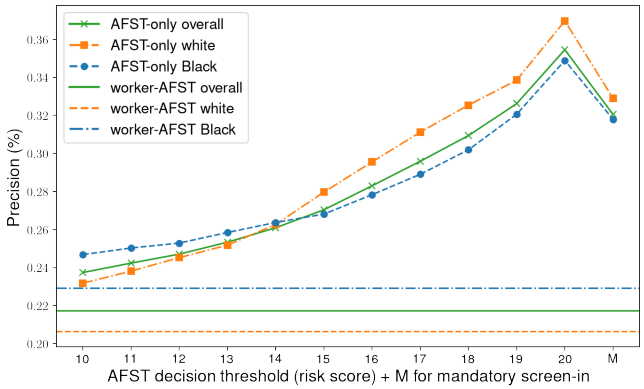
Table 5: Accuracy rates of AFST-only and worker-AFST policies binned by AFST score and mandatory screen-in (M) from August 2016 to May 2018 for Low risk referrals. Any AFST-only with threshold 10 or above would have screened out all referrals in this table, so the accuracy rates in the first column are equivalent to the percent of children who were neither re-referred within 2 months nor placed within 2 years. For example, of the 81 children with an AFST score of 1, 87.7% were neither re-referred nor placed in foster care. The rightmost column is the accuracy of workers' actual screening decisions.



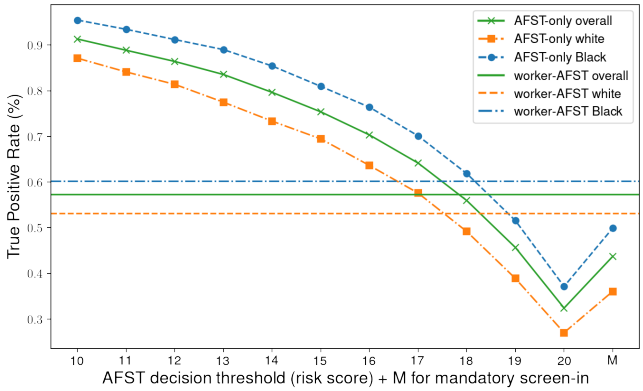
(a) Screen-in rate



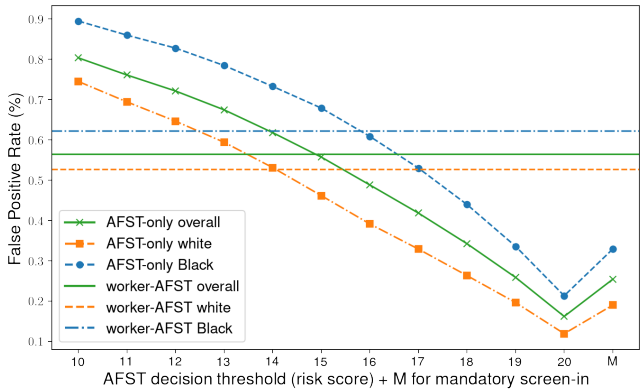
(b) Accuracy



(c) Precision



(d) True positive rate



(e) False positive rate

Figure 7: Comparisons of decision outcomes between worker-AFST and AFST-only decisions for referrals between 2016 to 2018. The x-axis is the various decision thresholds for the hypothetical AFST-only decisions (for a given threshold, AFST-only would screen in all referrals with that score or above, and screen out all referrals under).