TRYGFONDENS
BØRNEFORSKNINGSCENTER

Bring ideas to life
VIA University College

# Dokumentet er delvist forældet.

Juni 2022

Grundet ændringer i projektets design er dele af dette dokument forældet.

Det er ikke længere ambitionen at gennemføre afprøvning af Beslutningsstøtten på faktiske sager i dette projekt. Dermed er enhver beskrivelse af Pilottest 2 og Lodtrækningsforsøg forældet.

Baggrunden for ændringerne er beskrevet i "Notat om ændringer i projekt Underretninger i Fokus", som er tilgængelig på Trygfondens Børneforskningscenters hjemmeside https://childresearch.au.dk/udsatte-boern-unge-og-familier/projekter/underretninger-i-fokus og UC viden https://www.ucviden.dk/da/projects/underretninger-i-fokus.

Projektleder
Line Berg

AARHUS BSS
TRYGFONDEN'S CENTRE FOR CHILD RESEARCH
DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

VIA University College

TrygFonden

# Ethical considerations in relation to 'Focus on Notifications': A project on the use of predictive risk models in social work[*]

Michael Rosholm
PhD, Professor, University of Aarhus, TrygFonden's Centre for Child Research, and Department of Economics and Business Economics
rom@econ.au.dk

Simon Bodilsen
PhD, Assistant Professor, University of Aarhus, TrygFonden's Centre for Child Research, and Department of Economics and Business Economics
sibo@econ.au.dk

Anne Marie Villumsen,
PhD, Senior Associated Professor, VIA University College, Department of Social Work
amv@via.dk

*In this paper, we offer some ethical considerations regarding the research project Beslutningsstøtte ved underretninger and Underretninger i fokus (Focus on Notifications).*

**Keywords:** Ethics; Social Work; Machine Learning, Predictive Risk Modelling

**Date of this report: February 2021**

[*] We acknowledge financial support from TrygFonden. Helpful comments from Jacob Birkler, PhD in medical ethics and former chairman of the Danish Ethics Council (Etisk råd), as well as continuous discussions with our academic advisory board, consisting of professor Rhema Vaithianathan, professor Emily Putnam-Hornstein, and professor Brian Taylor, are gratefully acknowledged.

**February 2021 – working paper**

# Table of Contents

# 1 Introduction

The purpose of this paper is to discuss and reflect upon ethical issues concerning the introduction and testing of a predictive risk model (PRM) designed to assist child and family welfare services in detecting children at risk of failure to thrive as well as possible maltreatment. The PRM is specifically aimed at children regarding whom a notification of concern is sent to the municipal social authorities in Denmark.

Our aim with this paper is to openly and transparently discuss the ethics of the project and to try to address and discuss our ethical values in relation to the project. This implies **a)** presenting our ethical values relevant to the project, **b)** discussing ethical questions that arise, **c)** addressing ethical problems, and **d**) to discuss and weigh ethical dilemmas that arise. It also implies that we welcome any input to the following discussion of ethical values, questions, problems and dilemmas.

The predictive risk model is, at the time of writing, still in the developmental phase, **thus this ethical report should be regarded as a work in progress**, and it may be updated over the course of time it takes to finalize and test the tool, if e.g., the pilot phases reveal additional ethical problems or dilemmas to be addressed and weighed.

In the rest of this introduction, we first provide a brief overview of the project. Then we offer a discussion of the ethical values we want the activities and actions within the project to live up to, and offer perspectives on central ethical questions, problems and dilemmas that may arise as a consequence of the project.

## 1.1 Background: The research project

The research project aims to develop and test PRMs in Danish child and family welfare Services (CFWS henceforth). The project develops PRMs for detecting child maltreatment and children at risk in general using machine learning (ML) techniques. Moreover, the project involves (i) assessing ethical (this report) and legal aspects of employing such models in the CFWS, and to (ii) test their performance in two pilot tests, and finally (iii) to conduct a large scale randomized trial to assess the impacts of introducing such a tool on a number of outcomes ranging from processing time over interventions implemented to child well-being. During the course of the project, in addition to the quantitative assessments, qualitative data on experience using the tool from both social workers as well as children and families will be collected.

For developing the PRM, we used Danish administrative data to develop statistical models predicting the risk of whether a child is at risk of or suffers from maltreatment.[1] The analysis focuses on children for

[1] For more details on the most recent version of the PRM, see Bodilsen et al. (2021).

whom a notification of concern was sent to the CFWS in 2016 or 2017. Using subsequent removal and placement in out-of-home care as a proxy for child maltreatment, we estimated different models varying in their degree of complexity (linear probability, LASSO regularized logistic regression, random forest, and XGBoost models). Our final model has good predictive power as it is able to distinguish between maltreatment and non-maltreatment with a probability around 84% (area under the curve (AUC) ≈ 84%). Model predictions are also predictive of a wide range of other adverse child outcomes indicative of maltreatment (whether or not a child was diagnosed with a somatic symptom disorder, experienced a fracture, had damaged teeth, etc.). Together, these pieces of evidence suggest that the combined use of statistical methods and administrative data may identify children who are at risk of maltreatment with good precision. Furthermore, we showed that the predictions may reduce CFWS errors. Indeed, we found that 60% of the notifications for which nothing occurred during the investigation period, but which subsequently led to an out-of-home placement, belonged to the top two decile of the predicted risk distributions. Finally, we showed that our predictions may help to reduce social worker biases. Indeed, while we found no differences in the manner in which social workers treat children with similar risks of maltreatment but different ethnic background and gender, we found that they tend to treat children from different socioeconomic backgrounds differently.

The legal implications of various aspects of using this tool has been assessed by an external law firm: TrygFonden's Centre for Child Research or UC Viden.

The main conclusions from this assessment are the following:

1. *The residual risk when handling personal data is low to medium with a tendency towards low risk for the registered individuals.*
2. *The identified risks will be significantly reduced by the mitigating precautions already implemented or planned to be implemented, resulting in an acceptable risk for the rights of the registered individuals as well as their rights of freedom when handling their personal data.*
3. *The primary identified risks are de facto decisions in social work practice and the risk of errors in the statistical models.*

In order to test the tool, we have already conducted a pilot test in two municipalities, Hjørring and Silkeborg. This led to considerable revisions of the manner in which the information (the interface as well as the amount of information) was presented to the caseworkers. One of the main goals of these changes concerns transparency of the tool as well as its usefulness. Therefore, a second pilot test phase is planned to take place in late summer 2021.

The tentative forward looking time line for the project looks as follows:

| | 2020 | | 2021 | | 2022 | | 2023 | |
|---|---|---|---|---|---|---|---|---|
| | SPRING | FALL | SPRING | FALL | SPRING | FALL | SPRING | FALL |
| **Legal clarification** | ■ | ■ | | | | | | |
| **New Pilot** | | | ■ | ■ | | | | |
| **Preparing RCT** | | | | ■ | ■ | | | |
| **RCT** | | | | | | ■ | ■ | |
| **Evaluation** | | | | | | | ■ | ■ |

## 1.2 Ethical values, questions, problems and dilemmas in relation to the project

In the discussion of ethical values, questions, problems and dilemmas, we have been greatly inspired by Banks (2012), who discusses ethical values, questions, dilemmas and problems in relation to the specific context of social work, by the more general textbook introduction to ethics by Birkler (2019), as well as by three reviews on the use of predictive risk models in child protective services by Dare (2013), Dare & Gambrill (2016), and Drake & Jonson-Reid (2018).

The PRM offers an assessment of the risk that a child is maltreated, when a notification is received in the municipality. Hence, the model becomes visible in the meeting between social workers and families. Therefore, most ethical issues that arise also pertain to the use of the model during this meeting and in the subsequent handling of the case, including any decisions made.

We believe that central ethical values in social work - and in particular in relation to introducing the PRM into the decision-making process and the meeting between social workers and families - include

- Fairness: It is important that decisions are (and are perceived to be) fair in the sense that they are not biased or discriminatory
- Respect, trust, empathy and understanding: The environment in which the decision-making process takes place should, to the extent possible given the situation, build on respect, trust, empathy and understanding.
- Empowerment/autonomy: The involved families should be empowered to cope with the situation at hand

- Privacy/confidentiality: The affected families should feel that the information used by the social worker is treated confidentially and that their privacy is respected
- Transparency/explicability: The decision making process, and the factors involved in it, should be transparent and explicable to the affected families

The introduction of a PRM into the decision-making process naturally raises ethical questions whether some of these values become compromised as a result of the introduction of the PRM. If this is the case, we have an ethical problem that needs to be addressed. Moreover, it may raise some ethical dilemmas that need to be considered.

Drake et al. (2020) argue that ethical considerations are relative; the ethics of a predictive risk model should be seen in relation to the ethics involved in current best practices; *"PRM cannot be 'ethically sound' or 'ethically troubling' in any particular use case except compared to other best available practices."* (p.164)

**The overarching ethical dilemma** is that PRMs on the one hand offer an opportunity to exploit data and new techniques to improve the safety and wellbeing of children and families, e.g. by exploiting historical/statistical evidence to ensure that the right children receive care, by improving the efficiency and effectiveness of services provided to children and their families, and by providing an empirical foundation for systematic case-based judgement (thus increasing the fairness of the process by reducing any pre-existing caseworker biases). Søbjerg et al. (2020) discuss three reasons for using predictive risk models; a) reducing variability in human judgements, b) increasing accuracy, and c) reducing human bias. Coulthard et al. (2020) argue that big data approaches in relation to child protection decisions are an improvement ethically - including in relation to points raised in this report such as transparency - as existing approaches suffer from equivalent ethical issues. Taylor (2021) hence argues that statistical approaches need to be used alongside complementary human competencies within social work decision making.

On the other hand, many risks are also associated with the use of PRMs, such as the risk of perpetuating systemic biases and discrimination, reducing critical human and relational factors in decision making practices, thus reducing feelings of empowerment and autonomy, and the use of poor-quality outcomes. These risks are vividly demonstrated by Cathy O'Neil in her book "Weapons of Math Destruction'' (O'Neil, 2016). She paints a picture of a dystopic world, where decisions regarding hiring, obtaining loans, payments for insurance, etc. are increasingly made by non-transparent algorithms. The risk comprises the perpetuation and even amplification of existing biases and inequalities, with the risk of further institutionalizing discrimination against certain sub-groups, non-transparencies, lack of privacy, and loss of autonomy/human agency.

Other dilemmas comprise

1. Transparency vs privacy
2. Fairness vs the risk of reducing human and relational factors in the decision making process
3. Transparency vs the non-transparent nature of PRM

The ethical questions and problems we perceive in the current project – and which we try to address in this report – reflect concerns that some of our ethical values are challenged by the introduction of a PRM. These questions and problems comprise the following (some of which are related to the dilemmas listed above):

1. It is obviously an ethical problem if the PRM perpetuates existing discriminatory biases by caseworkers by integrating such biases into the model. This leads to the ethical question of whether we can reduce such biases and improve fairness overall.
2. An ethical problem also arises from PRMs potentially reducing critical human and relational factors in decision making practices, which raises the question of how we can ensure that families are met with respect, trust, empathy, and understanding.
3. If involved families feel that the use of a PRM oversteps some privacy borders, this is another ethical problem that should be addressed.
4. The use of PRM in the decision-making process may alienate families from the social worker and thereby alienation from accepting help or services provided by the social worker.

In our discussion below, we are - in addition to the studies already mentioned, inspired by two central documents; a recent and comprehensive review of ethics when using ML in social work with children (Leslie et al., 2020), and an ethics guideline for using artificial intelligence in general (High-Level Expert Group, 2019[2]). Hence, we have chosen to frame the more specific concerns around these two documents.

[2] We are aware of the Danish National strategy (2019) on AI. This strategy resembles and corresponds in many ways with the aspects and recommendations mentioned by the High-Level Expert Group (2019) which we have applied

In their discussion of the ethics of using ML in social work involving children, Leslie et al. (2020) specifically identify the following three problems:

1. The risk of reinforcing or even amplifying existing systemic bias and discrimination when using PRMs,
2. Deterioration of critical human and relational factors in decision making practices, and
3. Use of poor-quality outcomes owing to deficient data stewardship.

The High-Level Expert Group on AI (artificial intelligence) was established by the European Commission in order to draft ethical guidelines for trustworthy AI. The guidelines present the following seven key requirements that AI systems should meet in order to be trustworthy:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination, and fairness
6. Societal and environmental well-being
7. Accountability

The key requirements derive from fundamental rights stated in the EU treaties and the EU charter, which the High-Level Expert Group in the context of AI systems has narrowed down to principles or values of respect for human autonomy, prevention of harm, fairness, and explicability.

It is obviously a pre-condition that the tool meets the legal requirements, which the external law firm will stipulate. These have been investigated (se xx link/cite the legal document). Hence, in this note, we will not be addressing issues of the legality of using the predictive risk model but rather focus on the ethical questions and problems concerning the practical implementation of the tool.  In the legal document, robustness (High-Level Expert Group, 2019) of the tool is also addressed in relation to unintentional harm, reliability and prevention in relation to both technical and societal perspectives[3].

and discussed in this paper. In addition, the Danish National Strategy recommends AI as an adjunct to human decision-making as a way of exploring possibilities without compromising central values.

[3] Robustness is described as follows:  *"Even if an ethical purpose is ensured, individuals and society must also be confident that AI systems will not cause any unintentional harm. Such systems should perform in a safe, secure and reliable manner, and safeguards should be foreseen to prevent any unintended adverse impacts. It is therefore important to ensure that AI systems are robust. This is needed both from a technical*

### 1.3 Overview

In the following, we will first discuss artificial intelligence in relation to the tool developed in this project, that is, the ethical aspects of using ML models in social work. We will begin by discussing the extent to which the tool and the incorporated predictive risk model is an artificial intelligence system. Section 3 discusses the ethics of ML models in social work. In Section 4, we address the guidelines of the High-Level Expert Group mentioned above. Section 5 entails our conclusions.

## 2 Is the statistical tool an artificial intelligence system?

We will begin by discussing whether we consider our PRM to be an artificial intelligence system. There are two different perspectives on artificial intelligence. One perspective has to do with to rationality. Another points to AI as being a partly autonomous system.  In order to be characterized as AI system, the tool must be rational. According to High-Level Expert Group on AI (2018: 01):

> *"This refers to the ability to choose the best action to take in order to achieve a certain goal, given certain criteria to be optimized and the available resources. Of course, rationality is not the only ingredient in the concept of intelligence, but it is a significant part of it."*

On the one hand, seeing as the tool developed in our project does not point in any direction regarding actions to be taken by the social worker, it could be argued that it cannot be characterized as an AI system. On the other hand, the goal of the tool is to aid decision-making in social work and thereby support decisions on actions in social work. In this sense, it could be characterized as an at least partial AI system.

The second perspective regarding artificial intelligence refers to AI as a system with degrees of autonomy and ability to act:

> *"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-*

*perspective (ensuring the system's technical robustness as appropriate in a given context, such as the application domain or life cycle phase), and from a social perspective (in due consideration of the context and environment in which the system operates)."* (Ethics Guidelines For Trustworthy AI, p. 7)

> *based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."* (High-Level Expert Group on AI, 2018: 01).

The tool developed in this project does not have the capability to take actions or act with degrees of autonomy to achieve specific goals, nor is it intended to do so. Therefore, it cannot be characterized as AI in this sense. The tool is an algorithm developed *ex ante*, and predictions do not change over time, except in the case where we choose to update the model in light of additional information.

Overall, this leaves us with a statistical tool that does not meet the criteria for artificial intelligence. However, considering that rationality and supporting decisions on actions and thereby underpinning best course of action, it could be argued that elements of what could be characterized as an artificial intelligence systems are at play.

In conclusion, while the current project is not, in our opinion, a clear-cut example of AI, it has elements that might resemble artificial intelligence. Moreover, ethical questions arise irrespective of whether we consider the PRM to be artificial intelligence or not. In this paper, we will therefore address and discuss ethical questions, problems and dilemmas with inspiration from The High-Level Expert Group on AI (2018; 2019) (Definition of AI & Ethics Guidelines For Trustworthy AI ) and by the ethics review of ML in Children's Social Care (Leslie et al., 2020).[4]

## 3 Machine learning and ethics in social work

Leslie et al. (2020), in their ethics review of ML in children's social care, discuss the ethical considerations involved, when ML algorithms are used in social work involving children and families at risk. They open the review by pointing to the central ethical dilemma faced also by our project:

> *"On the one hand, it would seem a crucial moral imperative to use the growing array of applied data scientific techniques to foster the safety, wellbeing, and flourishing of children in need and their families and to bolster possibilities for optimal outcomes in family life."* (Leslie et al., 2020:08)

[4] Lindsay et al. (2020) offer an alternative distinction between 'explainable' AI, i.e. where the statistical methods (rationality) are or can be known to a human, and 'black box' AI, where the algorithms cannot be known to a human.

They explicitly mention the possibility of promoting evidence based insights to ensure the right children receive the right amount of care, improving efficiency and effectiveness of services provided to children and their families by public authorities, and provide an empirical foundation for case-based judgement. They also mention risks:

> *"On the other hand, the many risks associated with the use of predictive analytics in the provision of children's social care raise serious questions about where to draw boundaries in the utilisation of individual-targeting ML technologies in CSC."* (Leslie et al., 2020:09)

In particular, they mention the following three risks: (i) the risk of reinforcing or even amplifying existing systemic bias and discrimination when using PRMs, (ii) deterioration of critical human and relational factors in decision making practices, and (iii) the use of poor-quality outcomes owing to deficient data stewardship.

In the following, we will discuss each of these risks/problems in relation to the current project. To motivate this discussion, we begin by summarizing how the PRM may actually foster better quality decisions. If it does not, there is no ethical dilemma, because then we should not even consider using the model.

This is described in more detail by Bodilsen et al. (2021). They identify notifications for which the initial classification made by CPS appeared to have been erroneous and estimate the extent to which our PRM would have flagged them as being of concern. To approximate this error rate, they identify notifications for which no intervention and/or out-of-home placement occurred in the 4-month period after the notification was received. This period represents the maximum length of time during which CPS needs to conduct its investigation and implement adequate measures. Next, they estimate the share of these notifications for which either an intervention or an out-of-home placement occurred 4 to 8 months or 8 to 12 months after the initial notification was received.

This analysis suggests that there is room for improvement in social workers' decision-making process; overall, approximately 15% of the notifications identified as requiring no form of intervention were later reclassified as problematic. Focusing on cases that were identified as problematic (those which were reclassified after the end of the 4-month period), they show that a very significant share of these cases is estimated by the PRM to lie in the very top of the risk distribution. Strikingly, more than 60% of the notifications for which an out-of-home placement subsequently occurred receive a score of either nine or ten. This was also the case for approximately 25% of the notifications for which an intervention was subsequently implemented.

These results suggest that CFWS' identification of severe risk or maltreatment cases (or its speed) may indeed be improved and can potentially be improved by incorporating PRMs in the decision making process.

## 3.1 Risk of reinforcing or even amplifying existing systemic bias and discrimination

Since supervised ML models make accurate predictions based on past social and cultural patterns, whether or not these are discriminatory, there is a strong risk of perpetuating such systemic biases, and to the extent that it affects social workers' decision making in discriminatory direction, it may even amplify such biases. This causes an ethical problem, which has also been pointed out by our advisory board. It challenges one of our ethical values; fairness. Hence, we discuss the extent of the problem and ways in which we have tried to address it in the following.

First of all, we have analyzed the extent to which the PRM we have developed (see Bodilsen et al., 2021) reduces or enlarges any existing biases. In particular, we have looked at gender, ethnicity, and socioeconomic status (SES). Neither of these variables are included in the information set on which the algorithm can feed. Hence, it is the case that two notifications regarding children of different ethnicity, but where all other included factors are identical, will be given the exact same score by the tool (this is tested and verified). Bodilsen et al. (2021) analyze model predictions on the prediction sample split by ethnicity, SES, and gender, and find weak tendencies that caseworkers in the past may have discriminated slightly on ethnicity. However, we do find evidence of differential decision-making based on SES in the sense that children in high-SES families are much less likely to be removed (see Bodilsen et al., 2021). The PRM, however, appears to remove or at least reduce these biases, since it does not include any ethnicity- and SES-related information, and therefore assigns similar risks to otherwise identical notifications regarding children of different ethnicities and SES background.[5] We find no evidence of past discriminatory behavior by gender. Hence, the PRM does not discriminate by gender.

It is an integral part of the project to continuously monitor the predictions for any systematic biases and adjust the PRM accordingly; if/whenever such biases are detected or suspected. In addition, the project

[5] Of course, it might be that high-SES families have more resources (social and physical capital, such as good relationships with a supportive wider family and social network) on which to draw and make continuing care at home feasibly 'safe enough', even though these factors are not measured by the model. Thus the *correlation* with the 'discriminatory' factor may be a fact, but may not indicate *causation*. Hence, it may not reflect a bias as much as mediating factors, and a deeper knowledge of the wider context from the perspective of the professionals involved face-to-face in the situation (see also Søbjerg et al., 2018, for a discussion of this).

intends to introduce and instruct social workers carefully in the potentials and pitfalls in using the PRM in social work in the hope that we may guide them to using the PRM such that it promotes non-discriminatory behavior.

We thus perceive that the introduction of an appropriately designed PRM may actually reduce systemic biases and existing discriminatory tendencies by caseworkers.

## 3.2 Deterioration of critical human and relational factors in decision making practices

Obviously, the importance of interpersonal communication in meeting families respectfully and building trust, empowerment, empathy and understanding in the decision making process regarding children and families potentially affected by notifications, makes the human caseworker indispensable in social work. Using a PRM risk assessment as input to a decision-making process comes with a risk of de-personalizing the entire process, which poses an ethical problem. This may happen if caseworkers use the PRM as an "authoritative judge" to decision-making and use the model as a shield against the families ("the model says your child is a high risk of maltreatment, so I have no choice but to do such and such" – type of argumentation). This is discussed by Taylor (2021). It may also happen if the model is used as a mere guide to social workers (as intended) without its potentials and limitations being well understood by the social worker. A possible de-personalization or even alienation between social worker and family might also affect trust and understanding in such a manner that it becomes a challenge for the family to accept and collaborate on the possible help provided by the social worker.

To address these concerns, we will train social workers in the proper interpretation and responsible use and limitations of the risks predicted by the PRM, and we will point to ways in which the model predictions may be communicated to concerned families so as to preserve their dignity and feelings of understanding the rationale behind the decision made by the caseworker (Leslie et al., 2020). This would involve stating norms and ensuring mechanisms of implementation that preserve all the indispensable gains from human interaction and ensure that the PRM only adds integrity and evidence-basing to the decision making process.

Our project involves carefully training the social workers in understanding the use and in particular the limitations of the PRM tool. Moreover, the information available to the social worker includes, besides the risk assessment, a list of all information used in the risk assessment along with indications of which categories of information has increased and/or reduced the risk assessment. This should enable the social worker to enter a dialogue with the family and justify any decision, we believe, with a more reliable foundation than a subjective individual assessment alone would allow for. In addition, it provides transparency in the foundation of a decision because the categories of information are clear. Hence, we intend the tool to improve upon trust and empowerment issues, so that that de-personalization or alienation will not become an issue.

It is, nevertheless, an ethical problem which we acknowledge, and we will conduct qualitative research during the research project (e.g., observational studies and interviews with social workers as well as involved families) to investigate the extent to which this is perceived to be a problem. To the extent that it is, we will make efforts to address it in collaboration with the CFWS' and social workers. These assessments will be made during the second pilot, and any modifications of the model and the way it is presented to social workers and families will be made after the pilot and before the planned randomized trial. We will repeat the observational research during the planned randomized trial.

## 3.3 Use of poor-quality outcomes owing to deficient data stewardship

ML models are only as good as the data on which they are trained. Hence, data quality is a serious concern. At least two issues arise. First, the precision of the data on which the ML model feeds is important. Second, the outcome on which the model is trained should represent a good measure of what it is supposed to capture. The ethical value addressed here is trustworthiness.

We believe that the first issue is largely solved by relying on register-based data only. In the Danish context, such data is reputed for being accurate. Moreover, as mentioned already, all the data on which a specific risk assessment by the tool has been based is available to the social worker in an easily accessible manner, and it can therefore be verified by the social worker as well as the involved family.

The second issue is more problematic. We aim to detect children at risk and child maltreatment, but there is no readily available measure of 'maltreatment' in the administrative registers. Hence, we have used 'subsequent placement outside the home' as a proxy for (severe) maltreatment. This is definitely a debatable choice. We have also experimented with using information on less severe interventions (interventions applied in the family) and on future (severe) notifications. We have validated our preferred measure against alternative outcomes that we also believe are indicative of maltreatment, such as hospitalizations due to fractures, mental health diagnoses, mandatory well-being in school, dental quality measures, etc.

The predicted risks on our prediction sample are highly correlated with all these outcomes, also for children who are not subsequently placed outside the home (see Bodilsen et al., 2021). We take these results as a strong indication that our risk measure is at least closely related to maltreatment. Nevertheless, we acknowledge this risk, and we will continue to analyze its validity, and, if possible, improve upon it, as the project evolves.

# 4 Requirements from High-Level Expert Group

In this section, we examine the requirements from the High-Level Expert Group in relation to the proposed tool.  A treatment of each key requirement will be given in separate subsections.

## 4.1 Human agency and oversight

Fundamental rights are a central part of the requirements concerning human agency and oversight. In relation to human agency (High-Level Expert Group, 2019), it is highlighted that AI systems should support decision-making of the user and in no way deceit or manipulate users. In addition, users of AI systems should be able to make informed autonomous decisions regarding the system and they should be given the knowledge and tools to understand and challenge the system. Leslie et al. (2020) raise some of the same issues concerning ethics in ML.  Aspects such as cognitive biases that can influence how users interact with ML models in relations to how results are interpreted and a potential overuse, underuse, misuse or overreliance both individually and as part of organizational contexts. In short, our ML systems must not undermine human autonomy and attention should be paid to the use of the model in social work.

The PRM we have developed is a tool designed to assist social workers in their decision-making process. It does not in any way try to deceive them into making certain decisions. It provides a risk score, which essentially summarizes the information already available to the social worker in an index; an estimated risk that the child is maltreated represented by an integer score ranging from one to ten. In this way, it offers a replicable and potentially new perspective on the notification at hand. It does not in any way suggest actions. We plan to train social workers in the correct way of interpreting the risk score, what souces of information it builds upon. In addition, the interface will also show the information on which it builds, and how this information contributes to increasing or lowering of the risk score. This should give the needed transparency to the social worker and involved family, so that they may also challenge the score as well as the information used. The interface has been developed and tested in close collarobration with social workers in two pilot tests (the second is not conducted yet) and we will also monitor its usefullness in the randomized trial that is an integral element of this research project.

## 4.2 Technical robustness and safety

In relations to technical robustness and safety, again, we will only address ethical aspects. This means that aspects of data security or other GDPR related issues will not be addressed in this paper. As these are legal matters, they are addressed in the legal analysis (TrygFonden's Centre for Child Research or UC Viden). In this section, we will address accuracy of the tool, reliability and reproducibility. All of these are aspects of the scientific process in regards to quality such as reliability. However, it connects to ethical values because it is a matter of trustworthiness of the model.

### 4.2.1 Accuracy

The purpose of the tool is to provide social workers with a risk assessment of children who are at risk of maltreatment. A successful tool can enhance the decision making process of the social workers, such that the most appropriate preventive measure is being implemented more often. It is important to stress that the tool cannot replace social workers decision-making process, but is a decision support tool and offers a standardized manner of how information is processed at the CFWS.

To be useful in practice, as well as portraying a high level of trustworthiness, the tool has to deliver accurate predictions. In Table 1, we report a summary of the predictive performance of the four different supervised ML models, which have been considered in the developmental phase. The models listed in Table 1 are of increasing complexity, with the linear probability model as the simplest model and the XGBoost model (Chen and Guestrin, 2016) as the most complex model. As mentioned, the models aim to identify children suffering from maltreatment. Maltreatment, however, is not a directly observable variable. To overcome this problem, we follow the same approach as used to develop a similar tool in Allegheny County, Pennsylvania, and use out-of-home placements as a proxy for maltreatment. Despite this outcome having some limitations, we find evidence that its external validity suggests that it can be used to distinguish between children at low risk and at high risk of maltreatment.

In the ML literature, a wide range of accuracy metrics exists that can be used to evaluate the performance of a ML model. We have chosen to use the area under the receiver operating characteristics curve (AUC) as the evaluation metric. The AUC is a measure that ranges from zero to one, where a value of one indicates that the model can perfectly predict which children will be placed in the future. More generally, the AUC score can be regarded as the model’s ability to discriminate between positive instances (i.e., children experiencing out-of-home placement) and negative instances (i.e., children not experiencing out-of-home placement).  It follows from Table 1, that the AUC scores ranges from 82.60% to 84.26% depending on the choice of ML model.  According to Swets (1988) AUC scores above 80% is an indication of good predictive performance. The XGBoost model performs best and is therefore the ML algorithm that the PRM will be based on in the second pilot.

| METHOD | AUC (%) | 95% CONFIDENCE INTERVAL |
|---|---|---|
| | Outcome: Out-of-home placement within 365 days | |
| LINEAR PROBABILITY MODEL | 82.53 | 81.86 - 83.21 |
| LOGISTIC LASSO | 82.26 | 81.59 – 82.93 |
| RANDOM FOREST | 83.36 | 82.72 – 84.01 |
| XGBOOST | 84.08 | 83.45 - 84.71 |

*Table 1: This table summarizes the predictive performance of the considered ML models. The models are trained on a data set comprised of 120,395 notifications (representing 63,303 unique children), and evaluated on a sample of 52,649 notifications (representing 27,341 unique children) over the period from April 2016 to December 2017.*

As already mentioned, Bodilsen et al. (2021) also explore the relationship between the XGBoost model predictions and other adverse outcomes associated with maltreatment, such as mental illness, charges, and illegal school absence. Figure 2 in Bodilsen et al. (2021) shows that children identified to be at high risk as defined by the model predicting out-of-home placements, are also worse off when considering these alternative proxies for child maltreatment. Thus, children identified to be at high risk by the tool, are arguably the children that are most in need of help, whereas the children identified to be at low risk by the model are not different from the population of non-notified children in DK in terms of their prevalence to experience adverse outcomes.

We therefore conclude that the PRM has sufficient accuracy, and thereby trustworthiness, that we may test it in pilots and a randomized trial. Note, however, that accuracy will be further analyzed in the randomized trial and the model will not be recommended for social work practice without further very detailed analysis of this issue based on data obtained from the trial.

### 4.2.2 Reliability and reproducibility

In the previous section, we argued that the tool has good accuracy based on a sample of historical notifications and we find evidence that the tool is predictive of other adverse outcomes than the one it is designed to predict. To be useful and trustworthy in practice it is, however, a necessity that the tool works properly for any type of ongoing notification and in cases of missing input data. Furthermore, a careful evaluation of the tool requires that it is possible to make exact reconstructions of every prediction made by the tool.

With respect to the latter reproducibility requirement, this is automatically fulfilled due to the nature of the tool and the data it feeds on. Once the algorithm has been trained using a sample of historical data, the tool is deterministic in the sense that it will provide the *exact* same risk assessment for any two individuals with the exact same background characteristics. Thus, the tool does not learn continuously over the course of time. In order for the tool to provide a different risk assessment given a certain value for input of the tool, it requires that we go back to our developmental space on Statistic Denmark's server and recalibrate the model. Hence, by keeping track on which version of the tool a given risk assessment have been based upon, it will always be possible to reconstruct the assessment afterwards.

The model reliability criteria is also automatically taken into account by the choice of the ML method that the tool is based on. Our preferred model, the XGBoost model, allows for missing data among the input variables. This means that even in cases of incomplete information about a child in the databases of the municipalities, the tool will still provide a meaningful risk score based on the non-missing data inputs. This is a very attractive feature of the XGBoost methodology and a feature not easily accommodated by more standard methods such as the linear probability model and the logistic regression model also considered during the developmental phase.

In theory, the tool should be applicable for any type of notification. This includes also notifications for which it is the first time that a child or its family is involved in a notification and notifications regarding newborns. However, it could potentially be desirable to set an upper threshold for the amount of missing data that are allowed in order to use the tool. With limited information, it is plausible to believe that the tool can be imprecise or even misleading, and it might be unethical to attach the prediction made by the tool to the child under such circumstances.

We will build in safeguards against this situation, either by flagging predictions based upon limited information or by not calculating the prediction for such cases. Again, this should also enhance trustworthiness of the model in social work practice and for children and families at risk.

## 4.3 Privacy and data governance

An AI system must guarantee privacy and data protection throughout the system's entire lifecycle. In our case, the information upon which the tool is based is already used in the municipalities, which adhere to the GDPR rules and regulations. The tool uses this information to calculate a risk assessment, which will then be saved together with the case. The legality control (TrygFonden's Centre for Child Research or UC Viden) deals with this issue in more detail, and the judgement is that the tool is in accordance with the legal requirements.

The quality of the training and validation data sets at Statistics Denmark is considered very high. Once a risk assessment is calculated for a specific notification, a data sheet specifying the information that the assessment is based upon is available to the social worker, who can then confirm it with the involved family.

Moreover, only those persons handling the case would have access to the risk assessment. It is part of the user interface that the information is only available to those who have access to the case (the risk is saved in 'the case files').

We thus believe that data quality is high and verifiable and that data security is high as well; thereby assuring adequate privacy for children and families at risk. For more details on these issues, we refer to the legality assessment (TrygFonden's Centre for Child Research or UC Viden).

## 4.4 Transparency

As the use of ML methods in social work is still in its infancy, it is important that all processes of the present project have a large degree of transparency in order to increase the trustworthiness of the project as a whole and the PRM in particular. It is important that the data sources and the algorithm behind the tool are documented extensively, and that social workers are appropriately trained to understand what the tool can and cannot do.

As mentioned earlier, we have considered different supervised ML models of varying complexity in the developmental phase. The models we have considered are the linear probability model, a logistic regression model combined with least absolute shrinkage and selection operator (LASSO), the random forest model, and finally, the XGBoost model.[6] The two former models belong to the class of generalized linear models, which have been around for decades and are, among other things, popular model choices due to their simplicity and interpretability. The latter two methods are ensemble methods, where the output of several classification trees are aggregated in order to produce a model output. Trees as such are highly interpretable, however, this is not the case when multiple trees are combined. Hence, the predictions based on the random forest and the XGBoost models are not as easy to explain as the two other considered algorithms. On the other hand, these methods are typically found to deliver highly accurate predictions in many different contexts. A typical explanation for this phenomenon is the fact that these models are good at capturing non-linearities and interaction effects in high-dimensional datasets.

As it appears from Table 1, we also find that the more complex models deliver more accurate predictions in the present project. However, for these more complex models it is not as straightforward as for the simpler models to figure out how a given input factor contributed to the prediction made by the model. A non-transparent model might have the implication that social workers will be more prone to devalue the tool's risk assessments, if there is no accompanying information about which kind of information the model feed on and of these inputs relative importance for a given prediction. Moreover, a black-box risk assessment has the risk of reducing trust and feelings of empowerment for the social workers and the involved families.

In order to overcome this issue, we provide the social workers with additional information about the model in addition to a numerical risk score. Each time the tool is used to generate a risk assessment, it will be followed by an information sheet where all the model inputs are listed. This means that the social workers will have exact knowledge about which type of information the model relies upon, and thereby make it easier to infer the degree of additional information they might have about a specific case in comparison to the tool. This information sheet also highlights that the model does not contain any information, which the social workers do not already have access to in their own databases. The input variables in the model have been carefully selected in collaboration with the practitioners from the municipalities involved in the first pilot study, to make sure that the tool is based on information already available in their own systems.

Moreover, we will provide so-called *SHAP values* (Lundberg and Lee, 2017) each time the tool is used. SHAP (Shapley additive explanations) is a method that can explain the prediction of *any* ML model. By

[6] For a detailed description of the four models and the exact implementation details, we refer to Bodilsen et al. (2021).

using the SHAP methodology, it is possible to construct a list of model inputs ranked by their importance in generating a given risk assessment. We will use the SHAP values to provide the social workers with information about the ten model inputs that contributed most to a given prediction. By using the SHAP framework, it will be possible to investigate whether a given factor has an increasing or decreasing effect on the risk assessment provided by the tool. Our hope is that this addition to the tool will make it more visible for the user how the algorithm works and to challenge the predictions of the PRM.

Hence, we argue that, despite the 'black-box' nature of models such as XGBoost, it is still possible to make the predictions made by the model transparent by combining the method with the SHAP methodology and by explicitly making available to the social workers (and families) which type of information goes into the model, and which type of information has contributed to increasing and/or decreasing the risk score. All in all, this should increase both usefulness but also trustworthiness of the model. We will continuously investigate how the SHAP values are being used and interpreted by the social workers, and to which extent they find this new feature to be valuable. This is important since explainable ML methods, such as SHAP, still are in their infancy, and thus not much evidence exists on how these methods work in practice.

## 4.5 Diversity, non-discrimination, and fairness

In order for a statistical tool to be applicable in social work in practice, it is of utmost importance that the tool is fair and does not induce discriminatory behavior. We have already addressed this issue partly in Section 3.1. Here, we will elaborate slightly on these issues.

In order to avoid the discriminatory assessments that might arise from basing the risk scores on background characteristics such as race and ethnicity, we have decided that the PRM does not incorporate any information about the origin of the children and their parents.[7] By leaving out information on ethnicity amongst the set of maltreatment predictors, we can be certain that the differences in risk scores among a group of children cannot be attributed to differences in the origin of the children and their families. Our preliminary results furthermore suggest that leaving out information on ethnicity does not change the predictive performance of the model, nor does it appear that social workers in the past have discriminated against certain ethnic groups in terms of decisions made. The probability for the model to distinguish between cases of maltreatment to cases of non-maltreatment is statistically the same irrespectively of

[7] Relatedly, the model does not contain information about the number of emigrations and immigrations that the child has experienced during life, since this information might be highly correlated with ethnicity.

whether or not ethnicity-related information is a part of the model, which is reassuring from an ethical point of view.

Bodilsen et al. (2021) investigate the predictive performance of the PRM and further analyze whether or not the tool is discriminatory against disadvantaged population groups. Bodilsen et al. (2021) investigate whether children of non-western origin have higher prevalence of out-of-home placement once the predicted risk as generated by the preferred ML model is taken into account. Once the generated risk score is taken into account, there are no noticeable difference in the placement rates depending on the ethnicity of the children. It is only differences in the risk scores that drive the variations in placement rates, which is exactly the condition that a well-calibrated risk model has to satisfy, according to the fairness criterion put forward by Chouldechova et al. (2018).

In the same vein, Bodilsen et al. (2021) also consider the same for children of low- and high SES. In contrast to the numbers based on ethnicity, this figure shows that children from low SES families, who are assigned a risk score of ten by the tool, historically are much more likely to be removed from their home, compared to children from high SES families also receiving a risk score of ten. This suggests that the CFWS assess notification concerning low SES children differently than those concerning high SES children. The much higher historical placement rates for high-risk low SES children compared to high-risk high SES children could indicate discriminatory behavior along the socioeconomic scale. In such a situation, our tool has the potential to reduce such biases, since for any two individuals receiving the same risk score, there should, on average, not be any substantial difference in the likelihood to suffer from maltreatment. Thus, differences as the ones that are found along the socioeconomic dimension could hopefully be eliminated by appropriate use of data and PRMs.

## 4.6 Societal and environmental well-being

The project team does not foresee that the implementation of the tool will have negative environmental consequences.

Nor do we foresee any negative consequences pertaining to the general social relationship and attachment of neither the social worker nor the child and her family.

On the contrary, our hope is that the use of the PRM may improve the decisions made and thereby the well-being of children and involved families. Of course, this is only a working hypothesis for now, which is essentially the raison d’être for this entire research project; we want to investigate if a carefully designed

PRM, living up to all ethical and legal requirements as well as high statistical quality requirements, can improve upon decisions made, when made available to social workers. If this is not the case, there is no case for recommending use of the tool more widely. In addition, we want to explore whether the tool in any way negatively affects trust or a trusting relationship between social worker and family. If that is the case, again, there is no case for recommending use of the tool more widely

## 4.7 Accountability

Requirements regarding accountability covers different aspects. The first is auditability; the enablement of assessment of algorithms, data and design processes. We hope to achieve this by having a policy of publishing all documents produced in the project, by making available all details regarding the development and training of the PRM and the user interface. Finally, we aim to publish in scientific journals and hence have the designs assessed by fellow expert researchers in the field.

The next issue deals with minimization and reporting of negative impacts:

> "…the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, reporting and minimizing the potential negative impacts of AI systems is especially crucial for those (in)directly affected. Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI-based system." High-Level Expert Group (2019; 20).

We will discuss this issue with the involved municipal CFWS' to ensure the full openness. As this is a research project, we do not foresee the need for protection of whistleblowers, as all information pertaining to the use of the tool will be taken into account in the evaluation, and the tool is not at a stage of development where we intend to suggest it is rolled out to municipalities, nor may it ever be. This depends entirely on the outcome of the evaluation. If a whistleblower protection scheme is deemed necessary, we will discuss with our stakeholder board and the involved municipalities how to construct such a scheme.

Trade-offs: When a ML tool such as the one used in the current project is implemented, some trade-offs will inevitably occur. Whenever this is the case, we will closely discuss how to proceed and any ethical concerns will be given due considerations. If no ethically viable solution is available, the PRM will be altered in order to ensure an ethically sound employment of the tool.

Redress: When unjust adverse impact occur, there should be mechanisms in place to ensure adequate redress. In the current project, we are covered by the legal right of any citizen affected by a decision to complain (Chapter 11, Law of Social Service, DK).

### 4.7.1 Stakeholder involvement

Stakeholder involvement is an important part of developing a trustworthy AI system both during development and testing of the system as well as implementation in organizations (High-Level Expert Group, 2019).

Relevant stakeholders have been – and are still – contributing with valuable inputs during the process of developing the tool. Early on, a partnership was formed with two municipalities (Silkeborg and Hjørring) who assisted the project group in determining the needs for social workers employed in CFWS. This was to ensure that the tool was relevant to social workers in the municipalities.

In addition, a team of relevant participants have been invited into a stakeholder board during development, testing and implementation of the tool. The purpose of this board is to follow the project and offer different as well as critical perspectives on development, testing and the implications the tool. Perspectives on both usefulness and ethics are provided in relation to social work practice and children and families at risk. The board contains members from a variety of stakeholders who all are active in social work. The members of this board includes The National Board of Social Services, two NGOs focusing on the interest of both parents (ForældreLANDSforeningen) and children (Børns Vilkår), Local Government Denmark (KL), practitioners from the participating municipalities as well as the Danish Social Worker Association (DS). The board meets minimum once a year. All inputs gathered from these different stakeholders are discussed and applied in development of the tool.

## 5 Conclusion

In this document, we have discussed ethical questions, problems and dilemmas that arise in the development and trial of a PRM in social work with children and families at risk.

We have also presented and discussed steps we have taken to ensure that risks of unethical use of the tool and unethical consequences of the tool are as low as possible. We acknowledge that there are potentially ethical issues that we are not aware of at the present stage of this research project, and therefore this document is to be considered ‘ongoing work’. Any insights and comments to any problems we have not yet addressed are gratefully acknowledged. Our primary aim is to develop a tool that can be used and is perceived as useful and trustworthy by both social workers and families and that may actually improve decision-making in social work with children and families at risk.

## References

Banks, S. 2021. Ethics and Values in Social Work (5th ed). Macmillan International.

Birkler, J., 2019. Etik – en grundbog. Forlaget Munksgaard.

Bodilsen, S., Michel, B., Nielsen, S. A., Rosholm, M., 2021. Can Machine Learning Help Child Protective Services Predict Maltreatment? Manuscript in production.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 785–794.

Chouldechova, A., Benavides-Prado, D., Fialko, O., Vaithianathan, R., 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: Conference on Fairness, Accountability and Transparency. 134–148.

Coulthard, B., Mallett, J., Taylor, B.J., 2020. Better decisions for children with 'big data': can algorithms promote fairness, transparency and parental engagement? Societies, 10, 97. Available here.

Dare, T., 2013. The Dare Report: Predictive Risk Modelling and Child Maltreatment: An Ethical Review, Ministry of Social Development, Wellington, New Zealand. Available here..

Dare, T., Gambrill, E., 2016. Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County. The Allegheny County Department of Human Services. Available here.

Drake, B., Jonson-Reid, M., 2018. Administrative Data and Predictive Risk Modeling in Public Child Welfare: Ethical Issues Relating to California. Available here.

Drake, B., Jonson-Reid, M., Ocampo, M.G., Morrison ,M., Dvalishvili, D., 2020. A Practical Framework for Considering the Use of Predictive Risk Modeling in Child Welfare. The ANNALS of the American Academy of Political and Social Science, 692(1):162-181. Available here.

Finans- og Erhvervsministeriet (2019): National strategi for kunstig intelligens [National Strategy on Artificial Intelligence]. Read the National Strategy on AI here

Fluke, J.D., López, M. L., Benbenishty, R., Knorth, E.J., Baumann, J.D., 2021. Decision-Making and Judgment in Child Welfare and Protection. Oxford University Press.

High-Level Expert Group, 2019. Ethics Guidelines for trustworthy AI. European Commission.

High-Level Expert Group on Artificial Intelligence, 2018. A definition of AI: main capabilities and scientific disciplines. (Francesca Rossi, Member of the High-Level Expert Group acted as author. Brussels, 2018)

Leslie, D., Holmes, L., Hitrova, C Ott, E., 2020. Ethics review of machine learning in Children's social care. What Works for Children's Social Care (WWCSC). Alan Touring Institute. Oxford University, Rees Centre, Department of Education, UK.

Lindsay, L., Coleman, S., Kerr, D., Taylor, B.J., Moorhead, A., 2020. Explainable artificial intelligence for falls prediction. In Singh M, Gupta PK, Tyagi V, Flusser J, Ören T & Valentino G (Eds) Advances in Computing and Data Sciences. (Ch 8; pp 76-84) Singapore: Springer Nature.

Lundberg, S.M., and S.-I. Lee., 2017 A unified approach to interpreting model predictions. Advances in neural information processing systems..

O'Neil, C., 2016. Weapons of Math Destruction – how big data increases inequality and threatens democracy. Penguin Random House UK.

Swets, J. A., 1988. Measuring the accuracy of diagnostic systems. Science 240 (4857), 1285–1293.

Søbjerg, L.M., Taylor, B.J., Przeperski, J., Horvat, S., Nouman, H., Harvey, D., 2020. Using risk-factor statistics in decision making: prospects and challenges. European Journal of Social Work. Available here.

Taylor, B.D., 2021. Teaching and Learning Decision-Making in Child Welfare and Protection Social Work. In Fluke et al. (eds.), 2021.

Webpages:
TrygFonden's Centre for Child Research
UC Viden