The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool

Marissa Gerchick
American Civil Liberties Union
New York, NY, USA
mgerchick@aclu.org

Tobi Jegede
American Civil Liberties Union
Washington, DC, USA
tjegede@aclu.org

Tarak Shah
Human Rights Data Analysis Group
San Francisco, CA, USA
tarak@hrdag.org

Ana Gutierrez
American Civil Liberties Union
New York, NY, USA
agutierrez@aclu.org

Sophie Beiers
American Civil Liberties Union
Portland, OR, USA
sbeiers@aclu.org

Noam Shemtov
American Civil Liberties Union
New York, NY, USA
nshemtov@aclu.org

Kath Xu
American Civil Liberties Union
New York, NY, USA

Anjana Samant
American Civil Liberties Union
New York, NY, USA
asamant@aclu.org

Aaron Horowitz
American Civil Liberties Union
San Francisco, CA, USA
ahorowitz@aclu.org

ABSTRACT

The design decisions of developers and researchers in creating algorithmic tools — like constructing variables, performing feature selection, and binning model outputs — are sometimes cast as objective technical processes. In reality, these decisions are far from objective, and they are sometimes even made arbitrarily. In this work, we examine how algorithmic design choices can function as policy decisions through an audit of a deployed algorithmic tool, the Allegheny Family Screening Tool (AFST), used to screen calls to a child welfare agency about alleged child neglect in Allegheny County, Pennsylvania. We analyze design decisions in the AFST's development process related to feature selection, data collection, and post-processing, highlighting three values implicitly embedded in the tool through these decisions. By aggregating risk scores at the household level, the AFST effectively treats families as "risky" by association. In choosing to use training data from the criminal legal system and behavioral health agencies, the AFST prioritizes "making decisions based on as much information as possible," even when that information is potentially biased across race, disability, and other protected statuses. Finally, by including static features in the model that identify whether a person has *ever* been affected by the criminal legal system or relied on public benefits, the AFST chooses to mark families in perpetuity, compounding the impacts of systemic discrimination and foreclosing opportunities for recourse for families impacted by the tool. We explore the impacts of these decisions, individually and together, arguing that they function as policy choices that may have discriminatory effects and raise concerns about lack of democratic oversight.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594081

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Social and professional topics** → *Computing / technology policy*.

KEYWORDS

Algorithm, audit, policy, design, values, accountability

ACM Reference Format:
Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov, Kath Xu, Anjana Samant, and Aaron Horowitz. 2023. The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3593013.3594081

1 INTRODUCTION

In 2017, the creators of the Allegheny Family Screening Tool (AFST) published a report describing the development process for a predictive tool used to inform responses to calls to Allegheny County, Pennsylvania's child welfare agency about alleged child neglect.[1][2] In a footnote on page 14 of that report, the tool creators described their decisions in a key component of the variable selection process — selecting a threshold for feature selection — as "rather arbitrary" and based on "trial and error" [99]. Within this short aside lies an honest assessment of how the creators of predictive tools often view the development process: a process in which they have free rein to make choices they view as purely technical, even if those choices are made arbitrarily. But design decisions made in the development of algorithmic tools are not merely technical processes — they also

[1] **Disclaimer**: Our results are based on an analysis of the data provided to us by Allegheny County. In an effort to ensure our findings were based on a clear set of assumptions about the County's model development process, we reached out to the County for comment on our paper on January 23, 2023 and again on February 22, 2023. As of the time of publication of our paper, we have yet to receive comment from the County on our findings. In the event of comment from the County that provides material information that we were not previously provided, adjustments to our analysis may be made.

[2] The tool is not used to make screening decisions for allegations that include abuse or severe neglect, which are required to be investigated by state law [99, p. 5] [84, p. 7].

include ethical choices, value judgments, and policy decisions [70]. For example, the "rather arbitrary" threshold used in feature selection could have determined whether a family's behavioral health diagnoses or history of eligibility for public benefit programs would impact their likelihood of being investigated by the County's child welfare agency. When developers cast these kinds of design decisions as primarily technical questions [38, 92], they may disguise them as objective, even though they may be made arbitrarily, out of convenience, or based on flawed logic [88].

In this work, we demonstrate how algorithmic design choices function as policy decisions through an audit of the AFST. We highlight three values embedded in the AFST through an analysis of design decisions made in the model development process and discuss their impacts on families evaluated by the tool. Specifically, we explore the following design decisions:

- **Risky by association:** The AFST's method of grouping risk scores presents a misleading picture of families evaluated by the tool and treats families as "risky" by association.
- **The more data the better:** The County's stated goal to "make decisions based on as much information as possible" comes at the expense of already impacted and marginalized communities — as demonstrated through the use of data from the criminal legal system and behavioral health systems — despite historical and ongoing disparities in the communities targeted by those systems.
- **Marked in perpetuity:** In using features that families cannot change, the AFST effectively offers families no way to escape their pasts, compounding the impacts of systemic harm and providing no meaningful opportunity for recourse.

2 RELATED WORK

2.1 Algorithmic design as policy

As government agencies increasingly adopt predictive tools in areas ranging from healthcare [16, 36, 102], to education [24, 76], to the criminal legal system [63], and beyond, a growing body of work has focused on understanding the selection and use of algorithmic tools as encoding policy choices, including in the criminal legal system [27], health care policy [40, 89], in hiring and employment contexts [5], and in environmental regulation [10]. Mulligan and Bamberger [70] argue that, under the current paradigm of government agencies procuring and using algorithmic tools developed by third-party entities, government decision-makers abdicate important policymaking functions, allowing third-party developers' design decisions to function as policy choices. Creel and Hellman [23] examine how algorithmic decision-making systems — and the design decisions used to develop them — can be arbitrary and can have harmful impacts, and Keddell [52] explores how predictive analytics tools used in child protection can introduce or exacerbate bias and arbitrariness in decision-making. More broadly, efforts to interrogate technical systems and methods as value-laden artifacts, including in the context of use by government agencies, extend back several decades, spanning philosophy, information science, human-computer interaction, and other fields [30, 31, 61, 69].

Viewing design decisions as policy choices, a related body of work focuses on the actions and functions of data scientists, engineers, researchers, and other actors involved in the design of algorithmic systems, including in shaping the datasets used to build machine learning (ML) systems [11, 67, 68, 77], defining the measurements used to assess ML systems [44, 45, 65], and shaping how system outputs are communicated and explained [12, 14, 58]. Suresh and Guttag [95] outline the lifecycle of ML systems, describing how design decisions by developers throughout the lifecycle can contribute to downstream harms. Levy et al. [57] and Green [37] argue that data science work is inherently political, and Petty et al. [79] highlight that data extraction from communities and the related deployment of statistical tools can be dehumanizing and traumatic. As highlighted by Selbst et al. [92], algorithmic systems never exist in isolated environments; the practice of model development involves making choices that shape the social and political ecosystems in which they are deployed.

As the literature summarized in this section demonstrates, the use of algorithms by governments evokes a wide array of policy concerns. The idea that algorithmic design choices can function as policy choices is closely related to policy questions about governance of the algorithmic systems themselves. In examining how the technological affordances of the AFST may shape decision-making in the high-stakes policy context of child welfare screening, this paper is closely tied to related work considering questions of how tools like the AFST are governed or may be governed as technologies (see, e.g., [28, 39, 57]).

2.2 Predictive Analytics in the Child Welfare System

In recent literature, there has been a shift toward referring to the child welfare system as the family regulation or family policing system, based on the argument that the system disproportionately harms poor families and families of color and often responds to conditions of poverty with punishment rather than with supportive services [72, 86, 96, 101]. Throughout this paper, we use the term "child welfare" for clarity, while recognizing this literature and positing that to understand the use of predictive analytics tools in these contexts, it is important to understand the history and present dynamics of discrimination in the child welfare system [72, 86]. In the United States, Black families experience higher rates of poverty due to historic and ongoing oppression, and Black families have historically experienced the highest rates of removal of children from their families, similar only to Indigenous children in certain states [86]. More than one in two Black children in the United States will be subject to an investigation — and more than one in ten Black children will be separated from their parents and placed in the foster system — by the time they are eighteen [86]. A large body of work has demonstrated the harms of the child welfare system, including how it surveils, criminalizes, and separates marginalized families and attributes the impacts of structural inequality to individuals' failings [71, 72, 85, 86].

Advocates have long warned against the undue regulation of families by child welfare agencies and, in recent years, have raised concerns about how predictive analytics tools can perpetuate this broad family surveillance [1, 2, 29, 32, 33, 51, 52, 91]. Despite these concerns, child welfare agencies around the United States are increasingly incorporating predictive tools into various stages of their decision-making processes [6]. As of 2021, child welfare agencies in

at least 26 states and the District of Columbia had considered using predictive analytics tools, and jurisdictions in at least 11 states were actually using them [90]. These tools can vary in their application context, training data, and outcomes; for a survey and analysis of the different types of predictive analytics tools used in jurisdictions around the country, see Samant et al. [90] and Saxena et al. [91].

2.3 The Allegheny Family Screening Tool (AFST)

Developed by a team of researchers from institutions in New Zealand and the United States—in conjunction with the Allegheny County Department of Human Services (DHS)—the first iteration of the AFST, which we refer to as Version 1 (AFST V1), was launched in 2016 and was developed using data from past referrals to and investigations by DHS, medical records, and interactions with the juvenile probation system. The tool operates at the screening stage, when a call screening worker must decide whether to investigate an allegation of child neglect that comes through the hotline [75, 99]. The AFST — which is not used to make screening decisions for allegations that include abuse or severe neglect, which must be automatically screened-in for investigation under state law [99, p. 5] [84, p. 7] — estimates the probability that a child will be removed from their home by DHS and placed in foster care within two years of being referred to the agency. These probabilities are converted into risk scores between 1 and 20, which are further classified into risk "protocols" using policies developed by the County and the tool's developers. Since 2016, several additional iterations of the tool have been developed [20, 75, 98]; further details about the design of the tool and changes to the tool between iterations are described in a model card [64] we developed for the AFST, included as Appendix C. For a more detailed explanation of the County's screening process and how the AFST fits into that process, see, e.g., [99, p. 30].

Significant prior research has focused on evaluating the AFST's role in the County's child welfare processes, including examinations of its performance in deployment, its impact on racial disparities in the County's decision-making processes, and its alignment with the County's stated goals [29, 32, 94]. Several studies have examined interactions between human decision-makers and the AFST in deployment; De-Arteaga et al. [26] study a technical glitch in the deployment of AFST V1 that led to improperly calculated risk scores, using this data to retrospectively study the behavior of call screening workers interacting with the AFST. Cheng et al. [18] explore the call screening workers' adherence to the AFST's recommendations and find that, compared to the disparities that would have resulted from strict adherence to the recommendations, call screening workers' interventions reduce racial disparities in screen-in rates. Several recent works [49, 50] have explored call screening workers' interpretations of the AFST, including how they incorporate their historical knowledge of the tool and their experience in the field when making screening decisions.

To this landscape, we contribute a novel analysis of the AFST, extending a framework of algorithmic design as policy-making to the AFST's development and deployment decisions. We surface value judgments embedded in the processes used to build, deploy, and measure the tool, highlighting how these judgments effectively serve as policy decisions without meaningful democratic oversight. Viewed together and promulgated through the algorithm, these choices can exacerbate the harms of structural discrimination for already marginalized communities. We hope future work will expand upon this analysis to improve our understanding of the AFST as well as other algorithmic tools used in these and other contexts, including structured decision-making tools [7, 8, 34, 46] and other predictive tools used in the child welfare system (see [90] for a discussion of some of these tools).

3 METHODS

We analyze de-identified data produced in response to our data request by the Allegheny County Department of Human Services.[3] The data comprised approximately the same number of unique child-referral records from 2010 to 2014 as described in [98, p. 10]. Approximately the same number of records as described in [98, p. 10] had been screened in for investigation. Amongst those screened in, roughly 70% had been designated by the County as training data and 30% were designated as testing data. The data was very similar to the data used to train the version of the AFST described in [98] (we refer to the version described in [98] as AFST V2). The data differed slightly from the AFST V2 training data because of changes in the data that had occurred since the AFST V2 was developed. In particular, the data we were provided contained a few thousand child-referral records that were not used to train the AFST V2, and we were missing a small number of child-referral records that were used in the training process for the AFST V2. Both of these data sets contained roughly 800 variables, including, for each family, information about prior referrals and child welfare records, jail and juvenile probation records, behavioral health information, birth record information, and demographic information. These variables include indicators for whether a child on a referral is labelled as an "alleged victim" with regard to the referral to the hotline. Throughout our analysis, we use the term "alleged victim" to refer to individuals with this indicator, and we note that in the documentation for the AFST V1, the developers describe the County's labelling of which child or children on the referral are indicated as the alleged victim(s) as "somewhat arbitrary" because County staff are required to assess all children on a referral [99, p. 14].

In addition to this data, we were provided with weights and information about several versions of the AFST. This information included the weights for the model corresponding to Version 1 of the AFST (described in [99]) and the weights for Version 2 of the AFST (as described in [98]). We were also given the weights for the model in use at the time we received data (in July 2021), which was developed using the procedures described in [98], but differed slightly from the weights described in [98] because of updates to the model (and therefore the weights) in the time period between when [98] was written and when data was shared with us by the County. In this work, we refer to the iteration of the tool that we analyzed as AFST V2.1 (this is our term, not the County's, which we use to distinguish from AFST V2 described in [98]). We were also provided with three months of production data from early 2021 for AFST V2.1

[3] In connection with the data used in this analysis, we signed a Data Sharing Agreement with the Allegheny County, Pennsylvania Department of Human Services. The Data Sharing Agreement included a requirement that we give Allegheny County an opportunity to review our findings before publication.

and with information about the weights and design process for the version of the tool in use in Allegheny County at the time of writing of this paper (which we refer to as AFST V3). We did not have the training data or production data that corresponded to AFST V1 or V3, and accordingly, we conducted the vast majority of our analysis using the AFST V2.1 weights that were in use at the time data was shared with our team. During the time that the AFST V2 and 2.1 were in use in the County, the "protocols" associated with AFST risk scores changed several times (see [84, p. 7] for further details about the changes over time). In another paper, which has not yet been released, we explore the policy impacts of these changing protocols. For our analyses, we use the protocol most recently in place in the County associated with AFST V2 and V2.1 — and to our knowledge, also currently in place at the time of writing for AFST V3, which applies the:

- **"High-risk protocol"** to referrals where at least one person in the household has an AFST score of 18 or higher and there is at least one child on the referral under age 16. Referrals in this "protocol" are subject to mandatory screen-in, unless a supervisor overrides that default policy [98, p. 6].
- **"Low-risk protocol"** to referrals where all AFST scores in the household are 12 or lower and all children on the referral are at least 7 years old [84, p. 7]. For referrals in this protocol, screen-out is recommended.
- **"No-protocol"** to referrals that do not qualify as either "high-risk" or "low-risk." This protocol is not associated with explicit screening decisions; discretion is left to call screeners [98, p. 6]. There is no "medium-risk" protocol.

Though the County is using AFST V3 as of the time of writing, to our knowledge, several of the design decisions that we analyze in this work are still shaping this most recent version of the model. Further details about the tool and the training data are included in the model card [64] we developed for the AFST, included as Appendix C.

To understand the development and use of the AFST, we conducted an exploratory analysis of the training data used to develop the model, including an examination of the context of data sources, the processes used to construct features, and racial disparities in those features. Our analysis focused, in part, on racial disparities between Black individuals and households and non-Black individuals and households represented in this data, a grouping we used to align with the developers' reporting of results in [98]. We also conducted a review of documents related to the AFST to understand how policy choices like the screening recommendations associated with each "protocol" were made and promulgated in the context of the AFST. Based on this exploration, we selected three values embedded in the development of the AFST to highlight through the lens of design decisions related to data collection, feature selection, and the post-processing of model outputs. We analyze the impact of these design decisions on screen-in rates, racial disparities, and some of the metrics used in the development process to evaluate the AFST's performance.

Our use of various metrics — including Area Under the Curve (AUC), the Cross-Area Under the Curve (xAUC, explained in [47]) and False Positive Rates (FPR) — is not intended to suggest how the AFST should have been developed or to analyze whether the tool is fair. Reporting results for these purposes would require a complex understanding of the values that are embedded in these metrics. For example, how can we understand the tool's accuracy or evaluate when it makes errors when the tool predicts future agency actions, and the ground truth outcome upon which such results are based (whether a child is removed from their home) is informed by the tool, creating a feedback loop? We do not seek to answer such questions here. Rather, where possible, we use many of the same metrics that the County and development team used to justify the tool's creation and adoption to highlight how these design decisions have a significant effect on the tool's performance as assessed by its developers, even if we disagree with the values embedded in their assessments of the tool.

4 INTERROGATING VALUES EMBEDDED IN THE AFST

4.1 Risky by association

In creating the AFST, the developers of the tool made several consequential decisions about how to present risk scores to screening staff, ultimately transforming the model's outputs — predicted probabilities for individual children — into the format shown to call screeners — a single risk label or numeric score between 1 and 20 representing all children on a referral. In this section, we analyze this series of post-processing decisions [95] related to the aggregation and communication of the AFST's outputs. We argue first that these decisions are effectively policy choices, and that the AFST's method of grouping risk scores presents a misleading picture of families evaluated by the tool, treating families as "risky" by association, even when the risk scores of individual family members may be perceived as low. Viewing these decisions as policy choices, we highlight several additional ways these decisions could have been analyzed throughout the AFST's design and deployment process, which produce varying pictures of how the tool performs.

4.1.1 The AFST's method of grouping risk scores. In the first public development report about the AFST, the tool's developers wrote that "of considerable debate and discussion were questions surrounding how to present the risk scores to hotline screening staff" [99, p. 27]. Ultimately, the developers decided to transform the AFST's predicted probabilities into risk scores between 1 and 20, where each score represents five percent of the child-referral combinations in the testing data used to develop the model (i.e., using ventiles) [98, p. 10]. Perhaps worth noting is that in their 2018 analysis of the AFST, Chouldechova et al. characterize the choice of ventiles for the AFST as "not a principled decision" [20, p. 5]. For an example of how to interpret an AFST score, for a referral occurring in 2021, when the AFST V2.1 was in use, a score of 20 for a child indicated that the estimated probability of removal for that child was within the top five percent of probabilities in the testing data used to develop the model, which was based on referrals made between 2010 and 2014. Here, being in the top five percent is a relative measure — as highlighted in other analyses of the AFST [20], individuals who receive the highest possible risk score experience removal less than 50% of the time, a result that might differ starkly from intuitive interpretations of a score in the top five percent.

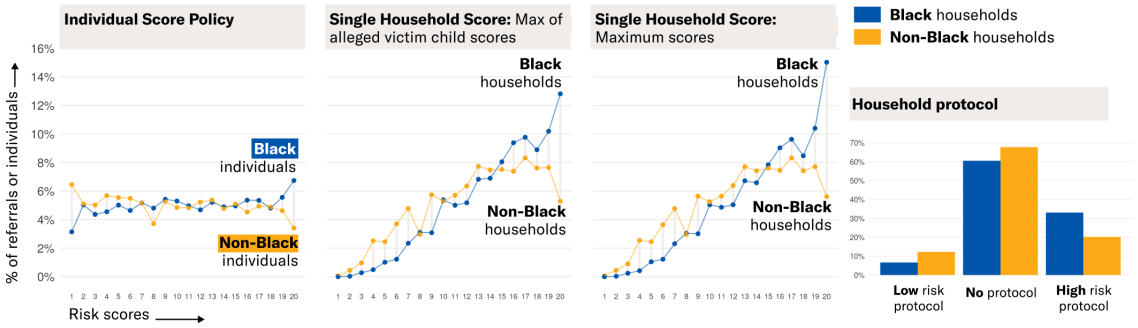

Figure 1: Distribution of Risk Scores by Race Under Different Scoring Policies. Under policies that assign a single score or protocol to the entire household, risk scores generally increase for all families relative to the individual score policy, and Black households receive the highest risk scores more often than non-Black households. Under the household protocol policy, 33% of Black households are "high-risk" while only 20% of non-Black households are "high-risk."

In addition to using ventiles for the scores, the AFST aggregates risk scores for all children in a household, presenting a single score or label that represents an entire family to call screeners. Though predictions are generated at the child level for each referral, call screeners either see only the maximum score across all children on a referral or a single risk label (e.g., "high-risk") that is determined in part by the maximum score of all children on the referral. Aggregation of risk scores sometimes occurs in other settings, including in the context of pretrial risk assessments in the criminal legal system, where researchers have repeatedly raised concerns about the combination of risk scores related to predictions of different outcomes for the same person [35, 62]. In the context of the AFST, the interpretation of scores for each child is somewhat complex before the household aggregation occurs; this interpretation is further muddied by the aggregation of scores at the referral level. A score of 20 for a referral means that, *for at least one child in the household*, the estimated probability of removal is within the top five percent of probabilities in the testing data from 2010 - 2014.

Imagine a referral related to a hypothetical family with three children, aged 5, 10, and 15 respectively, with AFST scores of 5, 10, and 18. One child, the five-year-old child with a risk score of 5, is labelled as the alleged victim by the County on the referral. How could this information be communicated to the call screener for the referral? As noted in Section 3, the County has a policy of evaluating all of the children on a referral when a call is received — not just those indicated as alleged victims — and this policy predates the AFST [99, p. 14]. But the existence of this policy alone does not answer this question of how scores are communicated to call screeners. For example, one option would be to show each child's individual score to the call screener, for a total of three scores. Or, with a constraint of only showing one score, the AFST could have displayed the score of the alleged victim (a score of 5), or the maximum score of all children (a score of 18), or a label such as "high-risk" for the entire household based on the score and the children's ages, akin to the County's current protocol policy. Under the policy that, to our knowledge, is currently in use in the County, this family would be grouped into the "high-risk protocol."

Each of these methods would have significant implications for the distribution of risk scores in the testing data used to develop the model. As highlighted in Figure 1, scoring policies that assign a single score to the entire household confuse the interpretation of the ventile scores.[4] Under the individual score policy (shown in the first panel of Figure 1), each numeric risk score generally corresponds to roughly 5% of the individuals in the data, and similar percentages of Black and non-Black families are assigned each numeric risk score (with the exception of scores 1 and 20). But under the policies that produce one score for each household (shown in the second and third panels of Figure 1), this distribution is heavily shifted upwards and disparately shifted for Black households.[5] For these policies, risk scores generally increase for everyone compared to the individual score policy. But racial disparities in the higher-score ranges are severely exacerbated by the household score policies — non-Black families are more likely to have scores below 10, and Black families are more likely to have scores above 10, with severe disparities at the numeric score of 20. Under a protocol policy like that currently in use in the County — where families are assigned to either the "high-risk protocol," "low-risk protocol," or "no-protocol" based on the maximum score in the household and the ages of the children on the referral (shown in the fourth panel of Figure 1) — 33% of Black households would have been labelled "high-risk," compared to 20% of non-Black households. Household size [6] does not account for these disparities; Black households are, on average, assigned higher risk scores than non-Black households of the same size. We discuss these results further in Appendix A.1.

4.1.2 Ways of measuring the AFST's performance. In the AFST's development reports, the tool's developers generally present results about the performance of the AFST using individual scores and measuring individual outcomes, examining whether, for each child,

[4] For the scoring policy shown in the second panel of Figure 1 (Single Household Score: Max of alleged victim child scores), we use the maximum household score for referrals where there is no one indicated as the "victim child" in the data. These instances were a very small portion (approximately .1%) of the overall data used in Figure 1.

[5] Here, we use the term "Black household" to describe households where at least one person on the referral is recorded as Black in the data given to us by the County; this approach follows the County's approach for defining Black households to measure racial disparities in screen-in rates.

[6] Throughout this paper, "household size" refers to the number of children associated with a household-referral in the data given to us by the County. We did not have information about the total household size associated with each referral.

Metric	Group	AFST Developer Reported Results for V1 and V2	Range of Possible Results for V2.1 Using Different Measures
Traditional AUC	Overall	0.7597 (for V2; see [98])	0.679 - 0.739
	Black families	0.7442 (for V2; see [98])	0.668 - 0.742
	Non-Black families	0.7735 (for V2; see [98])	0.672 - 0.731
Cross-AUC [47]	Black families	Not computed	0.566 - 0.703
	Non-Black families	Not computed	0.754 - 0.800
False Positive Rate	Overall	0.20 (for V1; see Table 1, [20])	0.20 - 0.44
	Black families	<0.25 (for V1; see Fig. 5, [20])	0.22 - 0.51
	Non-Black families	Not computed (for V1; see Fig. 5, [20])	0.17 - 0.37

Table 1: Performance results for the AFST V2.1 along various metrics, comparing reported results by the tool's developers for V1 and V2 and our estimations of the range of results each metric could take when using different methods of grouping scores and measuring outcomes at the individual and household level. For more granular results, see Appendix A.1.

a removal occurred within two years of a referral. These results inform key design decisions, including the modelling approach ultimately selected for the tool [98], but there is a mismatch between how the tool's results are reported (at the individual level) and how the tool is actually deployed — where each household receives only one score. To evaluate the impact of this mismatch, we define and analyze six different ways that risk scores could have been communicated and that outcomes could have been measured in the context of the AFST, which we refer to for this analysis as *policies*. These six policies — which could be shared and debated as formal policies about risk scores and household treatment — represent the possible combinations of several different score aggregation methods (using individual scores, maximum household scores, the alleged victim child's score, or household "protocols") and outcome measurements (measuring outcomes at the individual level or the household level, examining whether a removal occurred for *any* of the children on the referral within two years). The specifics of these combinations and a description of how scores and outcomes would be measured under each policy for the hypothetical family discussed in the previous section are included in Appendix A.1.

For each of these "policies," we evaluate the AFST's performance with metrics that were used in the tool's development reports and analyses about the tool, including the area under the receiver operating characteristic (ROC) curve (AUC) as in [98] and False Positive Rates (FPR) as defined in [20] for the policies that output ventile scores (the definition of a false positive for the policies that output protocols is included in Appendix A.1). We also generate results using the Cross-Area Under the Curve (xAUC) metric and associated Cross-Receivier Operating Characteristic (xROC) curve as proposed by Kallus and Zhou [47], which recognizes that predictive risk scores are often used for ranking individuals in settings with binary outcomes. The developers of the AFST and other researchers, such as Chouldechova et al. [20], make arguments about the fairness of the AFST in part based on race-specific AUC metrics. However, simply grouping by race and computing AUC for each group does not fully reflect the way that models like the AFST are used in practice. The AFST estimates the likelihood of a binary outcome: whether or not a child will be removed within two years. But the scores produced by the AFST are not just utilized in a binary manner: as Chouldechova et al. highlight in their visualization of the referral process [20, p. 12], the AFST informs workers' screening decisions as well as recommendations about service information and provision. As such, we can think of the AFST as seeking to rank children who will be removed above those who will not be removed, so we also present results for Black families and non-Black families using the xAUC metric [47].

Our results, summarized in Table 1 and broken down in more detail in Appendix A.1, indicate that *how* the tool is measured is consequential for our understanding of how the AFST performs. For each metric, we compute results for the AFST using each "policy," and demonstrate how these policies produce varying pictures of the AFST's performance by including the range of possible performance results generated for each metric in Table 1 (individual results for each policy and metric are included in Appendix A.1).

In our analysis, the AFST often produces the "best" results (e.g., with the lowest FPR and highest AUC) when it is measured as the County measured it: at the individual score and outcome level. But when we measure the AFST in a manner more closely aligned with how it is deployed — using a maximum score policy or a policy of assigning a "protocol" to each family — we sometimes see a lower AUC, a higher false positive rate, and greater racial disparities in performance results (see Appendix A.1 for further details). The cross-AUC analysis — across all of the policies — suggests significantly worse performance for Black people compared to non-Black people; additional detail is included in Appendix A.1.

4.1.3 Discussion. Our findings highlight that post-processing decisions about how to communicate and aggregate model outputs can be consequential. Determinations about how to measure an algorithmic tool's performance are not objective; they are subject to multiple alternatives, each with important consequences. These results support the emerging literature examining the importance of measurement in the design and oversight of AI systems, which posits that measurement is in itself a governance process and explores how harms stemming from algorithmic systems can sometimes be traced in part to measurement mismatches [44, 45]. Importantly, in this analysis, we do not impose normative judgments or recommendations about what values would represent an

acceptable result on each of these metrics. We also do not intend for this analysis of metrics to be used to argue about whether the tool is fair, or to advocate for the use of specific metrics to define fairness in this context, recognizing that debates about defining algorithmic fairness in the context of specific decision-points often fail to address the realities of how algorithmic systems operate in practice. Models that appear to be fair using these kinds of metrics can still operate in and exacerbate the harms of oppressive systems [38].

4.2 The more data the better

One of the County's goals in adopting the AFST, as stated in a 2018 process evaluation, was to "make decisions based on as much information as possible" [43]. This "information" included data from the County's "Data Warehouse" [22] such as records from juvenile and adult criminal legal systems, public welfare agencies, and behavioral health agencies and programs. Each of these databases contributed features that were used in the model, which was developed using LASSO logistic regression "trained to optimize for the [Area Under the ROC Curve] AUC" [98].[7] In this section, we explore the impacts of the inclusion of these features in the model, focusing on features related to behavioral health — which can include or be related to disability status — and involvement with the criminal legal system. We highlight concerning disparities related to these features and show that excluding these features from the model would not have significantly impacted the main metric the model was trained to optimize for — the AUC [98]. Ultimately, we find that the County's stated goal to "make decisions based on as much information as possible" [43] comes at the expense of already impacted and marginalized communities and risks perpetuating systemic racism and oppression.

4.2.1 Features from the criminal legal system in the AFST. Every version of the AFST has included features related to involvement with the criminal legal system, including incarceration in the Allegheny County Jail or interactions with juvenile probation (see [99] and [98] for details on these features in V1 and V2 respectively; these types of features are also used in more recent versions of the model). As part of our analysis of the AFST feature selection process — discussed further in Appendix A.2 — we examined whether there were features of the model that were disproportionately associated with Black children compared to white children. Some of the largest racial disparities we found were for features of the model related to juvenile probation, including indicator features for whether the alleged victim had ever been involved with juvenile probation or whether the alleged victim was involved with juvenile probation at the time of the referral (these features are discussed further in [98]). Black alleged victim children were almost three times more likely to have been or to currently be on juvenile probation at the time of the referral compared to white alleged victims. The AFST also includes features for whether other members of the household have ever been in the juvenile probation system (see [98]), regardless of how long ago or the details of the case. All of these features

[7] More precisely, the developers selected the model with the largest value for the regularization parameter λ among a sequence of candidate values, such that the resulting AUC was within one-standard error of the maximum observed AUC across all candidates [93].

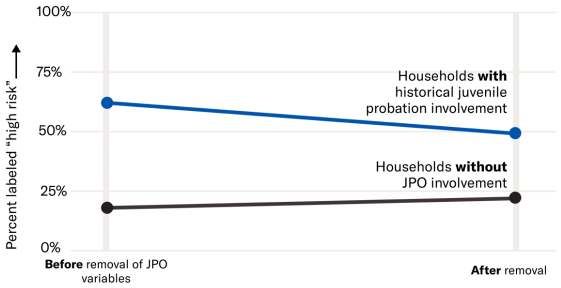

Figure 2: Households marked by any history with the juvenile probation system are overwhelmingly labeled "high-risk", although less so after removing juvenile probation (JPO) predictors from the training data.

can increase an individual's and household's AFST score, raising serious concerns that including these racially disparate features — which reflect the racially biased policing and criminal legal systems [4, 9, 60, 80, 83] — could exacerbate and reify existing racial biases. Including these variables has the effect of casting suspicion on the subset of referrals with at least one person marked by the "ever-in juvenile probation" flag, which disproportionately marks referral-households with at least one member whose race as reported in the data is Black. By this definition, 27% of referrals involving Black households in the county data have a member with the juvenile probation flag, compared to 9% of non-Black referral-households. Overall, 69% of referral-households with this flag are Black.

As shown in Table 2, removing the juvenile probation system data from the training data had a minimal impact on the County's own measure of predictive performance, AUC, which changed from 0.739 to 0.737 (the retraining process is described in more detail in Appendix A.4). While removing the variables related to the juvenile probation system would have impacted the performance of the model by a negligible amount, it would have reduced the percentage of families affected by juvenile probation labeled as high-risk by over 10%, as shown in Figure 2.

4.2.2 Disability-related features in the AFST. In developing the AFST, the County and the research team that developed the tool used multiple data sources that contained direct and indirect references to disability-related information. For example, the first version of the AFST [99] and the version currently deployed in Allegheny County as of the time of writing (AFST V3) include features related to whether people involved with a referral have recorded diagnoses of various behavioral and mental health disorders that have been considered disabilities under the Americans with Disabilities Act (ADA). Versions of the tool [99] have also included features related to public benefits — such as Supplemental Security Income (SSI) benefits — that may be related to or potentially proxies for an individual's disability status [3]. Some iterations of the tool have excluded some of the features pulled from public benefits or behavioral health sources (public reports indicate that the County's behavioral health agencies focus on services related to mental health and/or substance abuse [22]). For example, some public benefits data that was included in AFST V1 was excluded from AFST V2 and V2.1 because of changes in the data format [98, p. 4].

Importantly, these records come from public agencies, so families who access disability-related health care through private services are likely not recorded in these data sources. As highlighted in a 2017 ethical analysis [25] of the AFST commissioned by the County, the people whose records are included in these databases may have no way to opt out of this kind of data collection and surveillance.

We analyze the inclusion of three features directly related to disability status included in V2.1 of the AFST — an indicator variable for whether the "victim child" has any behavioral health history in the database, an indicator variable for whether the alleged "perpetrator" has any behavioral health history in the database, and, for a "parent" with a behavioral health history in the database, the number of days since they were last seen in behavioral health services (see [98, p. 4] for further discussion of these features). These three features are the only variables from the behavioral health data source with a non-zero weight in V2.1 of the AFST. However, disability status may have a larger overall impact on the model if other features in the model (like features related to eligibility for public benefits programs, involvement with the criminal legal system, or others) are proxies for disability status. Because of the weights assigned to these three features in V2.1 and the binning procedure used to convert probabilities to AFST scores, being associated (through a referral) with people who have a disability and access services related to those disabilities — as encoded through these variables — can increase an individual's AFST score by several points. This finding, discussed further in our Appendix A.2, is not just a theoretical possibility. In both the training data we reviewed and the production data from 2021, we identified several examples of individuals who had identical values for each feature considered by the model except for the indicator variable for whether the alleged "perpetrator" had any behavioral health history in the database. Among these matches, individuals with the behavioral health indicator had scores 0-3 points higher than those without the behavioral health indicator.

In addition, retraining the model with this behavioral health data removed as a source of potential feature candidates produces a model with a very similar AUC (see Table 2), raising the concern that the inclusion of these features may have an adverse impact without adding to the model's "predictive accuracy" as defined by the tool developers.

4.2.3 Discussion. Feature selection is one of many consequential decisions with policy impacts in the design of algorithmic tools. There are many different ways to perform feature selection, which is often focused on ensuring that only those variables that allow a model to perform best on a performance metric decided on by the model's developers are kept in the model [17, 53, 54]. One common approach for feature selection includes using some kind of accuracy maximization as a lone heuristic for feature selection, potentially based on a misplaced belief that there may be a single most-accurate model for a given prediction task [13, 63]. However, emerging research has highlighted the prevalence of model multiplicity — tasks or contexts where several different models produce equivalent levels of accuracy while using different features or architectures [13]. The development of the AFST was guided by AUC-maximization [98], an imperfect measure of accuracy [59], and one that is unable to meaningfully distinguish between models with and without disability-related and juvenile probation-related features. Given this model multiplicity in the context of the AFST, deciding whether to include or exclude variables from the juvenile probation system is fundamentally a policy choice. Even taking as given a prioritization of some measure of accuracy in the model development process, model multiplicity could allow tool developers and designers to prioritize considerations beyond accuracy [13] — including interpretability, opportunity for recourse, and fairness.

4.3 Marked in perpetuity

When algorithmic decision-making systems are deployed, impacted communities are often left without concrete protections and actionable resources to respond to those systems and harms that may stem from them [56]. Emerging literature focused on the concept of algorithmic recourse [48, 55, 82] explores opportunities for individuals affected by algorithmic systems to contest and challenge system outputs, potentially resulting in changes to the model's predictions or broader decision-making processes [15, 48, 78, 87]. In the context of the AFST, we examine recourse for call screeners who act on the tool's outputs and families who are evaluated by the tool, focusing on whether these groups have the ability to be aware of a risk score, understand the specific reasons and model features that led to the risk score determination, and contest both the risk score generated and the inputs to the model. We also explore the use of "ever-in" features — which indicate whether someone involved with a referral has *ever* been eligible for public benefits programs or been affected by the racially biased criminal legal and policing systems — through the lens of algorithmic recourse. We argue that including these static features in the model is a policy choice with serious impacts for families evaluated by the tool, including for opportunities for recourse and contesting the tool's predictions. We explore racial disparities in the presence of these features and examine whether "ever-in" features add "predictive value" as defined by the tool's creators. We find that the use of these features compounds the impacts of systemic discrimination and forecloses opportunities for meaningful recourse for families impacted by the tool.

4.3.1 "Ever-in" features and implications for recourse. The AFST includes several features that mark individuals or families in perpetuity — we refer to these collectively as "ever-in" predictors, as they are all defined in terms of a person ever having a record in a given data system, and "ever-in" is also used by the tool designers to refer to these features [98, 99]. These systems include whether a member of the household has ever been in the Allegheny County Jail or ever been in the Juvenile Probation system, as well as whether household members have ever been eligible for each of a range of public benefits programs administered by Pennsylvania's Department of Human Services, including Temporary Assistance for Needy Families (TANF) and SSI (see [98, 99] for details on these features). As highlighted in Section 4.2, some public benefits features used in earlier versions of the tool were excluded from this version of the tool. Because they are immutable, these predictors have the effect of casting permanent suspicion and offer no means of recourse for families marked by these indicators. If a parent in the household ever spent time in the Allegheny County Jail (regardless of the charges or whether that charge resulted in a conviction) or was

Model	AUC
Baseline — AFST V2.1	0.739
No juvenile probation (JPO) variables	0.737
No behavioral health (BH) variables	0.735
No "ever-in" variables	0.737
No JPO, BH, or "ever-in" variables	0.730

Table 2: AUC generated by retraining the model using the same procedure without juvenile probation (JPO), behavioral health (BH) or "ever-in" variables. Removing each of these sets of variables from the training data — or removing all of them — reduces the AUC on the test data of the resulting model by less than .01 in all cases.

ever eligible for public benefits (meaning they were enrolled in the state's managed Medicaid program, HealthChoices, whether or not they then used the benefits), they are forever seen as riskier to their children compared to parents whom these systems haven't reached.

Data stemming from criminal justice and policing systems is notoriously error-ridden and reflects the discriminatory practices and racially disproportionate harms of those systems [60, 83]. Our analysis of referrals from 2010 - 2014 showed that 27% of referrals of households with at least one Black member were affected by the "ever-in" Juvenile Probation predictor, compared to 9% of non-Black referral-households. Similarly, 65% of Black referral-households in the data were impacted by the "ever-in" jail indicator variable, compared to 40% of non-Black referral-households. Examining the public benefits features, we found that 97% of Black referral-households in the data were impacted by at least one of the "ever-in" variables coming from the public benefits data sources, compared to 80% of non-Black referral-households. As is the case with the behavioral health data discussed in the previous section, only families who access or establish eligibility for public services are represented in the public benefits data ingested by the AFST, meaning this data also reflects the historic and ongoing oppression and racial discrimination that contributes to higher rates of poverty for Black families compared to non-Black families and the historical racism that has shaped the benefits programs themselves [81, 86]. Public benefits databases are also not immune to serious data errors. For example, Colorado's state public benefits database, which is used as a data source for a similar predictive tool used by at least one child welfare agency in Colorado [97], has suffered from systematic errors for decades [42].

The use of these features in the AFST has the effect of creating an additional burden for families who have been impacted by public data systems. This additional burden is imposed not after political debate or an adversarial legal process, but quietly, through the decision to encode membership in a County database with a 1 or 0, and the related decision to use that encoding in a predictive model, despite a lack of demonstrable predictive benefit. When we re-trained the model with the "ever-in" variables excluded from the training data, we found that the baseline model had an AUC of 0.739, and the model without the "ever-in" variables had an AUC of 0.737 (see Table 2), a difference of only .002. The overall effect of this inclusion is compounded by the decision to aggregate risk scores by taking the maximum score across all individuals on a referral. Figure 3 shows that, as the number of people associated with a referral increases, the likelihood that at least one of them will have some history with the Allegheny County Jail and/or the HealthChoices program increases as well, and this is especially true for referrals associated with households who have at least one Black member. Overall, using the testing data from the model development process and the county's "high-risk protocol" to measure screen-ins that would have been recommended by the tool, we see referrals where at least one of the associated household members is Black (as recorded in the data) being recommended for screen-in at systematically higher rates than non-Black referral-households, and the disparity grows with the size of the referral (see Appendix A.3 for more detail). For referrals containing five or more individual records, 47.4% of Black referral-households are recommended for screen-in compared to 30% of non-Black referral-households, for a disparity of 17.4%. These disparities persist, but are reduced, after removing the "ever-in" variables before training the model. For instance, the racial disparity for referral-households with five or more people drops to 15.6% under the model without the "ever-in" predictors. We include additional detail and summaries in Appendix A.3.

4.3.2 Recourse in deployment. The notion of algorithmic recourse has been defined and operationalized in various ways, invoking different aspects of an algorithmic system's development and use and focusing on different types of agency that recourse could afford to individuals and groups affected by algorithmic systems [48, 78]. For example, some definitions suggest that algorithmic recourse is just the ability for affected parties to receive explanations about how an algorithm generated a particular output [48], while other definitions suggest that recourse is the ability to not only know how an algorithm reaches its decisions, but also to change decisions stemming from an algorithmic system [100]. Through its use of "ever-in" features, the AFST encodes a lack of recourse for families who have been affected by the criminal legal system or who have enrolled in public benefits programs.

We can also examine the context in which the tool operates to understand potential algorithmic recourse for people impacted by the AFST, including families and others who interact with the tool. Prior research and reporting about the AFST has touched on many elements of recourse related to the tool's deployment. For example, reporting about the AFST has highlighted how many affected families may be unaware that a predictive tool is even being used in the child welfare system in Allegheny County [41], and even when families do know a predictive tool is being used, emerging research suggests they are concerned about the tool's use and oversight, but lack a clear avenue for their concerns to be addressed [94]. Recourse for call screeners is also an important question; for instance, Cheng et al. [18] found that some call screeners think families with previous system involvement are assigned higher risk scores than other families, and to adjust for this perception, call screeners would sometimes disregard the AFST's risk assessment determination if they thought that the principal reason for a family's high risk score was their socioeconomic status. By challenging the outputs of the

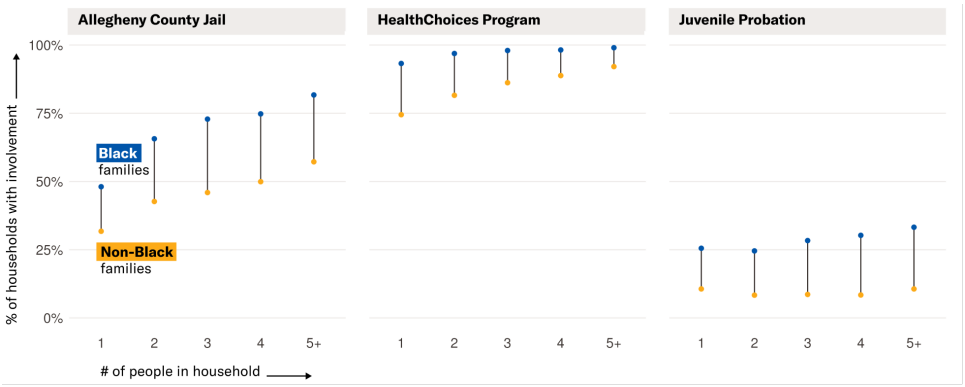

Figure 3: As household sizes increase, the likelihood that at least one member of the household will have some history with the Allegheny County Jail (ACJ), the HealthChoices program (HC), or juvenile probation (JPO) — as marked through "ever-in" features — increases as well. Black households are more likely to have involvement with these systems.

AFST, child welfare workers reduced the racial disparity in screen-in rates between Black and white children that would have resulted from complete adherence to the algorithm by 11% [18]. Much of the call screeners' work of looking for the reasoning behind racially and socioeconomically disparate risk score outputs from the AFST model may be guesswork [49], as call screeners are not shown the precise values of features that are used to generate scores.

5 LIMITATIONS

This work is one part of a large and growing body of work examining the design, deployment, and impacts of algorithmic tools. Our analysis focused on a limited set of design decisions related to a particular version of the AFST. To our knowledge, several of these design decisions are still shaping the deployed version of the tool, though we were only able to analyze the *impacts* of these decisions for V2.1 of the tool. While we think this work provides a useful case study with applications to predictive algorithms more broadly, the nuances of our findings may not extend to the design of similar predictive tools used in other contexts or to the use of different algorithmic tools used in the child welfare system itself.

As with any algorithmic tool, examinations of the tool should be holistic and should consider aspects of design and deployment beyond those considered here, including evaluations of the version of the AFST currently deployed, more detailed analyses examining outcomes during tool deployment, and other evaluations covering questions discussed in [18, 19, 26, 50] and other works. Future work should further explore the specific issues we highlight here and should consider similar questions in the context of structured decision-making tools [7, 8, 34, 46] and other predictive tools used in the child welfare system [90] and other contexts.

6 CONCLUSION

A 2017 ethical analysis of the AFST described "predictive risk modeling tools" in general as "more accurate than any alternative" and "more transparent than alternatives" [25, p. 4]. In its response to this analysis, the County similarly called the AFST "more accurate" and "inherently more transparent" than current decision-making strategies [74, p. 2]. But when tools like the AFST are created with arbitrary design decisions, give families no opportunity for recourse, perpetuate racial bias, and score people who may have disabilities as inherently "riskier," this default assumption of the inherent objectivity of algorithmic tools — and the use of the tools altogether — must seriously be called into question. Notwithstanding these concerns, the AFST's developers have continued to propose additional use cases for these kinds of predictive tools that rely on biased data sources. Similar tools created by largely the same team of researchers that created the AFST have recently been deployed in Douglas County, Colorado [97] and Los Angeles County, California [73].

In contrast to debates about how to make algorithms that function in contexts marked by pervasive and entrenched discrimination fair or accurate, Green [38] and Mohamed et al. [66] propose new frameworks that focus instead on connecting our understanding of algorithmic oppression to the broader social and economic contexts in which algorithms operate to evaluate whether algorithms can actually be designed to promote justice [38, 66]. In the years since the initial development of the AFST, impacted community members and others who interact with the AFST have expressed concerns about racial bias and suggested alternatives to the AFST, including non-technical changes to the County's practices such as improving hiring and training conditions for workers, changes to state laws that affect the child welfare system, and reimagining relationships between community members and the agency [94]. As Sasha Costanza-Chock poses in her 2020 book *Design Justice* [21], "why do we continue to design technologies that reproduce existing systems of power inequality when it is so clear to so many that we urgently need to dismantle those systems?"

ACKNOWLEDGMENTS

We thank the Allegheny County Department of Human Services for their transparency and time in providing information and data related to the AFST and in answering questions related to the AFST. We also thank several of our colleagues at the ACLU and HRDAG and our anonymous reviewers for helpful feedback.

REFERENCES

[1] 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.* 11, 4 (July 2021), 75–110. https://doi.org/10.52214/cjrl.v11i4.8741

[2] J. Khadijah Abdurahman. 2022. Birthing Predictions of Premature Death. https://logicmag.io/home/birthing-predictions-of-premature-death/

[3] Social Security Administration. 2023. Supplemental Security Income (SSI) in Pennsylvania. https://www.ssa.gov/pubs/EN-05-11150.pdf

[4] Michelle Alexander. 2012. *The New Jim Crow.* The New Press, New York, NY.

[5] Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data* 3 (2020), 17 pages. https://doi.org/10.3389/fdata.2020.00005

[6] Patricia Auspos. 2017. *Using Integrated Data Systems to Improve Case Management and Develop Predictive Modeling Tools. Case Study 4.* Technical Report. Annie E. Casey Foundation. https://www.aecf.org/resources/using-integrated-data-systems-to-improve-case-management-and-develop-predic

[7] Cora Bartelink, TA Van Yperen, IJ Ten Berge, Leontien De Kwaadsteniet, and CLM Witteman. 2014. Agreement on child maltreatment decisions: A nonrandomized study on the effects of structured decision-making. In *Child & Youth Care Forum*, Vol. 43. Springer, USA, 639–654.

[8] Cora Bartelink, Tom A Van Yperen, and J Ingrid. 2015. Deciding on child maltreatment: A literature review on methods that improve decision-making. *Child Abuse & Neglect* 49 (2015), 142–153.

[9] Frank R Baumgartner, Derek A Epp, Kelsey Shoub, and Bayard Love. 2017. Targeting young men of color for search and arrest during traffic stops: evidence from North Carolina, 2002–2013. *Politics, Groups, and Identities* 5, 1 (2017), 107–131.

[10] Elinor Benami, Reid Whitaker, Vincent La, Hongjin Lin, Brandon R. Anderson, and Daniel E. Ho. 2021. The Distributive Effects of Risk Prediction in Environmental Compliance: Algorithmic Design, Environmental Justice, and Public Policy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 90–105. https://doi.org/10.1145/3442188.3445873

[11] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[12] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable Machine Learning in Deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 648–657. https://doi.org/10.1145/3351095.3375624

[13] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 850–863. https://doi.org/10.1145/3531146.3533149

[14] Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. 2022. Post-Hoc Explanations Fail to Achieve Their Purpose in Adversarial Contexts. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 891–905. https://doi.org/10.1145/3531146.3533153

[15] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300271

[16] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (nov 2019), 24 pages. https://doi.org/10.1145/3359206

[17] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40, 1 (2014), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[18] 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 162, 22 pages. https://doi.org/10.1145/3491102.3501831

[19] Lingwei Cheng and Alexandra Chouldechova. 2022. Heterogeneity in Algorithm-Assisted Decision-Making: A Case Study in Child Abuse Hotline Screening. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 376 (nov 2022), 33 pages. https://doi.org/10.1145/3555101

[20] 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 *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, USA, 134–148. https://proceedings.mlr.press/v81/chouldechova18a.html

[21] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need.* The MIT Press, Cambridge, MA.

[22] Allegheny County. 2021. The DHS Data Warehouse. https://www.alleghenycounty.us/human-services/news-events/accomplishments/dhs-data-warehouse.aspx

[23] Kathleen Creel and Deborah Hellman. 2021. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 816. https://doi.org/10.1145/3442188.3445942

[24] Ying Cui, Fu Chen, Ali Shiri, and Yaqin Fan. 2019. Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Sciences* 120, 4 (2019), 208–227. https://doi.org/10.1108/ILS-10-2018-0104

[25] Tim Dare and Eileen Gambrill. 2017. Ethical analysis: Predictive risk models at call screening for Allegheny County. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf

[26] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[27] Jessica M Eaglin. 2017. Constructing recidivism risk. *Emory Law Journal* 67 (2017), 59. https://scholarlycommons.law.emory.edu/elj/vol67/iss1/2

[28] David Freeman Engstrom, Daniel E Ho, Catherine M Sharkey, and Mariano-Florentino Cuéllar. 2020. *Government by algorithm: Artificial intelligence in federal administrative agencies.* Technical Report. Administrative Conference of the United States.

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

[30] Mary Flanagan, Daniel C. Howe, and Helen Nissenbaum. 2008. *Embodying Values in Technology: Theory and Practice.* Cambridge University Press, Cambridge, 322–353. https://doi.org/10.1017/CBO9780511498725.017

[31] Batya Friedman. 1996. Value-Sensitive Design. *Interactions* 3, 6 (Dec 1996), 16–23. https://doi.org/10.1145/242485.242493

[32] Philip Gillingham. 2019. Can Predictive Algorithms Assist Decision-Making in Social Work with Children and Families? *Child Abuse Review* 28, 2 (2019), 114–126. https://doi.org/10.1002/car.2547

[33] Stephanie K Glaberson. 2019. Coding over the cracks: predictive analytics and child protection. *Fordham Urb. LJ* 46 (2019), 307.

[34] James P Gleeson. 1987. Implementing structured decision-making procedures at child welfare intake. *Child Welfare: Journal of Policy, Practice, and Program* 66, 2 (1987), 101–112.

[35] Lauryn P Gouldin. 2016. Disentangling flight risk from dangerousness. *BYU L. Rev.* 2016, 3 (2016), 837.

[36] Crystal Grant. 2022. *AI in Healthcare May Worsen Medical Racism.* Technical Report. American Civil Liberties Union. https://www.aclu.org/legal-document/aclu-white-paper-ai-health-care-may-worsen-medical-racism

[37] Ben Green. 2021. Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing* 2, 3 (2021), 17. https://doi.org/10.23919/JSC.2021.0029

[38] Ben Green. 2022. Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology* 35, 4 (2022), 1–32.

[39] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45 (2022), 105681.

[40] Sam Harper, Nicholas B King, Stephen C Meersman, Marsha E Reichman, Nancy Breen, and John Lynch. 2010. Implicit value judgments in the measurement of health inequalities. *The Milbank Quarterly* 88, 1 (2010), 4–29.

[41] Sally Ho and Garance Burke. 2022. An algorithm that screens for child neglect in Allegheny County raises concerns. https://apnews.com/article/child-welfare-algorithm-investigation-9497ee937e0053ad4144a86c68241ef1

[42] Benefits Tech Action Hub. 2022. Colorado Medicaid, SNAP, CHIP, and TANF Wrongful Denials. https://www.btah.org/case-study/colorado-medicaid-snap-chip-and-tanf-wrongful-denials.html

[43] Hornby Zeller Associates Inc. 2018. *Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation.* Technical Report. Allegheny County. https://www.alleghenycountyanalytics.us/wp-

content/uploads/2019/05/Process-Evaluation-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-4.pdf
[44] Abigail Z. Jacobs. 2021. Measurement as governance in and for responsible AI. https://arxiv.org/abs/2109.05658
[45] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901
[46] Will L. Johnson. 2011. The validity and utility of the California Family Risk Assessment under practice conditions in the field: A prospective study. *Child Abuse & Neglect* 35, 1 (2011), 18–28. https://doi.org/10.1016/j.chiabu.2010.08.002
[47] Nathan Kallus and Angela Zhou. 2019. The Fairness of Risk Scores beyond Classification: Bipartite Ranking and the XAUC Metric. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 309, 11 pages.
[48] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.* 55, 5, Article 95 (Dec 2022), 29 pages. https://doi.org/10.1145/3527848
[49] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 52, 18 pages. https://doi.org/10.1145/3491102.3517439
[50] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 454–470. https://doi.org/10.1145/3532106.3533556
[51] Emily Keddell. 2015. The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: Child abuse prevention or neo-liberal tool? *Critical Social Policy* 35, 1 (2015), 69–88.
[52] Emily Keddell. 2019. Algorithmic Justice in Child Protection: Statistical Fairness, Social Justice and the Implications for Practice. *Social Sciences* 8, 10 (2019), 22. https://doi.org/10.3390/socsci8100281
[53] Kenji Kira and Larry A. Rendell. 1992. A Practical Approach to Feature Selection. In *Machine Learning Proceedings 1992*, Derek Sleeman and Peter Edwards (Eds.). Morgan Kaufmann, San Francisco (CA), 249–256. https://doi.org/10.1016/B978-1-55860-247-2.50037-1
[54] Vipin Kumar and Sonajharia Minz. 2014. Feature Selection: A literature Review. *Smart Comput. Rev.* 4 (2014), 211–229.
[55] Himabindu Lakkaraju. 2021. Towards Reliable and Practicable Algorithmic Recourse. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 4. https://doi.org/10.1145/3459637.3482497
[56] Algorithmic Justice League. 2020. The Algorithmic Justice League's 101 Overview. https://www.ajl.org/learn-more
[57] 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), 309–334.
[58] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2022. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2103–2113. https://doi.org/10.1145/3531146.3534628
[59] Jorge M. Lobo, Alberto Jiménez-Valverde, and Raimundo Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17, 2 (2008), 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x
[60] Wayne A Logan and Andrew Guthrie Ferguson. 2016. Policing criminal justice data. *Minn. L. Rev.* 101 (2016), 541.
[61] Noëmi Manders-Huits. 2011. What values in design? The challenge of incorporating moral values into design. *Science and engineering ethics* 17, 2 (2011), 271–287.
[62] Sandra G Mayson. 2017. Dangerous defendants. *Yale Law Journal* 127 (2017), 490.
[63] Mikaela Meyer, Aaron Horowitz, Erica Marshall, and Kristian Lum. 2022. Flipping the Script on Criminal Justice Risk Assessment: An actuarial model for assessing the risk the federal sentencing system poses to defendants. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 366–378.
[64] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596
[65] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.
[66] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33, 4 (2020), 659–684.
[67] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300356
[68] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 323, 19 pages. https://doi.org/10.1145/3491102.3517644
[69] Deirdre K Mulligan and Kenneth A Bamberger. 2018. Saving governance-by-design. *California Law Review* 106, 3 (2018), 697–784.
[70] Deirdre K Mulligan and Kenneth A Bamberger. 2019. Procurement as policy: Administrative process for machine learning. *Berkeley Tech. LJ* 34 (2019), 773.
[71] Debbie Nathan. 2020. The Long, Dark History of Family Separations: How politicians used the drug war and the welfare state to break up Black and Native American families. https://reason.com/2020/10/13/the-long-dark-history-of-family-separations/
[72] Hina Naveed. 2022. *"If I Wasn't Poor, I Wouldn't Be Unfit" The Family Separation Crisis in the US Child Welfare System*. Technical Report. American Civil Liberties Union, Human Rights Watch.
[73] Children's Data Network. 2022. *Los Angeles County Risk Stratification Model: Methodology & Implementation Report*. Technical Report. Children's Data Network. https://dcfs.lacounty.gov/wp-content/uploads/2022/08/Risk-Stratification-Methodology-Report_8.29.22.pdf
[74] Allegheny County Department of Human Services. 2017. *Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County Response by the Allegheny County Department of Human Services*. Technical Report. Alleghany County Analytics.
[75] Allegheny County Department of Human Services. 2019. *Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions: Predictive Risk Package*. Technical Report. Allegheny County Department of Human Services. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf
[76] Mark A Paige and Audrey Amrein-Beardsley. 2020. "Houston, We Have a Lawsuit": A Cautionary Tale for the Implementation of Value-Added Models for High-Stakes Employment Decisions. *Educational Researcher* 49, 5 (2020), 350–359.
[77] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.
[78] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. 2022. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. arXiv:2203.06768 [cs.LG]
[79] Tawana Petty, Mariella Saba, Tamika Lewis, Seeta Peña Gangadharan, and Virginia Eubanks. 2018. Reclaiming our data. https://www.odbproject.org/wp-content/uploads/2016/12/ODB.InterimReport.FINAL_.7.16.2018.pdf
[80] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature human behaviour* 4, 7 (2020), 736–745.
[81] Eleanor Pratt and Heather Hahn. 2021. *What Happens When People Face Unfair Treatment or Judgment When Applying for Public Assistance or Social Services?* Technical Report. Urban Institute.
[82] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. 2020. Algorithmic recourse in the wild: Understanding the impact of data and model shifts.
[83] Rashida Richardson, Jason M Schultz, and Kate Crawford. 2019. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYU Law Review Online* 94 (2019), 15.
[84] Katherine Rittenhouse, Emily Putnam-Hornstein, and Rhema Vaithianathan. 2022. Algorithms, Humans, and Racial Disparities in Child Protective Services: Evidence from the Allegheny Family Screening Tool. https://krittenh.github.io/katherine-rittenhouse.com/Rittenhouse_Algorithms.pdf

[85] Dorothy Roberts. 2009. *Shattered bonds: The color of child welfare.* Basic Books, New York, NY.
[86] Dorothy Roberts. 2022. *Torn Apart: How the Child Welfare System Destroys Black Families–and How Abolition Can Build a Safer World.* Basic Books, New York, NY.
[87] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling Assumptions Clash with the Real World: Transparency, Equity, and Community Challenges for Student Assignment Algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 589, 14 pages. https://doi.org/10.1145/3411764.3445748
[88] David G. Robinson. 2022. The Kidney Transplant Algorithm's Surprising Lessons for Ethical A.I. https://slate.com/technology/2022/08/kidney-allocation-algorithm-ai-ethics.html
[89] David G Robinson. 2022. *Voices in the Code: A Story about People, Their Values, and the Algorithm They Made.* Russell Sage Foundation, New York, NY.
[90] Anjana Samant, Aaron Horowitz, Kath Xu, and Sophie Beiers. 2022. *Family Surveillance by Algorithm: The Rapidly Spreading Tools Few Have Heard Of.* Technical Report. American Civil Liberties Union. https://www.aclu.org/fact-sheet/family-surveillance-algorithm
[91] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376229
[92] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598
[93] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* 39, 5 (2011), 1–13. https://www.jstatsoft.org/v39/i05/
[94] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining New Futures beyond Predictive Systems in Child Welfare: A Qualitative Study with Impacted Stakeholders. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1162–1177. https://doi.org/10.1145/3531146.3533177
[95] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (–, NY, USA) *(EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, Article 17, 9 pages. https://doi.org/10.1145/3465416.3483305
[96] upEND Movement. 2021. Family Policing System Definition. https://upendmovement.org/family-policing-definition/
[97] Rhema Vaithianathan, Haley Dinh, Allon Kalisher, Chamari Kithulgoda, Emily Kulick, Megh Mayur, Athena Ning, Diana Benavides Prado, and Emily Putnam-Hornstein. 2019. *Implementing a Child Welfare Decision Aide in Douglas County: Methodology Report.* Technical Report. Centre for Social Data Analytics. https://csda.aut.ac.nz/__data/assets/pdf_file/0009/347715/Douglas-County-Methodology_Final_3_02_2020.pdf
[98] Rhema Vaithianathan, Emily Kulick, Emily Putnam-Hornstein, and D Benavides-Prado. 2019. *Allegheny family screening tool: Methodology, version 2.* Technical Report. Center for Social Data Analytics. 1–22 pages.
[99] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. 2017. *Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation.* Technical Report. Center for Social Data Analytics.
[100] Suresh Venkatasubramanian and Mark Alfano. 2020. The Philosophical Basis of Algorithmic Recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 284–293. https://doi.org/10.1145/3351095.3372876
[101] Emma Williams. 2020. 'Family Regulation,' Not 'Child Welfare': Abolition Starts with Changing our Language. https://imprintnews.org/opinion/family-regulation-not-child-welfare-abolition-starts-changing-language/.
[102] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468

A EXTENDED ANALYSIS

In this section, we provide more detail on our methodology and results for several parts of our analysis.

A.1 Section 4.1: Risky by association

A.1.1 Household Size by Race. In Section 4.1, we discuss the finding that Black families that are represented in the testing data have, on average, larger household sizes than non-Black families in the data. In the testing data, the average household size for Black families is 2.61 and 2.01 for non-Black families. However, household size does not explain the disparities in risk score distributions discussed in Section 4.1. Black households are, on average, assigned higher risk scores than non-Black households of the same size (see Figure 4).

A.1.2 The AFST's Risk Score and Outcome Policies. In practice, the AFST relates to both risk scores (derived from predictions from the model) and outcomes (whether a child was removed from the home within two years of a referral). To look at the intersection of score aggregation policies and outcome measurements, we define and analyze six different policies, representing different ways that risk scores could have been communicated and different ways that outcomes could have been measured. Table 3 summarizes these policies and how scores and outcomes would be measured under each policy for the hypothetical family discussed in Section 4.1.

In Table 1 of Section 4.1, we present a range of possible performance results for the AFST using several metrics — including the AUC, the Cross-AUC (see [47]), and the False Positive Rate — evaluated using the test data. In Table 4 below, we present AUC and Cross-AUC results for several of the policies. We exclude Policies 4 and 5, which assign households to one of three possible "protocols" rather than producing a score between 1 and 20. For all AUC and ROC computations, we use the predicted probabilities, not the risk scores, following the convention used in [98]. Corresponding ROC curves are included below (see Figure 5).

Table 5 presents false positive rates for each policy, breaking down the results in Table 1 of Section 4.1. For the policies that use ventile scores between 1 and 20, we consider a "positive" prediction to be one with a score of 16 or higher (following the definition used by Chouldechova et al. (2018) [20]), so we consider a false positive to be an individual or household with an AFST score at or above 16 where a removal does not occur within two years. For the polices that use protocols, a positive prediction is one classified as in the "high-risk" protocol, so a false positive is a household marked as "high-risk" where a removal does not occur within two years.

We see that false positive rates vary widely across policies, and the false positive rate is higher for Black children/households than for non-Black children/households across all of the policies. False positive rates and racial disparities in the rates are lowest under Policies 1 and 5, and highest under Policies 2 and 6. Policies 2 and 6 both assign one score for all children in the household and measure outcomes at the individual level, which closely tracks with the County's policy in practice. For these policies, the false positive rate for Black households is near 50% – meaning that across all children in households with at least one Black child who were not removed by the County within two years, half of the associated households would have been marked as "high-risk."

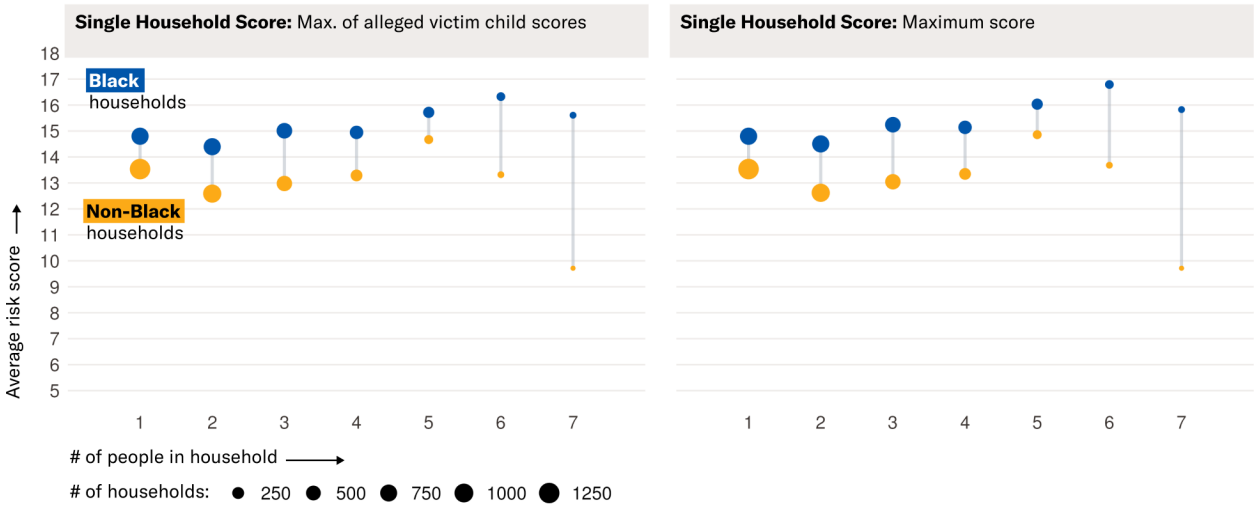

Figure 4: Risk Score by Household Size and Race. On average, Black households are larger than non-Black households in the data, and Black households receive higher risk scores than non-Black households of the same size under both score policies that assign a single score to the entire household.

Policy	How are risk scores communicated?	How are outcomes measured?	Meaning for example family described in Section 4.1
Policy 1	Individual scores for each person in the household	Individual outcomes measured for each person	Each child in the household gets their individually-generated risk score, and outcomes are measured for each child
Policy 2	Each household is assigned one score – the maximum score across the household	Individual outcomes measured for each person	Each child in the household gets the score of 18, the maximum score across all children, and outcomes are measured for each child.
Policy 3	Each household is assigned one score – the maximum score across the household	Outcomes are measured at the household level (e.g., was any child on a referral removed from the home within 2 years of that referral?)	The household gets the score of 18, the maximum score across all children, and outcomes are measured at the household level.
Policy 4	Each household is assigned to a "protocol" using the procedure in place in Allegheny County	Individual outcomes measured for each person	Each child in the household is assigned to the "high-risk protocol," because there is a score of 18 or higher in the household and there is at least one child younger than 16 in the household, and outcomes are measured for each child.
Policy 5	Each household is assigned to a "protocol" using the procedure in place in Allegheny County	Outcomes are measured at the household level (e.g., was any child on a referral removed from the home within 2 years of that referral?)	The household is assigned to the "high risk protocol," because there is a score of 18 or higher in the household and there is at least one child younger than 16 in the household, and outcomes are measured at the household level.
Policy 6	Each child in the household gets the score assigned to the "victim child" (taking the maximum score if there are multiple)	Individual outcomes measured for each person	Each child in the household gets the score of 15, the score of the "victim child," and outcomes are measured at the individual level.

Table 3: Possible AFST Risk Score and Outcome Policies Analyzed. Each row represents a possible combination of risk score aggregation (e.g., assigning individual scores vs. household scores) and outcome measurement (e.g., measuring individual outcomes vs. household outcomes) that could have been used for presenting AFST scores to call screeners and measuring the tool's performance.

Policy	"Traditional" AUC (Overall)	"Traditional" AUC for Black people	"Traditional" AUC for non-Black people	xAUC for Black people	xAUC for non-Black people
Policy 1	.739	.742	.731	.703	.767
Policy 2	.679	.668	.673	.566	.766
Policy 3	.733	.739	.699	.621	.800
Policy 6	.679	.670	.672	.580	.754

Table 4: Area Under the Curve Results on Test Data Using Different Score and Outcome Policies.

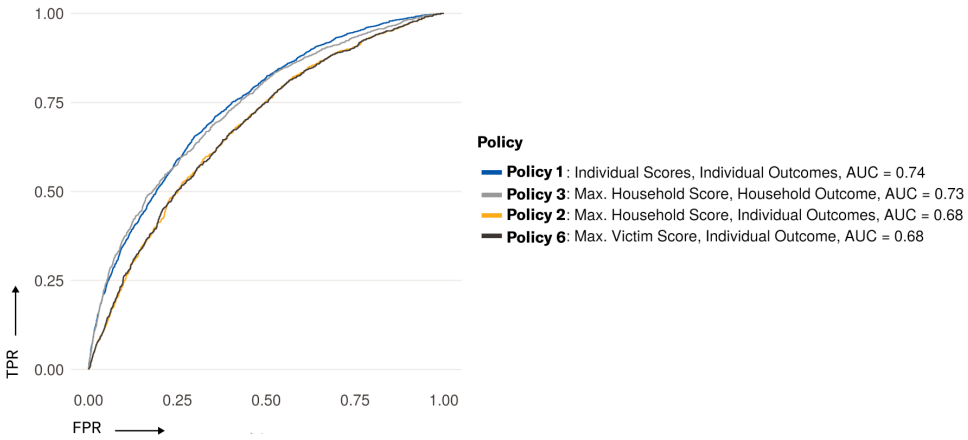

Figure 5: ROC Curve Using Different Score and Outcome Policies.

Policy	Overall FPR	FPR for Black people	FPR for non-Black people
Policy 1: Individual Scores, Individual Outcomes	.21	.22	.19
Policy 2: Max. Household Score, Individual Outcomes	.44	.51	.37
Policy 3: Max. Household Score, Household Outcome	.37	.43	.33
Policy 4: Household Protocol, Individual Outcomes	.26	.32	.21
Policy 5: Household Protocol, Household Outcome	.20	.23	.17
Policy 6: Max. Victim Score, Individual Outcome	.42	.49	.36

Table 5: False Positive Rates on Test Data Using Different Score and Outcome Policies, Overall and Grouped by Race.

A.2 Section 4.2: The more data the better

A.2.1 Disability-related features in the AFST. In Section 4.2, we discuss three weighted features in the AFST V2.1 that come from the behavioral health (BH) data source used to develop the AFST and that potentially relate to disability status. We highlight that the presence of these three behavioral-health related features can increase an AFST score up to four points. Here, we outline a demonstrative example of this finding. Relevant here, in the development report for AFST V2, the developers wrote that "the variable value is zero if the underlying data required to calculate the variable is missing." [98, p. 16].

We can decompose the model for the AFST V2.1 – LASSO logistic regression – using the logistic function. Namely, given $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ representing the p features in the model, a predicted probability $p(X)$ can be decomposed into:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Let

$$M = \sum_{i=1}^{p-3} \beta_i X_i$$

for all of the X_i that are *not* behavioral health features in the AFST V2.1, and let:

- X_v represent the indicator feature for whether the "victim child" has any behavioral health history in the database (note that this variable equals 1 if there is no behavioral health history for this person and 0 otherwise; see [98, p. 20])

- X_a represent the indicator feature for whether the alleged "perpetrator" has any behavioral health history in the database (note that this variable equals 1 if there is no behavioral health history for this person and 0 otherwise; see [98, p. 20])
- X_d represent the feature for the number of days since a parent on the referral was last seen in behavioral health services

Then we have:

$$p(X) = \frac{e^{\beta_0+\beta_v X_v+\beta_a X_a+\beta_d X_d+M}}{1+e^{\beta_0+\beta_v X_v+\beta_a X_a+\beta_d X_d+M}}$$

Consider a referral where the alleged perpetrator and the alleged victim both have a behavioral health history in County records (so $X_v = 0, X_a = 0$) and one of the parents has a record in the County's systems of using the behavioral health services (so $X_d > 0$). For this demonstrative example, let the value of X_d be the mean value of X_d for people with $X_d > 0$ in our data (i.e., the average number of days since a parent was last seen in behavioral health services for parents who have a non-zero record of using these services). Plugging in the intercept and the weights associated with these features, we have:

$$p(X) = \frac{e^{-3.157062+M}}{1+e^{-3.157062+M}} \quad (1)$$

Now consider a hypothetical referral where neither the alleged perpetrator nor the alleged victim have a behavioral health history in County records, and where neither parent has a record in the County's system of using behavioral health services (so $X_v = 1, X_a = 1, X_d = 0$). Then we have that:

$$p(X) = \frac{e^{-3.397846+M}}{1+e^{-3.397846+M}} \quad (2)$$

In Figure 6, we compare these two equations as functions of M. In Figure 7, we layer on the binning function that converts predicted probabilities into risk scores, to see what the effect of having these behavioral health features is on one's AFST score.

As highlighted by the dashed lines in Figure 7, having the values of the features described in the former example (where people on the referral have records of behavioral health history and service usage) can lead to an individual AFST score being up to four points higher than the AFST score for an otherwise-identical individual. In Section 4.2, we note that there were examples from the data we were provided of real families with this kind of circumstance – where they had identical feature values for all features not from the behavioral health sources, some variation in one of the behavioral health features, and different AFST scores.

A.2.2 Risk Ratio Calculation. In Section 4.2, we mention how one of the concerns with including variables from the criminal legal system in the model used for the AFST is that Black people are disproportionately represented in these variables. In this section,

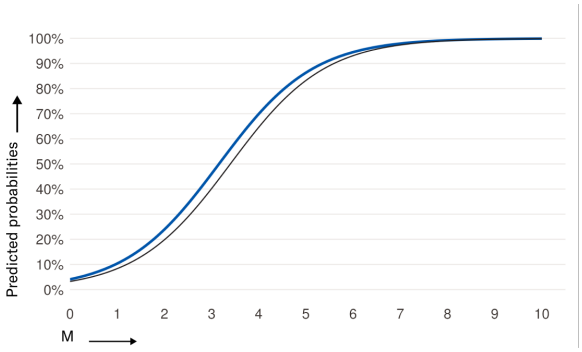

Figure 6: Equation (1) in blue and equation (2) in black as a function of M, showing that for hypothetical individuals with the same feature values except for the behavioral health variables, the person with values for the behavioral health features indicating behavioral health history and service usage could receive a higher predicted probability under the AFST V2.1.

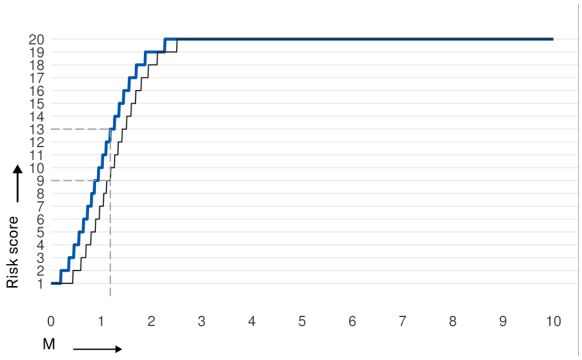

Figure 7: Figure 6 with the binning function used to convert probabilities to risk scores layered onto each equation. The dotted black line demonstrates one value of M for which two hypothetical individuals with the exact same feature values except for the behavioral health features would have a score differential of four.

we explain our methodology for determining which, if any, of the variables provided to us in the data given by the County were racially disparate.

To determine which variables were racially disparate, we first calculated an incidence ratio for Black and white people associated with each feature in the dataset. In this context, an incidence ratio takes the number of individuals of a given race with the specified condition and divides that number by the total number of people of that given race. For instance, the incidence ratio for the prevalence of Black individuals with a parent/guardian who was in juvenile probation at the time of the referral would take the number of Black individuals with parents/guardians in juvenile probation at the time of the referral and divide this by the total number of Black individuals in the dataset. We then divided the incidence ratio for Black individuals by the incidence ratio for that particular variable

for white individuals to get the risk ratio. If the risk ratio was greater than one, that meant that Black individuals were more likely to be represented in the given variable than their white counterparts. The formula we used to generate the risk ratios is as follows:

$$\frac{\frac{\text{Number of Black individuals with specified condition}}{\text{Total Number of Black individuals}}}{\frac{\text{Number of white individuals with specified condition}}{\text{Total Number of white individuals}}}$$

As an example, the risk ratio for the feature associated with the variable indicating whether the alleged victim was in juvenile probation at the time of the referral would be calculated using the following formula:

$$\frac{\frac{\text{Number of Black individuals with JPO\_NOW\_VICT\_SELF == 1}}{\text{Total Number of Black individuals}}}{\frac{\text{Number of white individuals with JPO\_NOW\_VICT\_SELF == 1}}{\text{Total Number of of white individuals}}}$$

In general, the variables that had the highest risk rations were the juvenile probation (JPO) variables.

A.2.3 Interpretability and Variable Importance. In our analysis of the dataset, one of the things that we wanted to understand was how to examine the features of the model – and coefficients associated with those features – in relation to one another. The scales of features in the dataset span from features on a binary scale to features on a scale of tens or hundreds, potentially hindering interpretation of the model coefficients. The county did not standardize the variables before training the model, and though this decision may have hindered interpretability of the coefficients, this decision had no impact on the features selected by the model or the model's predictions, because the R implementation used by the County standardized features by default.

For our exploration and feature analysis, we re-trained the model with normalized features to get a better understanding of how to interpret the outputs of the model and the weights generated by the model. This normalization process resulted in a change in the interpretation of the coefficients for many of the variables. After standardization, the coefficients of the normalized model could be interpreted as follows: a one standard deviation increase in X variable will increase or decrease the estimate of the odds that a child would be placed in foster care in the next 730 days by X amount. This normalization aided our interpretability of the model as we performed our exploratory data analysis of the tool. Because of the normalization, the values of several of the coefficients changed, including those associated with the Juvenile Probation (JPO) variables and some of the variables related to individual characteristics of the alleged perpetrator and victim.

A.3 Section 4.3: Marked in Perpetuity

In Section 4.3, we discussed how we removed "ever-in" variables that came from the jail, juvenile probation, and HealthChoices data sources from the training data before fitting the "no ever-in" model. Not all of the variables from these sources had non-zero weights in the baseline model, but we removed all of them from the training data prior to re-fitting.

We look at the impacts of including or excluding the "ever-in" variables in terms of hypothetical screen-in disparities, using the testing data from the model development process to measure screen-ins that would have been recommended by the tool. But these disparities are sensitive not only to the inclusion of the variables when training the model, but also to the household aggregation policy applied on top of the raw scores.

We first look at using the maximum of alleged victim scores with a threshold of 16 or higher to define "high-risk." The model without "ever-in" variables would have labelled referrals as "high-risk" more frequently, which primarily affects non-Black referral-households. This result has the overall effect of slightly reducing racial disparities in the percentage of referral-households that would have been labelled as high risk (see Table 6). We note that the decision to bin scores by ventile before any policy is applied constrains the overall variation in high score rates for any model, effectively guaranteeing that a certain proportion of referrals will be labeled high-risk.

Model	Pct. Black	Pct. Non-Black	Disparity
Baseline	0.511	0.363	0.148
No ever-in	0.516	0.372	0.144

Table 6: Percent of referral-households in testing data with maximum of the alleged victims' scores of 16 or higher for models with and without "ever-in" variables.

Applying the county's high-risk protocol — which, among other things, takes the maximum of all scores on the referral, rather than just considering the alleged victim scores — to the same data increases the difference in observed disparities between the models, again driven by the model without "ever-in" variables placing more of the screen-in burden on non-Black referral-households (see Table 7).

Model	Pct. Black	Pct. Non-Black	Disparity
Baseline	0.330	0.201	0.129
No ever-in	0.330	0.210	0.121

Table 7: Percent of referral-households in testing data labelled as "high-risk" under County risk protocols for models with and without "ever-in" variables.

These disparities are sensitive to household size; Figure 8 contains a full breakdown.

A.4 Model Replication Process

In order to measure and evaluate the impacts of specific data processing and modeling decisions in the development of the AFST, we considered an alternative to the decision that was made and fit a secondary model, then compared predictions from the two models on the test partition of the training data provided to us. In order to isolate the effects of the decision in question, we sought to replicate the other steps of model fitting as closely as possible.

To test how closely we had replicated the other model steps, we first used our model fitting scripts without altering any input decisions to see if we ended up with the same model as the county.

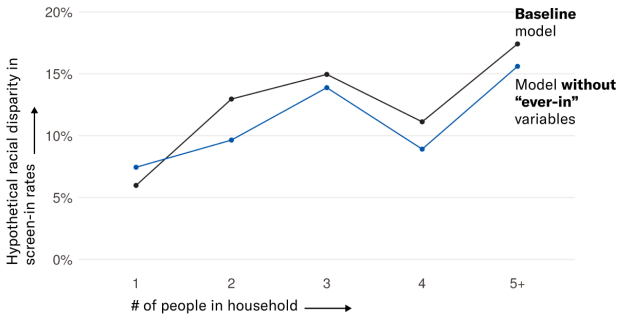

Figure 8: Hypothetical racial disparities in screen-in rates by household size, with and without "ever-in" variables in the model. Higher numbers indicate greater disparities in hypothetical screen-in rates, subtracting the hypothetical screen-in rate for non-Black referral-households from the hypothetical screen-in rate for Black referral-households.

From the county, we received training data which was partitioned into "train", "test", and "screen out" partitions. The train partition was further split into the 10 folds that the implementers used for cross-validation. We also received:

- A *codebook* with variable names and associated model coefficients for two versions of the ASFT LASSO model, labeled "lasso v1" and "lasso v2" (the county frequently refers to the "lasso v1" model as "AFST V2". To reduce confusion, in this paper we've referred to the lasso v1 and lasso v2, respectively, as "AFST V2" and "AFST V2.1"). The codebook describes more than 700 different variables, all of which matched column names in the referral data we received. The referral data includes some additional variables that are not used in the model, including variables related to the race of the different members of the household.
- A *coefficient table* containing model coefficient values for each variable in the model. These coefficients matched those reported as "lasso v2" in the codebook.
- A *predictions table*, which, for each row in the training data, contains a raw score as well as the "risk category" calculated by first binning the raw scores into 20 equal-sized bins.
- The scripts used to generate the coefficient table and the predictions table

We first confirmed that the coefficients in the coefficient table matched those reported in the codebook. We confirming our replication model had non-zero coefficients for exactly the same number of coefficients as in the codebook for the "lasso v2" model and that the coefficient values we calculated were within 10^{-15} of the coefficient values in the codebook. We then repeated the same comparisons with the values in the coefficient table, with similar success.

The R scripts we received use 10-fold cross validation to compare a range of candidate values for the λ regularization parameter, and select the largest value of λ for which the cross-validated estimate of AUC is within one standard error of the overall maximum [93]. We wrote code that would perform the same search, in a way that made it easier to provide alternate model specifications. We first applied our modeling code to the same set of features that were included in V2.1 of the model. We output the coefficient values of the resulting model, as well as raw and binned predictions. We compared the calculated coefficients to both the codebook and the coefficient table and confirmed that we got the same coefficients (to account for numerical precision issues, here we allowed for differences 10 to the negative 15th). Similarly, we confirmed that the replicated model produced the same raw predictions as ASFT V2.1 (once again, within one quadrillionth). Finally, we confirmed that both models output the same binned "risk level" for each record – here we had to be careful to exactly replicate the original method of rounding the scores when creating the risk bins, but not rounding scores when assigning them to bins – in order to exactly match the data we received from the county. We refer to this entire procedure, from the parameter search through defining and assigning bins, as "model fitting."

Note: In order to exactly replicate the model coefficients we had received from the county, we hard-coded the sequence of λ values used as candidates during the cross-validation search, rather than relying on the software defaults to generate a suitable sequence. However, we re-fit the alternate models using the software defaults to generate the candidate λ sequence, and the change did not alter any of our reported results.

Having convinced ourselves that we had successfully replicated the model fitting procedure described in the materials we received from the county, we proceeded to apply the procedure with different model specifications. In addition to the LASSO model coefficients, the codebook we received from the county includes an originating data source for each variable in the training data, describing the following data sources: Allegheny County Jail; Behavioral Health; Birth Record; Census; Child Welfare; Data Warehouse; HealthChoices; Juvenile Probation; Public Benefits. The different model specifications were:

- No JPO: For the "No JPO" version of the model, we removed any predictors from the "Juvenile Probation" dataset.
- No "ever-in:" We relied on regular expression searches of both the variable names and the variable descriptions to identify "ever-in" variables, which included eligibility status for the HealthChoices program as well as indicator variables for having been in the Allegheny County Jail or the Juvenile Probation system.
- No Behavioral Health variables: For the model without behavioral health variables, we removed predictors with "Behavioral Health" as the data source.

B DATA USE AGREEMENT

In connection with the data used in this analysis, we signed a Data Sharing Agreement with the Allegheny County, Pennsylvania Department of Human Services. The Data Sharing Agreement governed procedures for using and handling the data and included a requirement that we give Allegheny County an opportunity to review our findings before publication. A redacted version of this Data Sharing Agreement can be made available upon request.

C MODEL CARD FOR THE AFST

This model card is intended to be a non-exhaustive summary of some key aspects of the AFST. It was created using publicly available information by individuals at the American Civil Liberties Union (ACLU) and Human Rights Data Analysis Group (HRDAG) in connection with an algorithmic audit of the AFST. This model card has not been validated by Allegheny County. Last updated May 10, 2023.

Model Summary: The Allegheny County Screening Tool (AFST) is an algorithmic decision-making tool designed by researchers working with the Allegheny County Department of Human Services (DHS). The tool is used to inform responses to to calls to the County's child welfare agency about alleged child neglect.

Model Details. *Basic information about the model.*

- **Organization developing the model:** A team of researchers (see full list here), and Allegheny County DHS.
- **Model date(s) and version(s):** V1 Deployment: August 2016; V2 Deployment: December 2018; V3 Deployment: 2022. We do not discuss V3 in-depth, since there is no public report about V3 available as of writing.
- **Model type:** V1: Regression model, V2: LASSO Logistic Regression
- **Paper or other resource from the tool creators for more information:**
  - V1 Development Report (Published April 2017)
  - V2 Development Report (Published April 2019)
  - AFST Packet Created by Allegheny County

Training Data. *Details on the data used to develop the model.*

- **V1 Data description:** The model was developed using data from DHS's integrated data warehouse, consisting of data from various agencies in the county. See DHS Data Warehouse Report for more info.
- **V1 Outcome definition:** Tool developers wanted to design the model to "determine which reports of maltreatment involve children who are at greatest risk of: (1) future abuse and neglect, (2) future involvement with child protective services, and/or (3) future critical incidents (i.e., near-fatalities and fatalities)" (V1 Development Report, pg. 8). They ultimately chose to build a model using two outcomes: 1) Re-referral (when a call to the hotline about a child was initially screened out but another call about the same child is received within two years), and 2) Placement (when a call to the hotline about a child is screened in and ultimately results in the child being placed in foster care within two years).
- **V2 Data description:** V2 used similar data as V1. Some of the data had changed between the development of V1 and V2, so some features that were used in V1 were not used in V2 and vice versa (See V2 Report, pg. 3).
- **V2 Outcome definition:** In V2, the developers got rid of the re-referral model from V1 because it "did not have strong face validity" (See V2 Report, pg. 3) In V2, the developers built only one model, to estimate the likelihood that a child would be removed from their home within two years following a referral.

Training Process. *Details on the model development process.*

- **V1:** The developers conducted feature selection using the full training data set and then built the model using a 70/30 training/testing split. They explored non-parametric methods, including decision-trees, naïve bayes, random forest, adaptive boosting, and others, and ultimately selected a regression model.
- **V2:** While not explictly stated, the V2 Report implies that feature selection was performed through model building. The developers state they considered XG-BOOST, Random Forest, and SVM and ultimately chose LASSO logistic regression.

Validation Metrics. *Metrics used to measure model performance.*

- **V1 Performance measure(s) and results:** The developers used Area Under the Curve (AUC) overall and broken down by race. As discussed in Chouldechova et al. (2018), they made mistakes in these calculations that led to an improperly inflated AUC. They also evaluated the tool using external hospital data; see V1 Report for more info and specific results.
- **V2 Performance measure(s) and results:** AUC, calibration (referred to as "Rates of Placement Outcomes"), Positive Predictive Value (PPV) and True Positive Rate (TPR) and several other metrics were used. See V2 Report for more info and specific results.

Post-Deployment Evaluations. *Audits and evaluations.* There have been several external analyses of the AFST, including but not limited to:

- **Eubanks (2018),** examining issues with the tool's input data and deployment as part of the book *Automating Inequality*
- **Cheng et al. (2022),** finding that call screening workers' interventions reduce racial disparities in screen-in rates compared to disparities that would result from strict adherence to algorithmic recommendations
- **DeArteaga et al. (2020)**, studying a technical glitch in the deployment of the AFST that led to improperly calculated risk scores
- **Kawakami et al. (2022)** and **Kawakami et al. (2022)**, exploring call screening workers' interpretations of and interactions with the AFST
- **Wang et al. (2022)**, discussing several issues with the AFST — including related to target-construct mismatch, distribution shift, lack of contestability, and disparate performance — in a broader analysis of predictive tools
- **Gerchick et al. (2023)**, evaluating how design decisions in the development of the AFST function as policy decisions and have policy impacts (Note: this model card was created as part of this analysis)