Failing Our Youngest: On the Biases, Pitfalls, and Risks in a Decision Support Algorithm Used for Child Protection

Therese Moreau Hansen
Networks, Data, and Society (NERDS) group, IT University of Copenhagen
Copenhagen, Denmark

Roberta Sinatra*
Center for Social Data Science (SODAS), University of Copenhagen
Copenhagen, Denmark
Networks, Data, and Society (NERDS) group, IT University of Copenhagen
Copenhagen, Denmark
Pioneer centre for AI (P1)
Copenhagen, Denmark
ISI Foundation
Turin, Italy

Vedran Sekara*
Networks, Data, and Society (NERDS) group, IT University of Copenhagen
Copenhagen, Denmark
Pioneer centre for AI (P1)
Copenhagen, Denmark

ABSTRACT

In recent years, Danish child protective services have experienced increasing pressure, prompting the adoption of a decision-support algorithm to aid caseworkers in identifying children at heightened risk of maltreatment, named Decision Support. Despite its critical role, this algorithm has not undergone formal evaluation. Through a freedom of information request, we were able to partially access the algorithm and conduct an audit. We find that the algorithm has significant methodological flaws, suffers from information leakage, relies on inappropriate proxy values for maltreatment assessment, generates inconsistent risk scores, and exhibits age-based discrimination. Given these serious issues, we strongly advise against the use of this kind of algorithms in local government, municipal, and child protection settings, and we call for rigorous evaluation of such tools before implementation and for continual monitoring post-deployment by listing a series of specific recommendations.

CCS CONCEPTS

• **Applied computing**; • **Social and professional topics** → **Technology audits**;

KEYWORDS

Algorithmic audit, technological fairness, algorithmic decision making, algorithmic discrimination

ACM Reference Format:
Therese Moreau Hansen, Roberta Sinatra, and Vedran Sekara. 2024. Failing Our Youngest: On the Biases, Pitfalls, and Risks in a Decision Support Algorithm Used for Child Protection. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3658906

*These authors contributed equally to this research.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658906

1 INTRODUCTION

Algorithmic decision-making systems are increasingly being adopted by governments and public service agencies to make life-changing decisions. However, scientists [14], activists [21], policy experts [17], and civil society [47] have all voiced concern that such systems are being deployed without adequate consideration of potential harms, biases, disparate impacts, and accountability. Globally, controversial applications of machine learning systems include US counties estimating the probability for children to suffer from maltreatment [12, 16], Dutch tax authorities using risk classification models to flag fraud in childcare benefits [4], a World Bank funded automated social assistance system in Serbia [5], and an automatic system to recover presumed overpayments of social welfare in Australia [31], to name some. The limited evaluation of the issues of these algorithms can have far-reaching consequences not only for citizens but also for governments. For example, the mentioned Dutch and Australian algorithmic applications have since turned into scandals: the Dutch Government was forced to resign in 2021 after tax authorities unjustly accused more than 26,000 families from predominantly low socio-economic backgrounds of committing fraud, while the Australian federal government had to agree to a court settlement after it unlawfully raised 1.8 billion Australian dollars in debt.

Child maltreatment is a serious issue with severe and long-lasting consequences [29]. In Denmark, a study done in 2009-2010 interviewed a randomly selected sample of individuals born in 1984 and reported that, during childhood, 3.0% had experienced physical neglect, 5.2% emotional abuse, 5.4% physical abuse, and 3.4% sexual abuse [13]. In response, Danish social services have adopted a policy of being proactive rather than reactive regarding cases of child maltreatment. The primary method for child and family welfare services to identify abuse is by receiving notifications, which are legally mandated from a variety of sources, including public institutions and non-governmental organizations, whenever there are concerns about a child's well-being. In the period 2015-2019, the number of notifications regarding child or adolescent distress, neglect, or maltreatment received by Danish municipalities has

risen by 41.8% [39], due to legislative changes adopted in late 2013 in the area of child protection. In 2019 alone, the municipalities received a total of 138,088 notifications of concern regarding 79,024 children or adolescents [39, 40], which corresponds to roughly 6% of the total child and adolescent population in the country.

Since the introduction of the 2013 legal mandates, social workers are required to perform initial assessments within 24 hours of receiving a notification. This regulation, aimed at promptly aiding children in immediate danger, has significantly intensified the workload on Child Protective Services, surpassing the limits of their available human resources. It is against this backdrop that algorithms have been suggested as a possible solution for performing the required initial assessment within the first 24 hours. The assumption is that a predictive risk model can not only assist social workers in assessing the growing number of notifications in a timely manner, but it will also provide consistent risk assessment for the children being referred to Child Protective Services [2].

Inspired by the *Allegheny Family Screening Tool* from the US [12] and following the tendency of using algorithms within the Child Welfare System [38], a Danish research team developed *Decision Support* (DSS) [2]. Designed as a support tool for caseworkers assessing the risk of child abuse, DSS was developed and pilot-tested in collaboration with the municipalities of Silkeborg and Hjørring [34] from November 2018 to February 2019 on 208 cases. The model aimed to leverage the extensive personal data collected by Danish authorities about its citizens to build a predictive risk model. Using this information, the idea behind DSS is to build a predictive model that can provide qualified and consistent risk assessments for all children being referred, whether they had previously been in contact with child protective services or not [34].

Through a freedom of information request, we got partial information about the DSS algorithm and other relevant documentation, enabling us to conduct an audit of the machine learning model. Due to the sensitive nature of the data, direct access to the training data was not granted; instead, we received information about the model's structure, its coefficients, and some aspects of the training methods. Notably absent in the documentation, however, were a detailed explanation of all the pre-processing and training steps, and any information on model evaluation, including metrics such as t-statistics, p-values, R-squared values, or other performance metrics. Consequently, our work aims to audit the algorithm by focusing on the following: 1) Assessing methodological correctness within the model, and 2) comprehending its limitations through the use of counterfactual simulations [28]. This approach allows us to critically assess the model's functionality in the absence of direct data access.

2 THE DECISION SUPPORT MODEL (DSS)

In this section, we describe the DSS model, as reported in the documentation we accessed through the Freedom of Information request. This documentation is lacking some information though: for example, the pre-processing and training steps of the algorithm are not documented; however, these have been outlined by the Principal Investigator (PI) of the DSS project in an email correspondence [32]. While auditing the algorithm, we realized that more information was missing in the documentation and we reached out again to the PI. Yet, most of our requests, for example to inspect the code, were politely declined. In the rest of the paper, whenever we have to speculate about some aspects of the model or when something is the result of our considerations, we indicate so.

2.1 Proxies for maltreatment

DSS is a predictive model which estimates the risk of a child experiencing maltreatment. However, maltreatment is difficult to quantify, as there is no agreement on a single variable indicating whether a child is at risk. For this reason, the research behind DSS uses three proxies, defining maltreatment if one of the following three outcomes occurred within six months (180 days) of a notification being received by Child Protective Services:

(1) The child is placed in foster care or similar forms of out-of-home placements.
(2) A severe notification is received by the authorities. A notification is classified as severe if: a) physical or sexual abuse has been committed against the child, b) the child has committed a crime, or c) a parent has substance abuse.
(3) A preventive measure (defined by §52 in the Danish Social Services Act [15]) has been initiated. §52 measures range from families getting practical or financial help, to educational and pedagogical support, stays in daycare, children being offered spots in youth clubs, to children being put in foster care.

We observe one peculiarity of these definitions: proxy (1) is, in principle, included in proxy (3), and it is unclear from the documentation why these two proxies are both used, whether the overlap is intentional, or whether there could be different nuances in the definition of (1) and (3). Also, we do not know if all or only a subset of the possible preventive measures of (3) were used as a proxy for maltreatment. Arguably, for example, being offered a spot in youth clubs should not be considered a proxy for maltreatment.

To construct the labeled dataset, data from approximately 120, 000 notifications of concern submitted to social services between 2014 and 2015 was used (see Fig. 1). Note that, in principle, multiple notifications can involve the same child. The data comes from Statistics Denmark [1], which is the country's central authority on statistics and registry data. If any of the three above proxy-outcomes occurred within six months of notification, the child was labeled as being maltreated (i.e. 1 in the data), otherwise the child was labeled as well-treated (i.e. 0 in the data). Note that the model only uses information from children and families for which Child Protective Services have received at least one notification.

2.2 Pre-processing and algorithm training

Based on the email correspondence with the PI of the DSS project [32], these were the steps for the pre-processing and training of the algorithm, undertaken in the following chronological order: 1) delimitation of the data set, 2) standardization of variables, 3) randomized train-test (70/30%) split of the data, 4) estimation of model parameters using the training sample, 5) prediction and validation using the test sample. For the case of post-Lasso, feature selection was carried out on half of the training set while OLS was used to estimate the model coefficients on the other half.

2.3 Feature selection and model training

As model features, the developers of DSS initially included four kinds of information, namely about the notification (who reported it, the type of report, when it was reported), the child itself (age, gender, past history, place of residence), the parents (age, gender, origin, marital status), and information about siblings or other children living in the household (whether authorities have received notifications for them, how many, which types). Approximately 300 features were initially selected based on the notion that they are easily accessible and understandable to caseworkers.

To predict the risk of maltreatment the researchers tested three types of machine learning models: decision tree, random forest, and post-Lasso [6]. They found the models to have similar overall performance scores and concluded that the post-Lasso model was best because it was the most transparent. A post-Lasso model uses an ordinary Lasso model for feature selection followed by an ordinary least square (OLS) regression using the selected features [6]. The resulting linear model for DSS contains 9 features (see Table 1 for a description) and one intercept value, and predicts a risk score (rs) as:

$$\begin{aligned} rs = \; & 0.1175 \\ & + \; 0.0541 \cdot \frac{x_{\text{child age}} - 9.6308}{4.8251} \\ & + \; 0.0122 \cdot \frac{x_{\text{notifications past 90 days}} - 0.4919}{1.0474} \\ & + \; 0.0472 \cdot \frac{x_{\text{notifications past 180 days}} - 0.7724}{1.4455} \\ & + \; 0.0068 \cdot \frac{x_{\text{type 2 notifications}} - 0.0130}{0.1131} \\ & + \; 0.0212 \cdot \frac{x_{\text{type 7 notifications}} - 0.0208}{0.1428} \\ & + \; 0.0245 \cdot \frac{x_{\text{type 9 notifications}} - 0.0616}{0.2403} \\ & + \; 0.0116 \cdot \frac{x_{\text{interventions past 180 days}} - 0.0306}{0.1790} \\ & - \; 0.0002 \cdot \frac{x_{\text{placements past year}} - 0.0251}{0.1777} \\ & + \; 0.0094 \cdot \frac{x_{\text{placements past 5 years}} - 0.0680}{0.3364} \end{aligned} \tag{1}$$

Features are standardized by subtracting the mean of the population and dividing by the standard deviation (z-score normalization). For instance, the average age of children in the data is 9.6308 and the standard deviation is 4.8251. The weights of each feature (see Fig. 2A) are determined by the OLS after the features were selected by the Lasso model. The standardization implies that a raw, unstandardized input value of zero does not cause individual terms to cancel out. For example, if the number of past interventions ($x_{\text{interventions past 180 days}}$) is zero, the standardized value will be ~ -0.002, not 0, giving a negative contribution to the overall value of rs (from inputting the numbers into the 8th term of the equation). As such, we observe that DSS will interpret no prior interventions as a circumstance indicative of the child being at reduced risk, which could especially affect babies or young children, for which maltreatment has not been detected yet.

2.4 Definition of risk score

As DSS is a linear regression used for a classification problem, it runs into the issue of predicted risk scores rs being unbounded real numbers. To tackle this, rs was subjected to a monotonic transformation to convert values to [1; 10] (see Fig. 2B) with 1 being the lowest possible risk score and 10 being the highest. The following thresholds were used to transform rs into the bounded risk scores RS:

$$RS = \begin{cases} 1, & rs \leq 0.0094122 \\ 2, & 0.0094122 < rs \leq 0.0420653 \\ 3, & 0.0420653 < rs \leq 0.0649243 \\ 4, & 0.0649243 < rs \leq 0.0878977 \\ 5, & 0.0878977 < rs \leq 0.1103221 \\ 6, & 0.1103221 < rs \leq 0.1312139 \\ 7, & 0.1312139 < rs \leq 0.1450047 \\ 8, & 0.1450047 < rs \leq 0.1863387 \\ 9, & 0.1863387 < rs \leq 0.2367061 \\ 10, & rs > 0.2367061 \end{cases} \tag{2}$$

Equations (1) and (2) are the machine learning model which was pilot tested in two municipalities on 208 cases. Taken together, rs is the prediction from the regression model Eq. (1), while RS is the integer risk score, constructed based on binning rs into 10 classes (Eq. 2). During the pilot test it was RS that was presented to social workers. More specifically, a child with a predicted risk score of e.g. 0.24 will get a bounded risk score of 10 due to the transformation in equation (2). Why a predicted value of 0.24 should be indicative of high-risk is unclear, as the transformation has no direct intuitive interpretation. As explained in the documentation of the algorithm: *"The cut-off values are simply the empirical decile limits for the predicted values (so that each risk score corresponds to one decile)."* [32]. The bounds of the transformation function lead to bins of different lengths (Fig. 2B), such that predicted risk scores are evenly distributed, with 10% of the predictions falling within each bin on the 1-10 scale. In the model documentation, no evidence or sources are presented suggesting that the risk of maltreatment should be uniformly distributed. On the contrary, during the pilot test of DSS, the final risk scores provided by social workers followed a normal distribution with approximately 50% of the scores being 5, 6, and 7, and much fewer receiving values 1-2 and 9-10 (exact numbers are not provided in the documentation). As such, the predicted outcome of DSS, RS, does not represent the probability for a child to be at risk of maltreatment, rather it is a relative measurement. Yet, humans and, in particular, social workers could interpret a risk score as a probability of being maltreated, leading to undesired consequences for the screening based on RS only. Based on the above, the bounded predicted outcome of DSS cannot be recognized as a genuine risk score, although it is presented as an objective measure of risk in the documentation. Nonetheless, our audit will continue to use the term *risk score* when referring to RS. We do this to be consistent with the language used in the model documentation, and other documents that have been presented to the public, the DSS reference group, the DSS academic advisory board, and pilot municipalities.

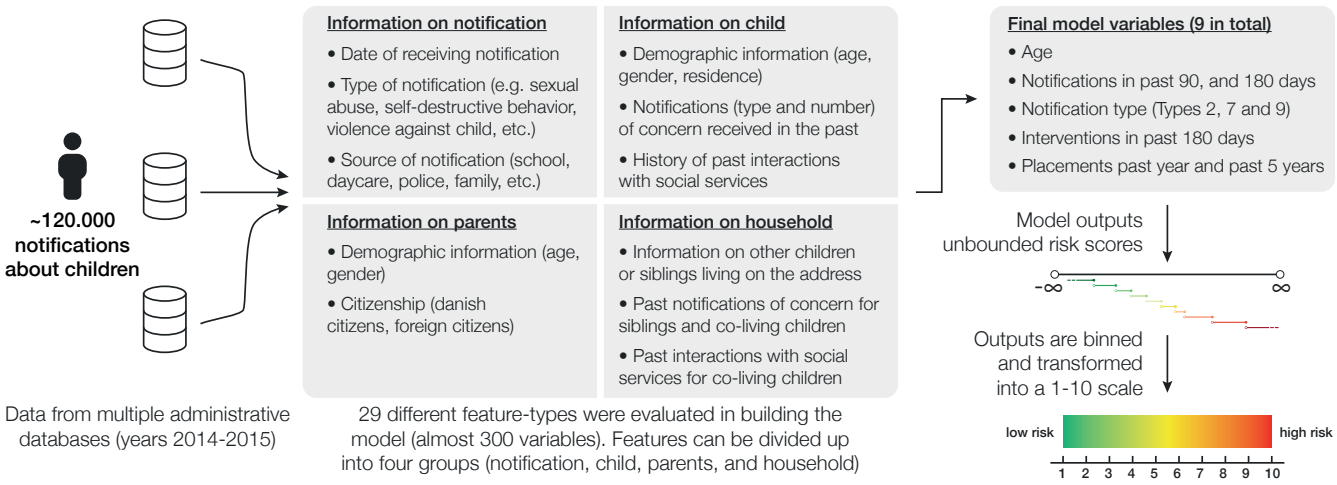

Figure 1: Overview of the Decision Support (DSS) model. Data from Statistics Denmarks' databases were merged to form a large dataset containing almost 300 variables. Nine variables were selected for the final model, based on a post-Lasso framework. The model outputs unbounded risk scores which are transformed to a 1-10 scale, where 1 denotes low risk of maltreatment, and 10 denotes high risk. The green-yellow-red color scale was used in the pilot test of the model, during which social workers were presented with model predictions encoded on this scale.

Variable	Domain	Description
$x_{\text{child age}}$	Integer	Age of the child
$x_{\text{notifications past 90 days}}$	Integer	Number of prior notifications received in the past 90 days
$x_{\text{notifications past 180 days}}$	Integer	Number of prior notifications received in the past 180 days
$x_{\text{type 2 notifications}}$	Binary	Is notification Type 2 (child has committed a crime)
$x_{\text{type 7 notifications}}$	Binary	Is notification Type 7 (child has suffered physical or sexual abuse)
$x_{\text{type 9 notifications}}$	Binary	Is notification Type 9 (substance abuse by a parent).
$x_{\text{interventions past 180 days}}$	Integer	Number of implemented interventions in the past 180 days
$x_{\text{placements past year}}$	Integer	Number of out-of-home placements in the past year
$x_{\text{placements past 5 years}}$	Integer	Number of out-of-home placements in the past five years

Table 1: Description of the 9 features used in the DSS model and the domain of the variables.

3 AUDITING THE DECISION SUPPORT MODEL

Due to privacy concerns, we do not have access to the training data of the algorithm, nor to how the model performance was evaluated. For this reason, we perform our audit by reviewing the methodology used for training the model, simulating cases and counterfactuals, studying disparate impacts, and highlighting how biases in the data generation process can skew results.

3.1 Our audit methodology

Our methodology, inspired by Obermeyer et al. [28], applies to scenarios where direct access to an algorithm — whether whitebox or blackbox — is available and can be queried for predictions. Below, we detail our methodology, bearing in mind that the applicability of some of the steps is conditional on the availability of information or data. First, we examine the methodology used by the developers of DSS for constructing the algorithm. This examination includes scrutinizing the data splitting and preprocessing methods, the choice of model, and the evaluation of model performance. Since we had access to a whitebox model, we also examined the weights of the features, to understand their importance in predicting the outcome variable. Second, regarding fairness, our audit focuses primarily on disparate treatment, determining whether individuals belonging to a specific population group are treated less favorably than other groups. Without having access to the labeled dataset, we focus on age-related disparate treatments, as age is one of the selected features of the DSS model. This allows us to make simulations and construct counterfactuals, to assess how the algorithm outputs risk scores for various control cases. Finally, we also speculate on potential disparate impacts related to other characteristics, such as socio-economic status, by leveraging publicly available statistical data.

Our approach is unfortunately limited by the available information, critically precluding a direct evaluation of how predictions can

propagate and potentially create feedback loops that self-validate the DSS model. Nonetheless, we discuss later the potential issues self-validation might introduce.

3.2 Methodological pitfalls of DSS

Several issues of the data pre-processing, detailed at the beginning of Section 2.2, merit closer scrutiny here. The first issue is about the ordering of steps 2 and 3, which indicates that information has leaked from the test set into the training process, or, put differently, that the test and training sets are not independent. This issue is widely recognized in the machine learning literature and can significantly impact the results [22]: the test set no longer provides a reliable measure of model generalization and performance is overestimated [23]. Unfortunately, we have no information about the performance of the model or how it was evaluated. We asked to see the evaluation documents and code as well as the evaluation results, but the research team politely declined the request. Therefore, it is difficult to quantify the real impact of the information leakage. However, we have documentation for a previous version of the model, developed prior to the one used in the pilot. This documentation states that the model achieved an average AUC (Area Under the Curve, where curve refers to the receiver operating characteristic curve) score of approximately 0.7492 (listed as 74.92% in the documentation) with confidence intervals [0.7388; 0.7595]. While we cannot provide an exact estimate of how information leakage has impacted the AUC estimate, studies on how methodological errors affect machine learning performance have demonstrated that similar methodological flaws can reduce AUC by up to 0.35 points [22] and can cause a bias in error rates of up to 6% [9]. For this reason, we ultimately expect the real performance of the algorithm to be below the stated, and expected, one.

The second issue is about how performance was estimated during model training. AUC evaluates the diagnostic ability of a binary classifier across all possible classification thresholds. While the DSS research team uses 3 different proxies to construct a binary classification problem (will a child be maltreated 6 months after a notification is received, see DSS model description above), the team uses post-Lasso, a regression model that outputs a numerical number, to predict maltreatment. As such, during model training, they need to transform the numerical prediction into a binary value, but how this is done is unclear. We surmise a threshold was used to binarize predictions from the regression model, but we cannot confirm it since the documentation contains no information about it, and our requests to see the code have been declined.

The third issue is about performance listed in percentages. AUC can under very specific conditions be viewed as a probability, where a model whose predictions are 100% wrong has AUC = 0, a model whose predictions are 100% correct has AUC = 1, and a model that is guessing at random has an AUC of 0.5. As such, AUC can be interpreted as the probability that a model ranks a random positive example more highly than a random negative example. This holds for cases where the classification problem is balanced, i.e. there are equally many instances of the positive and negative classes. For unbalanced problems, one needs to be more careful when interpreting AUC [36], as performance estimates can be overly optimistic [19]. Without access to the training data, it is difficult for us to know the exact class balance. Yet, we can get an estimate of the class balance by using publicly available data from Statistics Denmark [1]. Data from 2015, shows that child protective services received notifications for 56,541 children [40]. In the same year, interventions defined by proxy 3 as an indicator of child maltreatment were implemented for 38,391 children [41]. This means that the dataset has at least, not including the other proxies, an imbalance of approximately 70%/30% between positive and negative classes. How this exactly affects the model and its predictions is unclear, as label imbalance issues are not mentioned in the model documentation. Yet, the use of AUC raises concerns that the model performance has not been properly evaluated if AUC was the metric of choice.

The final and last methodological issue is related to the features and their corresponding weights (Eq. (1) and Fig. 2A), which exhibit unusual behavior. There are some obvious inconsistencies: for instance, if a child has been put in foster care within the last year (reflected in the feature $x_{\textbf{placements past year}}$), this lowers the risk score rs and potentially also RS, while if the child has been put in foster care within the last five years ($x_{\textbf{placements past 5 years}}$), rs increases. But, there is an even more problematic feature choice: the information used to construct the labels (proxy 2), is also being used as features. Type 2, 7 and 9 notifications are all considered severe notifications and therefore used as proxy-indicators to indicate maltreatment, the predicted variable. At the same time, these notifications are also used as model features to predict whether one more severe notification will be received, i.e. maltreatment, creating self-fulfilling prophecies.

3.3 Evidence of age bias

DSS is a mathematically simple white-box model. Still, the transformation to bound risk scores (Eq. (2)) makes it difficult to directly interpret the impact of its coefficients. To get an interpretation of the role of the various features (age, notifications, foster placements, etc.), we create various fictional children profiles (Fig. 3) and input the corresponding values into the model. We then obtain the algorithmic risk scores rs and RS and study how a slight change of a single variable affects the risk score when everything else is held constant. Here it is important to note that some variables are strongly correlated. For instance, the number of prior notifications in the past 90 days, and notifications in the past 180 days are strongly linked. A change in one variable will result in an equivalent change in the input value of the other variable. Similarly, Type 2, 7, and 9 notifications are conditioned on at least one notification being received, and the number of placements in the past year and five years are also linked. As such, when we change one of these variables, we have to change all other linked features as well.

Fig. 4 shows simulated risk scores for children aged 0-18 years, where we study how the risk score changes as a function of the type of notification received, and the number of notifications received. We compare this to a base risk profile (black line) estimated by setting all model variables except age to zero. The base risk indicates the score that all well-treated children (or children with no past history with social services) will have according to the DSS algorithm. We find, that depending on age alone, DSS scores children differently. For example, a well-treated 17-year-old (without any notifications, etc.) will have a base risk score of 8, while a 0-year-old

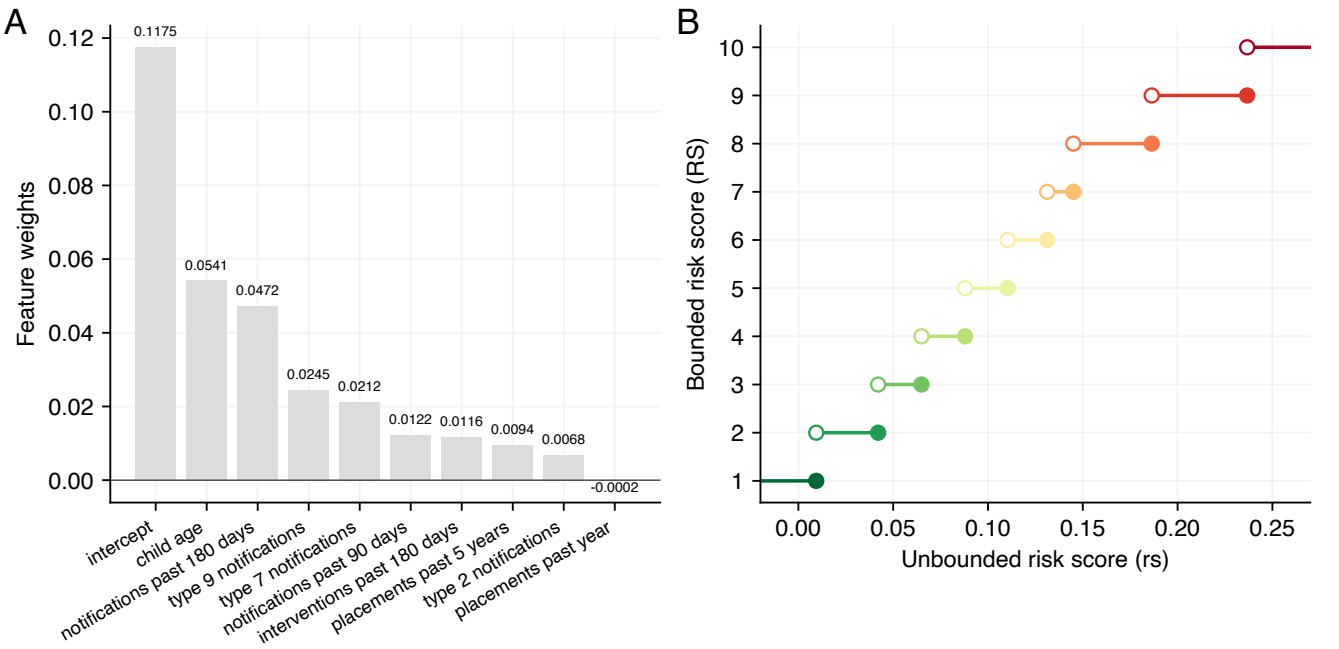

Figure 2: The Decision Support (DSS) model. A, Feature weights of the model, sorted according to magnitude. The weight for $x_{\text{placements past year}}$ is too small to be visually observed. B, Bounds used to transform unbounded risk scores (rs) to bounded risk scores (RS). Open circles indicate half-open intervals.

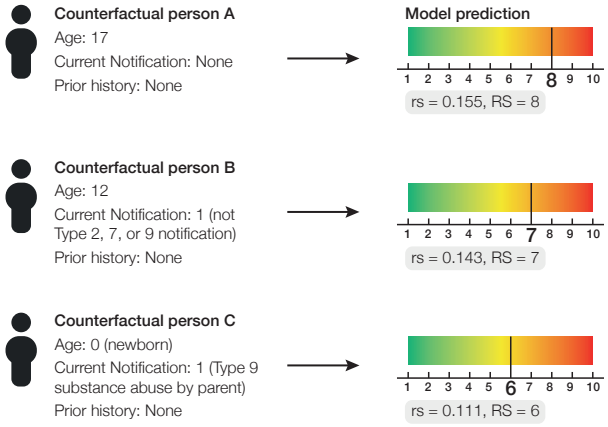

Figure 3: Examples of inconsistent risk scores for three simulated cases. A 17-year-old for whom the authorities have received no notifications will have a bounded risk score RS of 8 (unbounded risk score rs is 0.155). A 12-year-old for whom the authorities have received one notification (for example, for frequently skipping school) will receive a risk score of 7. Worryingly, a newborn whose parents have substance abuse problems and for which the authorities have received a notification receives a score of 6.

will have a risk score of 1. This behavior stems from the intercept and age components in Eqs. (1-2) and results in younger children systematically receiving lower risk scores, while older children get higher scores. Overall, DSS suggests that older children are at substantially higher risk of maltreatment. Any child above the age of 13, receives a risk score of a minimum of 6 solely because of their age. The magnitude of these predictions could perhaps be justifiable if there were a general welfare crisis among teenagers. However, no prior research or evidence suggests that to be the case. We believe this is an unintended and unmitigated consequence of the model, with potential age discrimination consequences if screening decisions are based on RS alone.

Assuming only one notification has been received, Fig. 4A shows the effect of receiving a Type 2 (crime committed by child), 7 (child has suffered physical or sexual abuse), and 9 (substance abuse by a parent) notification. Again, we find that age has a dramatic effect on estimated risk. For instance, well-treated 17- and 18-year-olds (about whom authorities have never received any notifications) are evaluated to be at equal risk as 1-year-olds who have been physically or sexually abused (type 7 notification). Similar issues appear for Type 2 and 9 notifications, with age playing a disproportionately large factor in estimating risk. Unfortunately, similar issues occur when it comes to the number of notifications received by authorities (Fig. 4B). Here well-treated 17-year-olds are ranked to be at equal risk as 1-year-olds about whom the authorities have received 4 notifications. The transformation in Eq. (2) amplifies the issues, but even without bounding risk scores, the algorithm will predict older children to be at higher risk than younger ones. This issue is present for all other features in the model, including the number

of placements and interventions. Fig. 3 provides examples of child profiles and intuitive inconsistencies in the risk scores *RS*.

While gender, parents' employment status, and reports of neglect have been determined to increase the risk of maltreatment [3, 18, 35], we have been unable to find evidence in the literature for age being a cause of increased risk. The strong impact of age on the algorithmic risk score could stem from many factors. We believe one of the major factors is that social services generally receive more notifications for older children (Fig. 5). If we look at the distribution of the number of notifications with regards to child age (Fig. 5A), we find a strong correlation (rank correlation, $r = 0.7$, $p < 10^{-4}$), and the correlation is consistent for statistical data for all years 2015-2020. Similarly, for all years 2015-2020, there is also a strong correlation between placements and age (rank correlation, $r = 0.97$, $p < 10^{-4}$). This indicates that proportionally more teenagers are being referred than younger children, which has a direct impact on the algorithm. One cause of this can be that younger children might not have the means or language to disclose abuse and maltreatment. In fact, research suggests that many children do not disclose abuse at all during childhood [24, 45]. Therefore, the absence of notifications does not mean the absence of maltreatment, highlighting the fundamental importance of choosing good proxies for the target variable [28]. One way to mitigate this issue would be to change the algorithm's training data. As it currently stands the data sample is biased as it only contains information for children about whom the social services have received at least one notification. This means that information about well-treated children (or whom social services have never received any notifications about) is not in the dataset used to train the model. Had these 'negative' cases (negative in the sense that they do not indicate maltreatment) been included in the model training, it would have been less likely that the algorithm had picked age as the most informative factor for child maltreatment.

3.4 Issues with self-validation

It is noteworthy that two out of the three proxies used to indicate maltreatment are directly affected by the social workers themselves, which renders DSS vulnerable to self-validation. For example, the social workers at Child Protective Services have the authority to initiate an out-of-home placement of the child. As such, the outcomes of the three proxy indicators of maltreatment (placement within 180 days, placements within 5 years, and interventions in previous 180 days) are directly affected by social workers who are taking into consideration the risk scores of DSS whilst assessing notifications of concern. This construction of the target variable implies that, for example, a very high-risk score, could nudge the social worker to perceive the immediate risk situation as alarming. If this perception causes the social worker to initiate an intervention or an out-of-home placement, then the target variable would become true, and thus, the model would be self-validating. This is a potentially critical issue because it renders the model's in-practice predictions difficult to evaluate. If the above-mentioned scenario occurs but the risk scores do not, in fact, reflect true risk situations, then an evaluation would give the impression of an accurate model, whilst in reality, the children would experience that their cases were handled excessively or insufficiently.

4 DISCUSSION AND THE FUTURE OF DSS

The DSS model was developed to be used by caseworkers of the Danish Child Protective Services. To convince caseworkers of the usability of the tool, DSS was presented as being (i) faster than humans at evaluating cases, (ii) more knowledge-based since it is based on thousands of previous cases, and (iii) able to streamline assessments by removing the 'bias' of individual caseworkers. However, this algorithm is unsafe to use, as we discuss below, and we urge all local governments, municipalities, and child protection organizations not to use the DSS algorithm or other similarly designed and validated algorithms. The primary and most important reason behind our recommendation is that the algorithm discriminates with respect to age, since it scores otherwise identical cases completely differently just based on the age of the child. Age is a protected attribute [10] and globally recognized as a ground for discrimination [26]. As such, avoiding automated discrimination based on protected attributes should be a prime concern, as DSS influences the lives of human beings, many of whom are vulnerable children. Secondly, the algorithm is trained using a flawed methodology, where its performance has been over-estimated, and where its target variables and features are based on questionable, and self-fulfilling, proxy values. Further, some of the indicators of neglect are direct proxies of poverty. For example, §52 interventions include families getting practical help from social services, however, wealthier individuals who might get the same support, just bought through a private entity, do not end up in the data. From the available model documentation it is unclear which §52 measures have been included in the training dataset by the DSS developers, or if there has been any selection at all. Further, there is some ambiguity on when these offers are initiated. The law states that the municipal board must decide to initiate a preventive measure when it is considered to be of significant value for a child's needs. The board must choose the measure(s) that can best solve the problems and needs that have been uncovered through a child welfare investigation. There is a correlation between the poverty levels of a municipality and the number of preventive measures that have been implemented by municipal boards (rank correlation $r = 0.48$, $p < 10^{-6}$), and while we cannot establish a causal link with the data we have access to, we expect this correlation to be reflected in the risk scores, with poorer families getting higher risk scores. Lastly, the ethical values encoded in the algorithm are dubious. For example, when it comes to Type 9 notifications (whether parents have issues with substance abuse), it is indisputable that no child, no matter their age, should grow up with parents suffering from substance abuse. However, is it reasonable that the immediate risk connected to parental substance abuse is lower for a one-year-old child who is deeply dependent on their parents, than for a seventeen-year-old child who has some prerequisites for managing themselves if their parents are not sober? We believe it is not. A similar questionable ethical concern arises from the fact that parents and children, during the pilot trial, were never notified that their case was evaluated by an algorithm, nor were they offered the option to opt-out **??**. This opt-out option is never discussed in the documentation; yet, we assume a lack of informed consent and opt-out because the documentation specifically states that *'The parents do not have to give consent if the tool is to be used, because*

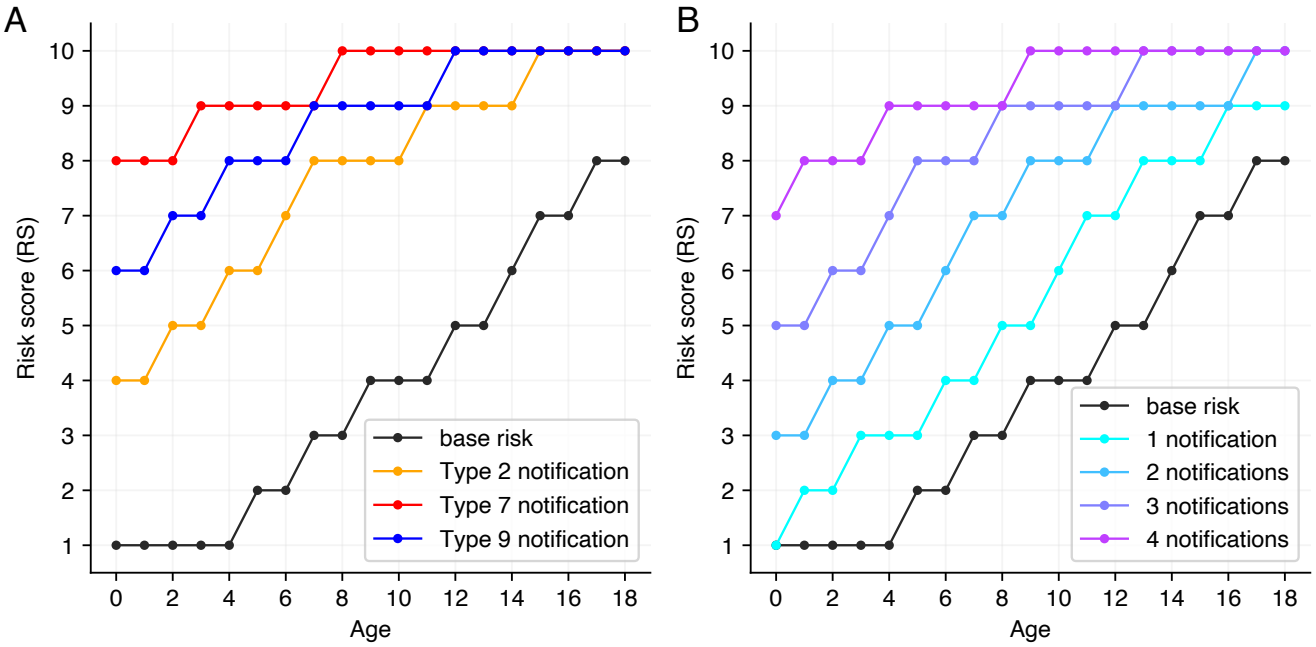

Figure 4: Simulated risk scores for individuals where one variables is changed. A, Difference between risk score for different notification types. Base risk is estimated by setting all features, except age, to zero. To calculate risk for type 2, 7 and 9 notifications we set the respective features to 1, and also assume that the number of received notifications (past 90 and 180 days) is 1. B, Risk scores for an increasing number of notifications. Here we set the number of received notifications in 90 and 180 to the same value.

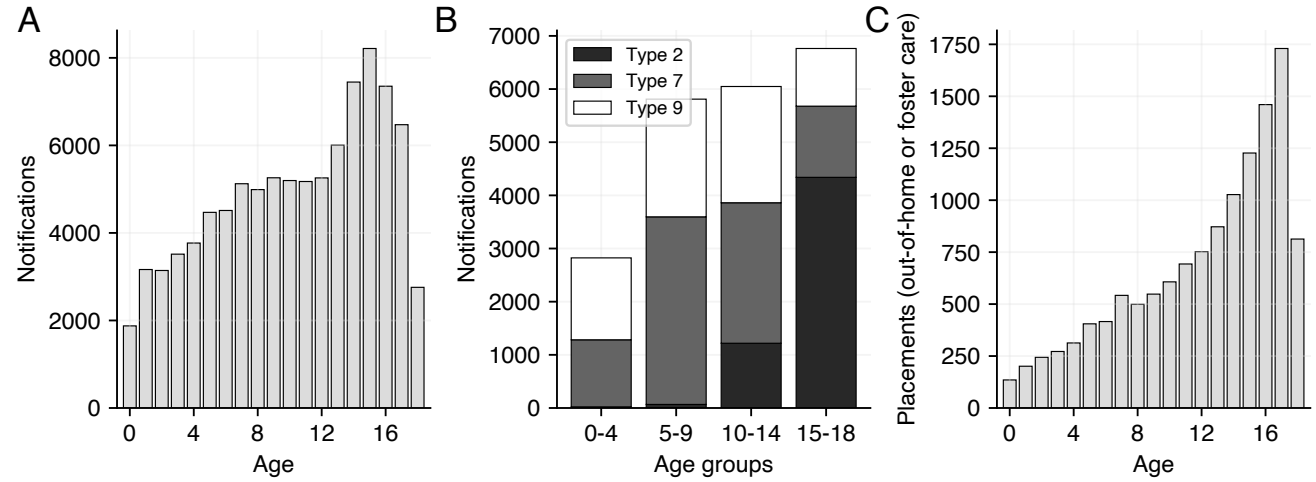

Figure 5: Correlation between age and model features. A, Number of notifications received in 2015 as a function of a child's age. Data is publicly available from Statistics Denmark [39]. B, Number of Type 2, 7, and 9 notifications received in 2015 split up according to 4 age groups [42]. C, Number of placements as a function of child's age [43].

then it is not usable. And it is no different than today. If we have to ask for consent, then it doesn't make sense. The legislation says that we must assess the notification.' (translated from Danish).

Nevertheless, DSS was piloted in 2018 and 2019 on approximately 200 cases in two municipalities, during which the model's predictions were compared to risk evaluations provided by social workers. One researcher from the DSS research team observed and later

interviewed the social workers who tested the algorithm, detailing different instances of how social workers handled assessments provided by the DSS risk scores. We report here a summary of these evaluations [25]. One example of evaluation comparison involves a sixteen-year-old referred with a type 2 notification (the child has committed a crime). According to DSS, the child got a risk score of 10. The initial risk score of the social worker on the case was 4, but the social worker chose to raise their score to 8 after being presented with the risk score of DSS. We do not know the true circumstances of the case, but we can reasonably assume that the social worker in this case was influenced by DSS's prediction. The most striking of DSS's predictions was the risk score of 1 given to a two-year-old child who was referred due to suspicion of neglect. The social worker initially assessed the risk score to be 9, indicating a high risk of vulnerability. After having been presented with the DSS score the social worker chose to maintain their initial assessment. No other information about the child or the notification is known to us. Yet, we can conclude that this DSS risk score was predicted solely based on the child's age as setting any other variable or combinations of variables to exceed a value of zero would have resulted in at least a risk score of 2. Even so, if the child's true conditions in any way resembled the risk assessment of the social worker, then DSS grossly underestimated the risk. In general, social workers adjusted their final risk score in 21% of cases [25] after being presented with risk scores of DSS, which points to some degree of trust in the model.

The mismatch between caseworkers and DSS can potentially stem from the fact that caseworkers were never involved in the development phase of the tool. The model was only presented to them after it had been developed. Following this, they were asked to evaluate how well the model worked, while not having the right information and training to evaluate it truthfully. Nonetheless, the feedback caseworkers provided indicates they could use some help to evaluate the high volume of notifications they are currently receiving, but that they are mindful of using such an algorithmic tool. The key concern of caseworkers, which they highlighted to the researchers during the interviews, is that in their daily work, they do not just look at one score, rather they asses the problem from multiple angles [25]. First, they evaluate whether the child is safe now, i.e. is it in immediate danger, and then what are the long-term consequences of not helping the child's family. There are different risk factors associated with these two assessments, and a single risk model cannot be used for both.

The results of the pilot and feedback from the caseworkers have not halted the development of a new version of DSS [33], which the research team plans to pilot in the near future in more Danish municipalities [27]. This new version is going to be described in a yet unpublished research article, which unfortunately the research team has politely declined to share with us [32]. During the development phase, we asked for additional details on the new model, which the researchers also declined to share with us [32]. As such, it is unknown whether any steps have been taken to fix the methodological issues, mitigate the age discrimination, address the self-validation loop, explain the new model's inner workings to social workers, and whether the children and families subjected to the next pilot will be informed of the existence and influence of DSS on their case this time around. Independent legal experts, however, have raised doubts about the legality of testing the algorithm on real cases, and for this reason, for now, any further pilot has been paused [7]. From the information we have available for the new version of DSS [33], the predictive part has been changed to use an XGboost model [11]. The performance of the model is stated as "*AUC: 83.95%*", although it is still a regression-type model that outputs unbounded scores, which are later transformed into 1-10 integer risk scores. The authors claim this new model is unbiased with respect to gender and ethnicity, but the authors never show any evidence to support these claims. Nonetheless, the authors do note that model predictions are very skewed with respect to socio-economic class [33].

Child maltreatment is an important, complex, and multifaceted issue that needs to be addressed [29]. However, we find that the DSS algorithm is not the right solution. The question is, can any algorithmic tool be used for this endeavor? In a recent mass collaboration study [37] hundreds of researchers, divided up into 160 teams, attempted to predict children's life trajectories using various machine learning techniques on a rich dataset regarding thousands of families. The study found that none of the 160 teams accomplished the task of making accurate predictions on children's life trajectories. In fact, the best predictions were only slightly better than those from a simple benchmark model. The study further questions whether it is even possible to use algorithms to predict life outcomes, and states that practical and predictive limits can exist.

5 IMPLICATIONS AND RECOMMENDATIONS

We recommend that the new version of DSS should not be piloted on any pending notifications until a new independent audit has been conducted, especially if this new version of DSS includes age, gender, ethnicity, or other sensitive attributes as variables. We further call upon policymakers and scientists to be careful when contemplating whether or not to develop and use predictive tools for social services. Machine learning and artificial intelligence tools work well on mathematically well-defined problems, in well-defined situations, with well-defined parameters [8, 44]. However, our world is incredibly complex, where data distributions constantly drift and evolve [30], and people change behaviors. As such, algorithms trained for this purpose must be constantly re-trained, re-evaluated, and audited. Training one algorithm and believing it will work indefinitely is a wrong assumption. The best solution to fix social problems is often not to use algorithms, but instead to invest resources to empower caseworkers and strengthen existing systems [20].

DSS is one of many algorithms currently being tested on issues relating to social aspects, especially for children's welfare. In addition to general recommendations about algorithmic systems being transparent, ethical, and respecting basic human rights [46], our recommendations are:

- It is vital to incorporate algorithmic audits during the development stage of models. One should not wait to do an audit until after model deployment when the system has already negatively impacted users. Once deployed, issues in the algorithm can become difficult or impossible to trace back to the original source.

- The training, testing, and implementation of high-risk systems should not be left to one team of researchers or practitioners. In the academic world, we have peer-review systems that are used to evaluate quality and pinpoint any issues. A similar system could have avoided the methodological pitfalls and other shortcomings of the DSS model.
- It is crucial to assess algorithms on all grounds of discrimination (or protected characteristics), even those that might not, at first, seem relevant. Even in the absence of explicit elements in the data, datasets may contain proxies that enable models to infer discriminatory grounds, e.g. gender, age, ethnicity, or socio-economic status, through these proxies.
- Algorithmic audits should cover an evaluation of both model outputs and inputs. I.e. it is vital to understand if the underlying data distributions are biased, or skewed in any form.
- Algorithms need to be constantly monitored, audited, and evaluated. As human behaviors evolve and change, algorithms might drift towards unsafe conditions, unless constantly maintained and retrained.

ACKNOWLEDGMENTS

R.S. acknowledges support from Villum Fonden through the Villum Young Investigator program (project number: 00037394). V.S. acknowledges support from DIREC Denmark (Explore Project, P25). We thank Tess Sophie Skadegård Thorsen and Michael Szell for helpful discussions and the reviewers for their constructive feedback.

REFERENCES

[1] [n. d.]. Statistics Denmark. https://www.dst.dk/en.
[2] 2023. Project desctiption: Notifications in Focus (in Danish only), https://childresearch.au.dk/fileadmin/childresearch/dokumenter/Underretninger/Projektbeskrivelse_august_2023..docx.pdf [Last accessed 2024-01-10].
[3] MA Al-Eissa, HN Saleheen, S AlMadani, FS AlBuhairan, A Weber, John D Fluke, M Almuneef, and KL Casillas. 2016. Determining prevalence of maltreatment among children in the kingdom of Saudi Arabia. *Child: care, health and development* 42, 4 (2016), 565–571.
[4] Amnesty International. 2021. *Xenophobic machines: DISCRIMINATION THROUGH UNREGULATED USE OF ALGORITHMS IN THE DUTCH CHILDCARE BENEFITS SCANDAL.* Technical Report.
[5] Amnesty International. 2023. *Trapped by Automation: Poverty and discrimination in Serbia's welfare state.* Technical Report.
[6] Alexandre Belloni and Victor Chernozhukov. 2013. Least squares after model selection in high-dimensional sparse models. *Bernoulli* 19, 2 (2013), 521–547.
[7] Line Berg. 2021. Notat om ændringer i projekt Underretninger i fokus (English title: Note on changes in project Notifications in focus). Only released in Danish, available at https://childresearch.au.dk/fileadmin/childresearch/dokumenter/Underretninger/20220620_Notat_aendring_i_projektdesign_juni_2022.pdf.
[8] Meredith Broussard. 2018. *Artificial unintelligence: How computers misunderstand the world.* mit Press.
[9] Gavin C Cawley and Nicola LC Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* 11 (2010), 2079–2107.
[10] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*. 339–348.
[11] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
[12] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. PMLR, 134–148.
[13] Mogens N Christoffersen, Cherie Armour, Mathias Lasgaard, Tonny E Andersen, and Ask Elklit. 2013. The prevalence of four types of childhood maltreatment in Denmark. *Clinical practice and epidemiology in mental health: CP & EMH* 9 (2013), 149.
[14] Kate Crawford and Ryan Calo. 2016. There is a blind spot in AI research. *Nature* 538, 7625 (2016), 311–313.
[15] Danish Social Services Act. [n. d.]. LBK no. 170 of 24/01/2022, https://www.retsinformation.dk/eli/lta/2022/170 [Last accessed 2022-04-05].
[16] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.
[17] European Commission. 2021. Proposal for a Regulation of the European Parliament and Of The Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final.
[18] Eveline M Euser, Marinus H van IJzendoorn, Peter Prinzie, and Marian J Bakermans-Kranenburg. 2010. Prevalence of child maltreatment in the Netherlands. *Child Maltreatment* 15, 1 (2010), 5–17.
[19] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. *Learning from imbalanced data sets*. Vol. 10. Springer.
[20] Seventy F Hall, Melanie Sage, Carol F Scott, and Kenneth Joseph. 2023. A Systematic Review of Sophisticated Predictive and Prescriptive Analytics in Child Welfare: Accuracy, Equity, and Bias. *Child and Adolescent Social Work Journal* (2023), 1–17.
[21] Human Rights Watch. 2020. *Automated Hardship.* Report.
[22] Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4, 9 (2023).
[23] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 4 (2012), 1–21.
[24] Kamala London, Maggie Bruck, Stephen J Ceci, and Daniel W Shuman. 2005. Disclosure of child sexual abuse: What does the research tell us about the ways that children tell? *Psychology, Public Policy, and Law* 11, 1 (2005), 194.
[25] Clara Siboni Lund. 2019. Algoritmer i socialfaglige vurderinger: En undersøgelse af socialarbejderes opfattelse af at anvende algoritmer til vurdering af underretninger. *Uden for nummer* 39 (2019), 20–31.
[26] United Nations. 1966. International Bill of Human Rights. Available at https://www.ohchr.org/en/what-are-human-rights/international-bill-human-rights.
[27] Notifications in Focus. [n. d.]. Invitation to participate in research project on notifications in focus. Aarhus University, VIA University College, Trygfonden. https://childresearch.au.dk/fileadmin/childresearch/dokumenter/Invitation_til_at_deltage_i_projekt_Underretninger_i_Fokus.pdf [Last accessed 2022-04-19].
[28] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[29] World Health Organization. [n. d.]. Jun 08 2020, "Child maltreatment", Fact Sheet, https://www.who.int/en/news-room/fact-sheets/detail/child-maltreatment [Last accessed 2022-04-19].
[30] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning.* Mit Press.
[31] Tapani Rinta-Kahila, Ida Someh, Nicole Gillespie, Marta Indulska, and Shirley Gregor. 2022. Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. *European Journal of Information Systems* 31, 3 (2022), 313–338.
[32] Michael Rosholm. 2020. Confirmed via email correspondence with the PI of the Decision Support project..
[33] Michael Rosholm, Simon Bodilsen, and Sanne Dalgaard Toft. 2021. Egenskaber ved den statistiske model i forskningsprojektet Underretninger i fokus (English title: Properties of the statistical model in the research project Notifications in focus). Only released in Danish, available at https://childresearch.au.dk/fileadmin/childresearch/dokumenter/Underretninger/20220617_Egenskaber_ved_den_statistiske_model_UiF.pdf.
[34] Michael Rosholm, Simon Bodilsen, and Anne Marie Villumsen. 2021. Ethical considerations in relation to 'Focus on Notifications': A project on the use of predictive risk models in social work. Working paper,https://childresearch.au.dk/fileadmin/childresearch/dokumenter/Underretninger/20220617_Ethical_considerations_Focus_on_Notifications.pdf.
[35] Bénédicte Rouland and Rhema Vaithianathan. 2018. Cumulative prevalence of maltreatment among New Zealand children, 1998–2015. *American journal of public health* 108, 4 (2018), 511–513.
[36] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 10, 3 (2015), e0118432.
[37] Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences* 117, 15 (2020), 8398–8403.

[38] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the US child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[39] Statistics Denmark. [n. d.]. UND1: Notifications of concern for children by administrative municipality, reporter (who is notifying), age and sex . https://www.statistikbanken.dk/UND1 [Last accessed 2022-04-19].

[40] Statistics Denmark. [n. d.]. UND2: Children who there are received notifications of concern about by administrative municipality, notification, age and sex. https://www.statistikbanken.dk/UND2 [Last accessed 2022-04-19].

[41] Statistics Denmark. [n. d.]. BUFF01: Children and young persons with preventive measures per 31st December by region and measure. https://www.statistikbanken.dk/BUFF01 [Last accessed 2022-04-19].

[42] Statistics Denmark. [n. d.]. UND3: Causes for notifications of concern for children by administrative municipality, cause, reporter (who is notifying), age and sex. https://www.statistikbanken.dk/UND3 [Last accessed 2022-04-19].

[43] Statistics Denmark. [n. d.]. ANBAAR2: Children and young persons placed outside of own home per 31st december by measure, age and sex. https://www.statistikbanken.dk/ANBAAR2 [Last accessed 2022-04-19].

[44] Rachel L Thomas and David Uminsky. 2022. Reliance on metrics is a fundamental challenge for AI. *Patterns* 3, 5 (2022).

[45] Sarah E Ullman. 2002. Social reactions to child sexual abuse disclosures: A critical review. *Journal of child sexual abuse* 12, 1 (2002), 89–121.

[46] UN Doc HR/PUB/11/04. 2011. Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.

[47] World Economic Forum. 2018. *How to Prevent Discriminatory Outcomes in Machine Learning*. White Paper. Global Future Council on Human Rights 2016-2018.