social sciences

MDPI

Article

Algorithmic Justice in Child Protection: Statistical Fairness, Social Justice and the Implications for Practice

Emily Keddell

Social and Community Work Programme, School of Social Science, University of Otago, Dunedin 9054, Aotearoa, New Zealand; emily.keddell@otago.ac.nz

Received: 5 August 2019; Accepted: 26 September 2019; Published: 8 October 2019

Abstract: Algorithmic tools are increasingly used in child protection decision-making. Fairness considerations of algorithmic tools usually focus on statistical fairness, but there are broader justice implications relating to the data used to construct source databases, and how algorithms are incorporated into complex sociotechnical decision-making contexts. This article explores how data that inform child protection algorithms are produced and relates this production to both traditional notions of statistical fairness and broader justice concepts. Predictive tools have a number of challenging problems in the child protection context, as the data that predictive tools draw on do not represent child abuse incidence across the population and child abuse itself is difficult to define, making key decisions that become data variable and subjective. Algorithms using these data have distorted feedback loops and can contain inequalities and biases. The challenge to justice concepts is that individual and group rights to non-discrimination become threatened as the algorithm itself becomes skewed, leading to inaccurate risk predictions drawing on spurious correlations. The right to be treated as an individual is threatened when statistical risk is based on a group categorisation, and the rights of families to understand and participate in the decisions made about them is difficult when they have not consented to data linkage, and the function of the algorithm is obscured by its complexity. The use of uninterpretable algorithmic tools may create 'moral crumple zones', where practitioners are held responsible for decisions even when they are partially determined by an algorithm. Many of these criticisms can also be levelled at human decision makers in the child protection system, but the reification of these processes within algorithms render their articulation even more difficult, and can diminish other important relational and ethical aims of social work practice.

Keywords: child protection; predictive analytics; rights; social justice; algorithms; decision making

1. Introduction

This article takes a critical perspective on the debates occurring in many nations in relation to the use of algorithms to assist with risk judgements in child protection contexts. Fundamental to discussions about the use of large, linked datasets to construct algorithms in this domain are the key ethical issues of fairness, transparency and accountability. Given recent developments, some scholars suggest this framework does not go far enough: that justice and rights are more effective concepts to analyse predictive tools, as they go beyond technical solutions, to consider broader social justice consequences (Gurses et al. 2019; Naranayan 2018). These debates should be of much interest to social work, given the professional commitment to social justice ideals in social work as a discipline, and the sharp uptake of predictive tools in child protection contexts where many social workers practice. This article discusses how the data used to create algorithmic tools affect their usefulness and create

important justice issues for both the families child protection systems work with and the social workers charged with their use. Justice issues are discussed with reference to statistical data fairness aspects such as the sample frame, the malleability of data points, and the feedback loop. Justice concepts are examined by connecting the potential for bias in algorithms with wider debates around rights for families in contact with child protection systems. Implications for transparency and implementation within the special context of the child protection system are discussed.

2. Setting the Scene: Algorithms in Context

The use of algorithms in child protection systems is expanding rapidly in the US and UK as well as other jurisdictions (Dencik et al. 2018; Whittaker et al. 2018). Alongside these technical expansions are important ethical and political discussions regarding their use in this context (Keddell 2015b; Dare and Gambrill 2016; Eubanks 2017). The child protection context is one already highly contested in terms of its aims, ideological underpinnings and institutional mechanisms. Whether a child protection system is based on a child protection-, child welfare- or child-focussed policy orientation, for example, will shape its philosophical basis, broad institutional structures, preferred priorities and methods of social work practice (Gilbert et al. 2011). In turn, these broad policy patterns intersect with political and economic structures, with a 'child protection' orientation finding an easy alliance with a neoliberal individualised approach to social problems and a residual state role (Keddell 2015a). Notions of rights are also contested in child protection, as the rights of children and parents have areas of convergence, as well as divergence. Consensus about the point at which they should diverge is often not clear in practice, as many studies of child protection decision making show (Benbenishty et al. 2016). Injecting algorithmic forms of decision making into this context adds a further dimension of complexity when considering justice and rights within a child protection system. As Veale and Brass (2019) note, the use of algorithms in public sector domains can brush over important political debates and the contested nature of policy aims. They note that "the literature on the governance of sociotechnical problems has similarly emphasised the intractability of 'unstructured' or 'semi-structured' problems where there is a lack of consensus around appropriate means and/or ends, and how participatory processes that open up rather than close down are required to socially reach more navigable issues" (p. 5). It would be difficult to find a social policy area with less consensus than child protection, where competing ideologies relating to the proper role of the state in family life, cultural considerations, and children's rights, needs and 'best interests' concepts are diverse and contested (Keddell 2017; Gilbert et al. 2011).

Implications for justice are made even more complex by the socio-technical context of use of algorithmic tools (Green and Chen 2019). There is not a single type of use, a single type of algorithm, uniform types of data, nor a single end user impacted by the use of algorithmic risk prediction tools in child protection. In terms of type of use, algorithmic tools can be used either to distribute preventive family support services, in child protection screening decision making, or in risk terrain profiling to predict spatially where child abuse reports might occur (Cuccaro-Alamin et al. 2017; Daley et al. 2016; van der Put et al. 2017). The type of algorithm selected categorises data in algorithm-specific ways to generate graded recommendations or binary flags and can include decision trees or regression methods amongst others, with varying levels of transparency or opacity. Data provenance or sources are from varied places depending on the national and local context, and have differing levels of representativeness, consent for use, ability for accurate linkage, biases, and 'explainability'. Children and parents involved in the child protection system have a set of complex and at times divergent rights and needs, complicating just who is considered the 'end user' and therefore who the algorithmic tool should be fair to. Finally, how an algorithm intersects with the multi-faceted social negotiations already underway in the child protection decision-making environment is important. How an algorithm is used determines its impact within the socio-technical institutional context that is a child protection organisation, yet discussions often focus on accuracy comparisons at the expense of considering the interacting nature of humans, social and institutional contexts, and algorithmic

tools (Green and Chen 2019). Tools are seldom used to automate decisions, but more usually are as an aid to human decisions, and can be made available to all or only some decision makers along the child protection decision making continuum (Allegheny County Department of Human Services 2017; Baumann et al. 2013).

Views on the use of algorithms in child protection can be polarised. As with any new technology or practice, there are important roles for both promoters and sceptics in the development of the debates. On the one hand, some argue predictive tools can contribute to the prevention of child abuse and neglect by efficient prediction of future service contact, substantiation or placement, through the triage of large linked datasets, drawing on more data than a human could rapidly and accurately appraise, and can select predictor variables based on predictive power in real-time (Cuccaro-Alamin et al. 2017). Particularly at system intake, when human decision-makers have limited information and time (particularly poor conditions for optimum decision-making), algorithms can quickly compute risks of future system contact (Cuccaro-Alamin et al. 2017). On the other hand, issues relating to class and ethnic biases in the data used, other sources of variability in the decisions used as data, data privacy implications, the issue of false positives, limited service user consultation and the lack of transparency of algorithmic processes are cited as serious challenges to the use of algorithmic tools in child protection, particularly where the recipients of services experience high levels of social inequalities, marginalisation, and lack of power in the state–family relationship (Keddell 2014, 2015a, 2016; Munro 2019; Eubanks 2017; Dencik et al. 2018).

While some tool developers have made inroads into responding to some of these criticisms, several issues remain unresolved (Chouldechova et al. 2018; Gurses et al. 2019). As Gurses et al. (2019) note, "addressing societal problems embedded in such computing systems may require more holistic approaches ... and they appeal to diverse theories, frameworks, and histories that challenge and expand the scope of FAT* studies" (Fairness, accountability and transparency). In what follows, I will discuss several aspects of these unresolved debates, specifically, the nature of the data used to construct such algorithms, the feedback loops algorithms rely on, and the contextual justice issues these create for both social workers and service users.

3. Predictive Tool Development

The use of number-based assessment tools in child protection contexts is not new. For some 30 years, various actuarial tools (that are in fact simple algorithms) have been used in child protection practice. The key aims were to establish consistency and the correct inclusion and weighting of specific risk factors, derived either from research or professional consensus (Gambrill 2005; Shlonsky and Wagner 2005). Actuarial tools have been the subject of debate. Key criticisms are that child abuse and neglect is inherently uncertain and is not directly amenable to prediction either by humans or statistically, as identified risk factors are neither necessary nor sufficient to predict abuse (Munro et al. 2014). Further critiques note that actuarial tools tend to reduce the professional discretion of social workers and do not include either the perspectives or cultural context of parents or children, nor the interpretive, social and relational elements of decision making (Gillingham 2011; Goddard et al. 1999; Munro et al. 2014). However, their accuracy in some studies are better than clinical prediction, while meta-studies show the majority are more accurate, though not all (Bartelink et al. 2015; van der Put et al. 2017).

Algorithmic predictive tools build on the actuarial tradition, arguing that they can improve on actuarial tools in several ways. They are able to draw on more variables derived from large administrative datasets, and weight them directly in relation to the outcome of interest. They can be updated with data in real-time or near real-time; they do not rely on a human to input data; and derive the predictive variables from the data itself, rather than relying on research or professional consensus to identify them (Cuccaro-Alamin et al. 2017). They can then be used to both direct limited resources to the most needy/risky families, or triage notifications to child protection services, serving utilitarian ideals of both demand management in a context of limited resources, and distribute fairly

based on need rather than more arbitrary methods of referral or child protection worker decision maker. Based on these logics, tools have been developed and are in use in numerous child protection contexts. One strand of development used linked administrative data to attempt to predict children at risk of future substantiation in the child protection system, originally for the purposes of selection for preventive, in-home visiting services in New Zealand (Vaithianathan 2012; Vaithianathan et al. 2013; Wilson et al. 2015). A later study in the same country trialled a similar tool at the intake centre of the national child protection service and compared the tool's accuracy to human decision makers—social workers—in that context (Rea and Erasmus 2017). Further research used a predictive tool to determine whether the predictions based on the child protection system data could also predict increased risk for hospital admissions and death by maltreatment (Vaithianathan et al. 2018). In the US, Allegheny county in Pennsylvania introduced a predictive tool in child protection screening—the Allegheny Family Safety tool (AFST)—to triage referrals at the point of notification to the child protection system (Vaithianathan et al. 2017). In Florida, tools were introduced by Eckerd Connect and its for-profit partner, Mindshare Technology, to predict child abuse harm, which was then adopted in Connecticut, Louisiana, Maine, Oklahoma and Tennessee and Chicago. Eckerd's tool, the 'rapid feedback safety program' tool was axed after a short time due to its inaccurate predictions. This latter tool rated children's risk of being killed or severely injured in the following two years, but the system was swamped with high-risk scores, including 4000 children deemed at 90% or higher of serious injury or death, while children who did experience serious harm were missed (Jackson and Marx 2017). In the UK, companies such as Xantura and others have been working with at least five local authorities to develop predictive tools to identify families likely to access services with preventive services before they present as high need (McIntyre and Pegg 2018).

In each instance, there have been concerns raised regarding issues such as data accuracy, the accuracy of the tool's predictions, profiling, stigma and data privacy, with varying levels of transparency regarding the tools from their progenitors. For example, a Freedom of Information Act request regarding the predictor variables of Xantura's tool was largely declined due to commercial sensitivity concerns (Sheridan 2018). The Allegheny tool has had much more transparency than others and considerable community involvement, a public technical report, and ethical and impact evaluations, but when asked to report the weighting of the actual variables used, Allegheny Department of Human Services declined to make this or the positive predictive accuracy of the tool—what percentage of those it accurately identifies at different risk levels—public (Eubanks 2018). In Aotearoa New Zealand, political concerns regarding the development of predictive tools in child welfare stopped further research on the tool's use and implementation due to concern about the experimental design (Kirk 2016). Concerns about the ethical issues associated with these tools have also been examined (Keddell 2015a, 2016; Dare and Gambrill 2016), yet there have been few examinations of algorithm use in child protection from the perspective of the growing movement on fairness, transparency and accountability, nor a consideration of the limits of technical solutions to the important justice issues they evoke (Chouldechova 2017).

What are the outstanding challenges to fairness, and the justice and rights issues they raise, when considering predictive algorithms in child protection? Key statistical biases in the data used to create child protection predictive models are as follows. Firstly, the lack of representativeness of the data sample frame, challenging the right to non-discrimination. The social malleability of the outcome an algorithmic tool is trained on, and the fundamental problems with how the algorithm 'learns' via its feedback loop both challenge the right to non-discrimination and equality of treatment, as they can result in people not being treated in a 'like for like' manner. They may also, through inequalities in the training data, reinforce classed and racialized inequity that is not reflective of actual differences in incidence. These claims are now examined.

4. Statistical Fairness and Social Justice

The concept of fairness has multiple definitions when used to analyse algorithms (Naranayan 2018). Definitions relate broadly to statistical fairness and social fairness or justice—but they are interrelated as discussed below. Statistical fairness issues include the sample frame (the boundaries defining the data sample), how feedback loops are constructed, and differences in predictive parity (how accurate the algorithm is for different groups). Social definitions of fairness are much broader, relating directly to social justice and rights. Moral conceptualisations of rights are relatively broad, while legal rights are more restricted and clearly defined. Rights-based definitions of fairness generally describe the right to equality of access, treatment and outcomes and the right to non-discrimination—that is, to not be unfairly targeted for restriction of rights. If rights are restricted, this should be proportionate to a person's harm or risk of harm, and non-arbitrary. Social definitions of fairness also encompass the right to due process—that is, the individual right to equality of treatment under the law and law-like systems, which is what the child protection system is. These are key aspects to consider when an algorithmic tool is used to define risk, as the potential for false positives within an algorithmic construction of risk (and other accuracy issues discussed below) can unfairly discriminate, and result in either an arbitrary or dis-proportionate response to the perceived level of risk. It can also challenge one's right to due process by resting on a risk of future harm potential and similarity to a group, rather than current individual behaviour (Hughes 2017). These processes therefore can create what McQuillan (2015) calls 'states of exception' when it comes to rights, with those calculated as high risk essentially losing rights claims due to their perceived 'high-risk' status (Keddell 2015b). Eubanks (2017) argues that in the child protection system due to the profound inequalities in the populations in system contact (often highly classed and raced), that algorithmic tools are more likely to be used because it is a 'low rights' environment to begin with, where few end 'users' have the power or resources to challenge tool use.

Finally, a social justice definition of fairness expects an algorithm not to reproduce or exacerbate social inequalities relating to group rights, such as those related to race, class or gender, a key claim of critics of algorithmic use in public services (Lepri Letouze et al. 2017; Barocas et al. 2017; Eubanks 2017; Keddell 2015b). In child protection systems, the rights of parents and children are both affected by algorithm use. Parents rights in particular are affected, and the emotive aim of protecting children from harm is often used to justify both data use without consent and the inevitable false positives (Dare 2013). Whether the child protection system is considered benign or punitive also affects views of rights considerations in child protection (Dare and Gambrill 2016). To be offered a voluntary, in-home support service or better housing has different ramifications than being investigated for child abuse. There are national differences in how child protection systems are constructed that shape the service offered in response. However, critical perspectives of child protection systems point out that in most Anglophone countries they operate as a key site of the reproduction of social inequalities, have both care and control functions and, while offering protection, can also promote normative parenting ideals mired in cultural and class specificities (Edwards and Gillies 2015). Furthermore, the mixed outcomes of child protection intervention for both children and their families point to the less positive implications of protection-oriented systems that rely on legal intervention within a broader residualist or neoliberal political and economic system (Gilbert et al. 2011; Featherstone et al. 2014).

5. Statistical Fairness and the Sample Frame

The first element of statistical fairness that relates to justice considered here centres on the sample frame, as this should capture the whole population it is attempting to predict an event within, or be a randomly selected, large subsample of that population. Without this, the algorithm "diverges from the population it attempts to represent" (Lepri Letouze et al. 2017, p. 7; Sloane 2018). This basic statistical principle ensures that predictions reflect accurate incidence base rates across an entire population and subpopulation groups. Without representational data of a phenomenon, an algorithm's predictions will become skewed, as it does not have an accurate picture of incidence, so it cannot find the predictor variables related directly to the phenomenon of interest (the outcome variable) (Barocas et al. 2017).

As it chooses predictor variables from the data it is provided with, skewed datasets may pick up confounding variables that predict the outcome it is trained on, but neither the outcome, nor the relationships between the predictors and the outcome of actual interest may be reliable.

This is especially pertinent in the child protection context, where proxies for child abuse in the child protection data algorithms are trained on may be only weakly related to abuse incidence across the population (Keddell 2016). This is because the sample frame is not representative of child abuse incidence, but instead reflects reported abuse. Much child abuse and neglect are never reported, and many families that are reported may be subject to both surveillance and personal biases. Via these processes, the sample frame available in data generated by child protection systems is not representative of the prevalence of child abuse across a population (Keddell 2015b, 2016; Daro 2009; McDonell et al. 2015; Swahn et al. 2006). For example, a study of a cohort of New Zealand children born in 1998 found that 10% were substantiated for child abuse within the child protection system, but a more representative longitudinal study in the same country found that while 10% of children had 'definitely' been abused, a further 27% 'probably had'. Even if only half of the 'probably had' group had been, that would be 23%, showing the higher rates of self-reported abuse compared to child protection system data (Danese et al. 2009; Rouland and Vaithianathan 2018).

In another example, a study examining the assumed intergenerational transmission of child abuse found that the greatest influence on the reported levels of intergenerational transmission was the role of detection bias as expressed in Child Protection Service reports (Widom et al. 2015). The extent of the intergenerational transmission of abuse and neglect as recorded in the system contact data "depended in large part on the source of the information used. Individuals with histories of childhood abuse and neglect have higher rates of being reported to child protection services (CPS)for child maltreatment but do not self-report more physical and sexual abuse than matched comparisons" (Widom et al. 2015, p. 1480). The way earlier reports to child protection systems affect later reports is another way in which child protection system data can become skewed. While an algorithm is likely to find earlier reports highly predictive of later reports, (see Wilson et al. 2015), the contribution of greater surveillance of families who have been reported before is not accounted for, and is unlikely to be acknowledged in an unthinking algorithmic process unable to understand these confounds (Keddell 2016). An algorithm views re-reports as unconnected by surveillance mechanisms.

If under-reported abuse was evenly spread across the population, there would be less concern about using child protection system data in predictive tools, as while some people would be missed, those identified would be an evenly spread subset of incidence. But there are substantial inequalities built into child protection system data that may not evenly reflect under-reporting (Bywaters 2015). For example, Swahn et al. (2006) compared self-reported abuse amongst detained youths with their child protection records. They found that official data generally "seriously underestimated the prevalence of maltreatment" (p. 415). Importantly, they also found that abused African Americans and girls were both more likely to have court records of abuse compared to Whites/Hispanics, and boys respectively. For example, 86% of White youths self-reported abuse, but only 12% had court records of abuse, 79% of Hispanics self-reported abuse, while just 7% had a court record. For African Americans, 83% self-reported abuse, while 19% had a court record. When converted to a ratio of self-report to court records, this ratio differed between groups, with 1:7 for whites, 1:11 for Hispanics, and 1:4 for African Americans, suggesting that African Americans are more likely to be reported for abuse than other groups, despite very similar self-reported rates of abuse. Inconsistencies between self-reported child abuse or incidence studies, and child protection system contact profoundly affect the ability of predictive tools to accurately identify those at risk of abuse, and lends weight to claims of racial and class-based discrimination built into such tools, as it may lead to some groups (for example, African-Americans) to be considered high risk by a predictive tool, and other ethnic groups to be calculated as low risk, when true incidence across these groups may be more similar than CPS reports suggest.

6. Issues in Predictive Parity between Racialized Groups—How Should Fairness Be Evaluated with a View to Justice?

Algorithmic prediction tools are used to generate risk scores, and at each risk level, the accuracy rates change—that is, the rates of positive predictive accuracy: what % of those identified have the outcome, as well as other indicators such as false positives, and true and false negatives (Chouldechova et al. 2018). One definition of statistical fairness in the use of predictive analytics is ensuring that the rates of true positives are equal between population groups, although this 'predictive parity' may change the risk level at which intervention of some kind occurs for each group, or change the rates of false negatives between groups (For example, the COMPAS debate over racial bias embedded in criminal justice tools was essentially over predictive parity (Whittaker et al. 2018)). Reducing one type of statistical bias may increase another. This means trade-offs are inevitable but, as discussed below, this is even more complex in child protection, as what is considered a true or false positive is mired in the social contexts they are interpreted in, producing many further confounding variables (Chouldechova et al. 2018; Whittaker et al. 2018). Further, the use of algorithmic tools takes place in a complex institutional and human environment where many factors contribute to decision outcomes, not the tool alone.

One important factor to know in order to judge the trade-offs between competing fairness criteria is the basic positive predictive accuracy of the model (the proportion of those identified as high risk that have the outcome of interest), together with what cut-off point recommendations will be generated at, and comparing them to what currently happens in a given context. It is only by knowing both predictive accuracy and comparing that to current practice that the inevitable 'weighing up' of harms and benefits can be done. The accuracy of algorithms in child protection is often reported as an 'area under the curve', instead of the more everyday understanding of positive predictive values. The few times positive predictive values are given in reports, they tend to be low. For example, one study shows 25% of those categorised in the top 10% of risk were true positives when measured over time (Wilson et al. 2015). A study of human decision-makers at the intake service of a child protection agency found that they were 60% accurate in their estimations of future harm, although the algorithmic tool used was 66% accurate (Rea and Erasmus 2017). Bartelink et al. (2015) in a meta-analysis of studies of decision-making tools in child protection found that the relative improvement of actuarial tools is around 10% improvement in accuracy (Bartelink et al. 2015). Their study also shows that the context of decision-making is important: in some studies, actuarial or predictive models are more accurate, in others clinical and predictive tools are about equal, and in a few, humans are more accurate, especially in complex decision contexts (Bartelink et al. 2015). These varied findings show the importance of actual comparisons with humans in context, rather than vague claims based on non-comparable research that 'algorithms are better'.

As mentioned, comparative accuracy studies or evaluations are rare. This because they are difficult to conduct, can be controlled by commercial interests, and are sensitive to public critique. One of the few to conduct a substantial independent evaluation of a child protection prediction tool and make it public, compared accuracy rates before and after implementation of the Allegheny Family Safety tool (AFST) (Goldhaber-Fiebert and Prince 2019). The tool does not mandate decisions but is used in conjunction with human decision makers as a decision support tool at the point of notification. The tool scores notifications from 0–20, and those scoring highly must be discussed with a supervisor. The case is then screened in or out. The evaluation was premised on a subsequent 'case open' decision by a case worker to judge whether the screen-in decision had been correct, and on a time limited re-referral rate to see if screen outs had improved. A caseworker is likely to have collected much more depth and quality of information and can weigh up the complex ethical issues at stake, so their judgement for these reasons are a good way to evaluate whether the algorithm-assisted decision was correct (Bartelink et al. 2015). On the other hand, the potential for bias still remains in the human decision and, as discussed below, particularly when considering racial equity, relying on a potentially biased human decision to test the accuracy of a potentially biased algorithmic one appears tautological. A recent

study, for example, found that humans using an algorithm to inform their decision still showed bias when combining it with their own judgement process (Green and Chen 2019).

Nevertheless, overall, the evaluation found an improvement of those children 'screened in' who then had a case opened, of 2.9%, from 43.7% to 46.6%. Of interest is that the improvement is small, (as were all changes) and that even with the tool, 53% of screen-ins are still considered false positives. The evaluation also showed the tool led to a small decrease in the accuracy of screen outs, down to 72.3% from 73.9% (based on those re-referred within a 6 month period), showing that those who were 'screened out' using the tool had a higher chance of re-referral, than before its introduction. On the one hand, it could be argued that the conclusion the tool reduced screen out accuracy is somewhat simplistic, as many community factors can affect rates of re-referral: policy changes, changes to the provision of preventive services, changes in poverty levels, and changes in social cohesion (Klein and Merritt 2014). On the other hand, a key claim of algorithmic tools is that they are able to reflect these rapidly changing social and organisational conditions in order to make predictions. This issue points to the difficulties of evaluating predictive tools in context, when so many confounding factors affect both algorithmic and human decision-making, as well as highlighting the limits of the data to begin with: without all relevant data, the tool cannot account for it in its computations.

The evaluation also measured the effect of the AFST on racial disparities. The evaluation shows that for those children screened out, the tool had no effect on the chances of re-referral for either Black or White children (p. 21). For those screened in, the percentage of Black children who were first screened in and then had a 'case opened' on them remained the same before and after the tool's introduction, at 47%, while the percentage of White children with a case opened increased from 39% to 46%. So while disparities between the groups of what could be considered 'true positives' reduced, this was because of a moderate increase of 7% in the perceived accuracy of White child screen-ins, not an improvement in the accuracy of Black child screen-ins, (who are in this case the protected class), nor of a decrease in the disproportionality of Black children (that is, the comparison to their population rate). As mentioned, this method of measuring accuracy relies on a human decision that could be subject to bias—that is, to 'open the case'. This is circuitous reasoning, as the tool's introduction is premised on its unbiased accuracy, yet a human, potentially biased, decision to open the case is used to judge it. While the case worker is not supplied with the score as a way to reduce the confirmation bias aspect of this process, nevertheless, the case worker knows that the case was screened in, in order for them to now be investigating it. This is an issue for the use of subsequent 'case opens' for all findings of the evaluation, but particularly so for evaluating changes in racial disparities.

Is there a more objective measure to add to this 'case open' measure in order to evaluate the claim that the tool 'reduces racial disparities'? An objective measure required to assess racial disparities is any change to the proportion of Black and White children notified, who were then screened in. This did not appear to be reported in the evaluation. This objective metric—how did the tool affect the proportion of screen-ins, by race—was absent. This example highlights the many nuanced issues of comparing racial groups when evaluating algorithmic tools, and in particular, what counts as equality of outcome.

The tool's evaluation raises the question: how should 'true positives' be measured when considering racial justice questions? If a subsequent 'case opening' is equated with a 'true positive', then the AFST tool reduces racial disparities by improving the accuracy of white child 'case openings'. But if equality is reducing the disproportionality of Black children being screened in compared to their population share, or disparities between White and Black children's rates of screen in, then this tool's effect is not known. But nor could either of these measures be taken solely as evidence of success—in order to consider that reduction successful, one needs to know the incidence of abuse across the population (see above). Reducing protection for Black children is an equally poor outcome, and changes in screen-ins could reflect increasing community need/incidence or bias in referrers. However, as above, algorithmic tools claim to be able to account for all these factors despite the limits of the data used. How an algorithm crystallises and reproduces disproportionality remains an

enduring issue, reproducing the disparities already within the wider interlocking child welfare systems, obscuring the ways they may be shaped by bias, and providing them with a veneer of objectivity. Interrogating these assumptions highlights not the creation of inequalities in child protection systems by algorithmic tools, but the reproduction and reification of existing ones.

7. The Social Production of Data and the Feedback Loop

In addition to the lack of representativeness of the sample, other data problems can also introduce bias and arbitrariness into predictive tools, challenging rights to equality of treatment. The outcome variables that tools are trained on are usually one of various decision points on the decision-making continuum (Baumann et al. 2013). These include decisions to refer, made by various people outside of child protection systems; decisions to substantiate abuse, made by child protection workers within child protection systems; decisions to remove children, usually made by a judge after a social work recommendation (Chouldechova 2017; Keddell 2016). Each of these decisions can be affected by factors other than the direct abuse or risk of abuse of a child that spring from the social context in which they are made, creating variability in responses to children in similar circumstances, and therefore the data recorded about them (Baumann et al. 2013; Keddell 2014).

For example, a recent study in England found that there were significant regional variations in care proceedings in the court system, suggesting different courts have different practices and thresholds for removal (Harwin et al. 2018). A study in New Zealand found that the site office of the national child protection service was the fourth most predictive variable out of 15 for abuse substantiation, even after other variables had been controlled for, suggesting something about the site office culture itself is at play (Wilson et al. 2015). Another UK study found that an 'inverse intervention law' seemed to be operating, where poor children in neighbourhoods surrounded by a larger less deprived neighbourhood were much more likely to have contact with the child protection system than equally deprived children surrounded by a highly deprived neighbourhood (Bywaters et al. 2015). Others point out that decisions to notify, substantiate and investigate child abuse can be shaped by multiple factors such as the values and beliefs of the social worker, experience, role type, perceptions of risk, available resources and professional or institutional cultures (Bywaters et al. 2018; Davidson-Arad and Benbenishty 2016; Fallon et al. 2013; Fleming et al. 2014; Fluke et al. 2016; Keddell 2014, 2016). Variations in specific ethnic groups' contact rates can be particularly affected by both bias and service factors. Indigenous children's system contact in Canada was more related to the variable provision of local culturally appropriate prevention services, than differences in case characteristics (Fluke et al. 2010). Māori children in New Zealand were perceived by social workers as more risky than non-Maori children in the same situation (Keddell and Hyslop 2019). A Spanish study found no bias related to ethnic group or socio-economic status (SES) in caseworkers, but in students about to become social workers, physical abuse in low SES families was perceived as more severe than in mid SES families, suggesting experience may influence bias in caseworkers (Arruabarrena et al. 2017). In each of these examples, statistical prediction tools that use decision points from the child protection system as the outcome the tool is trained on will reflect, and lend reification to, the many elements that contribute to system contact that have little to do with child abuse, and more to do with inequalities, individual practitioner values, location, decision-making variability and service supply and demand factors (Keddell 2014; McLaughlin and Jonson-Reid 2017).

Importantly, in none of the examples given did system contact reflect only the case characteristics of the actual families involved—all were shaped by other factors. These factors mean that hardly any of the decisions that become data points in predictive tools in the child protection context reflect objective outcomes or, rather, outcomes that represent what is assumed about them. Is a child protection office in a particular location more likely to substantiate cases of domestic violence (DV) as child abuse than another? That means that data derived from that office are likely to assign a higher risk score to children witnessing DV from that area in the future, and not those in the same situation in other areas. Is there classed or racial bias in the populations notified and/or investigated for child abuse? Then the use of, for example, parental contact with child protection systems (the third most predictive variable

in one study) (Wilson et al. 2015), or parental contact with the criminal justice system, is just as likely to reinforce existing ethnic inequalities in the system rather than reduce them, as it will erroneously identify ethnic or indigenous minorities as high risk compared to people from other groups.

In Aotearoa New Zealand, when the adults today were children, a ministerial inquiry found that the cause of ethnic disproportionality in New Zealand's child welfare system was due to "forms of cultural racism ... that result in the values and lifestyle of the dominant group being regarded as superior to those of other groups, especially Māori" (p. 9). Many US and UK examples are similar. Certainly, there is evidence of continuing bias that affects who is investigated and how similar parenting behaviours are viewed depending on the parent's socioeconomic and ethnic status (Roberts and Sangoi 2018; Wexler 2018). These processes show the embedded nature of ethnic or racial bias in the data algorithmic tools tend to draw on. As Whittaker et al. (2018) point out in the criminal justice context, "most assessment systems include several risk factors that function as proxies for race. One risk factor that is often used is "parental criminality" which, given the long and well-documented history of racial bias in law enforcement, including the over-policing of communities of color, can easily skew "high risk" ratings on the basis of a proxy for race." (p. 13). These same issues are pertinent to child protection data.

These data issues can skew an algorithm not only when it is created, but over time as data are fed back into it. If those data are inaccurate, then the algorithm's opportunity to correct itself is lost. Chouldechova notes that "A key challenge is that we do not get to observe placement outcomes for a large fraction of cases that are screened out. This makes it difficult to assess the accuracy of the models on the full set of referrals, not just those that were screened in" (Chouldechova et al. 2018, p. 12). Where cases are screened out, but abuse occurs again and is not re-reported, the algorithm learns it is correct, even though it was not. It adjusts its predictor variables accordingly. Likewise, other types of data that could significantly change a child's risk levels are not captured in administrative data, such as improved or decreased levels of informal social support—a key factor in child abuse development (Rostad et al. 2017). A parent might seek formal help from private providers for something that directly affects the likelihood of abuse, yet those data would not be included in the algorithm, nor will changed living arrangements or household membership, all of which could affect true risk (Eubanks 2017). Aradau and Blake note this, commenting that "Debates about big data problematize exactly the limitations of traditional statistical procedures, which ... do not capture the detailed relationships between individuals and groups as they exist and change in particular situations" (Aradau and Blanke 2017, p. 378).

It is these subtle ways that both bias and arbitrariness—equally destructive to notions of fairness—become 'baked in' to administrative data that predictor and outcome variables, and feedback loops over time, are drawn from. This results in consistently over-assigning risk to some people while understating it for others in algorithmic computations in child protection and assigning risks to individuals that may have little to do with their individual true 'risk' level. As Boyd states "Racism, sexism, and other forms of bigotry and prejudice are still pervasive in contemporary society, but new technologies have a tendency to obscure the ways in which societal biases are baked into algorithmic decision making" (Boyd 2014, p. 56).

This lends weight to claims that predictive tools result in the 'poverty profiling' of individuals that could impact negatively on their rights to fair treatment, especially given the marked socioeconomic inequities in child protection system contact (Eubanks 2017, 2018). The child protection context is not a benign one—while protecting children from abuse, unwarranted investigations are experienced by families as stressful, intrusive and stigmatising, while in colonial contexts, they operate as a key site of the ongoing reproduction of inequalities for Indigenous peoples (Healy et al. 2011; Keddell and Hyslop 2019). Choate and Lindstrom (2017), for example, examine the use of parenting capacity assessments that contribute to legal intervention in child protection systems. They find that these assessments rest on euro-centric definitions of family, have not been normed on Indigenous populations, and do not take into account the poverty, poor service access and intergenerational

trauma that frame the socio-political contexts of many first nations families. In light of how negative perceptions of Indigenous people can get into data that is then used as the basis of state intervention, there are many movements by Indigenous people that aim to reclaim data sovereignty, arguing that data about Indigenous people should be controlled by them, not the state (see the US Indigenous Data Network https://usindigenousdata.arizona.edu/about-us-0 or Te Mana Raraunga https://www.temanararaunga.maori.nz for examples).

Of increasing interest is who is missing altogether from data sources used to train algorithms. When there are groups who have hardly any contact with administrative systems, the ledger becomes even more unbalanced. The resulting algorithm will not only over-identify poorer and Indigenous or ethnic minority people but will miss risk amongst more affluent populations (Eubanks 2017). The relative intrusion this disparity creates in the lives of those least able to seek redress compared to the escape from scrutiny of wealthier populations challenges rights to non-discrimination, as some people are ascribed high-risk status while others escape surveillance altogether. Big administrative datasets essentially 'go easier' on some people compared to others. These examples highlight the problems in the sample frames used to generate data, and the potential for feedback loops to exacerbate this over time.

8. Consistently Biased?

Others argue that despite all this, the sheer amount of data allow for more consistent predictions and are more accurate as the algorithm self-selects the predictor variables—that is, which variables predict the outcome. Cuccaro-Alamin et al. (2017), for example, argue that "PRM (predictive risk modelling) as an approach is inherently more consistent than other risk assessment procedures. Variable selection, although limited by available data, is mathematical and there is no arbitrary selection of predictors ... they are learning models that continually adjust to new relationships inherent in the data" (p. 293, brackets mine). There are benefits to this, as theoretically speaking, unknown influences, if able to be captured in the data, could be identified by this process that may not be discernible to a human. Yet the limits of the available data are marked as the discussion above shows. Without accurate incidence, and other biases and non-relevant influences on data points, an algorithm's self-selection of variables can result in spurious associations and predictions that reflect bias of some kind, as in the example of parent's contact with child protection systems as a child discussed above. While an algorithm may indeed be a more consistent decision maker, consistently biased or skewed recommendations remain problematic. The algorithm self-selecting predictor variable means errors are also difficult to track, leading to reduced transparency.

Some argue that a predictive tool's function is only to predict, not suggest causes. What does it matter if it is predicting based on unknown associations? When the data are so varied (as above) and the associations generated unknown, the function of the algorithm may be calculating quite spurious associations and replicating 'ecological fallacies'. The invisibility of the function requires examination, including the underpinning assumptions that imply causality (Rowe 2019). A predictive tool may not *claim* causality, yet it is used in child protection as if it does imply causality, because high-risk scores generate information about specific families that can lead to intervention on those families (Pearl and Mckenzie 2018). This means that the use of an algorithm can identify families whose risk of child protection system contact may have little to do with their specific family relationships or behaviour, but results in an individualistic response that assumes there is something deterministic about them. As Mittelstadt et al. (2016) note: "Causality is not established prior to acting upon the evidence produced by the algorithm ... Acting on correlations can be doubly problematic ... Spurious correlations may be discovered rather than genuine causal knowledge ... Even if strong correlations or causal knowledge are found, this knowledge may only concern populations while actions are directed towards individuals" (p. 5).

Where data expected to affect predictions do not, this should cause developers to ask more holistic questions about why this is the case. For example, health data were included in one study, but as

they did not increase accuracy, were withdrawn from the model (Wilson et al. 2015). This could have indicated that the predictive tool was reproducing child protection system surveillance patterns (or something else), rather than predicting abuse across the population, as parental and child health issues (prematurity, poor mental health, substance abuse), are well known predictors of child abuse in the research literature (Munro 2002). Chouldechova et al. (2018) note this problem, stating that when considering why certain variables are highly predictive of the outcome variable in child protection algorithms, we "may nevertheless fail to offer a satisfactory answer to why more penetrating than that the particular values of input variables combined to produce a high-risk prediction. One may be able to understand the risk factors involved and how they combine in the model, but the models have no claim to being causal. The overall utility of such an understanding may be quite limited" (p. 11).

9. Improving the Feedback Loop or Reducing Justice?

This discussion so far explores statistical issues relating to fairness, highlighting the technical complexities of data production and evaluating tools, but as many authors highlight, broader discussions of justice must consider algorithmic use against a wider set of ethical and political concerns (Dencik et al. 2018; McQuillan 2015). Naranayan (2018) argues we need more focus on "connecting technical issues with theories of justice". When an algorithm has low accuracy related to problems with feedback, the technical solution focusses on improving the range of data (also called 'tech solutionism' (Gurses et al. 2019)). The larger justice issues around improving the data in the child protection context are not discussed. For example, after the cancellation of the Eckerd tool in Chicago child protection which overestimated risk, one of the solutions proposed was to reverse the current practice of expunging unsubstantiated cases from official records (and data) (Jackson and Marx 2017).

From a technical perspective, removing unsubstantiated cases introduced statistical bias into the data by excluding cases that were investigated, abuse was not found, but nevertheless may have had other similar characteristics as those that were. Their exclusion damaged the learning ability of the algorithm, as it reduced its ability to make fine-grained, more accurate differentiations between different cases. The fix proposed was to reverse expungement. But in some states, expungement exists to protect people's right to future due process, so that an earlier child abuse investigation does not taint future notifications and judgements. This method is used in some states to reduce racial and legal biases (Jackson and Marx 2017). Without data on unsubstantiated cases, an algorithm cannot learn whether it was correct in its earlier prediction, but to seek increasing amounts of unsubstantiated cases included in data could exacerbate the very biases predictive tools claim to reduce, by reproducing the racial and classed biases in referrals, even though they may be unfounded (Boyd 2014). This important justice issue should not be subsumed by the need to increase the accuracy of an algorithm.

This issue reflects the incidence issue described above. For those cases never notified to child protection services or notified but not investigated, neither of these situations means abuse has not occurred, just that no one saw and reported it. It is for these reasons that the 'moneyball' analogy often used to describe algorithmic tools in child protection is a bad one (Riley 2018). In baseball, all outcomes are easily observable and categorically definite. Player x scored a home run, and everybody saw it. When it comes to child abuse, neither the action itself is easily defined, nor who sees it. A child left in a bath slips and hits their head getting out. The parent was distracted by another child. Is this neglect? And what if no one saw it and reported it? Did it happen? Not in the data it did not. And what about if this situation occurred in a poor neighbourhood as opposed to a wealthy whiter one? Will it be viewed differently, and hence get into the data differently? Attempts to improve data lead to increasingly intrusive surveillance and challenges to legal equity. This means that while some claim a predictive model helps track and correct error rates, in fact error rates in real incidence are fundamentally unknown (see Dare and Gambrill 2016).

As is described above, attention to precisely what feedback the model is referring to in its selected outcome is important and challenges the ability of the model to learn accurately. It not only does not 'see' some abuse, the lack of consensus of the definition of abuse creates further distorted data,

as decisions to investigate or substantiate abuse can be highly variable (Doherty 2016; Munro 2002). In the classic Dawes et al. (1989) study of predictive abilities, the three outcomes they used to compare human and actuarial predictions were: major mental illness, the detection of brain damage, and the prediction of survival times relating to illness. All three, or at least the last two, are objective, concrete outcomes unaffected by earlier predictions about them occurring. Neither of these criteria apply to the child protection context, where the outcome is ill-defined and earlier decisions may affect later ones, as earlier substantiations may lead to greater surveillance, further interventions, and a greater likelihood of substantiating further referrals (Keddell 2016). This means that the future recurrence of the outcome required to 'train' the algorithm so it can improve on its predictive accuracy is sullied in the child protection context—it is trying to predict an outcome that is socially malleable and affected by the system that is at once intervening on it and recording data about it, creating confounding variables.

These inaccuracies in the feedback loop make it difficult for a predictive tool to become more accurate over time. As Gambrill and Schlonsky note in regards to actuarial tools: "The actual risk of recurrence cannot be explored in the absence of intervention by child protective service agencies. Given these limitations, obtaining an accurate base rate of maltreatment is probably impossible . . . Further, discovering the false positive rate is almost impossible. Once the risk has been responded to (i.e., child welfare services are provided) the likelihood of recurrence of abuse in the absence of intervention cannot be determined." (Gambrill and Shlonsky 2000, p. 823).

The damaging nature of the effects of poor feedback can be understood as an accuracy effect, and a justice effect. The distorted exposure to referrers of some groups of people relative to others, and other distortions in incidence reduces the effectiveness of the feedback loop by feeding back into the algorithm those cases that are likely to be a small proportion of child abuse cases across a population, and a skewed proportion of those total cases, overestimating the predictive power of factors such as low income, race and previous service involvement. Presented with only a proportion of future events as feedback, the algorithm then will correct its weighting of predictor variables, and this may reinforce not only inaccurate predictions, but may reproduce, not reduce, the biases within child protection systems. In conclusion, the poor sample frame and feedback issues create significant problems for the use of predictive tools in child protection. Now, I will turn to the implications for social workers using these tools and comment on rights implications.

10. Implications for Practice: Social Worker Responsibility and Family Participation

The use of predictive tools presents a number of implementation issues affecting both social workers and the 'users' of child protection services. One implementation issue for social work practice is the relative responsibility of the social worker for the decision made. Social workers have responsibilities to ethical principles and codes that assume they are solely responsible for the decisions they make, and they may also hold legal, statutorily defined responsibilities within child protection practice. For example, it is a 'social worker' in the Aotearoa New Zealand legislation who must 'form a belief' that a child is in need of 'care and protection' before legal proceedings can be undertaken (s14, New Zealand Government 1989), and decisions should uphold social justice and human rights obligations (International Federation of Social Work 2014). At the same time, within child protection there are numerous examples of social workers attracting blame, particularly for child deaths, even when there are extensive wider system failings (Lees et al. 2011; Munro 2011). The importation of new public management methods, within neoliberal economic and public service environments, heightens this individualising of blame, as notions of individual responsibility combine with an emphasis on audit and accountability, leaving social workers vulnerable to high levels of personal responsibility (Healy 2009).

In the context of algorithmic tools that may contribute to or even mandate decisions, a social worker who relies on such a tool, but cannot explain, nor understand its inner 'black box' workings, is vulnerable to approbation from professional regulators, managers within the system and families they are working with. They may be construed as the 'human in the loop' essential to limit the unrestrained

governance by the algorithm or retain human oversight. But in a high stakes environment and high blame environment, the opposite could also occur—the human actor may defer to the algorithm (Rahwan 2018). In practice, this can leave a social worker vulnerable to either excessive blame if they did not follow the risk score, or captured in a 'moral crumple zone' if they do. Elish (2019) describes a moral crumple zone as "how responsibility for an action may be misattributed to a human actor who had limited control over the behavior of an automated or autonomous system. Just as the crumple zone in a car is designed to absorb the force of impact in a crash, the human in a highly complex and automated system may become simply a component—accidentally or intentionally—that bears the brunt of the moral and legal responsibilities when the overall system malfunctions" (p. 20).

Studies of child protection workers' reactions to algorithmic tools illustrate some of these issues. While most practitioners surveyed as part of an Aotearoa New Zealand trial of a tool at intake were open to the predictive tool being used, a significant minority viewed the tool as a labelling device too focussed on risk while minimising protective factors. They were concerned that the data used did not contain accurate information, pointing out that "some information is not able to be put into the model even though it may include the most accurate information about a child or young person (and) some information is reported in the CYF database without being verified" (Rea and Erasmus 2017, p. 83). These concerns echo the more abstract concerns about data discussed above. In the Allegheny county trial, a quarter of practitioners using the tool chose to override its recommendations, suggesting some level of discomfort with the tool's scoring function, but also suggesting workers may not feel as pressured by the tool as I have suggested above (Chouldechova et al. 2018). The latter evaluation also showed that the tool did not improve consistency between case workers, suggesting many exercised discretion over how much they used the tool's recommendation. In a study by Bosk (2018) child welfare workers felt that the tendency of statistical tools to operate on the basis of the presence or absence of risk factors without an understanding of how they interacted in the lives of specific individuals led to an overestimation of risk in unfair ways, penalising people for demographic factors such as having more than three children, or children born close together. Thus, some frontline practitioners who often have the best 'practice-near' experience of the system, have reservations about data quality, and how the data shapes practice decisions (White et al. 2009). An ability to critically reflect on the tool may protect against the possibility of a tool becoming too powerful, and is a reminder that practitioners often utilise discretion even in highly constrained systems. How tools are implemented and managed are crucial to assisting workers to maintain ultimate control for decisions made.

There are also ethical issues of predictive tool use relating to families involved in the child protection system. Families should be able to expect decisions to be explained to them and have input into decision-making processes. Diminished input into decisions challenges many best practice and legal requirements to consider parent and child perspectives and include them in decision making (Healy and Darlington 2009). This lack of explainability affects social worker's ethical mandates. Where social workers themselves are unable to explain decisions to families, this lack of transparency may cause a sense of moral injury: where social workers know that transparency around decisions, and the implied ability for parents to contribute to and participate in decisions that affect them, is important, but are unable to implement this in their practice (Fenton and Kelly 2017). The marked lack of family involvement as stakeholders in tool development is also a justice issue. There is a call for citizens' councils or panels as ways to include a citizen voice when algorithmic tools are being developed in public service contexts. Yet the position of parents particularly as potential 'abusers of children' is a powerful unspoken reason for their rights as mandated 'users' to be diminished in both data consent and participation discussions. As people involved with child protection services come from the most marginalised populations, and due to the mandated nature of the 'service', the need for this inclusion is arguably even stronger than for other public service users (McQuillan 2018).

11. Human Rights and the Individual—The Right to Reasonable Inference, Consent and Relational Practice

Social work remains committed to respect for persons and human rights, as well as humane and relationship-based approaches to practice (International Federation of Social Work 2014). How are these aims affected by the limits of the data as described above? Algorithms rely on a process of grouping people according to their statistical similarities to others across a population, based on available data. Most legal conceptualisations of human rights, however, are intimately tied to the individual, and most nation states have undertaken international covenants to protect the rights of individuals. It is clear from the limits of the data in the child protection context that this right requires special protection in the context of the known biases in the child protection data algorithms draw on; the distal and incomplete nature of available data to accurately depict and predict human behaviour; and the 'networked' nature of biases in the data. Other adults in a household in addition to parents may also be risk scored in the prediction process. This can also exacerbate racialised bias, as people's personal networks are likely to reflect histories of racialized disadvantage and poverty, leading to 'networked bias', where parents may be considered high risk due to family members, neighbours or community location (Madden et al. 2017).

The conflation of personal with group risk challenges one's right to be treated as an individual in legal and law-like systems, where due process relies on individual consideration before the law (Barocas and Selbst 2016). This challenges the right to non-discrimination for individuals, as they are judged based essentially on their statistical similarities to others. While the General Data Protection Regulation (EUGDPR EU General Data Protection) enshrines the right to an explanation of algorithmic decisions, Wachter and Mittelstadt (2019) propose an alternative—a right to reasonable inference when using algorithmic scores. It is clear that given the problems with data in the child protection context, anything beyond the most tentative inference is highly contestable, and the 'inference' of high risk may be patently unreasonable. Given the power differential between the child protection system and many people coming into contact with it, their general observation may be exacerbated: "individuals are granted little control and oversight over how their personal data is used to draw inferences about them" (p. 3).

This control is heightened in contexts where there has been no or little consent to the use of data for this purpose. Some ethicists argue that the imperative to protect children from harm should attenuate data privacy concerns, especially when data or information is already shared without consent after a notification has been made to CPS. The highly emotive nature of the politics of child protection encourages this logic. However, the risk of extreme harm is very small, and false positives are high, while the size of the data net is vast, capturing many people for whom no harm will ever be found (the majority of reports in most countries). In this context, families have a right to know at least how their risk score has been interpreted, and the right to challenge it if they feel it is inaccurate. As Drefuss and Chang (2019) provocatively argue, in a 'no consent' or forced consent (to data sharing) environment, the right to challenge and correct is heightened. But whose knowledge is deemed 'correct' in such circumstances, and who controls perceptions of accuracy? The hierarchies of knowledge implicit in algorithmically produced forms of knowledge is that they are objectively correct and uncontestable, while service users' own views about their lives derived from lived experience is relegated to a lesser value, as are any other risk/protective factors not available in the data, nor social worker's own 'guilty knowledge' (that related to values, relationships, emotion and care) (McQuillan 2017; Weick 1999). This 'machinic neoplatonism' can evade due process considerations due to the opacity of algorithmic functions, and "appears to reveal a hidden mathematical order in the world that is superior to our direct experience" (McQuillan 2017, p. 1).

Along with knowledge, relationships themselves are also affected by the 'thoughtlessness' of algorithmic tools. The use of algorithmic tools adds a mechanistic and distant tone to practice relationships and prioritises the task of predicting future harm. In child protection, however, the prediction of future harm is only one of a variety of aims. To actually address the risk of harm, rather

than assess it, an engaged and collaborative relationship based on genuineness and trust is needed (Cameron et al. 2013; Spratt and Callan 2004). If people are aware that they have been identified as 'high risk' via predictive tools, perceptions of judgement and stigma may damage the quality of this necessary relationship (Spratt and Callan 2004).

12. Considering the Counter-Argument: Problems in Human Child Protection Decision Making

It is important, however, to note the mirroring of many algorithmic problems in human decisions in the child protection context. Many system inequalities reflect wider socio-structural conditions. Human decisions can also be arbitrary, biased and subject to institutional processes that lead to variations and inequities in decision outcomes (Keddell 2014). Humans can also rely on invisible heuristics and pattern matching that can lead to decisions based on a cognitive categorisation process, rather than a humane, relational approach that considers each individual person (Vedder 1999). Heuristics can become similarly biased when based on experiences that reflect a 'skewed sample' of particular kinds of cases compared to the general population. Other types of discretion-based assessment tools can be introduced to practice with little or no user involvement, direct the focus of practice in perverse or unintended ways dominated by risk, and can lack participatory processes for adults or children. Nor can Social Work claim that social work practice always embodies relationship-based ideals imbued with care, emotion and justice—social work's history is of both care and control, operationalizing many oppressive state agendas (Margolin 1997). The issue with algorithms is that due to the issues above, they replicate these problems, and through the appearance of objective 'science' and the difficulties in explaining them, add another layer of mechanisms that exacerbate rather than address these known injustices.

This leads to a consideration as to whether data can be developed to a point where at least the sampling issues could be corrected. In order to provide incidence data, an extreme level of surveillance would be required, as most abuse takes place in the home. Efforts to create large-scale population-wide databases of children have failed on this point. For example, the 'contact point' database in the UK was established in 2004 in the wake of the Victoria Climbie tragedy (a child who was killed despite many professionals being aware of her ongoing injuries and abuse) and aimed to contain information about all children in the UK. It was set up to try and improve data sharing but was shut down in 2010 due to concerns about both parents and children's rights to privacy, the increase in family surveillance and functionality of the database. The example of contact point highlights issues of class and surveillance more generally. The level of surveillance, data linking and targeting that results from these tools only gain traction because they initially affect people from low socioeconomic backgrounds in 'low rights' environments. When they are applied to middle class people via 'whole of population' approaches, there is often outcry (Eubanks 2017).

Another idea for improving data is to focus on more objective outcomes than those found in child protection agency data, such as child injury hospitalizations (see Vaithianathan et al. 2018). This removes the subjective nature of the outcome variable, and health data are becoming more complete in relation to the whole population, at least in countries where there is a national health system (as opposed to private providers who do not share data (see Eubanks 2017). While some have suggested child deaths, these are often in such small numbers that prediction is not possible, and to combine them with hospitalization data muddies the waters (Vaithianathan et al. 2013). Nevertheless, a focus on hospitalisations may correct the sample frame, improve feedback loops, and reduce some exposure surveillance effects, although the 'spurious correlations' issue discussed above still remain. Another way to consider research development is to focus on the counterfactual, for example focusing on predictive models trained on 'unsubstantiation' as Rodriguez et al. (2019) did, in order to see if protective factors could predict it. While they found accuracy was similar to other tools, they also found that system- and definitional-related factors—rather than family-related factors—were most predictive (such as abuse type, number of days between notification and investigation and if the child

was already in foster care). This suggests that the issue remains that some factors unrelated to real harm to children skew the data in significant ways.

13. Conclusions

Evaluating algorithmic tools in child protection by combining technical conceptualisations of fairness with social justice perspectives leads to a number of troubling conclusions. Without a database that reflects incidence, the racial and class disproportionalities within child protection system contact are likely to reproduce inequities that relate as much to surveillance biases as they do to differences in true incidence. Poor and ethnic minority families will therefore be subjected to higher rates of state intervention than real disparities in rates, while other children at risk may be incorrectly assumed to be low risk. Other elements of variability relating to poor outcome definition, and the inevitably messy context of decisions that become data points in databases may heighten inaccuracies and biases. Algorithmic tools can produce ecological fallacies, leading to spurious variable selection and prediction that reflect system factors rather than actual incidence risk. Statistical scores relating to group similarities are used to inform an individually focussed decision that may not reflect that specific person's risk level, but population level risk. As the child protection system is a law-like process with significant implications for people's lives, people should instead have the right to be treated as an individual and have decisions made about their lives that they are able to participate in and understand. Tool evaluations show very small changes to decision-making patterns compared to prior to tool introduction, suggesting the reproduction of existing decision patterns. Because the feedback loop requires information that is fundamentally not known, distortions in data may be exacerbated over time. Remedies to the feedback loop may also contain threats to justice, such as the retaining of unsubstantiated cases within highly racialized contexts. The broader justice implications for replicating, rather than remedying, both social inequities and the known problems with child protection decisions in child protection system data are only just beginning to be fully understood. The implications of algorithm tool use for both families and social workers require careful consideration, as the 'reasonable inference' services users have a right to, may be remarkably tentative in child protection. Social workers occupy a difficult position where they are responsible for upholding ethical codes and principles that may be challenged by using an algorithmic score. Further applied research on the ways social workers, families and algorithms interact as part of complex sociotechnical systems in child protection is needed, as well as ongoing interrogation of the sources and functions of data.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

Allegheny County Department of Human Services. 2017. *Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions*. Allegheny County: Allegheny County Department of Human Services.

Aradau, Claudia, and Tobias Blanke. 2017. Politics of prediction: Security and the time/space of governmentality in the age of big data. *European Journal of Social Theory* 20: 373–91. [CrossRef]

Arruabarrena, Ignacia, Joaquín de Paúl, Silvia Indias, and Mikel García. 2017. Racial/ethnic and socio-economic biases in child maltreatment severity assessment in Spanish child protection services caseworkers. *Child & Family Social Work* 22: 575–86.

Barocas, Solon, and Andrew Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 671: 62. Available online: https://ssrn.com/abstract=2477899 (accessed on 1 March 2019). [CrossRef]

Barocas, Solon, Elizabeth Bradley, Vasant Honavar, and Foster Provost. 2017. Big Data, Data Science, and Civil Rights. Computing Consortium White Paper. Available online: https://arxiv.org/abs/1706.03102 (accessed on 9 August 2018).

Bartelink, Cora, Tom A. van Yperen, and Ingrid J. ten Berge. 2015. Deciding on child maltreatment: A literature review on methods that improve decision-making. *Child Abuse & Neglect* 49: 142–53. [CrossRef]

Baumann, D., Jon Fluke, L. Dalgleish, and H. Kern. 2013. The decision-making ecology. In *From Evidence to Outcomes in Child Welfare: An International Reader*. Edited by Aron Shlonsky and Rami Benbenishty. New York: Oxford University Press, pp. 24–38.

Benbenishty, Rami, Bilha Davidson-Arad, Mónica López, John Devaney, Trevor Spratt, Carien Koopmans, Erik J. Knorth, Cilia LM Witteman, Jorge F. Del Valle, and David Hayes. 2016. Decision making in child protection: An international comparative study on maltreatment substantiation, risk assessment and interventions recommendations, and the role of professionals' child welfare attitudes. *Child Abuse and Neglect* 49: 63–75. [CrossRef] [PubMed]

Bosk, Emily Adlin. 2018. What counts? Quantification, worker judgment, and divergence in child welfare decision-making. *Human Service Organizations: Management, Leadership & Governance*. [CrossRef]

Boyd, Reiko. 2014. African American disproportionality and disparity in child welfare: Toward a comprehensive conceptual framework. *Children and Youth Services Review* 37: 15–27. [CrossRef]

Bywaters, Paul. 2015. Inequalities in child welfare: Towards a new policy, research and action agenda. *British Journal of Social Work* 45: 6–23. [CrossRef]

Bywaters, Paul, Geraldine Brady, Tim Sparks, Elizabeth Bos, Lisa Bunting, Brigid Daniel, Brid Featherstone, Kate Morris, and Jonathan Scourfield. 2015. Exploring inequities in child welfare and child protection services: Explaining the 'inverse intervention law'. *Children and Youth Services Review* 57: 98–105. [CrossRef]

Bywaters, Paul, Geraldine Brady, Lisa Bunting, Brigid Daniel, Brid Featherstone, Chantel Jones, Kate Morris, Jonathan Scourfield, Tim Sparks, and Calum Webb. 2018. Inequalities in English child protection practice under austerity: A universal challenge? *Child & Family Social Work* 23: 1365–2206. [CrossRef]

Cameron, Gary, Marshall Fine, Sarah Maiter, Karen Frensch, and Nancy Freymond. 2013. *Creating Positive Systems of Child and Family Welfare: Congruence with the Everyday Lives of Children and Parents*. Toronto: University of Toronto Press.

Choate, Peter, and G. Lindstrom. 2017. Parenting capacity assessment as a colonial strategy. *Canadian Family Law Quarterly* 37: 41–56.

Chouldechova, Alexandra. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5: 153–63. [CrossRef] [PubMed]

Chouldechova, Alexandra, Emily Putnam-Hornstein, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research* 81: 1–15.

Cuccaro-Alamin, Stephanie, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. 2017. Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review* 79: 291–98. [CrossRef]

Daley, Dyann, Michael Bachmann, Brittany A. Bachmann, Christian Pedigo, Minh-Thuy Bui, and Jamye Coffman. 2016. Risk terrain modeling predicts child maltreatment. *Child Abuse & Neglect* 62: 29–38. [CrossRef]

Danese, Andrea, Terrie E. Moffitt, HonaLee Harrington, Barry J. Milne, Guilherme Polanczyk, Carmine M. Pariante, Richie Poulton, and Avshalom Caspi. 2009. Adverse childhood experiences and adult risk factors for age-related disease: Depression, inflammation, and clustering of metabolic risk markers. *Archives of Pediatrics and Adolescent Medicine* 163: 1135–43. [CrossRef] [PubMed]

Dare, Tim. 2013. Predictive risk modelling and child maltreatment: Ethical challenges. In *Children in Crisis*. Hamilton: University of Waikato.

Dare, Tim, and Eileen Gambrill. 2016. *Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County*. Pittsburgh: Allegheny County Department of Human Services.

Daro, Deborah. 2009. The history of science and child abuse prevention: A reciprocal relationship. In *Preventing Child Maltreatment: Community Approaches*. Edited by Kenneth Dodge and Doriane Coleman. New York: The Guildford Press.

Davidson-Arad, Bilha, and Rami Benbenishty. 2016. Child Welfare Attitudes, Risk Assessments and Intervention Recommendations: The Role of Professional Expertise. *British Journal of Social Work* 46: 186–203. [CrossRef]

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. Clinical versus actuarial judgment. *Science* 243: 1668–74. [CrossRef] [PubMed]

Dencik, Lina, Arne Hintz, Joanna Redden, and Harry Warne. 2018. *Data Scores as Governance: Investigating Uses of Citizen Scoring in Public Services*. Cardiff: Cardiff University.

Doherty, Paula. 2016. Child protection threshold talk and ambivalent case formulations in 'borderline' care proceedings cases. *Qualitative Social Work* 16: 698–716. [CrossRef]

Drefuss, Suelette, and Shanton Chang. 2019. Uber surveillance in consumer markets. Paper presented at the Digital Citizen's Conference, University of Melbourne, Melbourne, Australia, July 24–26.

Edwards, Ros, and Val Gillies. 2015. Brain science and early years policy: Hopeful ethos or 'cruel optimism'? *Critical Social Policy* 35: 167–87. [CrossRef]

Elish, Madeleine Claire. 2019. Moral crumple zones: Cautionary tales in human-robot interaction. *Science, Technology, and Society* 5: 40–60. [CrossRef]

Eubanks, Virginia. 2017. *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. New York: St. Martin's Press.

Eubanks, Virginia. 2018. A response to Allegheny County DHS. Available online: https://virginia-eubanks.com/2018/02/16/a-response-to-allegheny-county-dhs/ (accessed on 28 March 2019).

EUGDPR (EU General Data Protection). 2019. General Data Protection Regulation. EU. Available online: https://eugdpr.org (accessed on 2 October 2018).

Fallon, Barbara, Martin Chabot, John Fluke, Cindy Blackstock, Bruce MacLaurin, and Lil Tonmyr. 2013. Placement decisions and disparities among Aboriginal children: Further analysis of the Canadian incidence study of reported child abuse and neglect part A: Comparisons of the 1998 and 2003 surveys. *Child Abuse & Neglect* 37: 47–60. [CrossRef]

Featherstone, Brid, Kate Morris, and Sue White. 2014. *Re-Imagining Child Protection: Towards Humane Social Work with Families*. Bristol: Policy Press.

Fenton, Jane, and Timothy Kelly. 2017. 'Risk is King and Needs to take a Backseat!' Can social workers' experiences of moral injury strengthen practice? *Journal of Social Work Practice* 31: 461–75. [CrossRef]

Fleming, Piers, Laura Biggart, and Chris Beckett. 2014. Effects of Professional Experience on Child Maltreatment Risk Assessments: A Comparison of Students and Qualified Social Workers. *British Journal of Social Work* 45: 2298–316. [CrossRef]

Fluke, John D., Martin Chabot, Barbara Fallon, Bruce MacLaurin, and Cindy Blackstock. 2010. Placement decisions and disparities among aboriginal groups: An application of the decision making ecology through multi-level analysis. *Child Abuse & Neglect* 34: 57–69. [CrossRef]

Fluke, John D., Tyler W. Corwin, Dana M. Hollinshead, and Erin J. Maher. 2016. Family preservation or child safety? Associations between child welfare workers' experience, position, and perspectives. *Children and Youth Services Review* 69: 210–18. [CrossRef]

Gambrill, Eileen. 2005. Decision making in child welfare: Errors and their context. *Children and Youth Services Review* 27: 347–52. [CrossRef]

Gambrill, Eileen, and Aron Shlonsky. 2000. Risk assessment in context. *Children and Youth Services Review* 22: 813–37. [CrossRef]

Gilbert, Neil, Nigel Parton, and Marit Skivenes. 2011. *Child Protection Systems: International Trends and Orientations*. Oxford: Oxford University Press.

Gillingham, Phillip. 2011. Decision-making tools and the development of expertise in child protection practitioners: Are we 'just breeding workers who are good at ticking boxes'? *Child & Family Social Work* 16: 412–21. [CrossRef]

Goddard, Chris R., Bernadette J. Saunders, Janet R. Stanley, and Joe Tucci. 1999. Structured risk assessment procedures: Instruments of abuse? *Child Abuse Review* 8: 251–63. [CrossRef]

Goldhaber-Fiebert, Jeremy, and Lea Prince. 2019. *Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office*. Pittsburgh: Allegheny County.

Green, Ben, and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision-making. *Proceedings of the ACM on Human-Computer Interaction* 3: 50–74.

Gurses, Seeda, Sita Gangadharan, and Suresh Venkatasubramanian. 2019. Critiquing and Rethinking Accountability, Fairness, and Transparency. *Our Data Bodies Project US: Our Data Badies Project*. Available online: https://www.odbproject.org/2019/07/15/critiquing-and-rethinking-fairness-accountability-and-transparency/ (accessed on 30 July 2019).

Harwin, Judith, B. Alrouh, S. Bedson, and Karen Broadhurst. 2018. *Care Demand and Regional Variability in England: 2010/11 to 2016/17*. Lancaster: Lancaster University.

Healy, Karen. 2009. A case of mistaken identity: The social welfare professions and New Public Management. *Journal of Sociology* 45: 401–18. [CrossRef]

Healy, Karen, and Yvonne Darlington. 2009. Service user participation in diverse child protection contexts: Principles for practice. *Child and Family Social Work* 14: 420–30. [CrossRef]

Healy, Karen, Yvonne Darlington, and Judith A. Feeney. 2011. Parents' participation in child protection practice: Toward respect and inclusion. *Families in Society: The Journal of Contemporary Social Services* 92: 282–88. [CrossRef]

Hughes, Tim. 2017. Prediction and Social Investment. In *Social Investment: A New Zealand Policy Experiment*. Edited by Jonathan Boston and David Gill. Wellington: Bridget Williams Books, pp. 179–202.

International Federation of Social Work. 2014. Global Definition of Social Work. IFSW. Available online: https://www.ifsw.org/what-is-social-work/global-definition-of-social-work/ (accessed on 1 August 2019).

Jackson, David, and Gary Marx. 2017. Data Mining Program Designed to Predict Child Abuse Proves Unreliable, DCFS Says. Chicago Tribune. December 6. Available online: http://www.chicagotribune.com/news/watchdog/ct-dcfs-eckerd-met-20171206-story.html (accessed on 22 February 2018).

Keddell, Emily. 2014. Current debates on variability in child welfare decision-making: A selected literature review. *Social Sciences* 3: 916–40. [CrossRef]

Keddell, Emily. 2015a. The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: Child abuse prevention or neo-liberal tool? *Critical Social Policy* 35: 69–88. [CrossRef]

Keddell, Emily. 2015b. Predictive Risk Modelling: On Data, Rights and Politics. Reimagining Social Work. Available online: http://www.reimaginingsocialwork.nz/2015/06/predictive-risk-modelling-on-rights-data-and-politics/ (accessed on 6 October 2019).

Keddell, Emily. 2016. Substantiation, decision-making and risk prediction in child protection systems. *Policy Quarterly* 12: 46–59. [CrossRef]

Keddell, Emily. 2017. Interpreting children's best interests: Needs, attachment and decision-making. *Journal of Social Work* 17: 324–42. [CrossRef]

Keddell, Emily, and Ian Hyslop. 2019. Ethnic inequalities in child welfare: The role of practitioner risk perceptions. *Child & Family Social Work*, 1–12. [CrossRef]

Kirk, Shelley. 2016. Children 'not lab-rats'—Anne Tolley intervenes in child abuse experiment. *Stuff*, July 30.

Klein, Sacha, and Darcey Merritt. 2014. Neighborhood racial & ethnic diversity as a predictor of child welfare system involvement. *Children and Youth Services Review* 41: 95–105.

Lees, Amanda, Edgar Meyer, and Jackie Rafferty. 2011. From Menzies Lyth to Munro: The Problem of Managerialism. *British Journal of Social Work* 43: 542–58. [CrossRef]

Lepri Letouze, Bruno, Jacopo Staiano, David Sangokoya, Emmanuel Letouzé, and Nuria Oliver. 2017. The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good. In *Transparent Data Mining for Big and Small Data, Studies in Big Data Series*. Springer: Cham, Switzerland.

Madden, M., M. Gilman, K. Levy, and A. Marwick. 2017. Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans. *Washington University Law Review* 95: 53–125.

Margolin, L. 1997. *Under the Cover of Kindness: The Invention of Social Work*. Charlottesville and London: University Press of Virginia.

McDonell, J. R., A. Ben-Arieh, and G. B. Melton. 2015. Strong Communities for Children: Results of a Multi-Year Community-Based Initiative to Protect Children from Harm. *Child Abuse & Neglect* 41: 79–96.

McIntyre, N., and D. Pegg. 2018. Councils use 377,000 people's data in efforts to predict child abuse. *The Guardian*, September 16.

McLaughlin, Michael, and Melissa Jonson-Reid. 2017. The relationship between child welfare financing, screening, and substantiation. *Children and Youth Services Review* 82: 407–12. [CrossRef]

McQuillan, Dan. 2015. Algorithmic states of exception. *European Journal of Cultural Studies* 18: 564–76. [CrossRef]

McQuillan, Dan. 2017. Data Science as Machinic Neoplatonism. *Philosophy & Technology* 31: 253–72. [CrossRef]

McQuillan, Dan. 2018. People's Councils for Ethical Machine Learning. *Social Media + Society* 4. [CrossRef]

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3: 2053951716679679.

Munro, Eileen. 2002. *Effective Child Protection*. London: Sage.

Munro, Eileen. 2011. *The Munro Review of Child Protection: Final Report, a Child-Centred System*. London: The Stationary Office Limited.

Munro, Eileen. 2019. *Predictive Analytics in Child Protection*. CHESS Working Paper No. 2019-03. Knowledge for use (K4U) Project. Durham, UK: Durham University.

Munro, Eileen, Julie Taylor, and Caroline Bradbury-Jones. 2014. Understanding the Causal Pathways to Child Maltreatment: Implications for Health and Social Care Policy and Practice. *Child Abuse Review* 23: 61–74. [CrossRef]

Naranayan, Arvind. 2018. *21 Fairness Definitions and Their Politics*. Youtube: Arvind Naranayan, Available online: https://www.youtube.com/watch?v=jIXIuYdnyyk (accessed on 1 August 2019).

New Zealand Government. 1989. *Oranga Tamariki Act*; Wellington: New Zealand Government. Available online: http://www.legislation.govt.nz/act/public/1989/0024/latest/whole.html (accessed on 1 August 2019).

Pearl, Judea, and Dana Mckenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Rahwan, Iyad. 2018. Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology* 20: 5–14. [CrossRef]

Rea, David, and Robert Erasmus. 2017. *Report of the Enhancing Decision-Making Project*; Wellington: Ministry of Social Development. Available online: https://mvcot.govt.nz/assets/Uploads/OIA-responses/Report-of-the-Enhancing-Intake-Decision-Making-Project.pdf (accessed on 15 September 2017).

Riley, Naomi. 2018. *Can Big Data Help Save Abused Kids?* Washington, DC: American Enterprise Institute, Available online: https://reason.com/2018/01/22/can-big-data-help-save-abused/ (accessed on 1 August 2019).

Roberts, Dorothy, and Lisa Sangoi. 2018. Black Families Matter: How the Child Welfare System Punishes Poor Families of Color. *Injustice Today*. March 26. Available online: https://theappeal.org/black-families-matter-how-the-child-welfare-system-punishes-poor-families-of-color-33ad20e2882e/ (accessed on 1 August 2019).

Rodriguez, Maria, Diane DePanfilis, and Paul Lanier. 2019. Bridging the Gap: Social Work Insights for Ethical Algorithmic Decision-Making in Human Services. *IBM Journal of Research and Development*, 1. [CrossRef]

Rostad, Whitney L., Tia McGill Rogers, and Mark J. Chaffin. 2017. The influence of concrete support on child welfare program engagement, progress, and recurrence. *Children and Youth Services Review* 72: 26–33. [CrossRef] [PubMed]

Rouland, Bénédicte, and Rhema Vaithianathan. 2018. Cumulative Prevalence of Maltreatment among New Zealand Children, 1998–2015. *American Journal of Public Health* 108: 511–13. [CrossRef] [PubMed]

Rowe, Michael. 2019. Shaping Our Algorithms Before They Shape Us. In *Artificial Intelligence and Inclusive Education: Speculative Futures and Emerging Practices*. Edited by Jeremy Knox, Yuchen Wang and Michael Gallagher. Singapore: Springer Singapore, pp. 151–63.

Sheridan, Ed. 2018. *Hackney Council Pays £360k to Data Firm Whose Software Profiles Troubled Families*. London: Hackney Citizen.

Shlonsky, Aron, and David Wagner. 2005. The next step: Integrating actuarial risk assessment and clinical judgment into an evidence-based practice framework in CPS case management. *Children and Youth Services Review* 27: 409–27. [CrossRef]

Sloane, Mona. 2018. *Making Artificial Intelligence Socially Just: Why the Current Focus on Ethics Is Not Enough*. London: London School of Economics., Available online: http://blogs.lse.ac.uk/politicsandpolicy/artificial-intelligence-and-society-ethics/ (accessed on 1 August 2019).

Spratt, Trevor, and J. Callan. 2004. Parents' Views on Social Work Interventions in Child Welfare Cases. *British Journal of Social Work* 34: 199–224. [CrossRef]

Swahn, Monica H., Daniel J. Whitaker, Courtney B. Pippen, Rebecca T. Leeb, Linda A. Teplin, Karen M. Abram, and Gary M. McClelland. 2006. Concordance between Self-Reported Maltreatment and Court Records of Abuse or Neglect among High-Risk Youths. *American Journal of Public Health* 96: 1849–53. [CrossRef] [PubMed]

Vaithianathan, Rhema. 2012. *Can Administrative Data Be Used to Identify Children at Risk of Adverse Outcomes?* Auckland: The University of Auckland.

Vaithianathan, Rhema, Tim Maloney, Emily Putnam-Hornstein, and Nan Jiang. 2013. Children in the Public Benefit System at Risk of Maltreatment: Identification via Predictive Modeling. *American Journal of Preventive Medicine* 45: 354–59. [CrossRef] [PubMed]

Vaithianathan, Rhema Nan Jiang, Tim Maloney, and Emily Putnam-Hornstein. 2017. Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation. In *Center for Social Data Analytics*. Auckland: Auckland University of Technology.

Vaithianathan, Rhema, Bénédicte Rouland, and Emily Putnam-Hornstein. 2018. Injury and Mortality Among Children Identified as at High Risk of Maltreatment. *Pediatrics* 141. [CrossRef] [PubMed]

van der Put, Claudia E., Merian B. R. Bouwmeester-Landweer, Eleonore A. Landsmeer-Beker, Jan M. Wit, Friedo W. Dekker, N. Pieter J. Kousemaker, and Herman E. M. Baartman. 2017. Screening for potential child maltreatment in parents of a newborn baby: The predictive validity of an Instrument for early identification of Parents at Risk for child Abuse and Neglect (IPARAN). *Child Abuse & Neglect* 70: 160–68. [CrossRef]

Veale, Michael, and Irina Brass. 2019. Administration by Algorithm? Public Management meets Public Sector Machine Learning. In *Algorithmic Regulation*. Edited by Karen Yeung and Martin Lodge. Oxford: Oxford University Press.

Vedder, Anton. 1999. KDD: The challenge to individualism. *Ethics and Information Technology* 1: 275–81. [CrossRef]

Wachter, Sandra, and Brett Mittelstadt. 2019. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review* 2: 1–84.

Weick, Ann. 1999. Guilty knowledge. *Families in Society* 80: 327–32. [CrossRef]

Wexler, Richard. 2018. Poor Kids End Up in Foster Care Because Parents Don't Get Margin of Error Rich Do. *Youth Today*, March 16.

White, Sue, Karen Broadhurst, David Wastell, Sue Peckover, Chris Hall, and Andy Pithouse. 2009. Whither practice-near research in the modernization programme? Policy blunders in children's services. *Journal of Social Work Practice* 23: 401–11. [CrossRef]

Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now Report 2018*. New York: AI Now Institute.

Widom, Cathy Spatz, Sally J. Czaja, and Kimberly A. DuMont. 2015. Intergenerational transmission of child abuse and neglect: Real or detection bias? *Science* 347: 1480–85. [CrossRef] [PubMed]

Wilson, Moira L., Sarah Tumen, Rissa Ota, and Anthony G. Simmers. 2015. Predictive Modeling: Potential Application in Prevention Services. *American Journal of Preventive Medicine* 48: 509–19. [CrossRef] [PubMed]