Katarina Fast Lappalainen

Protecting Children from Maltreatment with the Help of Artificial Intelligence: A Promise or a Threat to Children's Rights?

1 Introduction

In late 2020, a scandal was revealed that sent shockwaves throughout Sweden. A man, 40 years old, was reportedly found in a filthy apartment at the outskirts of Stockholm.[1] Bruised all over his body and in need of immediate care, he was malnourished, had no teeth and had difficulties expressing himself verbally. It turned out that he had not attended school since the age of 12 and had ever since been kept in isolation by his mother. Nevertheless, during the remainder of his childhood the boy remained enrolled at his school and continued to appear on the class lists and received incomplete grades, even though he was not attending school. His older sister, no longer living with the family, had notified the social services of the possible maltreatment of her brother. Despite

[1] Special thanks to the editors Mattias Dahlberg and Katja De Vries for invaluable comments and suggestions. I am also grateful to Paul Lappalainen, Swedish and American lawyer for important comments and language review. Finally, I wish to thank Sara Skoglund, social worker, systemic family therapist and court expert, U.K. and Pia Molander Wistam, Head of the Department of Children and Youth, Norrtälje Municipality, Sweden for valuable information and materials.

this, neither the school, nor the authorities acted upon the information available to them.[2]

Could this and other scandals where the community fails to protect children from maltreatment, have been prevented with the help of predictive tools using artificial intelligence (AI) for effective risk management?

Predictive tools for child protection based on AI have, with varying success, been developed in different parts of the world. Some examples are the *Vulnerable Children Predictive Risk Model* of New Zealand from 2012, the *Allegheny Family Screening Tool* used by Allegheny County in Pennsylvania in the U.S. since 2014 and the *Early Help Profiling System*, developed by Hackney County Council in London together with Xantura, a private company. In Scandinavia the *Gladsaxe-model* from Copenhagen, Denmark, seems to have been the first in the region as it was ready in 2018. In Sweden, the municipality of Norrtälje launched an AI tool to analyse cases based on preliminary warning referrals in 2020, to help in the detection of future cases of child maltreatment.[3]

It could be argued that such tools in general will help prevent maltreatment of children and enable social services to become more effective in their outreach work and thus the provision of support to children at high risk at a lower cost. The struggle to provide more effective child welfare and the reality of substantial funding cuts, common to the authorities in many European countries, increases the interest in such systems.

Contrary to the promise and hopes for AI tools is the fact that the use of such tools for child protection, comes with multiple risks from a children's rights perspective. This is certainly the case regarding the use of predictive risk modelling (PRM) in child welfare.

PRM can be developed using different techniques which can have somewhat different outcomes from a legal and ethical perspective. These techniques are described further in section 3. Models generally rely on a

[2] The case was reported in several media outlets, e.g., Sveriges Radio, https://sverigesradio.se/artikel/7613921 and Expressen 2020-12-01, *Släktingen hittade instängde mannen: det luktade ruttet* by Erik Wiman. In the end the mother was released as the prosecutor reportedly did not find evidence of any crime, since the son was not physically restrained from leaving the home. See e.g. *Åklagaren om instängde sonen: inga bevis för brott* by Niklas Eriksson, https://www.aftonbladet.se/nyheter/a/x3J3ln/aklagaren-om-instangde-sonen-inga-bevis-for-brott.

[3] Norrtälje Municipality website: https://www.norrtalje.se/ai_oro.

formalization of existing professional or actuarial expertise.[4] The variables in a model could, for example, be derived from the experience of child welfare practitioners who consider poor housing or single parenthood to be risk factors. Nevertheless, models based on former human experience or decision-making can be flawed, biased and out-dated.

A model can also be based on actuarial expertise showing the existence of a certain statistical relationship between a variable and child maltreatment, such as the personal history of child abuse of a care giver indicating a known statistical risk factor.[5] This means that the use of AI can provide result that are somewhat lacking in precision. AI models based on PRM,[6] meaning predictions based on as many variables as possible, even those that might seem irrelevant, are likely to be of limited precision and lacking context. A recent scientific mass collaboration, “The Fragile Families Challenge”, shows that machine learning models are not very accurate when it comes to predicting life trajectories, which give reason to question their usefulness in the context of child welfare.[7]

Consequently, these models have the potential to lead to wrongful decisions, resulting in interventions that should not have taken place (false positives), as well wrongful decisions leading to a failure to act (false negatives).[8] The efficacy of such models can therefore be questioned from a methodological point of view.[9]

[4] Bosk, E.A. *What counts? Quantification, worker judgment and divergence in child welfare decision making*, Human Service Organizations: Management, Leadership & Governance, 2018, 42:2, p. 205.

[5] Bosk 2018, p. 214.

[6] Cuccaro-Alamin, Foust, Vaithianathan, Putnam-Hornstein 2017 p. 293.

[7] Salganik M.J. et al. Measuring the predictability of life outcomes with a scientific mass collaboration, Proceedings of the National Academy of Sciences of the United States of America (PNAS), April 2020 117 (15) 8398-8403, 2020 March 30, https://doi.org/10.1073/pnas.1915006117.

[8] V. Eubanks, *Automating Inequality – How High-Tech Tools Profile, Police and Punish the Poor*, Picador, New York 2019, p. 157; Bosk 2018, p. 205; S. Cuccaro-Alamin, R. Foust, R. Vaithianathan, E. Putnam-Hornstein, *Risk assessment and decision making in Child protective services: Predictive risk modeling in context*, Children and Youth Services Review, 2017, p. 292, p. 295, see https://www.datanetwork.org/wp-content/uploads/PRM-CYSR-article.pdf.

[9] R. van Brakel, *The Rise of Pre-emptive Surveillance: Unintended Social and Ethical Consequences*, Chapter 14 in E.R. Taylor and T. Rooney, 2016, Surveillance Futures: Social and Ethical Implications of New Technologies on Children and Young People, Routledge, London, p. 194–196.

Predictive models based on actuarial or human expertise can increase the risk for unlawful or disproportionate interferences in several of the rights of the child, such as the respect for family life, prohibition of discrimination, and protection of personal data. They can also affect the positive obligations of the state to protect children from torture or inhuman and degrading treatment or even child fatality if an AI system fails to detect children at a high risk of maltreatment. Moreover, the lawfulness regarding the possible use of variables such as socioeconomic status, health status and cultural or religious background of the parents needs to be addressed.[10]

Risk estimation may also be carried out through text mining, such as natural language processing topic modelling (NLP/TP). The latter is used to analyse texts and do not rely on risk factors that are top-down listed by a human expert. Instead, in such models the relevant variables are identified bottom-up by algorithmically analysing earlier cases which might give a more objective outlook if factors such as poor housing and single parenthood indeed are risk factors.[11] Nevertheless, these AI models also risk entrenching bias related to historical data,[12] which has been depicted by Medvedeva et al. as "status quo bias"[13]: for example, if child care services are more inclined to investigate cases of child maltreatment in families with poor housing and single parents, the data model will learn that these variables are relevant whereas in fact the training data might be misleading as they do not include undetected cases of child abuse in wealthier households with two parents.

The purpose of this paper is to give a preliminary overview and analysis regarding the design and use of AI tools to identify children at high risk of maltreatment in relation to relevant children's rights. Are such child

[10] See e.g. G. Van Bueren, *Opening Pandora's Box – Protecting Children Against Torture, Cruel, Inhuman and Degrading Treatment and Punishment* in G. Van Bueren (ed.), Childhood Abused – Protecting Children against Torture, Cruel, Inhuman and Degrading Treatment and Punishment, Routledge (e-book) 2018, p. 85; J. Ennew, *Shame and Physical Pain: Cultural Relativity, Children, Torture and Punishment* in van Bueren 2018 p. 53.

[11] Harrison C.J. and Side-Gibbons C.J., *Machine learning in medicine: a practical introduction to natural language processing*, BMC Medical Research Methodology, 2021, 21:158, p. 3.

[12] L. Svensson, *Automatisering – till nytta eller fördärv?* (Automation – to benefit or ruin?) Socialvetenskaplig tidskrift 2019:3-4, p. 358–359.

[13] M. Medvedeva et al., *The Danger of Reverse-Engineering of Automated Judicial Decision-Making Systems*, ArXiv 18 December 2020, p. 4, https://arxiv.org/pdf/2012.10301v1.pdf.

protection tools aligned with children's rights as laid down in the UN Convention of the Rights of the Child (UNCRC), the European Convention of Human Rights and the EU Charter of Fundamental Rights? And to what extent?

An analysis of law in relation to government use of AI tools for child protection needs to be undertaken from different perspectives. Applying children's rights requires a child-centred approach, which takes its starting point in the idea that children are equal as human beings and independent rights holders. Child maltreatment is a complex and multifaceted problem as is decision-making that relates to it. The perspective applied here is therefore holistic and interdisciplinary, meaning that the law is one of many tools to make children's rights real.[14] To this end, sources regarding for example social work and computer engineering are therefore of vital importance.

More specifically, this analysis revolves around the use of technology for child protection and the interaction and interdependency of "rules and tools", which embodies consequences both for the child and for society, and adds to the holistic and interdisciplinary perspective of legal informatics, which operates at the intersection of law and information and communication technology (ICT).[15] An important part of legal informatics is furthermore to contribute to the design of emerging technologies, such as predictive tools based on AI for child protection, through the establishment of legal standards and frameworks based in law.[16]

The outline of this paper is the following. In sections 2 and 3 I discuss the significance of preventive measures regarding child maltreatment and the different models used to predict child maltreatment. In section 4 I assess them from a legal perspective (UNCRC and ECHR). In the fifth and final section I draw some preliminary conclusions about the use of predictive tools based on AI to prevent child maltreatment.

[14] M. Grahn Farley, *Barnkonventionen – en kommentar*, Studentlitteratur, Lund 2019, p. 13–14; Van Bueren 2018.

[15] S. Greenstein, *Elevating Legal Informatics in the Digital Age* in S. Pettersson (ed.) Digital Human Sciences: New Objects – New Approaches, Stockholm University Press (2021) p. 156; P. Seipel, *IT Law in the Framework of Legal Informatics*, Scandinavian Studies of Law (2004) vol. 47, p. 37 f.

[16] C. Magnusson Sjöberg, *Om rättsinformatik* in C. Magnusson Sjöberg (ed.) Rättsinformatik – Juridiken i det digitala informationssamhället, Studentlitteratur, Lund 2021, p. 21–22.

2 The significance of early intervention and the idea of child protection systems to prevent child maltreatment

Child maltreatment is a highly complex, multidimensional problem both at an individual and societal level as well as from a biological and psychological standpoint. In Art. 19 of the UNCRC it is defined as "all forms of physical or mental violence, injury or abuse, neglect or negligent treatment, maltreatment or exploitation, including sexual abuse". Child maltreatment is not only detrimental to the individual child but to society as a whole, and comes with astronomical costs. It is regarded as a major public health issue and is assessed to affect at least 55 million children in the WHO European region alone (53 countries).[17]

In the 1990s neurological research demonstrated that child maltreatment can cause permanent neurological effects on children's brains.[18] This means that child maltreatment not only can cause lasting physical damage as the result of abuse or neglect, but affects behaviour, emotional well-being, personal relationships and cognitive functions.[19]

These research findings were followed by a shift in social work towards early intervention and evidence-based practices.[20] Instead of focusing on reactive approaches, such as providing protection for children who may already have been maltreated, the goal is to prevent it from happening. A central part of this proactive approach is risk assessment, which can be performed in various ways, normally through "operator driven" clinical or actuarial assessments.[21] A more recent trend is the development of technological tools using PRM to identify children at high risk of

[17] D. Sethi, Y. Yon, N. Prakeh, T. Anderson, J. Huber, I. Rakovac & F. Meinck, *European status report on preventing child maltreatment*, World Health Organization, 2018, p. 3 f.; D. Glaser, *The effects of child maltreatment on the developing brain*, Medico-Legal Journal 2014, vol. 82(3), p. 98.

[18] D. Daro and A.C. Donnelly, *Reflections on Child Maltreatment Research and Practice: Consistent Challenges* in D. Daro A.C. Donnelly, L.A. Huang, B.J. Powell (eds.) Advances in Child Abuse Prevention Knowledge – The Perspective of New Leadership, 2015, Springer (e-book), p. 8.

[19] Glaser 2014, p. 97.

[20] D. Daro and A.C. Donnelly 2015, p. 8.

[21] H. Vannier Ducasse, *Predictive risk modelling and the mistaken equation of socio-economic disadvantage with risk of maltreatment*, British Journal of Social Work 2020, p. 2 f.

being maltreated.[22] However, using standardized risk assessment tools based on actuarial principles is not a new idea within the field of social work. Already in the 1980s, the prospect of using expert systems for decision-making regarding child interventions was proposed by Schoech et al., while also acknowledging that such a system "offers many legal and ethical challenges to the human service professions".[23]

3 Cases of artificial intelligence for child protection

Today several tools using artificial intelligence have been developed or are under development for predicting child maltreatment in various countries such as Denmark, the Netherlands, New Zealand, Sweden, the U.K. and the U.S. However, the success rate varies and many of the projects have been discontinued.[24] The focus of this paper is limited to some of the most well-documented, researched and debated tools, most of which are accounted for as being PRM tools based on principles of actuarial risk assessment,[25] with the exception of the Norrtälje NLP/TP model. Each of the different models however is used to build tools to predict the risk for child maltreatment.

Before the presentation moves on to artificial intelligence (AI) and notably PRM tools, as well as NLP/TP more specifically, for countering child maltreatment, the question is: what are they?

PRM is a form of "predictive modelling", which is a description for tools with the aim of making accurate predictions, such as "machine learning", "AI" and "data mining". Kuhn and Johnson define predictive

[22] P. Gillingham, *Predictive Risk Modelling to Prevent Child Maltreatment and Other Adverse Outcomes for Service Users: Inside the 'Black Box' of Machine Learning*, British Journal of Social Work, 2016, 46, p. 1045.

[23] D. Schoech PhD, H. Jennings, L. L. Schkade PhD & C. Hooper-Russell,1985, *Expert Systems – Artificial Intelligence for Professional Decisions*, Computers in Human Services, 1:1, 81–115, DOI: 10.1300/J407v01n01_06, p. 106.

[24] A. Møller Jørgensen, C. Webb, E. Keddel, N. Ballantyne, *Three roads to Rome? Comparative policy analysis of predictive tools in child protection services in Aotearoa New Zealand, England, & Denmark*, Nordic Social Work Research 2021, p. 2, https://doi.org/10.1080/2156857X.2021.1999846; P. Gillingham, *Decision Support Systems, Social Justice and Algorithmic Accountability in Social Work: A New Challenge*, Practice: Social Work in Action, 2019, Vol. 31, No. 4, p. 278.

[25] Gillingham 2016, p. 1045.

modelling as "the process of developing a mathematical tool or model that generates an accurate prediction".[26]

PRM, more specifically, can be defined as:

> a type of predictive analytics… a statistical method of identifying characteristics that risk-stratify individuals in a population based on the likelihood each individual will experience a specific outcome or event. The result of the model's mathematical algorithm is a risk score. Unlike model-building techniques traditionally used in risk assessment – in which variables are chosen on the basis of previously researched relationships with the specified outcome – in PRM, as many data points as possible are examined, even if there is no previously specified relationship with the outcome of interest.[27]

The model works through algorithms, i.e. an instruction to the computer with a series of steps or procedures to follow. The algorithms will be coupled with variables to create a mathematical model (machine learning).[28] The model can examine and learn from a large amount of data from a variety of sources such as administrative datasets, whereby hidden patterns, correlations, regularities etc. can be extracted, which in turn can help in making predictions of different kinds, such as predictions concerning future risks in fields as diverse as finance, health and meteorology.[29]

The result of the PRM tool is consequently a risk score that can be used to for example to support decision-making in child welfare.[30]

NLP/TP models can also be used for risk prevention. It can be described as a form of text-mining, which basically means that "a group of algorithms, reveal, discover and annotate thematic structure in a collection of documents". It has been used or tested for example in healthcare to predict disease risk, risk of hospital readmission or suicide.[31]

[26] M. Kuhn & K. Johnson, *Applied Predictive Modeling*, Springer, New York 2013, p. 1.

[27] Cuccaro-Alamin, Foust, Vaithianathan, Putnam-Hornstein 2017 p. 293.

[28] M. Broussard, *Artificial Unintelligence – How Computers Misunderstand the World*, MIT Press, Cambridge Massachusets 2018), p. 94.

[29] S. T. McKinlay, *Evidence, Explanation and Predictive Data Modelling*, Philosophy and Technology (2017) vol. 30, p. 462–464; S. Greenstein, *Our Humanity Exposed – Predictive Modelling in a Legal Context*, Stockholm University 2017, p. 22, p. 70 ff.

[30] Cuccaro-Alamin et al. 2017, p. 293 f.

[31] P. Kherwa and P. Bansal, *Topic Modeling: A comprehensive review*, EAI Endorsed Transactions on Scalable Information Systems 2019, p. 2 and 10; A. Rumshisky, M. Ghassemi, T. Naumann, P. Szolovits, V.M. Castro, T.H. McCoy and RH Perlis, *Predicting early psychiatric readmission with natural language processing of narrative discharge summaries*, Translational Psychiatry 2016, 6e921, doi:10.1038/tp.2015.182.

The use of these tools and techniques also presents certain challenges. First of all, the models are created by humans, which can be reflected in the design of a model. This can be both a strength and a weakness. The developers can be endowed with expert-knowledge regarding children at risk of maltreatment as well as experience and empathy. However, it also means that the developers of the tool have the potential to incorporate bias.[32] The developers might lack the necessary insight regarding the prejudices that can impact child welfare decision-making, such as the example of social workers that might be less likely to detect child maltreatment in wealthy, two-parent households.

These tools are also limited to what algorithms can actually process, including the availability of relevant data. The amount, quality and nature of data can be imperfect and incomplete, which can especially be the case regarding data concerning human behaviour.[33] Moreover, the processing of data within the tool is often referred to as a "black box", since it, unlike a human professional or expert, cannot provide any reasons for its predictions, meaning a lack of transparency.[34]

Another concern is that no such tool can be 100 percent accurate, which may result in results that are wrong. As stated by O'Neil:

> There would always be mistakes, however, because models are, by their very nature, simplifications. No model can include all of the real world's complexity or the nuance of human communication. Inevitably, some important information gets left out.[35]

The concern regarding accuracy therefore evokes important issues related to evidence and substantiation.[36] It can be discussed if assessments made by an AI tool should be used as evidence,[37] and more specifically what the probative value would be in a legal setting. Finally, the use of PRM and NLP/TP tools can raise various ethical and legal challenges, such as

[32] Broussard 2018, p. 67. O'Neil breaks down predictive modelling to the individual level and concludes that racism can be apprehended as a predictive model "whirring away in billions of human minds around the world. It is built from faulty, incomplete, or generalized data." See C. O'Neil, *Weapons of Math Destruction – How Big Data Increases Inequality and Threatens Democracy*, Penguin Books, USA, 2016, p. 22.

[33] McKinlay 2017, p. 463.

[34] Greenstein 2017, p. 73.

[35] O'Neil 2016, p. 20.

[36] Mcinlay 2017; Gillingham 2016, p. 1049–1052.

[37] McKinlay, p. 471–473.

racial discrimination and poverty profiling.[38] We can therefore not be certain that such tools will actually and effectively counteract child maltreatment.

3.1 The Vulnerable Children PRM

An initiative by the government of New Zealand in 2011 appears to be the first initiative in the world by a government to develop a PRM tool to predict child maltreatment, the Vulnerable Children PRM.[39] The initiative was part of a large-scale reform in child protection services, with a social investment approach,[40] which among other things included new legislation and the linking up of databases across public service systems.[41]

A team of researchers in economy, social work and ethics at the Centre for Applied Research in Economics (CARE) at the University of Auckland, New Zealand, was given the task of researching the question of whether it would be possible to use administrative data to identify children at high risk of maltreatment.[42] The team developed an algorithm drawing from a data set from public welfare benefit systems and child protection services. The children included in the analysis were children 1) identified with a family that had a benefit period, i.e., the length of time during which a family received some kind of social benefit between the child's birth and 2nd birthday, including pre-birth and pregnancy related periods and 2) born between January 2003 and June 2006, so that they would reach 5 years of age by the end of the sample period.[43]

The model made use of 132 predictor variables which were presented in five categories in the CARE report. The first two categories included

[38] Eubanks 2018, p. 158; Cuccaro-Alamin et al. 2017, p. 295.

[39] N. Ballantyne, *The ethics and politics of human service technology: the case of predictive risk modelling in New Zealand's child protection system*, Hong Kong Journal of Social Work, vol. 53, 2019, p. 15.

[40] Møller Jørgensen et al. 2021, p. 3; Ballantyne 2019, p. 18.

[41] Gillingham 2016, p. 1046.

[42] Ibid.

[43] CARE (2012), R. Vaithianathan, T. Maloney, N. Jiang, I. De Haan, C. Dale, E. Putnam-Hornstein, T. Dare, *Vulnerable Children: Can Administrative Data Be Used to Identify Children at Risk of Adverse Outcomes?* Centre for Applied Research in Economics, University of Auckland, New Zealand, p. 10 available at: https://www.msd.govt.nz/documents/about-msd-and-our-work/publications-resources/research/vulnerable-children/auckland-university-can-administrative-data-be-used-to-identify-children-at-risk-of-adverse-outcome.pdf.

variables related to the care, protection and benefit of the subject child and that of other children in the family. Some examples are findings of abuse and neglect, child protection notifications, court orders and proportion of time on a benefit. The third and fourth categories consisted of data relative to characteristics concerning the child's caregiver and the family at the start of the period. For example, the data included gender, age, level of education, whether the household consisted of single or dual caregivers, number of children, age of caregivers when the oldest and the subject child were born etc. The fifth and final category concerned the care and protection and benefits history of the subject child's caregivers before the age of 16 as well as benefit histories in adulthood.[44]

It was determined that the model could accurately predict maltreatment within an area under the receiver operating characteristic (ROC) curve of 76 percent (a performance measurement for classification problems) which is comparable to the rate found in digital mammography.[45] The team also outlined a "business case" discussing return on investment in the PRM tool, which would mean a great reduction of the costs per child.[46]

The ethical approach taken by the team has been described as consequentialist.[47] In sum, the conclusion of the ethical evaluation was that the PRM tool certainly gave rise to concerns regarding certain aspects such as the risk of false positives, the fact that non-beneficiaries are not risk assessed and privacy issues etc.[48] As long as these concerns could be significantly mitigated or ameliorated, they could be outweighed by the important potential benefits of the tool.[49]

When the Vulnerable Children PRM became known to the public, it met with great concern. The accuracy of the tool was questioned, as it would constitute surveillance of the poor and race discrimination against Maori families which are subject to a disproportionate rate of child removals.[50]

[44] CARE 2012, p. 10 f.

[45] Ibid., p. 15.

[46] Ibid., p. 19 f.

[47] Ballantyne 2019, p. 20.

[48] CARE 2012, p. 32–34. The report recommended a full ethical evaluation, which was later conducted by Dare in 2013, see the report, p. 35 and Ballantyne 2019, p. 21.

[49] Ballantyne 2019, p. 20 f.

[50] Eubanks 2018, p. 138.

When a new minister of social development took office (from the same political party as her predecessor, the New Zealand National Party) the project was stopped in 2015.[51]

However, some of the researchers from the CARE team had won a contract to develop a similar PRM tool on the other side of the world in Allegheny County in Pennsylvania in the U.S.[52]

3.2 The Allegheny Family Screening Tool

The Children and Youth Service (CYS) in Allegheny County had been the source of public scandals, garnering national attention, which was in part related to a policy of preventing cross-racial adoptions, the Baby Byron case, and the homicide of toddler Shawntee Ford by her father, who had a record of violence and substance abuse that was known to the CYS.[53] Over the years, the CYS was also struggling with budget cuts.

The CYS had taken several measures to deal with the mounting problems. One of these measures was to create a data warehouse which would serve as a central repository, integrating information collected by the Department of Human Services, other county agencies and state public assistance programs. The data warehouse, eventually containing over more than a billion digital records, later proved useful as the foundation for designing and implementing decision support tools and predictive analytics. One idea was to build an automated triage system to help in setting priorities and making better use of the resources available to the CYS.[54]

The CARE team from New Zealand was assigned to design a PRM tool, similar to the Vulnerable Children PRM, using the data warehouse to harvest data in order to make predictions about probable maltreatment of children residing in Allegheny County.[55] The Allegheny Family Screening Tool (AFST) is linked to the county child abuse and neglect hotline, the ChildLine. Formerly the staff at the CYS were required to manually access and analyse vast amounts of data. This can now rapidly be performed by the AFST, which will produce a risk score regarding the long-term probability of future involvement in child welfare. The AFST

[51] Møller Jørgensen et al. 2021, p. 4 f.

[52] Eubanks 2018 p. 138.

[53] Ibid., p. 133.

[54] Ibid., p. 135–136.

[55] Ibid., p. 136–137.

is combined with other traditionally gathered information. If the score reaches a certain level, the CYS is obliged to initiate an investigation. According to the information on Allegheny County's website, the use of the AFST does not replace a clinical judgment but is used as additional information.[56]

The AFST has been a source of inspiration to other counties in the U.S. Nevertheless, it is far from uncontroversial. Concerns, similar to those faced by the Vulnerable Children PRM in New Zealand, have been raised regarding the AFST.[57] Even though the AFST shows the same degree of accuracy as its New Zealand counterpart, 76 percent in the area under the ROC, there is a great risk of harm to children and their families when a false positive occurs.

The use of proxies in the AFST, such as that of re-referrals (abuse notifications) is problematic, meaning that re-referrals are a variable no matter the reason for them. This is for example the case if several referrals are made regarding the same child either by someone with the aim to harass a parent or a family or due to so called "referral bias", which is often racially grounded.[58] According to various studies there is a disproportionately greater number of referrals concerning black or biracial families in Allegheny County.[59]

Similarly, there is also a class-based disproportionality concerning children placed in foster homes as a majority of placements concern families receiving different benefits for families in need. In conclusion, the use of public services appears to be considered a risk factor, in the same way as the Vulnerable Children PRM. In this regard, the tool is not designed to protect children from all class backgrounds against maltreatment. Furthermore, it has been criticized for being a tool for poverty profiling, confusing "parenting while poor with poor parenting".[60]

In Europe, similar PRM tools have been introduced by local governments in several countries.

[56] Information on the Allegheny County website: https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx.

[57] Allegheny County has rebutted the critique by Eubanks on their website, although without specifying any inaccuracies. See https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx.

[58] Eubanks 2018, p. 143, p. 153, p. 156.

[59] Ibid., p. 153.

[60] Ibid., p. 157–158.

3.3 The Hackney Early Help Profiling System

Hackney County Council in London, U.K., introduced an *Early Help Profiling System* (EHPS) in 2018 to help identify children at risk of neglect or abuse as part of the policy and practice of the Troubled Families Programme. The system is based on a predictive risk model bringing together data from multiple agencies. The underlying idea is that the children will be identified at an earlier stage before they come into contact with social workers, which will reduce costs.[61]

Scandals such as Baby P and Victoria Climbié, where small children already known to the authorities had been tortured and murdered by their caregivers, made it painfully evident that failures to share and act on information by the social services can have lethal outcomes. These scandals led to the idea of introducing so-called early help profiling systems in some municipalities in the U.K.[62] The scandals also contributed to new legislation, in the form of the Children Act 2004, which enhanced the possibilities of data sharing between agencies and provided local authorities with better access to information about the services that children in their respective areas were in contact with and contact information regarding the professionals involved. This was to be ensured by the application and synchronization of public databases.[63]

Part of this development was the online database RYOGENS (Reducing Youth Offending Generic National Solution) developed by the British Government together with consulting firm Deloitte and some other private companies. RYOGENS enabled officials from different agencies, such as Education, Police, Health Services, Social Services, Youth Offending Team and Housing Services to share information regarding a child at risk by filling in a form including forty different risk factors. If a certain threshold of reported concerns was reached, the system would generate an alert, which would be handled by a RYOGENS management function.[64]

[61] L. Dencik, A. Hintz, J. Redden & H. Warne, *Data Scores as Governance: Investigating uses of citizen scoring in public services*, Project Report, December 2018, Data Justice Lab, Cardiff University, U.K., p. 56.

[62] Dencik, et al. 2018, p. 58.

[63] R. van Brakel, *The Rise of Preemptive Surveillance: Unintended Social and Ethical Consequences*, Chapter 14 in E.R. Taylor and T. Rooney, Surveillance Futures: Social and Ethical Implications of New Technologies on Children and Young People, Routledge, London 2016, p. 189.

[64] Van Brakel 2016, p. 190.

The EHPS can be seen as yet another initiative "to explore the application of 'big data' solutions" regarding early intervention practises.[65] However, the EHPS was also built in the context of yet another harsh reality for the child services of Hackney Council, namely the combination of drastic funding cuts and an increase in the number of children on child protection plans and entering care.[66]

The EHPS was developed together with the private company Xantura, and funded by EY and London Councils.[67] The model integrated data from multiple agencies to identify children at risk of neglect or abuse in order to "strengthen the triage and assessment process"[68] and was expected to provide social workers with monthly risk profiles with integrated information about families with the greatest need of early intervention. The EHPS was therefore expressly said to be designed not to be punitive, only to enable earlier intervention.[69]

Only pseudonymized data was used by the model, meaning that data would only be made identifiable to the professionals assigned to deal with alerts generated by the model indicating that a high-risk threshold had been passed.[70]

No systematic account of the predictive variables in the model seems to be publicly accessible, but datasets that have been identified in a research study as well as in the media relate to school attendance, exclusion data, housing association repairs, arrears data, police records on anti-social behaviour and domestic violence, names, addresses, dates of births, unique pupil numbers, children and adult social care, housing debt, council tax, housing benefits and substance abuse data.[71]

[65] Ibid., p. 189–190.

[66] L. Stevenson, *Artificial Intelligence: how a council seeks to predict support needs for children and families*, Community Care, 1 March 2018, available at: https://www.communitycare.co.uk/2018/03/01/artificial-intelligence-council-seeks-predict-support-needs-children-families/.

[67] Dencik et al. 2018, p. 55.

[68] Information on the website of Xantura: https://xantura.com/early-help-profiling-system/.

[69] Dencik et al. 2018, p. 56.

[70] Ibid., p. 58.

[71] Ibid., p. 60; N. McIntyre and D. Pegg, *Councils use 377,000 people's data in efforts to predict child abuse*, The Guardian 16 September 2018, available at: https://www.theguardian.com/society/2018/sep/16/councils-use-377000-peoples-data-in-efforts-to-predict-child-abuse. Vannier Ducasse has expressed that information about the English

Regarding the accuracy of the EHPS, it was reported that over 80 percent of Hackney households identified as most at risk by the model were also at risk in real life.[72] According to media reports the EHPS helped detect seven children in need of early help support of whom Hackney Council was earlier unaware of.[73] A study presented in 2020 by What Works for Children's Social Care shows that there is no evidence that machine learning works satisfactorily in terms of accuracy when it comes to identifying children at risk.[74]

As a whole, the development procedure of the model lacked in transparency due to references to Xantura's commercial interests, which was presented as the reason to why several freedom of information requests (FOI) by researchers were denied.[75]

The fact that no information as to how many families were wrongfully identified as high risk and how those situations were handled, for example concerning the possibilities of removing such wrongful assessments from the EHPS, has been the subject of criticism.[76]

The entry into force of the GDPR as well as the Cambridge Analytica Scandal surely played a role in highlighting the data protection concerns voiced in the media regarding the EHPS, especially as the persons targeted by the EHPS were not informed of the use of their personal data and that no opt-out options were presented to them.[77]

The EHPS came to a halt in 2019 when it was concluded that the expected benefits would not be realized, which was mainly due to the lack of accuracy and data.[78] Looking forward a local politician, Darren Martin of the Hackney Liberal Democrats, stated:

experiments regarding PRM tools for child welfare and early intervention is "meagre", see Vannier Ducasse 2020, p. 4.

[72] Denick et al. 2018, p. 62.

[73] E. Sheridan, *Town Hall drops pilot programme profiling families without their knowledge*, Hackney Citizen, 30 October 2019.

[74] Møller Jørgensen et al. 2021, p. 5; Turner, A, '*No evidence' machine learning works well in children's social care, study finds*, Community Care, 2020 September 10, https://www.communitycare.co.uk/2020/09/10/evidence-machine-learning-works-well-childrens-social-care-study-finds/.

[75] Møller Jørgensen et al. 2021, p. 5; Dencik et al. 2018, p. 59–60.

[76] Møller Jørgensen et al. 2021, p. 6.

[77] Vannier Ducasse 2020, p. 19; Dencik et al. 2018, p. 62.

[78] Møller Jørgensen et al. 2021, p. 5; Sheridan 2019.

> …In a future where algorithmic technology will be used more and more, people have to know exactly how their data is being used… What we need now is an assurance that any future trial of this nature needs to be put in a public consultation with full disclosure of exactly what data is collected and how it will be used.[79]

3.4 The Gladsaxe model

The Gladsaxe model, based on a predictive algorithm to identify children at risk, received a great deal of national attention in Denmark.[80] It was established by a municipality in the suburbs of Copenhagen, inspired by prior models developed in New Zealand and USA, with the aim of creating an early warning system for detecting vulnerable children before they showed any symptoms of dysfunction.[81] A clear advantage of the model was that it could provide an overall assessment of the situation of the child through the mining of data from different sectors, with the potential of serving as a valuable supplement to professionals. If the model identified a child, a specialist adviser would make a preliminary assessment. If the expert found that there was reason to proceed, the family would be contacted and offered help. If the family declined, the municipality would not take any further action.[82]

The point-based model used data about several risk indicators such as mental illness (3000 points), unemployment (500 points), missing a doctor's appointment (1000 points) or dentist's appointment (300 points). Divorce was also included in the risk assessment. The model extracted data from nine different public sources, for example, the employment system used by job centres, the central personal register, dentist journals, the day care system and notifications of concern received by public authorities.[83]

[79] Sheridan 2019.

[80] Møller Jørgensen et al. 2021, p. 7.

[81] Ibid. p. 8. Møller Jørgensen et al. points out that the cross-national influence of the Vulnerable Children PRM is evident in Rhema Vaithianathan's inclusion in one of the scientific advisory boards of the project.

[82] U. Andreasson and T. Stende, *Nordic municipalities' work with artificial intelligence*, Nordic Council of Ministers 2019, p. 22, available at: https://www.norden.org/en/publication/nordic-municipalities-work-artificial-intelligence.

[83] R.F. Jørgensen, *Data and rights in the digital welfare state – the case of Denmark*, Information, Communication & Society 2021, p. 8, https://doi.org/10.1080/1369118X.2021.1934069.

The model was supposed to be rolled out in relation to all families with children within the municipality, but there were problems related to the accuracy of the model and a relatively high error rate, mainly due to the lack of historical data. The municipality had also made a request to the data protection agency to be exempt from the data protection legislation in order to access data from different sources, which was denied.[84]

The problems did not end there. When the model became public knowledge through an article in the daily newspaper Politiken, it caused a public outcry. The model was depicted as a tool for mass surveillance of families with children and the idea of a point-based system went above and beyond what was deemed to be acceptable.[85]

Nevertheless, the Danish Government was ready to go through with a legislative proposal which would allow municipalities in Denmark to combine data regarding families with children and children in general as part of an overarching plan to fight parallel societies, also known as the "ghetto-plan", which would enable the scoring of neighbourhoods. If a neighbourhood scored high enough to be qualified as a ghetto, several measures would be put into place, such as applying automated risk assessment systems to families with children. The proposal was later withdrawn.[86]

The Gladsaxe model, as its focus was not only in relation to a part of the population receiving benefits but to the entire population, can be said to be an example of a model that generally had a broader reach than the Vulnerable Children PRM or the Allegheny Family Screening Tool.

3.5 The Norrtälje model

In 2020, the municipality of Norrtälje became the first in Sweden to develop a tool using a Robotic Process Automation system (RPA) involving AI to identify children at risk. The system would collect and analyse previous cases as a tool to help social workers make a decision concerning the initiation of a child protection investigation after receiving reports

[84] Ibid.

[85] B. Alfter, *Denmark* in *Automating Society – Taking Stock of Automated Decision-Making in the EU* – A report by AlgorithmWatch in cooperation with Bertelsmann Stiftung, supported by the Open Society Foundations, 1st edition, January 2019 p. 51; see also J. Sorgenfri Kjaer, https://politiken.dk/indland/art6365403/Regeringen-vil-overvåge-alle-landets-børnefamilier-og-uddele-point.

[86] Alfter 2019, p. 51; Andreasson and Stende 2019, p. 22.

of concern (*orosanmälningar*). The project was essentially funded by the municipality with some help from Vinnova, Sweden's innovation agency. Part of the background concerning the pilot project was the 50 percent increase in reports of concern between 2014 and 2018. There was an urgent need for support measures for the social services.[87]

The system basically works through a web-service for concern reports, where digital reports will be received from the mandatory reporters, such as schools, health care and police, in a structured manner. Next, the AI tool will read and analyse the information received via the web-service as well as registering it in the operating system. Finally, it will create a pre-assessment proposal (*förhandsbedömning*) via a predictive model tool for pattern recognition based on prior assessments. A child welfare officer will decide whether the pre-assessment proposal will be documented.[88]

The dataset includes anonymized administrative data related to all prior assessments regarding the initiation of a child protection investigation by the municipality. The model is designed to compare words in new reports with earlier reports and to make an assessment based on the latter reports.[89] Björn Preuss from the company *2021.AI* has provided the following explanation:

> We do not select any information manually or include any factors. We only use the text which is sent with every report. The only information which is prior to the model detected and filtered away is personal information like names, age, social number, etc. So the model cannot be biased towards a name, gender, age etc. All predictions are only based on historic text descriptions of cases and their statistical similarity, word and sentence patterns, etc.[90]

[87] See official statement by the IT-department at Norrtälje municipality regarding IT-investment, a platform for automation and decision support, 2019-07-17, available at: https://forum.norrtalje.se/welcome-sv/namnder-styrelser/kommunstyrelsens-arbetsutskott/mote-2019-08-28/agenda/tjansteutlatande-gallande-investering-for-plattform-for-automatisering-och-beslutsstodpdf-35012?downloadMode=open; *Larmet: 200 barn om dagen misstänks fara illa i Stockholms län*, SVT 17 February 2020, https://www.svt.se/nyheter/lokalt/stockholm/200-barn-om-dagen-orosanmaldes-under-2018.

[88] *Projekt för AI och robotisering av orosanmälan*, information on the Norrtälje municipality website, available at: https://www.norrtalje.se/ai_oro; P. Molander Wistam, PowerPoint Presentation 23 August 2021, RPA/AI Flödesbeskrivning.

[89] F. Adolfsson, *AI för Norrtäljes orosanmälan*, Voister 13 november 2019, available at: https://www.voister.se/artikel/2019/11/ai-for-norrtaljes-orosanmalan/.

[90] Quote from an e-mail from P. Molander Wistam, 24 August 2021. Also see M. With, Dansk IT 12 October 2020, *AI for the sake of the children*; the client case of 2021.AI:

After a legal review supported by the Swedish Association of Local Authorities and Regions, the municipality decided not to pursue the project.[91] Swedish law as such does not prevent the use of predictive modelling regarding historical data about individuals for the purposes of case management and for developing quality assurance within the social services.[92] There are, however, important limitations as to how this can be carried out. Search limitations according to the law, include, for example, automated data processing regarding reports of concern or pre-assessments which did not lead to a decision to initiate a child protection investigation, even if the child already had a case file at the social services.[93] This posed a problem regarding the deployment of the Norrtälje model, since it was necessary to use such pre-assessments as a source. The legal limits at hand have been the subject of debate for decades but the issue has repeatedly been dismissed as contrary to the right to the protection of privacy. However, the issue is not off the table and new legislative proposals are being considered by government authorities.[94]

The Norrtälje model does not seem to have been subjected to any research analysis thus far, but important research regarding automation in social services lay bare some of the important challenges that the use of historical administrative data, such as prior decisions, within an AI tool might entail. This concerns the possible cementing of former biases as well as the balancing of interests in individual cases, which is required by the principles of the rule of law.[95]

Applying AI to create more comprehensive, safe and accurate assessments of social service cases in Norrtälje Municipality available at: https://dit.dk/nyheder/2020/for-the-sake-of-the-children.

[91] A. Yanchur, G. Rosén Fondahn and S. Pilz, *A Swedish town bought an AI to spot children at risk, but decided against deploying it*, Algorithm Watch 10 August 2021.

[92] M. Nymark, *Användning av AI inom socialtjänsten*, report, Swedish Association of Local Authorities and Regions 2021-02-07, available at: https://skr.se/download/18.427140af179361c4e4616b7a/1620377226836/Anv_%20av_%20AI_%20inom_%20socialtjansten_%20rapport.pdf.

[93] Nymark 2021, p. 18.

[94] Socialstyrelsen (the National Board of Health and Welfare), *Att göra anmälningar som gäller barn sökbara*, Report May 2019, available at: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/ovrigt/2019-5-15.pdf.

[95] L. Svensson, *Automatisering – till nytta eller fördärv?* (Automation – benefit or harm?) Socialvetenskaplig tidskrift 2019:3-4, p. 358–359.

3.6 Concluding remarks

This preliminary overview of examples of the use of AI tools for child protection in social services reveals that most of them have been created in the context of struggles related to increases in caseloads, funding cuts and staff shortages in relation to social services as well as government ambitions to increase digitalisation in the public sector. Thus, the primary reasons for developing such tools mainly seem to be of a financial and administrative nature.

This overview also shows that many of the projects developing AI tools for child protection have been discontinued at the experimental stage, which mainly seems related to legal, ethical and public trust problems. Legal limits as well as state-structures can limit the amount of data that can be used in a model, which can render it more or less useless. The only tool that has survived so far is the Allegheny Family Screening Tool in the U.S.

The tools also differ in purpose and scope. Some of them use point-based systems related to individuals regarding certain characteristics or activities, while others use text-mining.

Nevertheless, this is an ongoing trend which is presumably here to stay.

4 Children's rights, child protection services and AI tools

4.1 Introduction to the children's rights system in Europe

AI-tools for child protection can have huge legal implications, in particular concerning children's rights, and have the power to severely impact the lives of children. In the end, however, it all comes down to how AI-tools are used and for what purposes. A framework for the use of AI-tools, at a minimum, needs to be developed that is in accord with children's rights. The child is an independent rights holder.[96] The focus of this paper is a European Human Rights perspective.

One of the most important children's rights instruments is the United Nations Convention of the Rights of the Child (UNCRC) adopted in

[96] W. Vandenhole, G. Erdem Türkelli and S. Lembrechts, *Children's Rights: A Commentary on the Convention on the Rights of the Child and Its Protocols*, Edward Elgar Publishing Ltd 2019, p. 15.

1989. It has been ratified by all the members of the UN except the U.S. It is binding for the signatory states. In states with monist systems, the UNCRC is directly applicable, whereas in states with dualist systems the applicability depends on whether that state has incorporated the UNCRC as a part of national law,[97] which is for example the case in Sweden since 2020.[98] The fact that the UNCRC is binding on the signatory states does not mean that it is enforceable by individuals. There is no international court of children's rights to turn to, and no other court unless a signatory state has decided to make the rights enforceable in a national court of law. Nevertheless, the Committee on the Rights of the Child has both a monitoring and advisory function.

More importantly the UNCRC is highly integrated into the European human rights system. The member states of both the Council of Europe and the EU are parties to the UNCRC, and the UNCRC has been described as "the touchstone for the development of European children's rights law".[99]

This development has mainly taken place within the European Convention on Human Rights (ECHR) framework. The ECHR from 1950 applies in most states in Europe, is a part of EU law, and provides an enforceable protection of children's rights through the European Court of Human Rights (ECtHR). The case law of the ECtHR has had an important practical impact on children's rights in Europe, including numerous cases regarding child protection,[100] even though the application of the principle of the margin of appreciation for the states has been a focus of criticism in cases related to "the best interest of the child".[101]

Inspired by the UNCRC,[102] children's rights are also regulated in Art. 24 of the EU Charter of Fundamental Rights (EUCFR) of 2009, but the scope is limited to certain cross-border situations, such as criminal law

[97] Vandenhole et al. 2019, p. 21.

[98] Prop. 2017/18:186; See also K. Åhman, P. Leviner, K. Zillén (ed.) *Barnkonventionen i praktiken – Rättsliga utmaningar och möjligheter*, Norstedts Juridik Poland 2020 p. 30–42 and Grahn Farley 2019 p. 26–28.

[99] *Handbook on European law relating to the rights of the child*, European Union Agency for Fundamental Rights and Council of Europe 2015, p. 26.

[100] Vandenhole et al. 2019, p. 18.

[101] R. Lamont, *Article 24 – The Rights of the Child* in S. Peers, T. Hervey, J. Kenner and A. Ward (eds), The EU Charter of Fundamental Rights – A Commentary, Hart Publishing 2014, p. 673.

[102] Lamont 2014, p. 674.

and immigration law. The reason for this is that the EU does not have any direct general competence regarding children's rights.[103] It will therefore not be further examined in this paper. Nonetheless it is noteworthy that the EU commission is actively working with children's rights and introduced a strategy on the rights of the child and the European Child in March of 2021. The strategy developed has been guided by the UNCRC with the purpose of securing access to basic services for vulnerable children. An important aim of the strategy is to break vicious cycles across generations related to child poverty and social exclusion.[104]

The analysis below will focus on the rights of most relevance to the use of AI-tools for child protection. This includes rights with a direct purpose of protecting children from maltreatment, the right to life and the prohibition of inhuman and degrading treatment, as well as the right to respect for family life in conjunction with the prohibition of discrimination. These tools have the capacity to both enhance and/or interfere with children's rights, which will be elaborated below.

4.2 The child's right not to be maltreated and the positive obligation for the state to make risk assessments

It can be said that the utmost duty to protect children rests upon the state. If parents or other legal caregivers are not able or unfit to take care of children, *i.e.* human beings under the age of 18 as prescribed in Art. 1 of the UNCRC, the state is required to intervene. In Art. 3.3. of the UNCRC this is expressed as:

> State Parties shall ensure that the institutions, services and facilities responsible for the care or protection of children shall conform with the standards established by competent authorities, particularly in the areas of safety, health, in the number and suitability of their staff, as well as competent supervision.

There is thus a *positive obligation* for the state to protect children from maltreatment, i.e., circumstances when a State has a duty to take action

[103] Lamont 2014, p. 662.

[104] Communication from the Commission to the European Parliament, the Council and the European Economic and Social and the Committee of the Regions, *EU strategy on the rights of the child*, Brussels 23.4.2021 COM (2021) 142 final.

in order to secure the protection of individuals within its jurisdiction,[105] which can involve complex risk assessments. When this turning point is reached is a delicate matter requiring a complicated balancing act which involves the human rights of both the child and the caregivers.[106] The idea that AI-based tools for risk assessments could be used to help in making such risk assessments therefore seems highly relevant.

If the state fails to protect a child a whole range of human rights come into play of both an absolute and relative nature, raising the question, to what extent are there exceptions to a right. In extreme cases such as Baby P and Victoria Climbié in the U.K. and the case of "Little heart" (*Lilla hjärtat*) in Sweden, where small children already known to the authorities, have died at the hands of their caregivers, the right to life laid down in Art. 2 of the ECHR and Art. 6.1 UNCRC, which is an absolute right, is applicable if the state did not act on the evidence or information available to them.

The ECtHR applies the so-called Osman-test to assess whether state authorities have taken the necessary preventive measures in cases where children are at high risk, i.e. when the positive obligation of the state is triggered:

> It must be established…that the authorities knew or ought to have known at the time of the existence of a real and immediate risk to the life of an identified individual or individuals from the criminal acts of a third party and they failed to take measures within the scope of their powers which, judged reasonably, might have been expected to avoid that risk.[107]

This threshold can be met in the case of domestic abuse against a parent who is known to the authorities, since this means that the child is at a high risk of maltreatment. For example, in the cases of *Kontrova v. Slovakia*[108] and *Talpis v. Italy*[109] where women had reported serious abuse and threats with lethal weapons by their partners to the police, the failure of the police to investigate and report to the social services led to the killing

[105] B. Rainey, E. Wicks, and C. Ovey, *Jacobs, White & Ovey – The European Convention on Human Rights*, 6th ed., Oxford University Press, U.K. 2014, p. 103.

[106] See e.g. *Z. and Others v. the United Kingdom*, n° 29392/95, Judgment (GC) 10 May 2001, § 74.

[107] *Osman v. The United Kingdom*, n° 23452/94, judgment (GC) 28 October 1998, § 115.

[108] *Kontrova v. Slovakia*, n° 7510/04, Judgment 31 May 2007.

[109] *Talpis v. Italy*, n° 41237/14, Judgment 2 March 2017.

of minor children in the family. The ECtHR found that the state had violated the right to life regarding the children. In the Talpis case the court concluded that:

> Article 2 of the Convention may also imply in certain well-defined circumstances a positive obligation on the authorities to take preventive operational measures to protect an individual whose life is at risk from the criminal acts of another individual.[110]

The ECtHR did not find that the authorities had made a correct risk assessment in the Talpis case. Adding to the Osman-test the Court stated that:

> In the Court's view, the risk of real and immediate threat must be assessed taking due account of the particular context of domestic violence. In such a situation it is not only a question of an obligation to afford general protection to society…but above all to take account of recurrence of successive episodes of violence within the family unit.[111]

The ECtHR found that the state had failed to live up to its positive obligations to take preventive operational measures to protect an individual whose life is at risk.

In some situations, however, it can be difficult or even impossible to foresee the killing of a child by his or her caregiver, such as in the case of *Penati v. Italy* where a father had killed his son and himself during a protected contact session between the father and son on the premises of the social services of a municipality. As long as the authorities have taken the necessary preventive measures that are available, they cannot be held liable for a violation of the right to life.[112]

The Grand Chamber judgement in *Kurt v. Austria*, where, following an escalating spiral of domestic violence involving both the mother and the children, the father shot his 8-year-old son to death at school, provides further clarifications regarding the Osman-test in the form of general principles. In this case, however, the dissenting opinion shows that the judges were not in agreement with each other on where to draw the line as to what can be demanded of the authorities when it comes to risk

[110] *Talpis v. Italy*, § 101.

[111] *Talpis v. Italy*, § 122.

[112] *Penati v. Italy*, n° 44166/15, Judgment 11 May 2021, § 188 (available only in French). The applicant has requested a referral to the Grand Chamber.

assessments regarding domestic abuse cases, and in particular the risk for lethal outcomes.

The court established that when there is a real and immediate risk to the life of a victim of domestic violence, the authorities have a duty to carry out a lethality risk assessment in an autonomous, proactive and comprehensive manner. Nevertheless, the Osman-test does not require that states use standardised risk assessments, such as standardised checklists based on criminological research, even though the court acknowledged that such assessments are useful.[113] The court also concluded that in the case where "several persons are affected by domestic violence, be it directly or indirectly, any risk assessment must be apt to systematically identify and address all the potential victims."[114] It also emphasised the importance of documentation, information sharing and coordinated support with other relevant stakeholders that come into regular contact with persons at risk, which in the case of children can be teachers. The authorities should also communicate the outcome of their risk assessment to the victims and, when necessary, give advice and guidance regarding different protective measures available to them.[115]

By ten votes to seven, the majority held that Austria had met these requirements in the Kurt-case and that there thus had been no violation to the right to life in this case.

The minority, however, found that the risk assessment was seriously flawed and that the State had breached the right to life. Among others, the minority pointed out that the authorities failed to make a separate risk assessment in relation to the children and did not treat the risk of domestic violence as one that impacted the family as a unit. This was particularly grave since the authorities had information which indicated a high risk to the children. Apart from statements given by the children themselves regarding physical abuse by the father, the authorities evidently downplayed the fact that the father had made explicit and repeated threats to the mother that he would kill the children.[116] The lack of standardized research-based assessment tools by the authorities was highlighted in this regard.

[113] *Kurt v. Austria*, n° 62903/15, Judgment (GC) 15 June 2021, §§ 168-171.

[114] *Kurt v. Austria*, § 173.

[115] *Kurt v. Austria*, § 174.

[116] *Kurt v. Austria*, Joint dissenting opinion by judges Turkovic, Lemmens, Harutynyan, Elósegui, Felici, Pavli and Yüksel, § 13.

A more common scenario is that if there is enough evidence to support that a child has been subject to torture, abuse or neglect, the authorities are obliged to act and thoroughly investigate such a case and, if necessary, take the appropriate measures. A failure to act, can constitute a breach of Art. 3 ECHR which includes the prohibition of torture or other inhuman or degrading treatment. Art. 19.1 of the UNCRC stipulates that:

> State parties shall take all appropriate legislative, administrative, social and educational measures to protect the child from all forms of physical or mental violence, injury or abuse, neglect or negligent treatment, maltreatment or exploitation, including sexual abuse, while in the care of parent(s), legal guardian (s) or any other person who has care of the child.

A crucial factor in this regard is the degree of maltreatment. The court has found that for maltreatment to fall within the scope of Art. 3 ECHR, the maltreatment must attain a minimum level of severity. To this end an overall assessment of the relevant circumstances of the case has to be conducted, taking into consideration, for example, the nature and context of the treatment, its duration, its physical and mental effects, and in some cases the sex, age and state of health of the victim.[117]

In cases regarding neglect or abuse, the need for authorities to act swiftly is crucial. The ECtHR has in numerous judgments criticised states for the failure to act on information available to them. In the case of *Z. and Others v. U.K.* repeated concerns had been reported to the social services about a family with four small children during a period of four and a half years. The children had been subjected to severe neglect and emotional abuse, where the parents kept the children locked up in their rooms which were extremely filthy, or locked them out of the home. The children were malnourished, dirty and were regularly caught stealing food from bins. It was not until the mother demanded that the social services put the children up for adoption and care, as she could not cope with them, that the children were taken in for emergency care. The Court found that, in the present case, it was not in dispute that the neglect and abuse suffered by the children reached the threshold of inhuman and degrading treatment. It was concluded that the authorities were under a statutory duty to protect the children and had a range of powers available to them, which included the removal of the children from their home. The Court acknowledged that the social services are faced with a diffi-

[117] *Costello-Roberts v. the United Kingdom*, n° 13134/87, Judgment 25 March 1993, § 30.

cult and sensitive task to balance the duty to uphold the countervailing principle of respecting and preserving family life and assessing the risk of maltreatment. Nevertheless, in the present case, there was no doubt as to the failure of the system to protect the children from serious, long-term neglect and abuse.[118]

In the case of *E. and others v. the U.K.* three sisters and a brother had been subjected to long-term, severe, physical and sexual abuse by their mother's partner. The partner was convicted of sexually assaulting two of the girls. When he came back to live with the family while on probation, the authorities failed to take the necessary steps to monitor and supervise the family and make the necessary risk-assessments, which meant that the abuse could continue for several years. The children suffered serious mental disorders as a result. The Court made the assessment that the state had not reasonably used the measures available. There was a clear pattern of a lack of investigation, communication and co-operation by the relevant authorities which would have had the possibility to avoid or at least minimize the risk of the damage suffered.[119]

A specific situation where the positive obligation of the state is normally triggered is, for example, when a head teacher reports concern of suspected maltreatment. This is especially the case since such a report presumably is reflective of teachers who have the child or children concerned on their watch on a daily basis. The authorities are hereby obliged to take the necessary precautionary measures, including a child maltreatment risk assessment.[120]

In sum, the case-law of the ECtHR provides us with general principles regarding the maltreatment risk assessment and lethality risk assessment. Suspicions of maltreatment and or risk for the child's life will trigger the immediate need for appropriate measures to be taken. The duty to trace child maltreatment is somewhat vague, but Art. 3 ECHR and Art. 19 UNCRC require legislative and administrative measures, as well as social and educational measures to be in place. Certainly, institutions such as schools and school health services play an important role in the detection

[118] *Z. and Others v. the United Kingdom*, n° 29392/95, Judgment (GC) 10 May 2001, § 74.

[119] *E. and Others v. the United Kingdom*, n° 33218/96, Judgment 26 November 2002, §§ 99-100.

[120] *Association Innocence en Danger v. France and Association Enfance et Partage v. France*, n° 15343/15, 16806/15, Judgment 4 June 2020 (available in French and German), § 161, § 167.

of child maltreatment. The right of the child not to be maltreated however does not at this point in time seem to encompass the prediction of child maltreatment in cases where there is no "smoking gun". States are however encouraged to use research-based, multidisciplinary risk assessment standards for the prevention and mitigation of child maltreatment.

Consequently, the use of AI-tools for child protection may have the potential to make the risk assessment process by the relevant authorities more effective, which in turn may enhance the protection of children from maltreatment and prevent death. Nevertheless, this requires that the system is legal, research-based and has a high degree of accuracy. In the light of Art. 22 GDPR it is also important that AI-driven child protection tools will be used in such a way that the experts will not solely rely on such tools. In a study conducted by Bosk, it was determined that one third of the social workers were positive to using a risk score, in part because it was seen as an important tool to prevent subjective decision making and perhaps more noteworthy in part because it "removed the responsibility (and terror) of making a mistake". If social workers would start to rely solely on risk scores, this could in practise constitute illegal automatic decision-making. Instead, they could serve as part of an elaborate method using several different tools. Moreover, it is important that such an AI-tool is developed in a proper manner including the examination of various risk factors, which means that issues regarding discrimination in particular must be assessed.

4.3 The child, the right to respect for family life and the prohibition against discrimination

If social services decide to take measures that can be more or less intrusive into the family life of the individuals involved or even separation of the family members, the right to family life stipulated in Art. 8 of the ECHR has to be considered.[121] This involves both the child and the caregivers, such as biological parents or foster parents.[122]

A primary consideration in this regard is the somewhat vague concept of "the best interest of the child" Art. 3.1 UNCRC, which is applied by

[121] *Strand Lobben and Others v. Norway*, n° 37283/13, judgment (GC), 10 September 2019, §§ 202-04.

[122] See e.g. *Kopf & Liberda v. Austria*, n° 1598/06, Judgment 17 January 2012 regarding the right of respect to private and family life of foster parents (Art. 8 ECHR).

the European Court of Human Rights as well as in many national legal systems. In this context it has the power to override the rights of the parents, since the aim pursued regarding child protection measures is the best interest of the child.[123]

As stated above, decisions regarding measures such as early intervention are a delicate matter. They involve issues such as what constitutes good or adequate parenting, which might give rise to discriminatory assessments based on factors such as socioeconomic status, the level of education of the parents, disabilities or illnesses, place of residence, race, religion, culture etc. Social services are thus required to work to prevent bias from being part of the decision-making process, which might prove particularly difficult when using AI driven tools. This is a hurdle that has to be overcome in an effective manner if such technology is to be used in the first place. A general prohibition of discrimination is regulated in Art. 14 of the ECHR and more specifically in Art. 2.1 of the UNCRC, which reads:

> State parties shall respect and ensure the rights set forth in the present Convention to each child within their jurisdiction without discrimination of any kind, irrespective of the child's or his or her parent's or legal guardian's race, colour, sex, language, religion, political or other opinion, national, ethnic or social origin, property, disability, birth or other status.

This article leaves room for a broad interpretation of what constitutes discriminatory treatment by the state. Article 14 ECHR is not applied independently but will be applied in conjunction with another right stipulated in the ECHR and is thus regarded as an ancillary right.[124] In the context of child welfare measures Art. 14 is often applied together with the respect to private and family life laid down in Art. 8 ECHR. Moreover, the protection against discrimination in Art. 14 is completed by Article 1 of Protocol No. 12 to the ECHR, which prohibits discrimination more generally, in the enjoyment of any right set forth by law. It is noteworthy that only 20 states among the signatory states have ratified Protocol No. 12.

[123] See e.g. *Vojnity v. Hungary*, n° 29617/07, Judgment 12 February 2013, § 43.

[124] *Guide on Article 14 of the European Convention on Human Rights and on Article 1 of Protocol No. 12 to the Convention – Prohibition on Discrimination*, Updated on 31 December 2020, Council of Europe/European Court of Human Rights, p. 6.

The risk of discriminatory assessments in this context is mainly related to the parents, which can include both characteristics and behaviour. To use characteristics as risk variables is therefore especially risky in regard to the prohibition against discrimination. It can also raise issues regarding so-called intersectionality, that is, the interplay of several grounds of discrimination at the same time, such as social background, sex, race, ethnicity, sexual orientation, disability, and age. In such cases there is a need for a more holistic and flexible approach, which cannot be satisfied by the use of single comparators.[125]

Several cases in the ECtHR case law are illustrative of this. The Court has criticised decisions to remove children from their parents solely on reasons of poor housing and poverty as contrary to the right to respect for family life. In some of these cases the measures notably targeted families where the parents had a certain ethnic background or disability.[126]

It has also been deemed contrary to Art. 8 and Art. 14 ECHR to base the withdrawal of parental rights or parental access rights solely on the ground of disability,[127] mental illness,[128] religious considerations[129] or sexual orientation of the caregivers.[130] The case law, however, does not indicate that factors such as social background, disabilities or religious conviction of the parents cannot be part of an overall assessment of the parent-child relationship in cases where there are other circumstances such as a risk of abuse and neglect.

The question is what predictive risk variables would be lawful or appropriate to use when developing an AI-tool that will be constructed on the basis of many risk variables which have the potential to provide an important overview of a child protection case. This also raises issues

[125] S. Atrey, *Comparison in intersectional discrimination*, Legal Studies, 2018, 38, p. 379–395.

[126] *Barnea and Caldararu v. Italy*, n°, Judgment 22 June 2017 (Roma origin); *Saviny v. Ukraine*, n° 39948/06, Judgment 18 December 2008 (blind parents); *Wallová and Walla v. Czech* Republic, n° 23848/04, Judgment 26 October 2006, §§ 71-72.

[127] *Kocherov and Sergeyeva v. Russia*, n°16899/13, Judgment 29 March 2016 and *Kutzner v. Germany*, n° 46544/99, Judgment 26 February 2002 (mental disabilities).

[128] *Cînta v. Romania*, n° 3891/19, Judgment 18 February 2020, §§ 47-57 (paranoid schizophrenia).

[129] *Vojnity v. Hungary*, application n° 29617/07, Judgment 12 February 2013 and *Hoffmann v. Austria*, judgment 23 June 1993 (Parents belonging to a Pentecostal Charismatic Church and Jehovah's Witnesses respectively).

[130] *Salgueiro da Silva Mouta v. Portugal*, judgment 21 December 1999 (homosexual parent).

regarding intersectionality, which in the context of AI-tools might prove to be a major hurdle, since AI-tools can only make automatic pre-assessments based on certain risk factors.

The use of predictive variables is especially problematic when it comes to measuring the predictive risk variables in relation to outcome variables. It does not seem that there is any data mining process, such as the statistical procedure referred to as *stepwise probit or logistic regression* (SPLR) used in the Vulnerable Children PMR, that is certain to produce meaningful correlations.[131]

Instead, there is a clear risk that such correlations may be exaggerated or irrelevant. SPLR for example does not take into account the distribution of factors related to maltreatment in the rest of the population. The fact that parents have learning disabilities or health problems does not in itself mean that they cannot provide good parenting. If 10 percent of this group of parents maltreat their children, there is still 90 percent that does not. The weight given to such factors therefore poses a problem. If five percent in the general population would be considered as maltreating their children, the weight given to learning disabilities or poor health would double the child's risk of maltreatment. However, in absolute terms the risk is much lower regarding children with parents facing such problems.[132]

Furthermore, an SPLR method may create misleading results, since "any factor that varies with maltreatment is taken to be theoretically suitable and to enhance" the PRM. It fails to assess the degree of these factors, as they do not occur only in abusive families. The use of such methods can therefore not be considered to encompass the complexity of a balanced assessment regarding child maltreatment[133] and has for this reason been labelled a "statistical fishing expedition".[134]

Considering the Vulnerable Children PRM and the Allegheny Family Screening tool, it is clear that there is a direct connection to the child's or the caregivers' social origins and property or lack of property. Indirectly there are issues related to, for example, race and/or ethnic origin, since these grounds are often linked to the fact that due to prior discrimina-

[131] Eubanks 2018, p. 144.

[132] Vannier Ducasse 2020, p. 7.

[133] Ibid.

[134] Eubanks 2018, p. 144.

tion, certain groups in society have been oppressed and as a result of that also belong to less advantaged socioeconomic groups.

AI-tools of this kind that only include children whose parents are on welfare benefits seem unlikely to be in accordance with the prohibition against discrimination.

In the light of the prohibition against discrimination, the use of several other risk factors are problematic concerning a PRM tool, regarding both the child and the caregivers. Even though, for example, a religious conviction might pose a risk for child maltreatment if the parents adhere to a religious sect,[135] it is hard to see how this would be handled within an AI-tool, with all the different dimensions that might have to be assessed.

In conclusion, the use of PRM-tools to prevent child maltreatment do not seem suitable for making decisions regarding pre-assessment in child protection cases. These are decisions which require empathy, flexibility and intuition.

The NLP/TP model developed by Norrtälje Municipality, however, does not seem to be as problematic as the PRM-models in relation to direct discrimination. However, there is a risk that the status quo bias in former decisions will be included, which may lead to the repetition of biased or unrepresentative decision-making amounting to unlawful discrimination. Moreover, having in mind the evolution regarding both research and values regarding the child-parent relationship of the past two decades, European perceptions of family have undergone important changes, not least regarding lesbian, gay, bi- and transgender families as well as the role of fathers in children's development. It is clear that the area of child-parent relations is a dynamic area, which will undoubtedly lead to different assessments regarding the best interest of the child and not least concerning child maltreatment assessments in the light of the principle of evolutive interpretation of the ECtHR.[136] This has to be accounted for when developing and using a predictive tool based on AI.

[135] See e.g. *Tlapak and Others v. Germany*, application n° 11308/16 and 11344/16, Judgment 22 March 2018 (practices of caning within a religious sect).

[136] Rainey, Wicks, and Ovey, 2014, p. 73–78.

4.4 A preliminary outline of the legal issues related to AI tools for child protection

The design and use of AI-tools for child protection raises several legal issues that can be identified from the discussion of the AI driven tools that relate to the rights of the child. They pose problems that need to be overcome or dealt with.

To start, it is important to note that the rights of caregivers can also both directly and indirectly affect the child, which is why it is not entirely possible to apply a child centred approach without involving the family to some degree.[137] There can also be a question of maltreatment outside of the family by other adults or other children in the child's vicinity.[138]

The AI tools can result in the profiling of families with children based on for example racial, socioeconomic- and health status, which directly or indirectly targets the child. It has been shown that statistical methods can lead to wrongful outcomes since the correlations that they produce can be both exaggerated and irrelevant. Furthermore, a tool can be constructed for screening of large parts of a population, such as families with children, which can amount to mass surveillance that can be invasive not only for the parents but also affect the child in a negative way. This may be contrary to the right to respect for privacy and family life as well as protection of personal data and can undermine public trust, with the effect that parents as well as children may avoid seeking help from the authorities when in need.

The AI tools will likely include the biases related to their developers, which often can be related to race, gender, culture and socioeconomic status. This can especially be the case if the tool is designed to target only the part of the population that receives welfare benefits. It also risks cementing such biases into future decision-making. Consequently, there is a risk that such outcomes will amount to unlawful discrimination. Furthermore, such tools include the risk of excluding children at high risk who can be found in other socioeconomic groups in society.

Concerns have been raised relating to the opacity of the AI tools, the "black box problem", which can cause difficulties in understanding the reasons for an outcome that might serve as the basis for decision-making. In this context there is a conflict with the right to a fair trial in Art. 6

[137] Van Bueren 2018, p. 86.

[138] Ibid., p. 84.

ECHR, including the right to a motivated decision. More importantly, how do you contest such a decision legally and who will make sure that the child will be represented and by whom?

It is also important to note that the automated decision-making can be problematic if, in reality, there is no meaningful human involvement and oversight. If there is a risk that the staff rely too heavily on the outcome of an AI-tool and do not undertake any further controls, there is a risk that the child is subjected to an automated decision-making against the law and particularly Art. 22 GDPR.[139]

AI tools are never a hundred percent accurate when it comes to identifying child maltreatment, and the law requires a certain level of proof particularly regarding child protection measures that are by definition an interference in the right to privacy and family life.

Transparency problems might also arise if a state or local authority develops models together with private, for-profit entities. Access to information regarding the tools can be at risk, due to commercial interests and intellectual property rights, which in turn might be necessary information in a court of law if a decision based on the tool is contested. Moreover, the lack of transparency poses problems concerning trust and a sense of fairness for the child, youth and her or his care givers, which might lead the child to turn against society.

Last but not least, who will be accountable and held liable when an AI tool fails? And what reparations in regard to the child can be expected?

5 Final remarks with a view to the future

Tools using AI to counter child maltreatment may have the potential to enhance risk-assessments and serve as valuable decision-making support regarding child maltreatment. This certainly needs to be further researched. There are also other issues such as how the tools are supposed to be used, what procedures are elaborated in relation to the use of such tools, who will be qualified to make assessments using such tools and how will evaluations be carried out etc.? This certainly requires comprehensive regulation.

[139] Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and profiling for the purposes of Regulation 2016/679*, 3 October 2017 as last revised and adopted on 6 February 2018, WP251rev.01, p. 21.

As has been shown in this paper, there are several legal concerns that have to be addressed before designing, developing and using AI-tools to detect and prevent child maltreatment. It can therefore be concluded that there is a need to develop a children's rights framework for the use of artificial intelligence for child protection, a framework that can be included in a broader strategy regarding sustainable use of AI.

Furthermore, government AI-tools for the prevention of child maltreatment will need to be "future-proofed". The European Commission introduced a proposal for an Artificial Intelligence Act (AI-Act) in April 2021.[140] At present, it is not clear if or when the AI-Act will be adopted, but it is most likely that some kind of regulation will be adopted in a not-too-distant future. This will set further limits on the use of AI-tools concerning public child protection measures, especially regarding data quality and procedures for risk management. Considering Art. 5 of the AI-Act, most of the tools analysed in this paper, would probably be at risk of being prohibited since they involve social scoring (recital 17) and/or will probably be defined as "high-risk", due to the risk of harm particularly in relation to the fundamental rights of individuals.

[140] Proposal for a Regulation of the European Parliament and of the Council, Laying down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Brussels 21.4.2021, COM (2021)206, 2021/0106 (COD).