CPH TECH POLICY BRIEF #6

NOVEMBER 2023

PUBLISHED BY THE CROWN PRINCESS MARY CENTER, COPENHAGEN CENTER FOR SOCIAL DATA SCIENCE & DEPARTMENT OF POLITICAL SCIENCE / UNIVERSITY OF COPENHAGEN

ALGORITHMIC FAIRNESS: LEARNINGS FROM A CASE THAT USED AI FOR DESICION SUPPORT

BY VEDRAN SEKARA[a], TESS SOPHIE SKADEGÅRD THORSEN[b], ROBERTA SINATRA[c]

This CPH Tech Policy Brief is based on a working paper by Vedran Sekara, Therese Moreau Hansen and Roberta Sinatra. The brief provides a small introduction to algorithmic fairness and an example of auditing fairness in an algorithm which was aimed at identifying and assessing children at risk from abuse.

a Assistant Professor, Department of Computer Science, IT University of Copenhagen
b PhD, Independent researcher and member of the advisory board to the Danish Pioneer Centre for AI
c Professor in Computational Social Science, Copenhagen Center for Social Data Science, University of Copenhagen, and Department of Computer Science, IT University of Copenhagen

OVERVIEW

Algorithmic decision-making systems are increasingly adopted by governments and public service agencies to make life-changing decisions. However, scientists, activists, policy experts, and civil society have all voiced concern that such systems are deployed without adequate consideration of potential harms, biases, disparate impacts, and accountability.

Taking its point of departure in a single case from two Danish municipalities, in which two of the three authors of this brief helped unearth a potentially risky and harmful use of algorithmic decision-support in the placement of children, this policy brief aims to explain and contextualize central issues around algorithmic fairness, bias, and auditing.

It is crucial to ensure that algorithmic systems work as intended, and work fairly. One vital type of check is to ensure that algorithms do not discriminate against any individuals, groups, or populations. Ensuring that algorithms produce 'outputs' of equitable quality, accuracy, and utility for different groups (e.g. men and women, old and young, abled and disabled, etc.), and for different intersections of them, is called algorithmic fairness.

WHAT IS ALGORITHMIC FAIRNESS?

Algorithmic fairness is the principle that algorithms, especially those used in decision-making processes, should strive to operate in a way that is impartial, unbiased, and does not perpetuate existing inequities or create new ones. In most cases, experts within algorithmic fairness emphasize that automated systems should treat all individuals equitably, regardless of their race, gender, socioeconomic status, or other grounds for discrimination[1] [2].

All algorithmic systems, from the simplest tools to complex artificial intelligence (AI) systems can create or reproduce biases. However, unlike traditional algorithms that follow explicit instructions, Machine Learning (ML) algorithms (sometimes called models, or AI models) "learn" patterns from data. This makes it more challenging to understand their inner workings and to ensure they are fair. Since data is often ingrained with different forms of biases these algorithms frequently reproduce them too.

Algorithmic fairness is often assessed through algorithmic audits. Audits can take on different forms depending on the type of algorithm, the code (open source or not), the availability of training data, access to the model (full or partial access), and the stage of development (e.g. the algorithm has been deployed or is in development). Audits can be undertaken by various parties, including external and internal experts, regulators, or researchers.

In the following, we show an example of algorithmic fairness research, by focusing on an algorithm that has been studied and audited by some of the authors of this policy brief, and provide general recommendations for policy makers about algorithmic fairness.

THE CASE OS "DECISION SUPPORT"

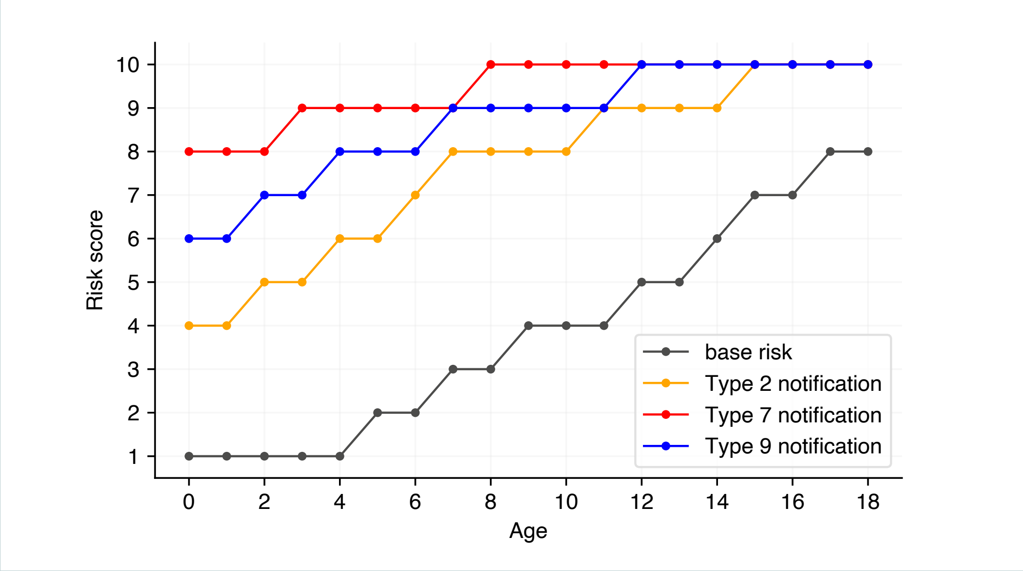
Algorithmic systems have been piloted in Denmark for identifying and assessing children at risk from abuse and neglect[3 4]. However, they have been mired in controversy. The usage of algorithms stems from the fact that Danish social services have adopted a policy of being proactive rather than reactive regarding cases of child maltreatment. The most common way for child and family welfare services to detect abuse is by receiving notifications. A variety of different persons and institutions, e.g. public institutions and non-governmental organizations, have a legal obligation to notify when they are concerned with the well-being of a child. Since by law, social workers must conduct an initial assessment within 24 hours of receiving a notification to ensure that children in immediate danger are aided promptly, the pressure on child protective services is immense and beyond the available human resources.

It is against this backdrop that algorithms were suggested[5 6] as a possible solution to support social workers by assigning a risk score to cases and therefore prioritizing their assessments. The vision was that an algorithm based on a predictive risk model could assist social workers in assessing the growing number of notifications by providing rapid and consistent risk assessment for children referred to Child Protective Services. Inspired by approaches from the US, such as the Alleghney Family Screening Tool[6 7], a Danish decision support solution (DSS) was developed, called "Notifications in focus"[5]. In collaboration with two Danish municipalities (Silkeborg and Hjørring), an ML algorithm was developed and pilot-tested in these municipalities. The model was pilot tested from November 2018 to February 2019 on 208 cases. The assumption behind this approach is that a predictive risk model can take all relevant protective and risk factors into account by leveraging the vast amounts of personal data that Danish authorities collect about its citizens. Using this information the idea behind DSS was to build a predictive model that could provide qualified and consistent risk assessments for all referred children, whether they had previously been in contact with child protective services or not.

WHITE BOX MODELS AND BLACK BOX MODELS

There are different types of machine learning models, often divided into white- and black-box models. White box models allow humans to gain insight into the algorithm's inner workings and easily interpret how it was able to produce its output and draw its conclusions. This is called explainability, meaning a human can explain how and why a model made a decision. Black box models are different; it is often very difficult, and sometimes impossible, to clarify how they came to their conclusions. As such, while a person can observe the input and the output, the inner workings of these models are harder to understand. Both types of models have benefits and drawbacks. White box models are interpretable and

FIGURE 1 Predicted risk scores for children of different ages and received notification types using the DSS algorithm



Note: Predicted risk scores for children of different ages and received notification types using the DSS algorithm. The base risk is estimated by setting all variables except age to zero. Risk scores for Type 2 (child has committed a crime), type 7(child has been abused), and type 9 (substance abuse by a parent) notifications are estimated from receiving 1 notification of that specific type.

explainable, but their overall accuracies generally tend to be lower. Black box models achieve higher predictive accuracy but are less transparent. In terms of auditing, white box models are considerably easier to audit and understand.

THE ALGORITHMIC MODEL

DSS is a predictive model which estimates the risk of a child experiencing maltreatment. However, maltreatment is difficult to quantify, as no single variable indicating whether a child is at risk exists. For this reason, maltreatment was said to have happened if one of these three things occurred within six months of a notification:

1. The child is placed in foster care or similar forms of out-of-home placements.
2. A severe notification is received. (i.e. physical, sexual, or emotional abuse or a parent has abused substances).
3. A 'severe' intervention (defined by §52 in the Danish Social Services Act13) is implemented. §52 interventions range from families getting practical help or children being offered spots in youth clubs, to children being put in foster care. Exactly what definition of 'severe' has been used for the algorithm is unclear, as the law does not use this terminology, nor does the documentation state what is considered severe.

120,000 notifications of concern submitted to social services between 2014 and 2015 were used as training data for the model. The data comes from Statistics Denmark, which is the country's central authority on statistics. The model included other variables, i.e. information about the notification (who reported it, type of report, when it was reported), the child itself (age, gender, past history, place of residence), and the parents (age, gender, origin, marital status). More than 200 variables were initially selected based on the idea that they are easily accessible and understandable to caseworkers.

AUDITING DECISION SUPPORT

Due to privacy concerns we do not have access to the training data of the algorithm. Instead, our audit reviewed the methodology, evaluated model variables and weights, simulated cases, and studied disparate impacts, highlighting how biases in the data generation process can skew results. Here, we do not describe how the audit is performed, as it would need to cover methods and technicalities in detail. Instead, we provide a summary of the issues that the audit uncovered.

WHAT ARE THE ISSUES?

Evidence of age bias. DSS is a mathematically simple model. To understand it, we input values for fictional children, and study how a slight change of a single variable affects algorithmic risk scores when everything else is constant. Fig. 1 shows estimated risk scores for children aged 0-18 years. We tested the model for age bias by inputting identical cases, where the only variable we changed was age. We found that depending on age alone, DSS will score children differently. For example, a well-treated 17-year-old (without any notifications) will have a base risk score of 8, while a 0-year-old will have a risk score of 1. Overall, the model suggests that older children are at substantially higher risk of maltreatment, everything else being equal. Any child above 13, receives a risk score of a minimum of 6 solely due to 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 bias in the model.

The strong impact of age on the algorithmic risk score could stem from many factors. Perhaps more reports are made for older children, or perhaps more placements are made when children accumulate multiple reports, which automatically occurs through accumulation of time. One cause of this could 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[8]. Regardless, one way to alleviate this issue would be to change the algorithm's training data.

Invisible children. The data sample is biased as it only contains information for children which the social services have received at least one notification about. This means that information about children who are well-treated (and for whom social services have never received any notifications) is not in the dataset used to train the model. These children are invisible to 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 of child maltreatment. In other words, the training dataset reflects only instances where maltreatment has been reported, rather than providing a comprehensive view of both reported and unreported cases. Put differently, this corresponds to trying to teach an algorithm how to detect good apples from bad apples by only showing examples of bad apples. As such, the algorithm is not getting the full picture, which can lead it to make incorrect judgments.

Risk of poverty bias. Some of the chosen indicators of neglect are direct proxies of poverty. For example, §52 interventions include families getting practical help from social services. However, wealthier individuals who get the same support, just bought through a private entity, do not end up in the data. We expect this to be reflected in the risk scores, with children from wealthier families being more invisible in the data, and poorer families getting higher risk scores.

Issues with self-validation. It is noteworthy that two 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 are directly affected by social workers. This 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.

IMPLICATIONS

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 framed to be (1) faster at evaluating cases, (2) more knowledge-based since it is based on thousands of previous cases, and (3) able to streamline assessments by removing the 'bias' of individual caseworkers. However, as our audit identified this algorithm suffers from multiple issues, which invalidate these claims. DSS has been piloted in 2018 and 2019 on approximately 200 cases in two municipalities, where its predictions have been compared to risk evaluations given by social workers. We bring here an extract of these comparisons. One example involves a sixteen-year-old who had been referred with a type 2 notification (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 conclude that the social worker in this case might have been influenced by DSS's prediction. Nonetheless, 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 would result 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. The mismatch between caseworkers and DSS can potentially stem from the fact that case workers were never involved in the development phase of the tool. The model was only presented to them after it had been developed.

Child maltreatment is an important, complex, and multifaceted issue that needs to be addressed. However, we find that the DSS algorithm is not the right solution. The question is, can any algorithmic tool be used for this endeavor? Recent research has raised questions whether it is even possible to use algorithms to predict life outcomes[9].

POLICY RECOMMENDATIONS

DSS is one of many algorithms currently being tested on issues relating to social or human aspects. For instance, a predictive algorithm has been used to predict unemployment risk[10]. We 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. However, our world is incredibly complex; people change behaviors, as a result the underlying data distributions constantly drift. As such, algorithms trained for a specific purpose will have to be constantly re-trained, and re-evaluated. Training one algorithm and believing it will work indefinitely is a wrong assumption. As such, when aiming to fix social problems, it is important to seriously weigh the consequences of building, maintaining, auditing, and using algorithms, compared to, for instance, dedicating resources to empowering case workers and strengthening existing systems.

In addition to general recommendation about algorithmic systems being transparent, ethical, and respect basic human rights[2]; our recommendations about fairness are:

- 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.
- 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.
- It is vital to assess algorithms based on all grounds of discrimination (or protected characteristics), including ones which might not, at first, seem relevant. Even if certain elements are not present in the data there might be proxies in a dataset which make it possible for models to infer ground for discriminations (e.g. gender or age) 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.
- Audits should not only be about the performance and bias of an algorithm, but also about the algorithm's impacts on human rights. An algorithm can be unbiased, yet still cause adverse impacts.

REFERENCES

1 UNITED NATIONS, INTERNATIONAL BILL OF HUMAN RIGHTS (1966). AVAILABLE AT https://www.ohchr.org/en/what-are-human-rights/international-bill-human-rights.

2 GUIDING PRINCIPLES ON BUSINESS AND HUMAN RIGHTS: IMPLEMENTING THE UNITED NATIONS 'PROTECT, RESPECT AND REMEDY' FRAMEWORK, UN DOC HR/PUB/11/04 (2011). AVAILABLE AT www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

3 KJÆR, JAKOB S. "GLADSAXE INDSTILLER ARBEJDET MED OMSTRIDT OVERVÅGNING AF BØRNEFAMILIER." POLITIKEN, 6, FEBRUARY, 2019. AVAILABLE AT https://politiken.dk/indland/art7023935/Gladsaxe-indstiller-arbejdet-med-omstridt-overv%C3%A5gning-af-b%C3%B8rnefamilier.

4 KULAGER, FREDERIK. "KAN ALGORITMER SE IND I ET BARNS FREMTID?" ZETLAND, FEBRUARY, 3, 2021. AVAILABLE AT https://www.zetland.dk/historie/s8YxAamr-aOZj67pz-e30df/.

5 "UNDERRETNINGER I FOKUS", AARHUS UNIVERSITY. AVAILABLE AT https://childresearch.au.dk/projekter/familie-og-forebyggelse/projekter/underretninger-i-fokus.

6 CUCCARO-ALAMIN, STEPHANIE, ET AL. "RISK ASSESSMENT AND DECISION MAKING IN CHILD PROTECTIVE SERVICES: PREDICTIVE RISK MODELING IN CONTEXT." CHILDREN AND YOUTH SERVICES REVIEW 79 (2017): 291-298.

7 EUBANKS, VIRGINIA. AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR. ST. MARTIN'S PRESS, 2018.

8 LONDON, KAMALA, ET AL. "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.

9 SALGANIK, MATTHEW J., ET AL. "MEASURING THE PREDICTABILITY OF LIFE OUTCOMES WITH A SCIENTIFIC MASS COLLABORATION." PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 117.15 (2020): 8398-8403.

10 MOREAU, THERESE AND KULAGER, FREDERIK. "VI HAR SKILT JOBCENTRENES ALGORITME AD." ZETLAND, 10, JUNE, 2021. AVAILABLE AT https://www.zetland.dk/historie/sOMVZ7qG-aOz9m93B-a30b8.