What Works for Children's Social Care

The Alan Turing Institute

Rees Centre

Department of Education

University of Oxford

ETHICS REVIEW OF MACHINE LEARNING IN CHILDREN'S SOCIAL CARE:

January 2020

What Works *for* Children's Social Care

Acknowledgements

The authors of this report would like to thank What Works for Children's Social Care for commissioning this project and for providing continuous support as we carried out our research and writing. In particular we would like to acknowledge Michael Sanders, Louise Reid, and Vicky Clayton for their patience and commitment to making a difference. We are also incredibly grateful to the family members with lived experience of children's social care who shared their views and experiences with us and to the Family Rights Group for making this outreach possible and for helping us ground our research in the reality of social work practice. Lastly, we are thankful for the stakeholders who shared their experiences and for the support for ethical innovation that we found in local authority representatives, researchers, industry leaders, and social care and family rights interest groups.

Authors

Leslie, D., The Alan Turing Institute, **Holmes, L.**, Rees Centre, University of Oxford, **Hitrova, C.**, The Alan Turing Institute and **Ott, E.**, Rees Centre, University of Oxford.

Funding

Department for Education, England.

About What Works for Children's Social Care

What Works for Children's Social Care aims to seek better outcomes for children, young people and families by bringing the best available evidence to practitioners and other decision makers across the children's social care sector. We generate, collate and make accessible the best evidence for practitioners, policy makers and practice leaders to improve children's social care and the outcomes it generates for children and families.

About The Alan Turing Institute

The Alan Turing Institute (the Institute) is the national institute for data science and artificial intelligence. Established in 2015, the Institute brings together thirteen UK universities. The Institute researches data science and AI applications to tackle some of the biggest challenges in science, society and the economy. The Institute's Public policy programme works alongside policy makers to explore how data-driven public service provision and policy innovation might solve long running policy problems and to develop the ethical foundations for the use of data science and artificial intelligence in policy-making. The goal of the programme is to enable technology to have a positive impact on the lives of as many people as possible.

About the Rees Centre

The Rees Centre (the Centre) at the University of Oxford produces research evidence to improve policy and practice in the areas of children's social care and education. The Centre aims to improve the life chances and particularly the educational outcomes of those who are, or have been supported by children's social care services. In 2017, the Centre and the Thomas Coram Research Unit at UCL established the Children's Social Care Data User Group which brings together academics, local authority data managers, analysts, charities and funders with a shared vision that administrative data from children's social care and other relevant agencies in England can be analysed and fed back into policy and practice to improve the way that children's social care services respond to children, young people and their families.

If you'd like this publication in an alternative format such as Braille, large print or audio, please contact us at: **wwccsc@nesta.org.uk**

CONTENTS

FOREWORD

Whether you're a sceptic or an enthusiast, it is hard to avoid recognising that the use of machine learning in children's social care is growing, in some places very rapidly. For advocates of the tool, the use of advanced analytics has the potential to improve services and outcomes for young people and their families by helping to rapidly find patterns in complexity. For their opponents, these tools risk dehumanising families, ingraining patterns of discrimination, and compromising the professional judgement of social workers while increasing unwanted intrusion into family life.

Both groups are right. Used well, in some circumstances, there is little doubt of the power of these tools to help professionals to make positive changes. Used poorly, or in the wrong contexts, it has the potential to be useless, or to actively cause harm. The consideration of what "well" and "poorly" mean, and what the right and wrong circumstances are, is both a question of effectiveness - how well do the tools actually work - and one of ethics.

It is for this reason that I'm pleased that What Works for Children's Social Care has commissioned the Rees Centre at the University of Oxford, and The Alan Turing Institute, to conduct this review of the ethics of using machine learning in children's social care. By bringing a combination of rigorous academic research, a practical focus, and the thoughts and experiences of practitioners, professionals, researchers and families to bear, they have produced this report and the recommendations within it, which we are proud to publish.

This is not the final word on this topic, and it does not aim to be. Instead, I hope that it will contribute to a much needed debate and when and where machine learning is appropriate, and what safeguards need to be in place to ensure ethical practice.

Michael Sanders

Executive Director
What Works for Children's Social Care

EXECUTIVE SUMMARY

The promises and perils of machine learning in children's social care

There could not be a more important time to think about the role that ethics should play in the context of using machine learning (ML) technologies in the domain of children's social care (CSC). Across the press, academia, and the worlds of policy and practice, concerns abound about the possible impacts of the growing use of ML in CSC on individuals, families, and communities. Many express legitimate worries about how the depersonalising and de-socialising effects of trends toward the automation of CSC are harming the care environment and negatively altering the way frontline workers are able to engage with families and children. Others raise concerns about how these data-driven ML systems are merely reinforcing, if not amplifying, historical patterns of systemic bias and discrimination. Others, still, highlight how the mixed results of existing ML innovations are signalling widespread conditions of poor data quality and questionable data collection and recording practices.

While these trepidations are valid and are helping to sharpen society's focus on the salient ethical issues that most demand concerted attention, they perhaps tell only one side of a more complicated story. In less than a generation, the explosive growth of ML, and of applied data science more generally, has become a transformative social, political, and economic force the world over. By helping researchers, analysts and practitioners to identify and draw insights from complex patterns extracted from large datasets, ML models have found useful applications in bolstering evidence-based decision making across a growing variety of sectors from healthcare, education, and transportation to agriculture, energy, and environmental management. With its capacities to assist the public sector in improving the personalisation of services, the prediction and analysis of trends, organisational functioning, and resource allocation, ML technologies hold the potential to significantly advance public welfare and the social good.

Keeping both these promises and perils in mind, how then can society responsibly harness the immense salutary potential of ML innovation in the realm of CSC? Provided that the way to such an unlocking of ML potential could be found, using it to foster the safety, wellbeing, and flourishing of children in need and their families would be a compelling prospect. Such innovations could, for example, be used to craft interventions that safeguard the dignity of child and family alike by focusing on outcomes that optimise family functioning, health, safety, and child development. They could empower families through the data-driven crafting of humane, informative, and strengths-based interventions that provide support for the achievement of their own self-defined goals. They could also provide insights at the organisational and institutional levels, improving the effectiveness and adeptness of service provision and providing empirical information for policy-formation.

The project at a glance

It is against this backdrop that What Works for Children's Social Care (WWCSC) requested this report on the research question "***Is it ethical to use machine learning approaches in children's social care systems and if so, how and under what circumstances?***".

The findings we present here take some preliminary steps to providing an answer. They are aimed at data scientists, policy makers, local authority (LA) children's services departments, civil servants, and citizens. Where possible, we have tried to avoid extensive technical discussions, and we have attempted, where necessary, to provide plain language definitions of specialised terms and background information to aid the non-technical reader.

This research is informed by a range of methods – a literature review, an integrative examination of existing ethical frameworks in social care and ML, a stakeholder roundtable with 31 participants, and a workshop with 10 family members who have lived experience of children's social care.

While our results are preliminary and still in need of further consultation, we offer, in what follows, a **three-tiered framework for thinking about the ethics of ML in CSC**. In order to make the ethical stakes and practical implications of the difficult and multi-level question posed to us by WWCSC as clear as possible, we have broken it down into three further ones around which these three tiers are organised.

The first tier asks: ***Should we be doing this?*** Here, we take an *external point of view*—a view from outside of existing practices of using ML models in CSC—which refrains from assuming that its use is legitimate *per se* so that the bigger picture issue of the very justification of that use can be tackled head on. The point here is to examine the ethical criteria that would make the use of ML in CSC justifiable if they were satisfied in real world settings and then to examine the problematic contexts in which such criteria might not, in actuality, be met. In this section, we bring together existing frameworks in the ethics of social work and the ethics of machine learning and artificial intelligence (AI) in order to formulate an integrated ethics of ML in CSC. We then use these ethical criteria to consider whether there are empirical factors intrinsic to the wider system in which CSC is situated (including historical patterns of inequity, the context of austerity, and conditions of system, organisation, and participant readiness) which may prevent the justified application of ML in CSC.

The second tier poses the question: ***Can we do this right?*** It takes an *internal point of view*, which assumes that the use of ML in CSC can, in fact, be justified so that we can identify and explore responsible practices of ML innovation in CSC from the inside of the design and production of the technologies themselves and internally to their processes of implementation. In this section, we present standards for best practice across ML's design and deployment lifecycle, paying special attention at each step of the way to the CSC context.

The third tier poses the question: ***What is to be done?*** It takes a *forward-looking point of view* that is focused on the potential of data scientific insights to transform the future of CSC for the better. It fleshes out recommendations for optimising the capacity of future data scientific research, community- and family-based collaboration, and deliberate innovation intervention to produce tangible societal benefits and advance individual, familial, and public wellbeing.

First-tier findings

Primary among our findings in answering the first-tier question is the integrated ethical framework for the use of ML in CSC that we present in detail in the full report. In summary form, its basic elements are as follows:

Ethical values that set the direction of travel for the responsible use of ML in CSC

- Respect the dignity of individual persons, empower them, and value the uniqueness of their aspirations, cultures, contexts, and life plans
- Connect with each other sincerely, openly, and inclusively, and prioritise trust, solidarity, and interpersonal collaboration
- Care for the wellbeing of each and all, and serve others with empathy, selflessness, and compassion
- Protect the priorities of social justice and the public interest by ensuring equity, recognising diversity, and challenging discrimination and oppression

Practical principles that establish the moral justifiability of the integrated practices of social care and ML innovation

- Fair, sustainable, ever-improving social care
- Social care that supports and empowers
- Transparent, responsible, and accountable social care

Professional virtues that establish common principles of professional integrity shared by social work and responsible ML innovation

- Be sincere, honest, and trustworthy
- Uphold ethical values and best practices
- Lead by competence and example
- Maintain appropriate professional boundaries
- Make considered professional judgments
- Be professionally responsible
- Be objective and impartial in making professional judgments
- Use evidence-based reasoning when rendering decisions

Our ultimate aim in setting out an integrated ethics of ML in CSC is to put its resulting values and principles into an actionable form. Such a form should support practice and bring together all stakeholders involved in the complex, multi-level and collaborative processes of conceptualising ML applications and projects. It should help them to

cooperatively define their objectives, and it should assist them in designing, deploying, and monitoring their applications responsibly.

What is needed for this is a vehicle of common commitment—a way for all those who are dedicated to doing good through the responsible design and use of data scientific applications to continuously coalesce around a mutual recognition of the ethical motivations, practical principles, and professional standards of conduct that should motivate, direct, underwrite, and steer responsible practices of ML innovation in the field of children's social care. We will call this living document a ***Commitment to care, collaboration, and understanding*** and provide in the full report a preliminary mapping of what this might look like.

The final task we undertake in the first tier is to look closely at how the ethical values that lie behind the responsible use of machine learning in children's social care might provide a **critical yardstick** of sorts against which the application of this kind of technology in the sensitive and demanding domain of CSC can be measured. To do this, we consider several empirical factors, which might call into question the justifiability of using ML in CSC. In particular, we examine and analyse three such factors:

- Public management in the context of austerity
- System, organisation, and participant (SOP) readiness
- Social inequality and cycles of poverty and discrimination

Second-tier findings

In the second tier of this review, we respond to the question: ***Can we do this right?*** We investigate how the practical principles that we have articulated in the ethical framework might help to provide guardrails for responsible conduct. We also examine how such principles might give shape to best practices from a point of view internal to the boots-on-the-ground activities of ML innovation and use. To do this, we move step-by-step through the design and implementation pipeline of the production and use of ML models in CSC, paying special attention to domain-specific needs and potential pitfalls of the CSC use case. In outline form, here is what we cover in this second tier:

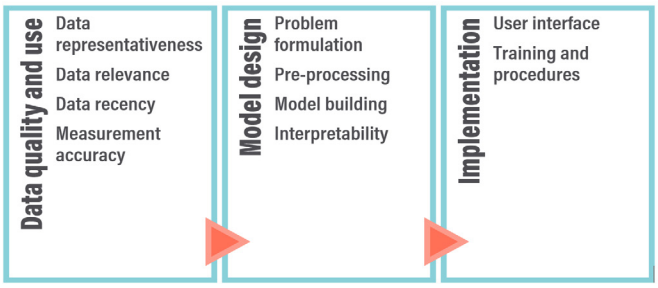

Figure 1. The stages of responsible machine learning innovation covered in this review

Third-tier findings

The final section of this review responds to the third-tier question, What is to be done? It provides some preliminary recommendations for steering the present direction of the use of ML in CSC, both in its application to practical, real-world problems and as a medium for research insight and discovery. It presents eight such recommendations:

1. ***Mandate the responsible design and use of ML models in CSC at the national level.***
2. ***Connect practitioners and data scientists across local authorities to improve ML innovation and to advance shared insights in applied data science through openness and communication.***
3. ***Institutionalise inclusive and consent-based practices for designing, procuring, and implementing ML models.***
4. ***Fund, initiate, and undertake active research programmes in system, organisation, and participant readiness.***
5. ***Understand the use of data in CSC better so that recognition of its potential benefits and limitations can more effectively guide ML innovation practices.***
6. ***Use data insights to describe, diagnose and analyse the root causes of the need for CSC, experiment to address them.***
7. ***Focus on individual- and family-advancing outcomes, strengths-based approaches, and community-guided prospect modelling.***
8. ***Improve data quality and understanding through professional development and training.***

While we conclude this review with recommendations, which are outlined in depth in the main report, we would also like to highlight from the outset that this study is primarily intended to help clarify some of the most substantial and complex ethical issues that arise in the context of the real-world application of ML in CSC. For this reason, the report should be utilised both as a means to reflect on external questions about the appropriateness and justifiability of using ML applications in CSC (both for specific use cases and in general) and as a preliminary guide for developing internal processes of data scientific innovation and implementation that incorporate ethics considerations at multiple points throughout the development and deployment lifecycle.

I. INTRODUCTION

Machine learning in society

In less than a generation, the explosive growth of machine learning (ML) has become a transformative social, political, and economic force the world over. For better or worse, the pervasiveness of networked computing, of digital interconnectedness, and of ubiquitous data extraction together with rapid progress made in computing power and algorithmic techniques are now presenting stakeholders across veritably every sector of society, and at every socioeconomic level, with unprecedented opportunities as well as significant challenges.

The opportunities may well seem boundless. Data-driven insights generated by ML innovations have already started to advance crucial dimensions of human wellbeing and improved prospects for a more sustainable future. In the field of healthcare, for instance, biomedical ML applications are allowing doctors to better target cancer drugs, to detect diseases earlier and more effectively, and to carry out surgical procedures with unprecedented precision. In many other future-critical fields too, from environmental science to energy management, ML applications are combatting climate change and deforestation, supporting biodiversity, catalysing agricultural productivity, producing 'smarter,' more efficient cities, and helping to provide new possibilities for the democratised distribution of essential goods and services.

However, in a networked digital world, where connected devices containing countless sensors and sites of behavioural measurement intermingle with omnipresent ML systems, seemingly intractable challenges also abound. In this networked reality, algorithmically personalised services can reach into the unwitting private lives of targeted data subjects and have an active curatorial hand in the formation of their identities. At the same time, when left to their own devices, opaque algorithmic methods of relevance-ranking, popularity-sorting, and trend-predicting can produce calculated digital publics bereft of any sort of participatory social or political choice (Gillespie, 2014; Ziewitz, 2016; O'Neil, 2016; Bogost, 2015; Striphas, 2015; Beer, 2017; Cardon, 2016). More troubling still, such ML-enabled capabilities for hyper-personalised targeting, anticipatory calculation, and algorithmic administration at scale are manifesting in intrusive hazard-pre-emption regimes (O'Grady, 2015) ranging from data-driven border control (Amoore, 2009; Amoore & Raley, 2016) and predictive policing to commercial surveillance. They are also enabling digital autocracies to engage in population-level behavioural monitoring and disciplinary control.

As these technologically induced risks have come into clearer view, critics have begun to voice legitimate concerns about the dangers to personal and social freedom posed by the rapid proliferation of ever more computationally powerful applications of ML. Some have focused on the 'big tech' driven political economy of surveillance capitalism wherein the bald exploitation of behavioural patterns serves as a tool for consumer manipulation and corporate profit (Zuboff, 2015, 2019). Others have observed how widespread government use of digital tracking and automation-supported decision-making can—when carried out uncritically, irresponsibly, and without inclusive community involvement—function to reinforce deep-seated patterns of poverty, inequality, and marginalisation (Eubanks, 2018).

Ethical risks for machine learning in children's social care

The problem of the ethics of machine learning in children's social care (CSC) is situated at the centre of this difficult socioeconomic, political, and cultural constellation. On the one hand, it would seem a crucial moral imperative to use the growing array of applied data scientific techniques to foster the safety, wellbeing, and flourishing of children in need and their families and to bolster possibilities for optimal outcomes in family life. The descriptive, prescriptive, and analytic tools afforded by ML can assist data scientists, policy-makers, and frontline workers in marshalling evidence-based insights to make sure children receive care when and at the level needed. These tools can potentially be used to safeguard the dignity of child and family alike by focusing on outcomes that optimise family functioning, health, safety, and education. And they can empower families through the data-driven crafting of humane, informative, and preventive interventions that provide support for the achievement of their own self-defined goals, thereby fostering their autonomy and wellbeing (What Works for Children's Social Care, 2018). They can also provide insights at the organisational and institutional levels, improving the efficiency and effectiveness of service provision for troubled families

and providing empirical information for policy-formation and case-based judgment.

On the other hand, the many risks associated with the use of predictive analytics in the provision of children's social care raise serious questions about where to draw boundaries in the utilisation of individual-targeting ML technologies in CSC. Three such risks are of fundamental concern in the context of the ethics of ML and will factor heavily into this report. These are:

- the potential reinforcement/amplification of systemic bias and discrimination in the use of predictive analytics
- the potential deterioration of the critical human and relational factors in children's social care practices, and
- the potential generation of poor-quality outcomes owing to deficient data stewardship

Predictive analytics and bias amplification

First, the widespread use of predictive risk modelling in frontline ML applications that assess hazards to the safety of individual children could reinforce or augment social dynamics of bias and discrimination. Because they draw insights from existing distributions of data, supervised machine learning models, when they work reliably, make accurate out-of-sample predictions by replicating the social and cultural patterns of the past—regardless of whether these patterns are inequitable or discriminatory.

This is a deep and seemingly unavoidable problem in the children's social care sector inasmuch as the correlation of child neglect and maltreatment with historical patterns of poverty and deprivation make the feedforward of these patterns in effective data mining all but inescapable. Such patterns, moreover, are further reinforced as the predictive recommendations yielded by the ML systems in use become part of future data distributions and hence of future training and testing data for other ML models to come. The problem is made worse by the skewed nature of the data used for these predictive purposes. That is, much of it is acquired by public authorities who primarily work with individuals from low income families. Not only does this vicious cycle of deprivation and data capture lead to an amplification of discriminatory configurations whereby impoverished individuals are as such likely to be assessed as having a higher risk profile, it creates blind spots in the ML system's capacity to accurately predict outcomes for children from other socioeconomic backgrounds.

Predictive analytics and human/relational factors

A second important risk has to do with the critical human and relational factors that play such a big part in the ethical provision of CSC. From the standpoint of affected individuals and frontline practitioners, the role of interpersonal communication, empathetic understanding, individual empowerment, reciprocal trust, and dialogically honed professional judgment make a human-centred approach to care provision a vital necessity. This raises the question of how frontline workers will be able to effectively manage the potentially depersonalising consequences of integrating automated decision support systems into their relationships with families.

The fear of 'reducing human beings to percentages' (Binns et al, 2018) comes from both parents and practitioners, and a question remains as to whether or not sufficient resources can be dedicated to train users to responsibly implement algorithmic decision support systems. Such training would position ML tools as important but subsidiary aids to evidence-based judgment while preserving due regard for the dignity of affected individuals. This would involve the demanding task of putting automation in its proper place. Such a task could be accomplished by setting up norms and mechanisms of implementation that preserve the continued priorities of interpretive charity, dialogically informed opinion formation, humane situational awareness, reasonableness, and context-specificity.

Predictive analytics and outcome and data quality

The third crucial risk originates in the relationship between the quality of the data used to train and test ML systems and the quality of the results they produce when operating in real-world environments. The so-called 'garbage in, garbage out' rule (Surden, 2014; Lehr & Ohm, 2017; Zarsky, 2017; Glaberson, 2019)—that algorithmic models are only as good as the data on which they are trained, tested, and validated—has especially significant ramifications for the safety-critical use case of ML risk modelling in CSC where decision outcomes can have life or death consequences. It is simply not the case that the voluminous data collected and held in administrative systems by local authorities can be gathered, linked, and fed, without further ado, into ML models that then churn out actionable insights.

Inaccuracies, contested information, and systemically consequential errors can enter into a dataset at multiple points along the extraction, collection and

consolidation workflow. All of these trouble spots must be diligently patrolled and scrutinised to ensure the global quality of inputs incorporated into the set. When done properly, this is an arduous and labour-intensive enterprise, which involves domain expertise, technical know-how, and wide-angled vision. Making sure that a dataset is sufficiently representative of the population it covers, that it contains relevant, recent, and accurately recorded information in sufficient amounts, and that its contents are appropriate, reasonable, and interpretable is a tall order to fill, especially in an area where misrecorded, biased, error-prone, missing, and outdated information is prevalent.

Ethical concerns for machine learning in children's social care

The need to respond to these three kinds of risk (namely, the potential deterioration of critical human factors, the potential generation of poor-quality outcomes, and the potential amplification of bias and discrimination) will motivate and inform the report on the ethics of machine learning in CSC that follows.

Such types of risk, in fact, line up with three sets of ethical concerns that should be placed front and centre in any extended consideration of responsible ML design and deployment in the field of CSC:

- **concerns with HUMAN AGENCY and SOCIAL INTERACTION:** Can we *protect the social value of human agency and interpersonal connection* to enable individual empowerment and inclusive social solidarity? As we will see, these concerns align with the ethical values of **RESPECTING** the dignity of individual persons and **CONNECTING** with each other sincerely, openly, and inclusively.

- **concerns with WELLBEING and PUBLIC BENEFIT:** Can we *foster wellbeing and advance tangible societal benefits* through the humanely oriented and democratically navigated pursuit of data scientific discovery and innovation? As we will see, these concerns align with the ethical value of **CARING** for the wellbeing of each and all.

- **concerns with SOCIAL JUSTICE and EQUITY:** Can we *ensure social justice* in the face of existing discrimination and societal inequity? As we will see, these concerns align with the ethical values of **PROTECTING** the priorities of social justice and the public interest.

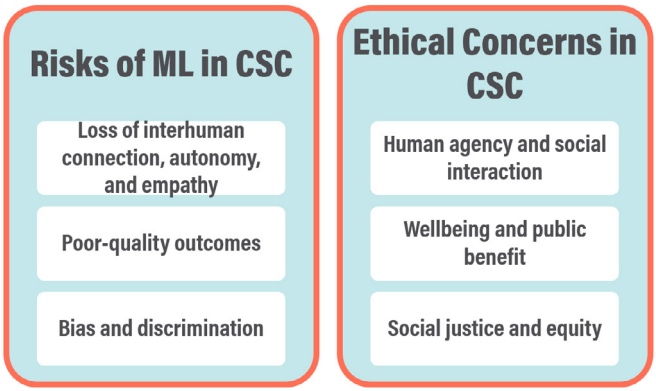

Figure 2. Risks and ethical concerns of machine learning in children's social care motivating the structure of the present report

By framing the design and use of ML in CSC through these normative lenses of justice, agency, interaction, and wellbeing, we are better positioned to answer the question for which What Works for Children's Social Care commissioned this tailored review, namely: "***Is it ethical to use machine learning approaches in children's social care systems and if so, how and under what circumstances?***"

Addressing this question through the normative lenses just mentioned will involve using them as bifocals. That is, it will involve looking at possibilities for the ethical use of ML approaches in CSC from both *external and internal points of view*. From an *external point of view*—a view from outside of existing practices of using ML models in CSC—we will refrain from assuming that its use is legitimate *per se* so that the bigger picture issue of the very justification of that use can be tackled head on. The point here is to examine the ethical criteria that would make the use of ML in CSC justifiable if they were satisfied in real world settings and then to examine the problematic contexts in which such criteria might not, in actuality, be met. From an *internal point of view*, we will start by assuming that that use can, in fact, be justified so that we can look at the practices of using ML in CSC from the inside of the design and production of the technologies themselves and internally to their processes of implementation. From this internal perspective, we will look at how these practices of designing, producing, and implementing ML in CSC can be carried out equitably, ethically, and responsibly.

In keeping with this distinction between internal and external ways into the question of the ethics of using ML in CSC, the following sections of this report will be structured around three further questions, the last of which demands that we also assume a third, forward-

looking point of view that is focused on the potential of data scientific insights to transform the future of CSC for the better. Summary outlines for each section are provided:

1. Should we be doing this? Big picture considerations in the ethics of ML in CSC

This section begins with a review of existing perspectives on the practical ethics of social work and on the practical ethics of artificial intelligence (AI) and machine learning. By finding common ground among the ethical values and practical principles that are prevalent in these two fields and integrating unique but crucial features of both, it endeavours to provide some serviceable guidance for considering the ethical issues surrounding the use of ML in CSC. It then explores how these steps toward an applied ethics of ML in CSC might inform the justification of its use given several external, empirical factors, which might call into question the justifiability of the application of this kind of technology in the sensitive and demanding domain of CSC.

2. Can we do this right? Existing practices, pitfalls, and prospects

This section focuses on the actual activities involved in the design and implementation of ML systems in CSC. It outlines best practices in responsible ML innovation, while at the same time moving through the use case of the ML production and deployment pipeline in CSC (from problem formulation and data collection, curation, and pre-processing to model-building and implementation). It critically assesses that use case based on the standards set by those best practices.

3. What is to be done? Recommendations for the future of ML in CSC

The final section fleshes out eight recommendations for optimising the capacity of future data scientific research and innovation in CSC to produce tangible societal benefits and advance individual, familial, and public wellbeing.

It should be noted, at the start, that this tailored review is intended to help clarify some of the thorniest ethical issues that arise in the context of the use of ML in CSC. As such, the report should be employed as a means to reflect on external questions about the appropriateness of using ML applications in CSC (both for specific use cases and in general). It should also be utilised as a preliminary guide for developing internal processes of innovation and implementation that incorporate ethics considerations at multiple points throughout the development and deployment lifecycle.

It must be stressed that considerations of the ethical issues highlighted in this report are to be undertaken on a case-by-case basis by researchers, data scientists, managers, practitioners, and affected individuals and, when possible, that such deliberation should occur in open and inclusive conversation with all others whose interests are impacted. Every ML application is unique, not just technically but also by virtue of the context in which it is envisioned and developed, who envisions and develops it, who it impacts and in what ways, and how it is designed and used. The ethics of ML in CSC should, in this respect, not be seen as a one-off checklist, nor as an appendage to already created ML models. As this review will demonstrate, determinations about whether or not and how to use ML applications as well as choices made throughout their design, development, and deployment can incorporate specious assumptions, unintentional errors, and deep-seated biases into the research and innovation process. These determinations and choices demand active moral reflection and self-criticism. The complexities and vulnerabilities involved in the CSC sector serve only to enhance the importance of such continuous ethical consideration.

II. BACKGROUND

While the fact-based perspective that motivates research in data science has been adopted broadly across the public sector in the UK for quite a long time, many of its local and national authorities have been slow to harness the full power of ML technologies to deliver more efficient, more equitable, and more responsive public services (Margetts & Dorobantu, 2019). In less controversial use cases, where the benefits of the capability of ML models to generate population-level insights and macro-scale mapping, planning, and forecasting are more straightforward (as in public health, education, and emergency management), the path to rectifying this sluggish uptake may be as simple as greater investment of intellectual capital and organisational resources. However, where a degree of personalisation or individual targeting is involved, the utilisation of ML technologies by public authorities triggers a host of concerns about privacy invasion, governmental overreach, technological solutionism, dehumanisation, and discrimination. In the particular case of using ML decision-support tools in CSC, concerns are magnified in virtue of the fluid and often fraught relationship between justifiable public intervention to safeguard children and the private life of the family. These concerns are exacerbated by the potentially safety-critical impacts of this sort of predictive risk modelling (children's lives and/or long-term wellbeing are at stake).

Reservations about the use of predictive analytics by local and national governments to create efficiencies and to ease budgetary burdens can create barriers to innovation. Such barriers are more or less justifiable, depending on whether the ML applications of interest will actually equitably improve service delivery, lifting up individuals, families, and communities in need, and providing better resources for objective, evidence-based decision making. This targeted review will attempt to clarify the issues surrounding the legitimacy of such public sector barriers regarding the use of ML in CSC.

In this background section, some preparatory information is provided to orient the reader to several of the socio-historical, conceptual, and technical complexities that underlie the ethics of using ML in CSC. After providing context about the methods of this study and children's social care in England, we will explore some of the practical, conceptual, and societal challenges faced by social workers in this field. We will then provide broad stroked overviews of the controversial role of risk assessment in CSC and of the technical basics of machine learning to prepare our non-technical audiences for the discussions to come.

Methodology

This report was prepared by The Alan Turing Institute and the Rees Centre at the University of Oxford at the request of the What Works for Children's Social Care. The research behind its results has been comprised of a literature review, an integrative examination of existing ethical frameworks in CSC and ML, a stakeholder roundtable that took place at The Alan Turing Institute in London in July 2019 (31 attendees) and a workshop with 10 family members who have lived experience with children's social care, which took place in December 2019 and was facilitated in partnership with Family Rights Group.

The stakeholder roundtable included attendees from government, local authorities, academia and research, non-profit organisations and stakeholder groups, and companies. The discussion was first grounded by participants' examples of using data science and machine learning in the context of children's social care and went on to more broad conversations about the ethics of ML in CSC.

The family workshop focused on ascertaining the perspectives of family members with lived experience about the potential use of ML in CSC and the associated ethics considerations. Emerging recommendations from our literature review and roundtable were also reviewed and revised as part of the workshop.

Combined, the literature review, the framework integration, the roundtable and workshop aimed to collect a breadth of information for examining the framing questions mentioned above. These methods did not aim to be exhaustive.

Children's social care in England

In England, local authorities (LAs) have a statutory duty to provide children's social care services to all children identified as being in need. The term 'in need' is defined in the Children Act (1989) as being a child or young person who is 'unlikely to achieve or maintain,

or have the opportunity of achieving or maintaining, a reasonable standard of health or development without the provision for him/her of services by a local authority' or if his or her 'development is likely to be significantly impaired, or further impaired without the provision of such services' or 'if he or she is disabled.' Every LA has the duty to safeguard and promote the welfare of children within their area who are in need and, so far as is reasonably consistent with that duty, to promote the upbringing of such children by their families, by providing an appropriate range and level of services for those children. The Common Assessment Framework describes three interrelated dimensions: children's developmental needs, the capacity of parents and carers to respond appropriately to these needs, and the impact of wider family and environmental factors on parenting capacity and children (Children's Workforce Development Council, 2009; Holmes, McDermid, Padley & Soper, 2010; HM Government, 2018).

Where an LA receives a referral and has some reasons to be concerned that a child may be suffering, or likely to suffer, significant harm, section 47 of the Children Act 1989 requires it to undertake an investigation. Some 'children in need' receive support from CSC while remaining at home with their families. Others become looked after via a voluntary agreement (section 20) or a care order (section 31) made by the court to place the child in the care or supervision of a designated local authority. The majority of looked after children are placed in foster placements (73%) (Department for Education, 2019a), including with approved relatives and friends, or in secure units, children's homes, or semi-independent living arrangements (11%). A minority of children are placed with parents (6%), for adoption (3%), or in other settings (3%). The process for assessing children in need and their families is described in Working Together to Safeguard Children (HM Government, 2018).

Wider context to CSC in England

The CSC system has faced an increase in demand for its services alongside austerity limiting the resources available to LAs. There is evidence of increased pressures placed on CSC and these may account for the challenges in recruitment and retention of social workers (House of Commons, 2019). In 2017 the Local Government Association (LGA) noted that children's social care is 'being pushed to breaking point, with 75% of councils' overspending for children's services (Local Government Association, 2017) and 89% of directors of children's social care services reporting in 2016-2017 that they found it increasingly challenging to fulfil their statutory duties to provide support to children in need due to the limited available resources at their disposal (All Party Parliamentary Group for Children, 2017).

According to the Care Crisis Review (2018), the number of care order applications was at a record level in 2017 and the number of looked after children was at its highest since the introduction of the Children Act (1989) (Care Crisis Review, 2018). The number of section 47 safeguarding investigations has also risen steadily, growing by 151% between 2006-2007 and 2016-2017 (House of Commons, 2018). This growth in child protection intervention rates has been viewed as a dimension of expanding social inequity whereby 'children and/or their parents face unequal chances, experiences or outcomes of involvement with child welfare services that are systematically associated with structural social dis/advantage and are unjust and avoidable' (Bywaters et al., 2015).

The importance in understanding the effectiveness of the system and reaching the right children

While service demand has increased, the incidence of child mortality by homicide or assault and the number of people guilty of child cruelty or neglect have dramatically decreased in England and Wales between 1893 and 2016. However, the number of child protection registrations and children entering care increased between 2000 and 2016, as did cases of child maltreatment, such as sexually-motivated crimes or emotional abuse. (Degli Esposti et al., 2019). One possible interpretation of this data could conclude that CSC services may be effective in reducing abuse and neglect. A deeper consideration of the statistics, however, may reveal that the types of crimes against children could have shifted in nature.

An important factor in considering the effectiveness of CSC services is whether they are working with the right children and families. Recent analysis published by the Department for Education (2018) indicated that of the 1.5 million children referred to CSC between 2014-15 and 2016-17 nearly a third (0.4 million) were deemed not to be in need. There may also be children who are receiving a service that is insufficient or, conversely, too intensive for their level of need. This has been highlighted in recent research by Forrester (2017) whereby the difficulties of assessing the 'right families' and proportionate involvement of CSC are considered. Forrester argues that we cannot evaluate outcomes without addressing the issue of proportionality and whether CSC are working with the right families.

Data use in children's social care

The nature and availability of data to inform and support decision-making in CSC

Research has highlighted that data submitted by CSC to government departments as part of national statutory returns, such as the SSDA903 (Department for Education, 2019) and Children in Need (CiN) Census (Department for Education, 2018) constitute a small proportion of the data held and utilised within local authority children's services departments (Holmes & McDermid, 2012; Ward, Holmes, & Soper, 2008).

In a recent Research in Practice Change Project (Bowyer, Gillson, Holmes, Preston, & Trivedi, 2018) work was undertaken with nineteen LAs to explore their data usage at a local and regional level to inform strategic and operational planning and decision making. The project identified a range of practices and initiatives whereby local area data sets are linked and matched, either between agencies or across different parts of the children's social care system. The study identified a particular paucity of data about the services that children (and their families) received. Data was often not recorded in a systematic way across the local authorities, and instead was recorded in separate, non-centralised databases, and/or spreadsheets. These findings show that many of the difficulties associated with the availability of data highlighted by McDermid (2008) still remain.

Appropriate use of data

In addition to the availability of data, a fundamental consideration is the effective and appropriate use of data. Beninger, Newton, Digby, Clay, and Collins (2017) have highlighted the use of performance monitoring data for internal auditing purposes. The need for a rigorous and strategic approach to self-appraisal coupled with an open and honest response to feedback and inspections has also been underscored (Bryant, Parish, & Rea, 2016). The intelligent use of data to support better decision making has additionally been emphasised in the recent Care Crisis Review (2018).

CSC administrative data is often designed for performance monitoring and is useful in showing trends amongst populations that interact with children's social care. The use of the data for other purposes may encounter issues including appropriateness and the ecological fallacy of applying population-level data to individual cases. Context matters, and whilst trends can help inform decisions, it is not appropriate to use these data for decision-making at an individual or case level based solely on aggregate level analysis.

Using data and information to support decision-making in the context of CSC

Understanding and acting on the evidence that is held by different agencies involved in CSC can be a challenge (Reder, Duncan, & Gray, 1993; Brandon et al., 2008; Brandon et al., 2009; Brandon et al., 2012). CSC must balance relationship-based practice, standards for 'good enough' care, limited resources, evidence on possible outcomes, and the complex set of circumstances that come into play when determining which children and families receive support services and what type of support to offer.

In addition to making practical trade-offs, CSC practitioners face the complex task of considering and weighing all available, interrelated, and relevant information for each case (Simon, 1997; Stevenson, 2007). Even with standardised, nationally mandated procedures in place for CSC, studies have shown that decision-making in CSC may only consider partial information. For instance, decision makers may not holistically take into account a family's history or sources of support when responding to recent circumstances (Brandon et al., 2008) or may focus on a particular type of child maltreatment, e.g. neglect, whilst not considering others (Brandon et al., 2009). CSC practitioners may focus on information readily available to them rather than looking at all available information and give more weight to memorable, vivid, emotion-arousing information. Individuals often focus on the first or last piece of information they receive and are slow to change their views. Consequently, social workers may be more critical of evidence that does not support their existing opinion of a given family (Munro, 1999; Stewart & Thompson, 2004). This may be because decision-making in children's social care is complex and decisions are not only affected by the availability of data that exists (information, data, and professional expertise), but also by time and workload pressures, the timeframe of children and guidance on timelines, decision fatigue, and a range of beliefs or behaviours that can unconsciously influence decision-making (such as looking for evidence that confirms pre-existing views and judging based on relative rather than objective merits) (Brown & Ward, 2013; Cuccaro-Alamin, Foust, Vaithianathan & Puntam, 2017).

Opportunity for algorithmic interventions or cause for concerns?

The above sheds light on some of the context of the current CSC system as well as decision-making complexities faced by social workers and

other professionals. This could be construed as an argument for an increased presence of ML models in the field as such tools are often promoted with promises of efficiency, objectivity, and improved resource allocation. However, the above also raises considerable concerns about whether the currently existing system and its decisions, as recorded and reflected in the data, should be used to train ML models that may replicate these patterns. Moreover, it gives rise to the question of whether the use of ML decision-support tools is allowing the avoidance of rather than tackling the structural reasons giving rise to increasing demands for CSC services. These questions deserve a public dialogue and in-depth consideration by policy- and law-makers who have democratic mandate to steer the direction of the CSC field in general.

Risk assessment and machine learning in children's social care

Some of the first risk assessment approaches came from the field of public health and epidemiology, as researchers attempted to anticipate and predict health conditions in entire populations (Garrison, 2012). There are many differences between public health and CSC that influence the ethics of using risk assessments. Public health deals with population-level data to identify and help prevent health risks, whereas child welfare deals with reports, referrals, and assessments to identify and investigate individual cases of child maltreatment and provide services for families (Glaberson, 2019). The complexity and nuances of CSC, and the wider environment within which they operate also need to be considered (Association of Directors of Children's Services, 2018; La Valle, Hart, Holmes & Pinto, 2019).

Depending on the method of their creation, such risk assessment models are broadly referred to as consensus-based (developed together with experts on the basis of the state of research and the expert's knowledge, experience) or actuarial-based (developed on the basis of empirical research that seeks to identify statistically relevant predictive risk factors that are weighted and compiled within a single assessment) (White & Walsh, 2006). Actuarial-based risk assessment sought to do what ML promises today—to find and map the predictive relationship between different variables on the basis of empirical data.

Recent examples of the use of risk assessments and ML within the English context have focused on the development of tools and methods as part of the Troubled Families Programme (TFP) (Ministry of Housing, Communities & Local Government, 2019). The Programme targeted intervention services at families experiencing multiple problems, including crime, truancy, unemployment, domestic abuse, and mental health. As part of the Programme, some local authorities turned to ML models to help identify families at risk and target support services more effectively. The role of ML in the Programme is further discussed below in the Section Applications of predictive analytics in social care.

Rationales for systematic assessments and machine learning

The key rationales that encourage the adoption of systematised assessment tools in CSC are not too different than those used to justify the use of ML in the present. These are the desire for (White & Walsh, 2006):

- rational reasoning for decisions affecting children and families
- greater consistency and objectivity across cases and between social workers
- the choice of more effective intervention approaches
- more efficient resource allocation
- guaranteeing the accountability of public bodies
- supporting social workers, including when they may lack sufficient training and support

During our roundtable, stakeholders highlighted a few reasons why they began developing or exploring ML tools in CSC:

- to identify families in need of additional support
- initially to explore how the general system of CSC functioned but, ultimately, to help personalise service for individual families in light of the 'knowledge' or insight derived from ML methods
- to support efforts for community safety that were already being developed in the context of policing.

These reasons focussed on ML as a tool to aid professional decision-making in CSC, but not as actuarial risk assessments themselves (developed on the basis of empirical research that seeks to identify

statistically relevant predictive risk factors that are weighted and compiled within a single assessment).

Roundtable participants also raised the question of whether ML models were now looked to as a solution in CSC to reach efficiencies in the context of austerity measures when local authorities are faced with insufficient budgets, staffing and rising case-loads.

Risks of actuarial risk assessments and ML

Actuarial risk assessments have been recommended by some in the English context as an aid to professional judgement. In the US, such risk evaluation tools have rivalled and sometimes outperformed other methods such as consensus-based risk assessments or individual decision-making by social workers in terms of accuracy (Ruscio, 1998; Dawes, Faust & Meehl, 1989; Leschied, Chiodo, Whitehead, Hurley, & Marshall, 2003). Despite their success, actuarial risk assessments have been criticised for their inability to meet the inherent methodological challenges of assessing risks in the complex, dynamic, and sensitive field of children's social care and child welfare. Some such risks include (Gambrill & Shlonsky, 2000; Hart, 1998; Doyle & Dolan, 2002):

- the reliability and validity of the variables and measures used to define and quantify risks
- challenges of setting appropriate thresholds for interventions for important concepts, such as 'child in need' in the English context
- the changes that can take place in the real world that may not be reflected by the model
- the fact that actuarial risk assessment tools are optimised to predict a specific outcome often within a specific subpopulation and therefore lack generalisability and applicability to different conditions
- the lack of base rate data about the prevalence of children's maltreatment among the general population which hinders assessing the accuracy of the tools
- bias in the data and over-representation in the base data of certain minority groups based on patterns of contact with children's social care

ML models, like all statistical techniques that rely on correlations to identify patterns in complex data distributions, are subject to these same risks and more. When these risks materialise, they can manifest in poor performance and inaccuracy. Consequences of inaccuracy when identifying children and families in need or at risk of abuse and neglect can include the generation of 'false positives' that stigmatise families, cause stress, waste time, and interfere with the right to privacy and 'false negatives' that miss children in need of protection.

During the roundtable discussion, some highlighted the hazard that public servants and the general population do not have a good enough understanding about the limitations of ML models, such as the uncertainty that underlies statistics in general. They warned against idealising ML-based solutions instead of being realistic about the limited insights that they can extract from data in the CSC environment of practice.

What is machine learning (ML)?

ML is a general approach in computer science that allows algorithms to carry out tasks on the basis of data without being explicitly and completely pre-programmed by designers.

Types of machine learning

There are a few different approaches to ML, each with their own set of strengths, weaknesses, and most appropriate applications. Below we distinguish between supervised, unsupervised, and reinforcement learning. There are, however, a range of possible mixtures between these approaches, such as semi-structured or semi-reinforcement learning.

Supervised learning models are algorithms that are trained on a dataset which contains labelled data. 'Learning' occurs in these models when numerous examples are used to train an algorithm to map input variables (often called features) onto desired outputs (also called target variables). On the basis of these examples, the ML model 'learns' to identify patterns that link inputs to outputs. ML models are then able to reproduce these patterns by employing the rules honed during training to transform new inputs received into classifications or predictions. Classifications, in this context, determine whether inputs fall within one of a set of known classes, and predictions or regression tasks calculate the value of an unknown point on the basis of a set of inputs (Murphy, 2012). Supervised learning requires access to accurately labelled 'ground truth' data. The performance of these algorithmic models can then be tested on labelled data to determine the algorithm's accuracy.

Unsupervised learning algorithms are trained on a dataset without explicit instructions or labelled data. These models identify patterns and structures by drawing inferences from the densities or similarities of data points in the dataset. Such algorithmic models can be used to cluster data (grouping similar data together), to detect anomalies (flagging inputs that are outliers compared to the rest of the dataset), and to associate a data point with other attributes that are typically seen together. Due to the lack of 'ground truth' data the accuracy of unsupervised learning algorithms is difficult to test and assess.

Reinforcement learning algorithms learn on the basis not of existing data, but of their interactions with a virtual or real world. Reinforcement learning 'agents' search for an optimal way to complete a task by taking a series of steps that maximise the probability of achieving that task. Depending on the steps they take, they are rewarded or punished. These 'agents' are encouraged to choose their steps so as to maximise their reward. They 'learn' from past experiences, improve with multiple iterations of trial and error, and may have long-term strategies to maximise their reward overall rather than looking only at their next step. This type of ML model has application in areas where it can interact with a virtual environment (as in games) or with the real environment (as in autonomous vehicles).

The availability, accuracy or lack of certain data, the intended objective of the system, and the domain of its application all play a role in informing the choice of a type of ML model pursued. This decision must also consider ethical imperatives that arise as a result of the envisioned use of the model.

The role of designers and developers in the training of a ML model is very important, regardless of the type of ML model approach chosen. Humans can bake-in their own assumptions, biases, or errors in the design of the model.

With **supervised** learning, designers will make assumptions when determining what the model's target should be, how it should be measured, and what inputs an algorithm should be trained on and in what way. Much literature has focused on the multiple steps along a model's development that provide entry points for human biases.

In **unsupervised** learning designer assumptions and biases may enter the models through the choice of the datasets used and the data features and variables included in the algorithms' training. Moreover, designers may prioritise the finding of certain types of patterns over others and may tweak the optimisation parameters of the algorithm, e.g. the threshold after which clusters are separated. Biases could enter also when users or data scientists interpret the outputs of the model.

Finally, with **reinforcement** learning developers determine the incentive structure for the algorithm. They build the environment in which the algorithm 'learns' on the basis of feedback given, determined by particular outcomes achieved. The creation of the incentive structure for the algorithm itself will represent the assumptions made by designers and developers about what is desirable and how it should be measured.

From our review, it is evident that the most widely used type of ML in the field of CSC is supervised learning, among other reasons because it allows for some level of accuracy and performance testing of the model against existing data. For this reason, the rest of this report will focus mainly on supervised ML models.

Types of analytics

There are different ways in which ML (and other statistical approaches) can be applied to extract insights from data. Depending on the specific orientation of these insights, the above-mentioned ML capabilities (classification, regression, clustering, anomaly detection, etc.) can be used for descriptive, predictive, and prescriptive analytics. What distinguishes these three analytics applications of ML is their degree of complexity, 'autonomy,' and potential impacts, dictated by their increasing sophistication.

Descriptive analytics can help explain past events so as to better inform the present. ML and other models can be used for data aggregation and data mining to highlight patterns and relationships from data and help humans find explanations about past events. By summarising data and presenting it into a human-interpretable manner, descriptive analytics can help humans learn from past experience and better understand data.

Predictive analytics are used to identify possible future outcomes on the basis of inputs and to estimate the likelihood of such outcomes. Predictive analytics involves the use of advanced statistical models and ML models trained on large datasets of examples. ML models used for predictive analytics do not necessarily need to be so complex as to be uninterpretable. They can be, for example, decision trees, which could be both classification decision trees (classifying an input into a set of predetermined

categories) or regression decision trees (calculating the value of an output based on a set of inputs).

Prescriptive analytics use optimisation and simulation algorithms to identify a best set of outcomes given certain inputs and can provide options and suggestions about how to achieve desirable outcomes or mitigate risks. These analytics are a relatively recent development and combine predicting possible future events, as well as their underlying reasons. Prescriptive analytics provide decision support by simulating and quantifying the impact of potential decisions on the future. Prescriptive analytics are complex to implement and involve different analytics techniques, including ML and computational modelling.

Each one of these different analytics can be used in conjunction or at different stages of a model's development. For example, descriptive analytics could be used to understand the relationship between different sets of data—inputs and outputs—which could then be used to identify the most influential input factors for predicting a certain output (James, Witten, Hastie, & Tibshirani, 2017). These analytics could also be used in parallel in complementary yet different tools.

Applications of predictive analytics in social care

What has captured much of the attention of researchers and the general public has been the use of predictive analytics in children's services. This societal application of ML has raised questions *inter alia* about the way assessment technologies impact human autonomy and interpersonal relationships, about the ability of ML models to 'pass judgment' about individuals' future actions, and about the possibility of accurately and completely representing the circumstances of an individual person's life in terms of a set of population-based data and statistical inferences. Indeed, according to a 2016 report by the Behavioural Insights Team (BIT), there are doubts as to whether it is possible to successfully use ML systems for predictive analytics in individual cases of CSC (Tupper et al., 2016). There are a large number of factors that could be predictive of outcomes in CSC and individual cases are likely to be a unique combination of such factors interacting (Tupper et al., 2017). Still, the BIT saw potential for the use of such analytics to provide supplementary information to social workers, e.g. by highlighting the likely outcomes of different cases and prompt deeper considerations by social workers.

Despite ethical concerns, the potential positive impacts of ML have attracted the attention of authorities both in the UK and abroad. The use of risk scoring in particular has been the topic of an extensive report by the Cardiff Data Justice Lab which revealed that local authorities across the UK use such tools in diverse ways, including to target children and families in need of additional support (Dencik, Hintz, Redden & Warne, 2018).

Multiple LAs used the Troubled Families Programme to fund innovative data science projects for better targeting of social care to families with multiple challenges. The Troubled Families Programme (phase 1: 2012-2015, phase 2: 2015-2020) supports the provision of holistic social care services to support families facing multiple significant challenges, including challenges impacting on their ability to care for their children. The programme has resulted in LAs implementing locally tailored programmes. In total, 248,528 families and 864,205 individuals have taken part of the programme (Ministry of Housing, Communities and Local Government, 2019).

Research initiatives around the UK are also taking place. Notably, the Data Lab, a body established as part of Scotland's Innovation Centres programme, is looking to extract useful insights from data in a variety of fields, including children's wellbeing. The Lab is moving towards data for children in collaboration with UNICEF and the research community. Although in its early stages, the Lab is working to establish research projects to identify patterns from data on a larger scale and to use these patterns to help inform policy and practice.

III. SHOULD WE BE DOING THIS?

BIG PICTURE CONSIDERATIONS IN THE ETHICS OF MACHINE LEARNING IN CHILDREN'S SOCIAL CARE

In this section, we explore how a practical ethics of machine learning in children's social care might give us critical leverage to assess the wider question of whether ML technologies, especially predictive risk modelling tools, should be used in the CSC sector in the first place. We will begin with a general discussion of the role of practical ethics in social work and in machine learning. We will then explore the relationship between the two and examine the possibility of integrating them into an applied ethics of ML in CSC. Finally, we will consider how an integrated ethics of ML in CSC might shed light on the question of the moral justifiability of using these technologies to address some of the wider societal and contextual issues that CSC services face—problems such as resource scarcity and austerity, system readiness, long-term cycles of poverty, and deep-rooted public management tendencies to over-rely on technology for the provision of streamlined solutions to deeper and more intractable political and socioeconomic issues.

A wide-angled view of practical ethics

Posing the question, 'Should we be doing this?', about the use of ML in CSC demands that we take a step back for a moment from the day-to-day empirical problems involved in social work and ML innovation and think carefully about the moral purposes and motivations of these practices. 'Should' questions are *normative* questions. They ask us if we 'ought' to be engaging in a particular activity. They call upon us to reflect on whether such engagement is the morally right or wrong thing to be doing.

Answering such a normative question involves making clear one's ethical point of view. Ethical concepts, values, and beliefs put moral demands on our practices, and they constrain what kinds of activities and involvements are justified in circumstances where human actions impact others and wider society. When we reflect on our ethical values and beliefs, we are not only making sense of the principles and rules that make our actions morally justifiable, we are also orienting ourselves to act on those values and beliefs. Ethics, as such, is **both** about *justifying morally correct conduct* **and** about motivating and *setting a direction of travel for that conduct*.

This additional burden of motivating and orienting moral action means that, beyond merely helping us to **justify our practices**, ethics must accomplish two further tasks:

1. It must help us to get an idea of what sort of society we ideally want to live in and what kind of life is morally meaningful and worth living. Ethics must help us reflect on the good: **on living a good life for each and on fashioning a good society for all**.

2. Ethics must help provide us with signposts for those human qualities and morally valuable characteristics that each of us should aspire to obtain and that should give shape to each of our inner lives and moral characters. Ethics must help us reflect on **virtues**: on those **dimensions of character**, like honesty, humility, and courage, that **set our everyday activities and routines on a morally well-purposed and well-directed track**.

As we delve into the specifics of and integrate the ethics of social work and of machine learning, being able to flesh out the role of these three components of ethics writ large will be crucial. In other words, we will need to examine in detail: (1) the action-motivating and direction-setting **ethical values** that support and underwrite practices of CSC and ML innovation, (2) the **principles of conduct** that govern the moral justifiability of these practices, and (3) the **virtues or traits of moral character**—in our case *professional virtues*—that guide upright and appropriate actions and interactions in the context of those practices.

These three elements of ethical values, practical principles, and professional virtues (or standards of behaviour) will provide the mantle for an integrated ethics of ML in CSC that can be put into an actionable form, a form that can be inclusively employed in real-world practices of CSC. Our ultimate aim is that the product of this integration of social work ethics and ML ethics not be just another abstract framework of practical ethics. Rather, we hope for it to be used as a living, convening, and participatory document that signals a common commitment to the shared purpose of using ML technologies in CSC in ways that advance public wellbeing and benefit society—especially the most vulnerable, marginalised, and disempowered of its members. We will call this

document a ***Commitment to care, collaboration, and understanding***.

Before setting this out we provide some context about the interrelated origins of the ethics of social work and of ML.

Overlapping origins of the ethics of social work and of machine learning

Broadly speaking, the ethics of social work and the ethics of machine learning (and, more generally, the ethics of artificial intelligence) share common origins in the imperatives to protect the vulnerable and to champion issues of justice, human dignity, individual worth, and social equity. For the ethics of social work and the ethics of ML/AI alike, these two sets of priorities have been primarily influenced by the traditions of **bioethics and human rights** (Reamer, 1985, 2014; Levi, 2008; Loewenberg & Dolgoff, 1982; Rhodes, 1986; Cowls & Floridi, 2018, 2019; Latonero, 2018).

In a significant sense, both of these traditions emerged out of concerted acts of public resistance against violence done to disempowered people. Whereas human rights has its origins in efforts to redress the well-known barbarisms and genocides of the mid-twentieth century, in the case of bioethics, its emergence tracked the public exposure in the 1960's and 1970's of several atrocities of human experimentation, where it was discovered that members of vulnerable or marginalised social groups had been subjected to the injurious effects of institutionally run biomedical experiments without having knowledge of or giving consent to their participation (Kuhse & Singer, 1998).

The most consequential of these horrors of experimentation came to be known as the Tuskegee experiment—a forty-year study on syphilis carried out by the US Public Health Service in which treatment was withheld from impoverished African American males suffering from the disease in order to study the natural history of its long-term effects (Katz et al., 2008; Reverby, 2009; Luna & Macklin, 1998). As a result of acts of whistleblowing and public outrage about this study, the US Congress formed the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research in 1973. This Commission ultimately produced what is now considered to be one of the founding documents of bioethics, *The Belmont Report*, which laid out the basic principles that were felt to be essential for guiding and governing the treatment of research subjects: **respecting each individual's capacity for autonomy and self-determination, protecting people from harm, looking after the well-being of others, and treating all individuals equitably and justly**.

While these tenets of bioethics (respect for persons, beneficence, and justice) largely stress the ethical values that underlie the safeguarding of individuals in instances where the exercise of technological and administrative power puts them in harm's way, the human rights tradition mainly focuses on the set of social, political, and legal entitlements that are due to all human beings under a universal framework of juridical protection and the rule of law. Responding directly to political acts of violence inflicted on people at the societal level, human rights discourse is anchored in a set of universal principles that build upon the idea that all humans have an equal moral status as bearers of intrinsic human dignity. As rights-bearing individuals, all people are therefore entitled to equal freedom and moral regard under the law, to the protection of their civil, political, and social rights, to the universal recognition of their personhood, and to the right to free and unencumbered participation in the life of the community.

It is easy to see why the combination of the weight bioethics places on the care and protection of vulnerable individuals and the weight the human rights discourse places on social justice and on the protection of civil, social, and political rights has formatively appealed to both social work ethics and the ethics of ML/AI.

History of ethics in social work

Moral values such as dignity, autonomy, uniqueness, respect, justice, and equality have informed the practice of social work since before the 1950's (Reamer, 2014; Biestek, 1957; Cabot, 1973; Hamilton, 1951; Joseph, 1989). The deep engagement of social work professionals with practical ethics and with ethical codes of conduct did not occur, however, until after the explosion of interest in applied and professional ethics that was spurred in the 1970's and 1980's by the expanding influence of bioethics. Not only would this field have a profound impact on many of the related aspects of social work's focus on caring for and ensuring the safety and health of families (Reamer, 1985; Levi, 2008), but also the basic principles of bioethics have remained central features of ethical codes of conduct in social work across the globe.

In a similar way, the normative perspective of human rights discourse has enabled social work ethics to transformatively reframe the societal

problems of poverty, inequality, and deprivation in terms of a rectification of the injured dignity of service recipients, making the provision of social care and support a restorative, constructive, and natively ethical pursuit. It should be noted, though, that—beyond the mainstream ethical orientations of bioethics and human rights—the influences of the **ethics of care tradition** and of **progressive thinking on social justice** have led the ethics of social work to put special emphasis on the importance of empathy, of interpersonal solidarity, and of the rights of the oppressed and powerless (Tronto, 1993; Lonne, Harris, Featherstone & Gray, 2016; Featherstone, 2001; Gray, 2010; Hugman, 2005; Meagher & Parton, 2004; Orme, 2008; Parton, 2003; Marion-Young, 1990; Ife, 1997; Healy, 2001; Skegg, 2005; Mapp, 2008).

History of ethics in machine learning

For the more recently developed ethics of machine learning (ML) and artificial intelligence (AI), the widespread appeal to the principles of bioethics (Jobin, Ienca & Vayena, 2019; Zeng, Lu, & Huangfu, 2018; Floridi & Clement-Jones, 2019) has emerged organically from the exposure of individuals affected by the 'decisions' and behaviours of ML and AI applications to the potentially harmful but often subtle, depersonalised, and unseen exercise of technological power. This vulnerability of decision subjects to injuries inflicted by such automated systems has been punctuated by the inherent accountability gap that arises in the use of such systems (Leslie, 2019). In 'making decisions' and performing tasks that have previously required the thinking and reasoning of responsible humans, ML systems are increasingly serving as practical fiduciaries or trustees of humanly impactful decision-making without at the same time being accountable, in any non-metaphorical way, for the outcomes of such processes. Whereas human agents can be called to account for their judgements and decisions in instances where those judgments and decisions affect the interests of others, the statistical models and underlying hardware that compose ML systems are themselves not responsible in a morally relevant sense (EPSRC, 2011; Bryson, 2017). As inert and program-based machinery, ML/AI systems are not legally or morally accountable agents. An appeal to the principles of bioethics has, in large part, been an attempt to fill this gap by bringing the human-centred values of respect for persons, interpersonal care, and equitable treatment to the forefront.

ML/AI ethics has also drawn upon the normative resources of human rights discourse in order to remedy such an accountability gap. In articulating the social and political conditions necessary for the universal recognition of human dignity, the frame of human rights is intended to provide direction for the governance constraints that should be placed on the design and use of ML systems, so that these conditions can be realised. When starting from a human rights perspective, ML technology producers and users are guided to place appropriate parameters on the behaviour of their artefacts so that values of equality and non-discrimination as well as universal social and political rights can be prioritised and safeguarded.

Beyond this navigating and ballasting function, the incorporation of human rights principles into the practical ethics of ML/AI has been more chiefly motivated by a need to respond to the set of difficult macroscale societal problems that the rapid and global proliferation of these technologies is now presenting. Quintessential among these is the question of the inequitable distribution of the societal benefits that may be reaped from ML/AI innovation. Existing patterns of inequality inevitably shape people's access to the benefits of emerging technologies. This is especially true when the means of production of those technologies are largely controlled by monopolistic syndicates—as is arguably the situation in today's big tech-driven 'platform' political economy (van Dijck, 2018; Gillespie, 2018; Bratton, 2015).

In the case of ML/AI technologies, the prospect of the inequitable distribution of their societal benefits is a crucial issue, because, as they continue to improve with the availability of data and the expansion of computing power, these systems are increasingly developing into non-substitutable supports for the provision of critical social goods and services across all sectors of society. Inasmuch as such systems are, in fact, thus becoming gatekeepers for the advancement of vital, humanity-level public interests, the asymmetrical or skewed distribution of the benefits they yield at *both local and global levels* will have greater and greater social impacts (especially on those who are most vulnerable, powerless, and liable to economic displacement) and will potentially generate ever more widescale harms. Human rights discourse has been a good fit for the normative redress of such universally consequential issues.

Integrating the ethics of social work and machine learning /AI ethics

Having explored some common motivations and overlapping moral foundations of social work ethics and ML/AI ethics, we are better positioned to find shared ground among the values, principles, and professional virtues that are prevalent in these two fields in order to provide some serviceable guidance for considering the ethical issues surrounding the use

of ML in CSC. Such an integration task will involve two steps.

First, we will draw upon a wide range of the existing codes of conduct, principles statements, and ethical frameworks in both fields to provide a detailed picture of the resonances and points of convergence between them. We will employ the three-tiered structure of the components of practical ethics outlined above (i.e., ethical values, practical principles, and professional virtues) to map out these resonances and convergences.

We will then use the results of this mapping to provide a preliminary view of what their integration into an applied ethics of ML in CSC might look like. The central goal of such an integration will be to put the applied ethics of ML in CSC into an optimally useable and accessible form, so that it can be actively adopted by all affected stakeholders as a vehicle of common commitment to the shared purpose of using these technologies exclusively in ways that advance public wellbeing and benefit society.

Ethical values: Setting the direction of travel and motivating moral action

The table on the next page presents a summary comparison of the core ethical values of social work and ML/AI. It aligns the values they share row-by-row and consolidates them in the left column into four basic values that motivate and set the direction of travel for the responsible use of ML in CSC:

- **Respect** the dignity of individual persons, empower them, and value the uniqueness of their aspirations, cultures, contexts, and life plans
- **Connect** with each other sincerely, openly, and inclusively, and prioritise trust, solidarity, and interpersonal collaboration
- **Care** for the wellbeing of each and all, and serve others with empathy, selflessness, and compassion
- **Protect** the priorities of social justice and the public interest by ensuring equity, recognising diversity, and challenging discrimination and oppression

Taken together, these core ethical values form the motivational and navigational centre of an applied ethics of ML in CSC. They provide us with a principled starting point for setting the direction of travel in applying ML innovation to CSC, and they give us an idea of the standards against which the choice to use ML applications in CSC can be assessed and justified.

Table 1. Aligning and integrating the ethical values of social work and of AI and machine learning ethics

Ethical values of social work and ML/AI		
Values for the ethics of ML in CSC	**Values from the ethics of social work**	**Values from the ethics of ML/AI**
Respect the dignity of individual persons, empower them, and value the uniqueness of their aspirations, cultures, contexts, and life plans	Uphold and promote human dignity (1, 3, 4, 5, 8)	Ensure individuals' abilities to make free and informed decisions about their own lives (9, 10, 12, 14, 15, 17, 18, 19)
	Empower people (1, 3)	Safeguard autonomy, the power of self-expression, and the right to be heard (9, 11, 12, 14, 15, 16, 18)
	Value each individual person in the uniqueness of their goals, contexts, passions, and life plans (1, 3, 4, 5, 7, 8)	Secure each individual's capacities to contribute to the life of the community (12, 15, 18)
	Respect the right to socially responsible self-determination (1, 2, 3, 4, 5, 7, 8)	Support each person's ability to flourish and to pursue their passions and talents according to their aspirations (9, 10, 12, 14, 15, 16, 17)
	Identify and develop individual strengths (1, 4, 5)	
Connect with each other sincerely, openly, and inclusively, and prioritise trust, solidarity, and interpersonal collaboration	Promote the right to participation (1, 3, 4, 5, 6, 7, 8)	Safeguard interpersonal dialogue, meaningful human connection, and social cohesion (11, 12, 15, 18)
	Treat each person as a whole within the family and community (1, 4, 8)	
	Prioritise relationships with families (4, 6, 7, 8)	Prioritise participation, diversity, inclusion, and consideration of all voices (10, 11, 12, 15, 18, 19)
	Engage people as partners in the helping process (1, 3, 4, 6, 7, 8)	
	Practice social care work relationally and dialogically (3, 4, 5, 6, 7, 8)	Use the AI and ML technologies to enable bonds of interpersonal solidarity (10, 12, 14, 15, 18)
	Work in solidarity (1, 4, 6, 7, 8)	
	Strive to build the trust and confidence of families (1, 2, 6)	Use AI and ML technologies to reinforce trust, empathy, reciprocal responsibility, and mutual understanding (9, 12, 13, 14, 15, 16, 17, 18, 19)
Care for the wellbeing of each and all, and serve others with empathy, selflessness, and compassion	Uphold and promote wellbeing (1, 4, 8)	Design and deploy AI and ML systems to foster the welfare of all stakeholders (9, 10, 12, 13, 14, 17, 19)
	Practice empathy, compassion, and care (1, 4, 8)	Do no harm with these technologies and minimise the risks of their misuse or abuse (9, 10, 11, 12, 13, 14, 15, 16, 17, 19)
	Serve others above self-interest (1, 5, 7, 8)	Prioritise the safety and the mental and physical integrity of people when conceiving of and deploying AI applications (9, 11, 13, 14, 15)

Protect the priorities of social justice and the public interest by ensuring equity, recognising diversity, and challenging discrimination and oppression	Advocate for the vulnerable and oppressed **(4, 6, 7, 8)**	Treat all individuals equally and protect social equity **(10, 11, 14, 15, 16, 19)**
	Distribute resources fairly, according to need **(1, 4)**	
	Pursue social change **(1, 4, 5, 7, 8)**	Use digital technologies as a support for the protection of fair and equal treatment under the law **(10, 11, 12, 13, 14)**
	Exercise power with others for the collective good **(1, 2, 3, 6, 8)**	Prioritise social welfare, public interest, and the consideration of the social and ethical impacts of innovation in determining the legitimacy and desirability of AI technologies **(10, 11, 13, 14, 15, 17)**
	Challenge the abuse of human rights **(1, 3, 4, 5, 8)**	
	Challenge discrimination **(1, 2, 3, 4, 5, 7, 8)**	Use AI to empower and to advance the interests and well-being of as many individuals as possible **(10, 12, 13, 14, 15, 17)**
	Challenge unjust policies and practices **(1, 2, 3, 4, 6)**	
	Recognise diversity **(1, 3, 4, 5, 6, 7, 8)**	Think about wider impacts on community, society, and biosphere **(10, 12, 14, 15, 16, 17)**
	Promote sensitivity to oppression and cultural and ethnic diversity **(2, 3, 4, 5, 6, 7, 8)**	

Select social work ethics documents:

1. British Association of Social Workers – The Code of Ethics for Social Work: Statement of Principles (2012)
2. General Social Care Council – Codes of Practice for Social Care Workers (2010)
3. Northern Ireland Social Care Council – The National Occupational Standards for Social Work (2003)
4. International Federation of Social Workers – Global Social Work Statement of Ethical Principles (2018)
5. Canadian Association of Social Workers – Code of Ethics (2005)
6. National Association of Social Workers – NASW Standards for Social Work Practice in Child Welfare (2013)
7. NASW – NASW Code of Ethics (2017)
8. Australian Association of Social Workers – AASW Code of Ethics (2010)

Select AI and ML ethics documents:

9. IEEE – Ethically aligned design, 1st Edition (2019)
10. EU High-level Expert Group on AI – Ethics Guidelines for Trustworthy AI (2019)
11. Access Now and Amnesty International – Toronto Declaration (2018)
12. UK House of Lords, Select Committee on Artificial Intelligence - AI in the UK: Ready, willing and able? (2018)
13. DCMS – Data Ethics Framework (2018)
14. AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations (2018) (Floridi et al., 2018)
15. University of Montreal – Montréal Declaration: Responsible AI (2017)
16. ACM US Public Policy Council – Statement and Principles on Algorithmic Transparency and Accountability (2017)
17. Future of Life Institute – Asilomar AI Principles (2017)

18. The Royal Society – Machine learning: The power and promise of computers that learn by example (2017)
19. FAT/ML – Principles for Accountable Algorithms and a Social Impact Statement for Algorithms (2016)

Note that links made between the ethical values of ML/AI ethics and social work ethics and the selected documents highlight commonalities and shared trends in the development of these values. The wording in each separate document is, however, different and should be treated as uniquely meaningful in each separate context of use.

Practical principles: Establishing the moral justifiability of the practices of social care and ML/AI innovation

Building on the above ethical values, practical principles of applied ethics of ML in CSC orient and constrain justified conduct. They answer the more practice-based question: How can we do this right?

The table below presents a summary comparison of the core practical principles of social work and ML/AI. It aligns the principles they share row-by-row and consolidates them in the left column into three big picture goals of practice that are shared in common by social work and responsible ML innovation and that therefore support and reinforce the objectives of best practices for the use of ML in CSC:

- **Fair, sustainable, ever-improving social care**
- **Social care that supports and empowers**
- **Transparent, responsible, and accountable social care**

Table 2. Aligning and integrating the practical principles of social work ethics and of AI and ML ethics

Principles for social work and machine learning / AI		
	Practical principles of social work	**Ethical principles for social work and mahcine learning / AI**
Fair, sustainable, ever-improving social care	Uphold public confidence in social care services (1, 2, 3)	Produce sustainable ML innovations that are safe from a technical and operational point of view and ethical/ morally justified in their outcomes and wider impacts (9, 10, 12, 13, 14, 15, 16, 17, 18) Ensure that the ML systems you design and implement are fair, equitable, and do not do harm through bias or discrimination (10, 11, 12, 13, 14, 15, 19)
	Establish and maintain the trust of families (2, 3, 5)	
	Facilitate and contribute to evaluation, research, and improvement (1, 3, 5, 6, 7, 8)	
	Develop professional relationships and learn from others (1, 3, 5, 6, 8)	
	Help others develop and teach others (3, 6, 7, 8)	
	Be prepared to whistle-blow and protect the interests of the vulnerable and powerless (1, 2, 3, 4)	

Social care that supports and empowers	Act with the informed consent of families, unless required by law to protect that person or another from risk of serious harm **(1, 2, 3, 5, 8)**	Ensure that algorithmically supported outcomes are applied to the individual's life that they affect with appropriate sensitivity to the specific circumstances of that life and to the unique qualities of that individual's identity, context, and relationships **(9, 12, 14, 15, 17)**
	Provide information clearly and understandably **(1, 3, 8)**	
	Strive for objectivity and self-awareness in professional practice **(1, 3, 5, 8)**	Ensure that algorithmically supported outcomes are interpretable and can be made easily understandable to affected parties **(9, 10, 11, 12, 14, 15, 16, 18, 19)**
	Maintain confidentiality and explain policies about confidentiality to families **(1, 2, 3, 5, 6, 8)**	
Transparent, responsible, and accountable social care	Be accountable for work quality and take responsibility for improving professional knowledge and skills **(1, 2, 3, 5, 8)**	Be transparent: by ensuring that design and implementation processes are open, accessible, and justifiable through and through. **(9, 10, 11, 13, 15, 16)**
	Maintain clear and accurate records **(1, 2, 3, 6, 8)**	
	Take responsibility for one's own practice and continuing professional development **(1, 2, 3, 5, 6, 7, 8)**	Be accountable and answerable for the part you play across the entire ML design and implementation workflow and make sure that the results of this work are traceable/auditable from start to finish. **(10, 11, 13, 14, 15, 16, 17, 18, 19)**
	Share information appropriately **(1, 3, 5, 8)**	
	Report resource or operational hurdles to helping people **(1, 2, 3, 4, 6)**	

Select social work ethics documents:

1. British Association of Social Workers – The Code of Ethics for Social Work: Statement of Principles (2012)
2. General Social Care Council – Codes of Practice for Social Care Workers (2010)
3. Northern Ireland Social Care Council – The National Occupational Standards for Social Work (2003)
4. International Federation of Social Workers – Global Social Work Statement of Ethical Principles (2018)
5. Canadian Association of Social Workers – Code of Ethics (2005)
6. National Association of Social Workers – NASW Standards for Social Work Practice in Child Welfare (2013)
7. NASW – NASW Code of Ethics (2017)
8. Australian Association of Social Workers – AASW Code of Ethics (2010)

Select AI and ML ethics documents:

9. IEEE – Ethically aligned design, 1st Edition (2019)
10. EU High-level Expert Group on AI – Ethics Guidelines for Trustworthy AI (2019)
11. Access Now and Amnesty International – Toronto Declaration (2018)
12. UK House of Lords, Select Committee on Artificial Intelligence - AI in the UK: Ready, willing and able? (2018)
13. DCMS – Data Ethics Framework (2018)
14. AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations (2018) (Floridi et al., 2018)
15. University of Montreal – Montréal Declaration: Responsible AI (2017)
16. ACM US Public Policy Council – Statement and Principles on Algorithmic Transparency and Accountability (2017)
17. Future of Life Institute – Asilomar AI Principles (2017)
18. The Royal Society – Machine learning: The power and promise of computers that learn by example (2017)
19. FAT/ML – Principles for Accountable Algorithms and a Social Impact Statement for Algorithms (2016)

Note that links made between the practical principles of ML/AI ethics and social work ethics and the selected documents highlight commonalities and shared trends in the development of these principles. The wording in each separate document is, however, different and should be treated as uniquely meaningful in each separate context of use.

Professional virtues: Establishing common principles of professional integrity shared by social work and responsible ML/AI innovation

The professional virtues of a given domain of activity guide upright and appropriate actions and interactions in the context of that activity. They indicate the path of least resistance to carrying out that activity with integrity, and they thus set the everyday practices and routines that form the basis of that activity on a morally stewarded track. The table below provides a consolidation of the common principles of professional integrity shared by social work and responsible ML/AI innovation:

Table 3. Common professional virtues of social work ethics and of AI and machine learning ethics

Uphold ethical values and best practices
Be sincere, honest, and trustworthy
Lead by competence and example
Maintain appropriate professional boundaries
Make considered professional judgements
Be professionally responsible
Be objective and impartial in making professional judgments
Use evidence-based reasoning when rendering decisions

Bringing it all together in a *Commitment to care, collaboration, and understanding*

The three elements of an integrated applied ethics of ML in CSC that we have just presented (ethical values, practical principles, and professional virtues) are meant to provide the mantle for the responsible and well-purposed use of ML technologies in the CSC context. However, it is important to avoid the trap of treating such an ethical framework as independent from the warm-blooded processes of technological innovation and the concrete practices of CSC that make up the responsible design and implementation of ML in CSC.

To the contrary, the ultimate aim of setting out an integrated ethics of ML in CSC must be to put it into an actionable form that brings together all stakeholders involved in the complex, multi-level and collaborative processes of conceptualising ML applications and projects, defining their objectives, and designing, deploying, monitoring, and implementing them. Ethical reflection and inclusive, ethically-informed dialogue must be involved at every step of this process. What is needed for this is a vehicle of common commitment—a way for all those who are involved in the responsible design and use of data scientific applications to continuously coalesce around a mutual recognition of the ethical motivations, practical principles, and professional standards of conduct that should motivate, direct, underwrite, and steer responsible practices of ML innovation in the field of CSC.

One way to accomplish this is to concretise and specify such a common commitment in a living document that signifies a joint dedication to the shared purpose of using ML technologies in children's social care in ways that advance public wellbeing and benefit society—especially the most vulnerable, marginalised, and disempowered of its members. We will call this document a ***Commitment to care, collaboration, and understanding***. On the basis of our research and drawing upon the input gathered from the participants at our roundtable and workshop, we offer here a preliminary mapping of what this might look like:

Commitment to care, collaboration, and understanding

for the responsible use of data science in social care

We come together as citizens, social workers, data scientists, and civil servants to affirm our shared commitment to fostering human flourishing, to sustaining care, connection, and support, and to building a better, more inclusive, and more just society through our actions and through the technologies we design and implement.

We commit:

To Respect the dignity of every person, to empower them, and to value the uniqueness of their aspirations, cultures, contexts, and life plans;

To Connect with each other sincerely, openly, and inclusively and to prioritise trust, solidarity, and interpersonal collaboration;

To Care for each other's wellbeing and serving others with empathy, selflessness, and compassion; and

To Protect the priorities of social justice and the public interest by ensuring equity, recognising diversity, and challenging discrimination and oppression

We commit to collaborate in pursuing these common ends by listening to each other carefully, patiently, and thoughtfully, and by striving for mutual understanding with candidness, honesty, and sincerity. We commit to being objective and impartial in our judgments and determinations, while remaining humane and empathetic in our efforts to understand.

We together affirm that our practices of care and any tools we design and use to support them will be fair, sustainable, supportive, consent-based, and transparent. And we commit to one another and to all others affected by such practices and such tools that we will be accountable and take responsibility for the consequences of our decisions and behaviours. We further affirm that if we find our tools are not helping us to achieve our common ends, we will improve them so that they do or else not use them at all.

Finally, we together affirm that, in determining whether to innovate as well as the direction of our innovation, we will actively seek the counsel of all affected people—especially of the historically marginalised, powerless, and voiceless—and endeavour to remain worthy and faithful stewards of the shared life of the community, of humanity, and of the biosphere as a whole.

Should we be doing this?: Applied machine learning ethics in children's social care as a critical yardstick

So far, this study has been putting into place building blocks for providing an answer to the first part of the question put to its authors by WWC, namely, *"Is it ethical to use machine learning approaches in children's social care systems, and if so, how?".*

To this end, the first two levels of the integrated ethics of ML in CSC outlined in the previous sections (i.e. the levels of ethical values and practical principles) provide critical and analytical leverage for responding to each of the two respective parts of the question. The ethical values address the first part of the question: Is it ethical to use ML in CSC? They key us in to the normative standards against which the choice of using ML applications in CSC can be criticised, judged, and justified. The practical principles address the second, 'how' part of the question. They allow us to hold the actual design and implementation of the ML systems to moral-practical criteria such as fairness, accountability, and transparency, assuming, of course, that these systems have been deemed ethically permissible in the first place.

In this section, we look closely at how the ethical values that we have identified as lying behind the responsible use of machine learning in social care might provide a critical yardstick of sorts for the use of ML in CSC. We examine several external, empirical factors, which might call into question the justifiability of the application of this kind of technology in the sensitive and demanding domain of CSC.

In the ensuing sections on data use, model design, and implementation, we investigate how the practical principles might help both to provide guardrails for responsible conduct and to give shape to best practices from a point of view internal to the boots-on-the-ground activities of ML innovation and use.

Continuing legacies of austerity and New Public Management

As previously mentioned, one of the main empirical concerns raised by roundtable participants pertained to the possibility that the use of predictive analytics in CSC by local authorities was driven by the need to find effective and efficient solutions in the context of austerity measures, rising demands, and understaffing. Fears that the use of ML in CSC marks a continuation of longer term trends of the 'technicisation of social work,' in the wake of a greater 'bureaucratisation of organisations and response of social work professionals' (Otway, 1996; Howe, 1992) originate in the depersonalisation and de-communalisation of social care that has been part of the governance techniques of New Public Management. Such techniques have largely taken root as effects of post-financial crisis austerity strategies.

Though largely aimed at creating economic efficiencies and increasing institutional effectiveness by decentralising administrative structures and quantifying performance measurement (Dunleavy & Hood, 1994; Banks, 2011), New Public Management has also had negative effects. It has, from a critical perspective, functioned to create hyper-proceduralised regimes of administrative control and monitoring, thereby chipping away at the human relationship-centred practices and professional judgement-based service delivery that have traditionally been at the core of social care (Bywaters et al., 2017; Veale & Brass, 2019). Indeed, as austerity pressures have endured, there has been a growing sense that these trends towards the quantification of performance, risk, and accountability and the overemphasis on output measurement have increasingly made the adoption of ML technologies by local authorities (whether or not these technologies are actually effective) a support for the dehumanising end of administrative efficiency.

From an applied ethical point of view, insofar as the incorporation of ML systems into mechanisms of service provision operates, in practice, to devalue and to attenuate human-centred approaches to CSC that stress community- and family-based social care, they are not ethically permissible. This is not to say that ML systems *as such* are to be avoided in CSC, but rather that, when these systems buttress the exercise of administrative power in ways that displace the values of respecting human agency, supporting interpersonal connection, sustaining wellbeing through care, and advancing social justice for the vulnerable, they are not morally justifiable. Moreover, when the use of these data scientific tools enables output-centred, risk-framed, and measurement-based attitudes of public management to avoid reflexive and critical interrogation of the structural causes behind the societal problems being 'solved' by the tools themselves, they are liable to unethically perpetuate the dynamics of domination and inequity that underlie those very same problems.

System, organisation, and participant readiness

A related, but wider-scale, empirical concern about the justifiability of integrating ML innovation into the current practices and community settings of CSC was expressed by participants of our Family Engagement workshop. Here, families pointed out that the present conditions of the social care system (including

prevailing organisational cultures, attitudes, skill levels, and resources) might not be conducive to the effective adoption of responsible ML innovation, especially in cases of individual-impacting frontline decision support. This apprehension about system, organisation, and participant (SOP) readiness raises a set of substantive questions that have a direct bearing on the justifiability of using ML in CSC, given existing circumstances: To what degree can current systemic conditions, organisational contexts, and the cognitive and psychological antecedents of participant action support the ethically-guided design and use of ML systems in CSC? To what degree can they enable and promote the effective integration of values-based innovation into practices and community settings?

Being able to answer these questions about SOP readiness is crucial inasmuch as this insight will allow us to understand constraints on the feasibility and potential uptake of ethically-informed ML projects in CSC. As Ghate observes, following Fixsen et al., 'All new interventions are required to take their place within a wider and preexisting system of care. If the existing system is already stressed, if it rejects, obstructs, or even simply fails actively to support a new intervention, and if it does not act as a good "host," the most promising innovations can easily become marginalized and fail to sustain their promise' (Ghate, 2016; Fixsen, Naoom, Blasé, Friedman, & Wallace, 2005). In this respect, being able to grasp the state of SOP readiness in the current children's social care environment is vital, because it would offer us access to the ecological, cultural, and attitudinal impediments that may need to be cleared away for justifiable ML applications to be sustainably integrated into practices of care in CSC settings.

First and foremost, such an understanding requires knowledge of the current empirical determinants (barriers and enablers) of the effective integration of responsible ML innovation practices into the social care environment. While more programmatic research must still be undertaken to explore the real-world barriers and enablers of responsible ML innovation in the domain of CSC, preliminary steps in this direction have already been initiated. Drawing on work of McKinsey, the thematic analysis of the Department for Education Children's Social Care Innovation Programme (Round One) identified '10 main barriers to radical improvement and innovation in children's services in England' (Sebba, Luke, McNeish & Rees, 2017). The McKinsey analysis organised these barriers into structural and practical components:

Main barriers to radical improvement and innovation in children's services in England (From Sebba et al., 2017)

Structural barriers	Practical barriers
▪ LA's can lack the ***critical mass and organisational competence*** to commission for radical improvement and innovation in services ▪ Service provider organisations aren't always ***incentivized to radically improve and innovate*** their services	▪ Frontline social care staff don't always have the ***time, skills, or confidence*** to radically improve and innovate services ▪ LA ***leadership can lack capability and incentives*** to radically improve and innovate services ▪ There are some ***legal barriers***, and many ***perceived barriers***, to radically improving and innovating services ▪ New innovations often fail to '***spark***' and proven innovations often fail to ***spread effectively*** across LA's ▪ Poor ***data quality and availability*** makes it hard for social workers, LA's, regulators, and central government to drive radical improvement or innovation ▪ Current ***performance management*** system tends to promote ***compliance*** rather than radical improvement and innovation ▪ Challenges in the ***collaboration at the interface*** of different agencies limits innovation, particularly for child protection ▪ Culture of resistance to change and ***risk aversion***

While extremely valuable in helping to draw attention to some of the crucial determinants of the effective integration of responsible ML innovation into the care environment, this snapshot of barriers does not yet provide a full view of where they are situated in broader systemic and organisational contexts and of how these determinants are interrelated. From such a wider-angled standpoint, beyond considering any one of these various obstacles to innovation in isolation, attention must also be paid to how such obstacles are embedded in the broader social, cultural, economic, legal, and political contexts of CSC. In pursuing this more holistic approach, we would be better positioned to identify the underlying infrastructure of systemic, organisational, and psychological/motivational factors that operate separately and in concert to determine the readiness of participants in children's social care settings to accept the changes brought about by potentially disruptive innovations (Metz & Albers, 2014; Ghate, 2016).

This broader ecological view of the innovation environment is important, because it provides a useful way to organise the positive and negative determinants of innovation intervention outcomes. In particular, it enables us to arrange the determinants in a systematic manner, which may then allow for the development of a more deliberate and logical approach to identifying and anticipating them. It consequently works toward clearing possible pathways to capacity-building.

Much of this research into understanding the effective integration and sustainability of innovation and evidence-based interventions in organisational and community settings has already been undertaken in the fields of implementation science, organisational theory, social psychology, and sociology, among others (For helpful surveys: Tabak, Khoong, Chambers, & Brownson, 2012; Nilsen, 2015; Leeman et al., 2015; Strifler et al., 2018). As a whole, these approaches have scrutinised (1) the design and implementation processes behind effectively or ineffectively translating research into practice, (2) the specific types of structural and agential determinants that influence the success of implementation outcomes, and (3) the aspects of implementation outcomes that can be identified as indicative of success or failure (Nilsen, 2015).

While a significant amount of effort among researchers, who adopt this ecological perspective, has been dedicated to various health and social care settings (for example, Griffith, Zammuto, & Aiman-Smith, 1999; Klein & Knight, 2005; Rizzuto & Reeves, 2007; May, Mair, Dowrick, & Finch, 2007; Aarons & Palinkas, 2007; Powell et al., 2019; Moullin, Dickson, Stadnick, Rabin, & Aarons, 2019) programmatic investigation of this kind in the context of the use of ML in CSC is still in its nascent stages. Be that as it may, in considering the degree to which existing conditions of SOP readiness support the ethically-guided design and use of ML systems in CSC, it may be helpful, following insights from implementation research and organisation theory, to explore a preliminary mapping of the interconnected structural, organisational, and participant contexts which have, in that literature, been identified as determinative of the success or failure of innovation projects and evidence-based interventions. Building largely off of the scaffold of factors and contexts provided by the Exploration, Preparation, Implementation, Sustainment (EPIS) framework (Moullin et al., 2019; Powell et al., 2019), such a mapping should include both the specific SOP contexts and the innovation factors that cuts across them:

Preliminary mapping of some key determinants of success for SOP readiness	
Contexts/factors and their variables	**Significant determinants of effective integration**
Innovation factors ▪ Innovation-values fit ▪ Innovation-needs fit ▪ Innovation-knowledge fit	**Innovation-values fit:** Success of an innovation intervention will be affected by the degree to which its characteristics align with the values, beliefs, purposes, and mission of the innovation producers, users, and individuals affected by its implementation (Klein & Sora, 1996; Glisson & Schoenwald, 2005; Aarons, Hurlburt, & Horwitz, 2011). **Innovation-needs fit:** Success of an innovation intervention will be affected by the degree to which its characteristics align with the administrative and practice needs of users and the service needs of individuals affected by its implementation (Klein & Sora, 1996; Aarons et al., 2011; Moullin et al., 2019). **Innovation-knowledge fit:** Success of an innovation intervention will be affected by the degree to which its characteristics align with users' cognitive needs, adaptability, skill levels, and capabilities; organisations' commitments to training and development; and the cognitive participation, sense-making, and informed acceptance of users and individuals affected by its implementation. (Zahra & George, 2002; Murray et al., 2010; Proctor et al., 2011; Finch et al. 2013).
System/outer context ▪ Socioeconomic and political setting; service environment; regulatory and policy environment ▪ Interorganisation relationships and networks ▪ Leadership characteristics of those outside of intervening organisations who are in a position to support innovation ▪ Support and funding that is external to organisations and agencies	**Developing inclusive participation, supportive relation-ships, and alignments between relevant stakeholder groups:** Success of an innovation intervention will be affected by the degree to which there is inclusive participation and co-creation in development and implementation processes across governmental and non-governmental organisations, communities, and affected individuals. This includes family-based and participatory decision-making processes in the service environment (Glisson & Schoenwald, 2005; Crampton, 2007; Aarons et al., 2011; Ghate, 2016). This also includes the dynamic sharing of knowledge and experience between relevant innovation designers, users, and affected stakeholders so that solutions are optimised and rendered acceptable to all affected by them (Aarons, Fettes, Sommerfeld, & Palinkas, 2012; Moullin et al., 2019). **Partnership building:** Success of an innovation intervention will be affected by the degree to which meaningful partner-ships can be formed and cultivated between organisations, community groups, and affected individuals, so that the innovation is cooperatively shaped and collectively monitored for quality (Mendel, Meredith, Schoenbaum, Sherbourne, & Wells, 2008). In knowledge-based innovation environments, community-academic partnerships are especially important (Aarons et al., 2014) **Interorganisation cooperation:** Success of an innovation intervention will be affected by the degree to which meaningful and continuous collaborations are undertaken between relevant organisations. These collaborations are recursively interactive: there is a reciprocal responsiveness to feedback and input between actors, which enables organisational learning (Becan et al., 2018). **System-level leadership competence:** Success of an innovation intervention will be affected by the degree to which governmental leadership at all levels establishes affirmative goals as well as actively supports and promotes the innovation (Akerlund, 2000; Mancini & Marek, 2004; Moullin et al., 2019). **External support and fidelity monitoring:** Success of an innovation intervention will be affected by the degree to which external support, beyond training and education provided by intervention developers, is available for users. Initial training and internal oversight by organisations are often not sufficient to guarantee implementation fidelity and sustainability (Sabalauskas, Ortolani, & McCall, 2014; Powell et al., 2019). **Addressing service-level resource barriers:** Success of an innovation intervention will be affected by the degree to which limited resources cause logistic impediments to service provision such as lack of appointment availability or inconvenient service locations. Likewise, financial resource limitations on the service-user side can create impediments to the penetration and sustainability of innovation due to the inability of disadvantaged service users to engage in the intervention (Powell et al., 2019).

Organisational/inner context ▪ Organisational culture and climate; institutional receptivity to change ▪ Policies, procedures and past experience ▪ Leadership characteristics at every organisational level and stage of design and implementation ▪ Fidelity monitoring and quality assurance ▪ Resource availability from exploration to sustainment	**Absorptive capacity**: Success of an innovation intervention will be affected by the degree to which an organisation is able to build upon a strong knowledge and skills base and assimilate new knowledge into existing practices and capabilities. This assimilative ability is often supported by established mechanisms for sharing and disseminating knowledge throughout the organisation (Damanpour, 1991; Ferlie & Shortell, 2001; Grol, Bosch, Hulscher, Eccles, & Wensing, 2007; Aarons et al., 2011). Such an ability may be challenged by excessive workloads, high levels of variation between workers in their training and educational background, and non-specialised roles that demand completion of multiple tasks (Yoo, Brooks, & Patti, 2007; Ebert, Amaya-Jackson, Markiewicz, Kisiel, & Fairbank, 2012; Gleacher et al., 2011; Lang, Franks, Epstein, Stover, & Oliver, 2015; Nadeem & Ringle, 2016; Wenocur, Parkinson-Sidorski, & Snyder, 2016; Powell et al., 2019). **Change Readiness:** Success of an innovation intervention will be affected by the degree to which an organisation's members share confidence in their efficacy to implement change, value change as important and beneficial, reject institutional inertia, and share a resolve to initiate, persist, and co-operate in carrying out innovation (Weiner, 2009; Gleacher et al., 2011; Aarons et al., 2011). **Receptive context:** Success of an innovation intervention will be affected by the degree to which the norms and shared expectations of an organisation create conditions of openness to change and lower the burdens of compliance and opposing demands. A receptive context is enabled in organisational environments that encourage ingenuity, demonstrate tolerance to novel or unconventional ideas, and accept conceptual risk-taking (Ash, 1997; Aarons et al., 2011). **Organisation-level leadership:** Success of an innovation in-tervention will be affected by the degree to which members in leadership positions steward a cultural environment that is amenable to innovation adoption and take ownership over end-to-end best practices and responsible innovation (Edmondson, 2004; Aarons et al., 2011). **Fidelity monitoring and quality assurance:** Success of an innovation intervention will be affected by the degree to which an organisation has a well-defined and effective support structure for quality assurance and fidelity oversight (Hoagwood et al., 2007; Ebert et al., 2012; Murray et al., 2014; Nadeem & Ringle, 2016; Powell et al., 2019) **Innovation appropriate resource availability:** Success of an innovation intervention will be affected by the degree to which an organisation's resource availability is sufficient for the development, implementation, and sustainability demands of the specific innovation they are producing and deploying. Resource shortage may cause a deterioration of service quality and a reduction in availability, which then leads to decreased service initiation and completion (Weiner, 2009; Murray et al., 2013, Wenocur et al., 2016; Powell et al., 2019).
Participant/agent context ▪ Psychological and motivational antecedents of adoption and reception ▪ Attitudes, perceptions, and beliefs that enable or obstruct adoption and reception ▪ Cognitive abilities, skills and investments that enable or obstruct adoption and reception	**Positive attitudes about the innovation that is linked with perception of the need for change:** Success of an innovation intervention will be affected by the degree to which participants have pro-innovation attitudes and a strong belief in the role that an innovation intervention will play in bringing about a needed change. Buy-in about the transformative utility of an innovation from implementers leads to more consistent decisions to adopt the innovation and undergo training in prepara-tion for its use (Nadeem, Jaycox, Kataoka, Langley, & Stein, 2011; Murray et al. 2014; Sigel et al., 2013; Powell et al., 2019). **Coherence of the intervention:** Success of an innovation intervention will be affected by the degree to which the implementation of the innovation makes sense to its users, who are then able to invest it with meaning. An easy-to-understand and easy-to-describe innovation that has a clear purpose for relevant participants has a better chance of effective integration into community settings (May, 2006; May et al. 2009; May & Finch, 2009; Murray et al., 2010; Finch et al., 2013). **Cognitive participation:** Success of an innovation intervention will be affected by the degree to which relevant partici-pants are able to justify and to see the legitimacy of an innovation intervention. When participants grasp that an innovation is a good idea, they are more likely to invest it with commitment (May, 2006; May et al. 2009; May & Finch, 2009; Murray et al., 2010; Finch et al., 2013).

Reflexive monitoring and engagement: Success of an innovation intervention will be affected by the degree to which relevant participants are able to reflectively engage, evaluate, and sense check the intervention over time. When users and affected stakeholders are able to proactively assess the effects and impacts of an innovation, they are better able accept the legitimacy of its continued use (May, 2006; May et al. 2009; May & Finch, 2009; Murray et al., 2010; Finch et al., 2013).

Keeping this plotting of the markers of SOP readiness in mind, let us return to the question of whether existing systemic, organisational, and cognitive-psychological conditions support the justifiability of using ML systems in CSC. While a thorough answer to this question lies in empirical research that is yet to be done, it is clear from our literature review and from the results of both the stakeholder roundtable and the family engagement workshop that, as a whole, the current state of the SOP readiness of relevant organisations, communities, and participants presents significant challenges to the feasibility of developing and deploying individual-impacting ML decision support systems in a responsible and ethically justifiable way. A premature pursuit of this kind of innovation in the care and community settings of CSC may lead to failures of effective integration and implementation that have negative impacts on the wellbeing of affected individuals and communities.

Nevertheless, what our preliminary mapping of determinants shows is that many of the enabling conditions of successful innovation intervention (such as developing the innovation-values fit for a particular ML project, building an organisational culture characterised by absorptive capacity, change readiness, and innovation receptivity, and developing local, system-level relationships where inter-organisational networks, inclusive multilateral participation, and supportive partnerships prop up and sustain innovation) is context-based, use-case specific, and actionable when carried out deliberately and with forethought. From the point of view of applied ethics, this latter element of actionability is most significant, for it suggests that, guided by ethical values and principles of responsible practice, the deliberate and ecology-aware development and implementation of ML innovation may be possible, provided that systemic, organisational, and attitudinal impediments to SOP readiness are cleared away by purposeful, intentional, and reflective innovation intervention practices.

Social inequality and cycles of poverty and discrimination

A third empirical factor, which has direct bearing on the justifiability of the application of ML technologies in CSC, has to do with the correlation between the involvement of families and at-risk individuals with child welfare services and socio-historical patterns of poverty, deprivation, and inequality. It is widely acknowledged that severely disadvantaged people have a disproportionately high level of participation in CSC services (Bywaters, Brady, Sparks, & Bos, 2014; Pelton, 1989), causing some to assert that children's social care might best be viewed 'within an oppression framework' (Lonne et al., 2016; Curtis & Denby, 2011). It has similarly been observed that due to patterns of overpolicing and higher levels of visibility, disadvantaged groups can be subjected to regimes of digital tracking and automation that reinforce deep-seated patterns of inequality, and marginalisation (Eubanks, 2018).

The robust association between deprivation, CSC rates, and disadvantaged ethnic categories suggests, moreover, that expanding relations of social disadvantage and repeating cycles of poverty are directly linked to racial demographics (Bywaters et al., 2017). Furthermore, the topic of the relationship between inequality and the social determinants of CSC participation (Wilkinson & Pickett, 2009; Featherstone, Morris, & White, 2014) has largely been sidestepped in public policy debates about child welfare, where, more often than not, interrogation of underlying structural factors is avoided (Bywaters, 2015). This has led to a lack of adequate public dialogue about how to address this troubling relationship between poverty, race, and CSC involvement.

In this connection, the major problem that arises regarding the use of predictive analytics in children's social care is that the features that are indicative of social disadvantage and deprivation, and that are simultaneously linked to socioeconomic and racial discrimination, are also highly predictive of the adverse outcomes used to measure child maltreatment and neglect.

For predictive risk assessment systems to be effective—taking into account the contemporary structures of inequity and discrimination that give shape to the underlying data distributions they model—such systems must operationally reproduce and reinforce precisely those structures in generating their outputs. From the perspective of an applied ethics of ML in CSC, the inescapability of this algorithmic reproduction of patterns of social injustice should caution against an unreflective reliance on such technologies. While the deeper configurations of poverty, inequality, and deprivation

that are instantiated as demographic features of decision subjects are predictive-because-correlated, such configurations *do not of themselves* cause child maltreatment or neglect. As a consequence, when statistical inferences that derive from a predictive ML model's fit to such configurations are employed to gain a causal understanding of the risks that are to be imputed to a specific decision subject, they should be scrutinised accordingly and handled with extreme interpretive caution and care.

Though ensuring the safety and wellbeing of an endangered child is an inviolable priority with no simple cure-all solution, technological or otherwise, the role that predictive ML systems should play in this difficult safeguarding task, given the potential for its use to fortify historical patterns of injustice, is less than clear. This legitimate hesitation with regard to the moral justifiability of using predictive risk modelling in CSC should perhaps redirect the energies of responsible data science in this domain towards more socially transformative purposes that are in line with ethical values of respect for individual dignity, interpersonal connection, care, and protecting social justice. As a rule of thumb, it should be stressed that:

When there is a strong correlation between unjust, inequitable, or discriminatory social factors and input features that map onto a predicted outcome of interest, the responsible use of applied data scientific methods may be to readjust its focus:

1. ***To shed empirical light on the causal influences and historical reasons behind those factors so as to inform policy change and the transformation of society in the present instead of contributing to the reinforcement and potential amplification of these inequitable factors in practice, and***

2. ***To redefine outcomes/target variables in a way that aims to improve family functioning as well as the developmental, physical, cognitive, emotional, and social conditions of the lives of children in need and thereby to directly and constructively address the problems of social injustice at their sources and on the ground.***

As we will see in the next part of this report, this latter call to constructively and holistically redefine outcomes for children and families in need is a challenge that largely lies ahead for applied data science in CSC. In the final section of this review, 'Where do we go from here? Recommendations for the future of ML in CSC,' we will explore in detail some other suggestions for optimising the capacity of future data scientific research and innovation in CSC to produce tangible societal benefits and advance public wellbeing. For now, let it suffice to say that, as a general-purpose technology and as a medium of insight and discovery, machine learning holds the promise of affording humans countless opportunities for moulding the contours of a better and more equitable future for each and all.

Deciding what shape that society of tomorrow will take will involve determining, democratically and inclusively, how to steer the values and motivations that are currently driving the gathering energies of technological advancement in machine learning. This section has presented a brief sketch of how this crucial human commission can be undertaken with ethical purpose and in the spirit of care, collaboration, and understanding.

IV. CAN WE DO THIS RIGHT?

EXISTING INNOVATION AND IMPLEMENTATION PRACTICES, PITFALLS, AND PROSPECTS

In the previous part of this report, we focused on the external question of whether ML technologies, especially predictive risk modelling tools, should be used in the CSC sector in the first place. Integrating the ethics of ML and the ethics of social work, we presented an ethical framework that serves both (1) as a critical measure and normative guidepost to help policy-makers and practitioners to assess the justifiability of using ML systems in CSC and (2) as a three-level platform for responsible innovation, which provides guidelines for thinking about the ethical values, practical principles, and professional virtues.

In this part of the report, we endeavour to build on the second of these. In particular, we concentrate on moving from principles to practice. By looking closely at data use, model design, and system implementation, we investigate how the practical principles derived in Section III might help both to provide guardrails for responsible conduct and to give shape to best practices from a point of view internal to the on-the-ground activities of ML innovation and use.

It should be noted from the outset, however, that the question 'Can we do this right?' is purposefully agnostic regarding the actual feasibility of the responsible design and delivery of ML systems in CSC. The necessities of good data quality, careful pre-processing, deliberate model design, diligent testing and validation, and sufficient training for trustworthy implementation are exceptionally demanding, especially with regard to individual-impacting risk assessment. Each of these dimensions of the ML production and delivery lifecycle must be realised for the use of the system as a whole to be ethically justifiable.

Our goal here is to move through important parts of the ML lifecycle, keeping in mind at each step along the way, how practical principles such as fairness, sustainability, empowerment, transparency, and accountability can be realised in the responsible design and use of ML in CSC.

Data quality and use

Policy-makers have used algorithms – in the form of what has traditionally been called software or code, mathematical formula for instructing a computing machine what to do with pieces of data – since the 1950s, when computers first entered government for administrative processing of tax and benefits in the UK and the US. These algorithms used to rest on the top-down application of logical statements and rules alone, such as 'if this person earns above this threshold, then deduct this percentage as tax.' In contrast, ML models use the availability of large-scale digital data to 'learn' in a bottom-up, inductive way, distinct from the deductive rules-based algorithms of these earlier systems. In other words, ML models extract information from training datasets to 'learn' how to carry out a certain task instead of being explicitly programmed to follow a set of pre-defined rules. ML models' reliance on data to 'learn' makes it extremely important that the data used to train an algorithm is capable of presenting a complete view of the problem that the ML model is designed to resolve. To this end, the data used to train an ML model should be representative, relevant, recent, and accurate (Leslie, 2019, p.15-16).

The representativeness, relevance, recency, and accuracy of data should be assessed at the start of the development of an ML project to inform the planning of the project itself and decisions such as:

- what data to acquire or collect initially
- what data to select as a training dataset for an ML model
- what steps to take in pre-processing of the data to mitigate any potential problems relating to the quality of the data
- how to account for data shortcomings while tuning a model's parameters
- what information to provide to intended users of the algorithmic tools about the possible limitations of the model, given issues of data quality

Data assessments should be made in a transparent manner, recorded and justified, with sufficient forethought about their consequences. Technical and domain experts should work together to assess the data and data sources by considering them in the context of their origins. It will be up to the project team to implement the procedures needed to ensure that these requirements are fulfilled. The table below seeks to summarise some of the questions that should be asked with regard to data.

Table 4. Overview of data quality considerations necessary to ensure responsible and ethical machine learning innovation derived from (Leslie, 2019)

Data requirements	
Representativeness	▪ Is the training data representative of the population under consideration? ▪ Have you planned how to balance the dataset you will use, so that it appropriately and equitably reflects sub-populations? ▪ Have you thoroughly considered risks of under- or over-representation in the dataset?
Relevance	▪ Are the chosen data sources relevant to and capable of providing a reasonably comprehensive and balanced view of the phenomenon to be modelled? ▪ Where data to provide a reasonably comprehensive and balanced view of a phenomenon is lacking, have you considered how to amend the purpose of your model to appropriately utilise the data available?
Recency	▪ Is the data you plan to use an up-to-date reflection of the phenomenon and populations you are trying to model? ▪ Have any large-scale reforms, policy changes, or changes in methods of data recording taken place that affect whether the data you want to use accurately portrays phenomena, populations, or related factors in an accurate and up-to-date manner? ▪ Is the timeliness of the data you plan to use sensitive to small or minor shifts that may take place within neighbourhoods, cultures, or operational policies? If so, have you properly established that your use of such data meets the challenges of these shift sensitivities?
Measurement accuracy	▪ Are elements of subjective bias or human error potentially involved in any aspects of data collection across your dataset? If so, have you diligently established that such risks have been addressed and mitigated, so that your dataset is sufficiently sound and reliable? ▪ Have appropriate methods for recording data been used? ▪ What information has been lost in the data recording, how valuable is it, and what are the implications of not having it?

Data representativeness

The dataset used to train an ML model should be representative (Barocas & Selbst, 2016; Lehr & Ohm, 2017; Glaberson, 2019; Leslie, 2019). It should suitably mirror the make-up of its underlying population as it relates to the predictive target of the model. We are taking the term 'population', here, in its statistical sense, as the total set of observations that can be made for a group,as a whole.

Representativeness is essential, because a dataset, which does not accurately reflect an underlying population in its relation to a distribution of concern, will not generalise when it processes new, out-of-sample inputs. Over- or under-representation of groups of the population, for example groups with protected characteristics, can lead to skewed ML outputs, potentially to the disadvantage of vulnerable groups. If an ML model is designed to assess the risk of child abuse or neglect in particular populations, the tool may overestimate the risk for members of an over-represented group and may underestimate the risk for members of an under-represented group. Moreover, systemic errors in measurement, that disparately affect a vulnerable sub-population, may lead to an inaccurate representation of that group in proportion to others, which are more reliably measured.

The necessity of representativeness applies to all sources of data that compose a dataset, as each different data source can contribute its own different form of sampling bias (Glaberson, 2019, p. 346). Whereas in social scientific research, where sampling is randomised to ensure representative datasets, in domains such as CSC, each source of data presents its own risks of under- or over-representation. CSC datasets sometimes draw from public assistance records, criminal justice records, behavioural health records, and social services records—all of which may be plagued by socioeconomically influenced selection biases. Families in poorer communities are often disproportionately represented in the records of CSC upon which some depend on a continuous and long-term basis, and the members of these same families are often also disproportionately represented in criminal justice records due, in part, to the over-policing of their neighbourhoods. The accrual in a dataset of these patterns of over-representation often, in turn, leads to higher rates of identification of child neglect or other maltreatment among these groups (Dingwall, Eekelaar & Murray, 2014; Stone, 1998, p. 92; Hay & Jones, 1994). Concern about the selection biases which arise as a result of the families that interact with – and hence appear in the data held by – public authorities was echoed in the stakeholder roundtable and workshop discussions as well, with participants highlighting the lack of sufficient data on 'the rich' in the context of CSC.

To train a predictive ML model, data scientists have to take measurable outcomes as their target variables – what the model will seek to predict or classify. In the case of CSC this often means training a model on cases of risk of potential harm where an initial decision to investigate a case, such as instances where referrals end up with 'screen-ins', is used as such a measurable outcome (Chouldechova, Putnam-Hornstein, Benavides-Prado, Fialko, & Vaithianathan, 2018). In these instances, however, the portion of the population that is screened out will not be included in the training data, because their outcomes are unknown. This causes what has been called the 'selective label' bias (Lakkaraju, Kleinberg, Leskovec, Ludwig, & Mullainathan, 2017) where the dataset loses representativeness and generalisability due to the fact that observed outcomes are not available for a significant fraction of the underlying population.

Tackling the problem of representativeness requires both that solid domain knowledge of the underlying population is obtained and that mitigation measures are taken to rectify and balance skewed samples where problems of over- or under-representation arise. These strategies, however, face serious challenges. Accurate and reliable measures of the demographic makeup of an underlying population as it relates to the target variable of interest is necessary to fix issues related to representativeness, but it is often lacking. In many instances, there is a scarcity of consistent information available about the particular demographic characteristics of families that interact with CSC. Moreover, information about other features which could reveal the socioeconomic inequalities that are often at the root of over- or under-representation, such as 'parental income and wealth, housing conditions, educational background, health, age, marital, or employment status' is frequently also lacking (Bywaters et al., 2017, p. 56).

On top of this, significant challenges to mitigating issues of representativeness may arise from the overlap of sampling biases and historically embedded inequities—a dynamic that may lead to a trade-off between balanced datasets and model performance. The data collected and used for training predictive risk models in CSC arise from the real-world provision of public services across communities where sociohistorical patterns of over- or under-representativeness may be entrenched and, hence, where a dataset cannot be balanced or rectified simply by collecting more data or by targeted resampling of under-represented groups. While other mitigation techniques may be used to balance a dataset (Chawla, 2010; Brownlee, 2015), where extant formations of societal discrimination are replicated in the selected sample, there is no clear-cut technical solution available to create optimally representative datasets.

Data relevance

When considering which data sources to use for the development of an ML model, researchers should give careful consideration both to their relevance and to the reasonableness of including them. The answers to questions such as 'what are the most appropriate and reasonable data sources' or 'what types of data should be used' differ on a case-by-case basis and depend on what each ML model is designed to measure and in what context. Solid domain knowledge is crucial for answering these questions. At a very basic level, the selected data sources should incorporate factors that are significantly related to a model's target variable (See Problem formulation below). When deployed appropriately, ML systems in CSC should be utilised as supports for evidence-based reasoning. The data used to fit them to the distributions of interest should therefore be rationally related to those distributions. The systems should provide the user/implementer with objective and empirically-anchored insights that are understandable and that can be easily incorporated into their wider deliberations about the

real-world contexts to which the system's statistical generalisations are being applied.

With this in mind, concentrated efforts to ensure data relevance in the use of ML in CSC should be bias-aware. **Bias awareness**, in this regard, involves safeguarding the **comprehensiveness, representativeness, and inclusivity of the dataset** used for training a model in order to counteract the distortive effects of patterns of socioeconomic deprivation reflected in the data. Data collected and procured should provide a comprehensive picture of the population under consideration. Frequently choices made about which kinds of data to include in training ML models in CSC carry explicit and implicit biases that lead to a focus on impoverished and disadvantaged subpopulations and that consequently overlook less deprived or wealthier demographics. This approach excludes potentially relevant data. When trained, models built on the incomplete and relevance-deficient datasets that may stem from this bias will augment the discriminatory cycling-in of children and families in poverty or experiencing other sociodemographic disadvantage. Similarly, models will potentially contain dangerous blind-spots where more affluent but equally imperilled children may slip through the cracks.

Another dimension of bias-awareness that is crucial for ensuring data relevance involves the task of **gaining a more comprehensive working picture of the families and children affected by the use of ML systems**. Existing predictive analytic tools in CSC tend to focus on negative factors rather than reflecting positive, prospective factors (Glaberson, 2019, p. 343-344, Vaithianathan, Putnam-Hornstein, Jiang, Nand, & Maloney, 2017, p. 37-43), so that intervention concentrates exclusively on the aspect of protecting children from mistreatment. However, uses of ML systems in CSC that fail to account for paths of improved family life and that do not consider the positive aspects of a child's developmental experience present an incomplete picture, which neglects the support and empowerment dimension so fundamental to ethical social care practices.

During our stakeholder roundtable, it was observed that a focus on negatives and risks runs through the CSC system. However, an LA noted that its social workers, who used a tool that highlighted risk factors, requested that the tool also present the strengths of families in order to represent a more holistic image of the family. Such a request demonstrates the legitimate need to balance risk-centred data with data indicative of family strengths, wider protective factors and possible positive outcomes oriented to improving family functioning and advancing the wellbeing, happiness, and development of children and their families.

Much more research is needed in this area of holistic and prospect-oriented data collection and modelling. This may involve investigating how to craft a portfolio of data analytic tools that draws on a fuller vision of families and children in need. Researchers might, in this respect, consider which data sources may be relevant for obtaining a more holistic and coherent understanding of the families and children about whom statistical inferences are being generated. Such research should include an examination of ways to protect vulnerable and historically over-surveilled groups from more potentially harmful data collection. Even in cases where such data gathering is intended to extract information about positive and developmentally supportive factors, it is conceivable that such data may be misused and raise data protection issues. Consideration should be given to ways that the use of such data can be rigorously defined/limited and that affected data subjects can be actively involved in the process as a whole as well as optimally protected.

Data recency

Data used to train ML models has to be recent enough to be able to capture and represent the phenomenon and/or population being modelled. Datasets represent a snapshot of a particular time period. Both large changes and smaller incremental changes taking place over a longer period of time could make data less representative of reality. In the context of CSC, the degree to which data is recent enough to be valid can be influenced by reforms and amendments to underlying laws or procedures (i.e. changes to legal thresholds or definitions), development of neighbourhoods, population shifts and manners of practicing social work. Once deployed, an ML model should be regularly reviewed and updated with recent data, which reflects the changing situation 'on the ground' (Klingele, 2016, p. 576). Where predictions rely on outdated data or on inferences learned from a context that no longer exists, models will produce inaccurate results.

Likewise, where reforms have taken place, the performance of models, whose fit to the distribution is based on prior/outdated social and legal structures, could potentially even undermine such reforms (Koepke & Robinson, 2018, p. 1730). Changes and improvements in CSC services can invalidate the predictive qualities of data from the past in many ways. For example, in the US, a parent's placement in foster care as a child (Glaberson, 2019, p. 344 discussing Williams & Monroe, 2017) and a child's prior contact

with child services (DePanfilis & Zuravin, 1999; Fluke, Shusterman, Holinshead & Yuan, 2008; Sledjeski, Dierker, Brigham & Breslin, 2008) have both been found to be predictive of child abuse or neglect. The predictive power of such features rests on the assumption (backed by the numbers) that foster care and child services interventions have been ineffective in the past. Successful reforms that improve the effectiveness of these services would diminish the predictive power of these variables, so a model that retains the inferences from the prior data will end up identifying risks inaccurately and inequitably to the affected decision recipient(s).

Similarly, the thresholds for access to CSC, as applied by different LAs, are often adapted to the changing demand for services and to resource availability. They are also influenced by the central government's policy considerations and strategies, which are often external to the LAs providing CSC services. Furthermore, they are influenced by the outcome of Ofsted inspections. All of these factors are subject to change. An ML model, especially one used for risk assessments, should be updated whenever policies, procedures, and practices are changed if the model is to remain useful and accurate.

Researchers should work with domain experts and stakeholders to identify and understand these changes within CSC as well in other related areas from which data may be sourced (e.g. criminal justice system, mental health programmes, public benefits policies). They should draw upon existing learning and analysis of such changes. This understanding would allow them to identify the limitations of the available data. It is important that data recency is ensured for all data sources.

Information about the recency of data used to train an ML model should also be recorded and provided to users of the model to inform them about the potential limitations of the tool. Information about how regularly a model is retrained should similarly be recorded and shared. In cases where local authorities are engaging with third-party firms, who are supplying and maintaining the ML model in use, rigorous standards of transparency and reporting (regarding these aspects of data recency and model training) should be agreed upon and codified in advance. Local authorities should also aim to develop in-house data evaluation expertise so that, if they do rely on third party suppliers, they can play an active role in oversight and monitoring of their own ML systems.

Measurement accuracy

Data used to train ML models should be accurately measured and recorded. Errors can affect the quality of data at the moment of their collection through human mistakes, poor measurement instruments, or faulty recording methods. Inaccurate or incomplete data will not fully represent the complex factors that they are meant to capture. When inaccurate data are used to fit models, it will affect the inferences and correlations an ML model 'learns' and undermine the model's performance. The data used in an ML model should originate from 'suitable, reliable, and impartial sources of measurement and sound methods of collection' (Leslie, 2019, p. 15).

The accuracy of the data recorded in CSC can be affected by human biases. Like all of us, social workers are subject to biases, such as confirmation and framing biases that can colour their decisions. For example, when studying reports of child abuse in Britain between 1973 and 1994, Munro found that the evidence used in risk assessments was 'often faulty' due to inaccuracies in reporting caused by biases, dishonesty or errors (Munro, 1999, p. 745). The issue of data and recoding accuracy was a key feature of the workshop with family members with lived experience, who highlighted the issues around contested information being included in their records.

Measurement inaccuracies can also arise. First, humans entering information into administrative system can make mistakes. Second, linking information across datasets can erroneously connect separate individuals or fail to connect all data about a single person. These gaps and disconnects can be exacerbated by organically changing information, such as surnames that are altered for reasons of marriage or divorce. Finally, self-reported information can also be misleading, and its accuracy can depend on the context in which it was collected.

Measurement accuracy could also be affected by the way in which the data were recorded. Social workers are often required to input data in a particular IT system, selecting appropriate categories and sub-categories from a drop-down menu, for example. This way of recording data can oversimplify situations, lumping together cases and people who perhaps do not belong together, and removing the underlying subtleties of the situations being recorded. When subjective judgments are hidden behind such categorisations and groupings, they cannot be easily retrieved or revised (Capatosto, 2017, p. 5).

Stakeholders during the roundtable and workshop also raised the point that datasets prepared by local

authorities can be biased by the incentives behind their collection or by the desire to make those commissioning the data collection look better. LAs and other organisations providing CSC services are often inclined to demonstrate that their work is effective, and they have motivations to do so under pressure from management or from Ofsted inspections. Furthermore, programmes like the Troubled Families Programme relied on a 'payment-by-results' model for funding, which further incentivised positive outcomes in the data. During our family roundtable, moreover, family members shared that they felt pressure to provide positive service feedback, when asked, in order to ensure they receive other services in the future.

Finally, some factors and aspects related to CSC do not lend themselves to measurement. This is not only the case for potentially positive aspects of a family's life, such as the dedication and love of a parent or of extended family members, but also with regard to negative child experiences. Child abuse or neglect are difficult to measure in general due to under-reporting or under-recording. Not all children who are in need recognise themselves as such or are identified as such by their caretakers, social workers, or other members of their communities. CSC or police statistics do not reflect occurrences of child abuse that are not reported to them.

Even when information about potential child abuse or neglect reaches social work agencies, it may not be recorded if the concerns reported do not reach the necessary threshold or if there is insufficient evidence for the child abuse or neglect (NSPCC, 2018, p. 1). Police can also decide not to record a report. A UK audit in 2014 found that 19% of reported instances to the police that should have been recorded were not. For example, sexual offences and violence with or without injury were not recorded 26% and 33% of the time respectively (Office for National Statistics, 2019).

When assessing data accuracy, researchers and domain experts should consider the data and its origins holistically, including the measurability of the underlying phenomena, the individuals recording the data, and the methods of measurement and recording used. The context in which data was recorded could also affect the reliability of data, especially for self-reported data. Any available metadata should also be considered in such assessments.

When data has already been categorised or labelled, researchers should ensure that they understand what the categories and labels represent or conceal. The role of individual judgment in recording or framing data is, however, an inescapable part of CSC services. This is further discussed in the section on pre-processing below.

Model design

On the basis of the available data and its qualities, project teams come together to make the first design choices about the future ML model. Each element involved in an ML model's design can be influenced by the decisions made by the model's designers and developers. As such, it can incorporate their biases or assumptions at each phase of the production process. This makes it extremely important to consider the impacts of each design decision and to ensure that it is made in an informed and unbiased way.

The table below is intended to provide a snapshot of some of the significant decisions that have to be made during the process of model development.

Table 5. Overview of design decisions throughout machine learning development with ethical implications

Overview of design decisions with ethical implications	
Problem formulation and outcome definition	▪ What outcome or instance will the model assess, classify, or pre-dict? ▪ How will this outcome be represented by a specific target variable or its measurable proxy? ▪ Is the target variable a reasonable and justifiable representation of the objective of the model? ▪ Might the choice of the target variable have an inequitable impact on different groups of the population?
Pre-processing	▪ How is the training data labelled, annotated, and organised? ▪ How should inaccurate or missing variables be identified and treated within the training and testing data? ▪ What impacts might these decisions have on the performance and explainability of the model? ▪ What are the relevant attributes and features that will serve as input variables to the model? ▪ How are the feature engineering tasks of binning, aggregating, extracting, or decomposing attributes being carried out? How are they checked, and controlled for biases?
Model building	▪ Against which formal criteria and benchmarks should the performance and fairness of a model be assessed and optimised? ▪ Are the inferences, significant correlations, and proxies within the model reasonable and justifiable? Are there any that are potentially unjust, unreasonable, or inequitable? ▪ Which standards of transparency, interpretability, and explainability should the model conform to? Which kinds of models are appropriate to support the chosen degree of interpretability and explainability?

At every step of the way, the potential impacts that these design decisions may have on vulnerable individuals and groups should be considered. Both technical and domain expertise are necessary for making these choices in an optimally informed manner. Additionally, dialogue with affected stakeholders should occur at critical points, such as the initial stages of problem formulation and outcome definition. These conversations may contribute valuable insights for understanding what direction an ML project team should move in across the design lifecycle. Collaborative and inclusive deliberations within the project team and beyond about the ethical impacts of such choices are crucial for establishing their justifiability.

Problem formulation and outcome definition

A supervised ML model needs to start with the clear formulation of a problem and with a clear definition of an outcome that reflects the intended solution to that problem. It must 'estimate something, and the first step of any analysis is to define what that something should be and how it should be measured' (Lehr & Ohm, 2017, p. 672-673). That 'something' is the target variable and training an ML model will involve feeding an algorithmic structure data in order to shape it into a reliable mechanism for mapping a range of input variables to the target output. The ML model will then use the inferences and correlations 'learned' to predict values of the target variable on the basis of new inputs. To ensure that the model is useful and effective, it is important to choose the goal of that model carefully, because the target variable will be the measurable result that the ML model is optimising for, and what it will predict and calculate.

As a preliminary point, we should highlight that child maltreatment is a broad term, encompassing issues like neglect, abuse, emotional or physical violence, sexual abuse, and many other factors (Herrenkohl, 2005). Some have argued that due to the distinct nature of each of these maltreatments, they deserve separate examination (English, Bangdiwala & Runyan, 2005, p. 442). Developing an ML model that seeks to predict maltreatment in this differentiating way would either require developing different ML models for different 'types' of child maltreatment or more sophisticated modelling approaches, such as the developing field of multi-target ML (for an overview see Waegeman, Dembczyński & Hüllermeier, 2018).

Choosing an appropriate proxy variable to represent the model's target is another difficult task. Due to the complexity of the CSC field and CSC data, it is recommended that ML researchers work closely with domain experts to understand the possibilities and limitations of the available data (Bromfield & Higgins, 2004) and to choose, where possible, appropriate proxies to represent the target of their ML model. The proxy variables should be quantifiable, measurable, and chosen:

- So that they appropriately reflect a legitimate goal, e.g. risk or optimal outcome identification and justifiable thresholds or standards for it
- With consideration of the human and organisational biases that affect them
- With an understanding of what these variables represent in reality
- So that they are not directly affected by the outputs of the model in order to avoid feedback loops

Potential pitfalls of identifying risk and choosing appropriate proxy variables

Using risk identification as an ML model's objective can be challenging for at least two reasons. Most importantly, there is no clear agreement or definition of what constitutes risk or of what thresholds for assessment or intervention should be (Welbourne, 2002, p. 345, 346, 352; Rose & Meezan, 1996), nor when a child should be considered to be in need. In practice, such decisions are context-dependent, require professional judgements and are based on interpersonal relationships of frontline social care workers with children and families. The way these concepts are applied may differ by individuals, location (considering the differing thresholds for access to CSC services between LAs in England), or by culture (Straus & Kantor, 2005). Due to the contested nature of defining child abuse or neglect, some argue that objective certainty about what constitutes it may not be possible and that ongoing democratic practices of evaluation and re-evaluation may always be a necessary component (Taylor & White, 2001, p. 54).

Secondly, even if a definition of the risk of child abuse or neglect could be agreed to, it is unlikely that concrete data points exist that pinpoint precisely when it happens (Glaberson, 2019, p. 342). While this makes the critical job of formalising outcome definition in the statistical frame almost impossible, ML project teams have no choice but to try to select appropriate proxy variables to represent outcomes of interest for the families and children in involved in CSC. This fraught task of deciding on apt proxy variables brings its own set of challenges.

Human and organisational biases are among the more difficult features to control for when proxy variables are being identified and incorporated into model design. Some often-used proxy variables for models that seek to assess risks to children are the outcomes from key points of decision-making within CSC, such as referrals of children to the system (prior to any examination about whether such referrals are substantiated), substantiation of referral reports, and rates of re-referrals or placements outside the home. When not based on substantiated or validated outcomes, such variables, however, may 'inherit the formalizations involved in pre-existing assessment mechanisms' (Barocas & Selbst, 2016, p. 680), which is to say that they may incorporate the subjectivities and biases of such prior decision-making mechanisms. When choosing an appropriate proxy, researchers should be mindful of what the variables they are using represent.

To be sure, human biases can affect many points of decision-making within CSC services, and this in turn will affect the validity of potential proxy variables. For example, referrals and re-referrals can be impacted by biases of surrounding communities and the relationships of families with neighbours. These referral biases are more likely to affect parents with lower socio-economic status and certain geographic characteristics indicative of disadvantage (The Allegheny County Department of Human Services, 2017). The rates and makeup of substantiated reports can also be affected by the rates of referrals for different groups of the population (Glaberson, 2019, p. 342), as well as by the thresholds applied by LAs and subjective biases of social workers (Garrison, 2012, p. 25-26).

Organisational biases can also affect variables. In an English context, section 47 (Child Protection) investigations that may follow a referral, may be affected by the risk-averse dimensions of the culture of social work. The escalation of a 'culture of blame, shame and fear' among both social workers and families was noted in the 2018 Care Crisis Review and has led to children or families referred to LAs being increasingly likely to undergo a section 47 investigation (p. 4, 17). In 2015-2016, 26% of referred cases were investigated, while in 2009-2010 the number was 15%. Yet, the proportion of substantiated investigations fell by 20% in the meantime (Bywaters et al., 2017, p. 54).

One should additionally consider what proxy values represent as a whole. Questions may be raised as to the legitimacy of reducing complex and multidimensional outcomes to partial and ultimately incomplete measurements. For example, even using 'substantiation' to represent child abuse or neglect may not be suitably reflective of the target concept in general (Gillingham, 2016, p. 1049-1052; Cross & Casanueva, 2009) and some researchers have even suggested that the data should be disregarded for research purposes (Kohl, Jonson-Reid & Drake, 2009). Focusing only on the label of 'substantiation' fails to capture what exactly was substantiated, who was affected by it, and what the reasons were behind substantiating a report. Researchers should be clear about what the proxy values they choose actually represent, and they should remain realistic about the limitations of proxies to capture the complexities and nuances of the social phenomena that their ML systems are modelling.

A final aspect to consider when choosing a proxy variable is avoiding feedback loops. Feedback loops can occur when the target variable of a model, or another significantly influential feature of a model, is affected by the outputs of that model or how they are utilised in practice. For example, imagine a model is used to provide preventive services to families that are considered in need on the basis of a few variables, including their previous interactions with public services. If a family is engaged preventively with supportive CSC services, the model would identify this increase of the family's engagement with public services as a sign of its increasing need and continue recommending more engagement. When undetected and reintegrated continuously into service provision, feedback loops can ultimately impact the validity and accuracy of an ML model.

In the case of the Allegheny Tool, an assessment model developed to support call screeners evaluating referrals in Allegheny county, Pennsylvania, researchers chose to use placement rates as a proxy value. This approach was found to effectively prevent these kinds of feedback loops since the value of the target variable (i.e. whether a child is placed or not) does not directly depend on the decisions made by the intended users of the tool (call screeners). The risk of a feedback loop between the algorithm's outputs and its optimisation was considered mitigated (Chouldechova et al., 2018, p. 4).

Toward more justifiable outcome definitions

Instead of focusing exclusively on identifying children at risk, ML models in CSC may also be used to help improve outcomes for children and families by providing insights about services best suited to ensure optimal family functioning and to foster a child's behavioural, emotional, cognitive and social development as well as educational success. This wider-angled approach would be in line with recent efforts to build out non-reductive, rights-based, and well-rounded outcomes frameworks in CSC by the Department for Education's Children's Social Care Innovation Programme, What Works for Children's Social Care, and Oxford University's Rees Centre (Sebba et al., 2017; What Works Centre, 2018; La Valle et al., 2019) These efforts build on the earlier insights of the Munro Report, which criticises the common focus in CSC services on processes, performance indicators, and targets over the quality and effectiveness of support provided to children (Munro, 2011, p. 6).

Such attempts at a more holistic redefinition of outcomes in CSC are supported by the results of a recent qualitative study of an affected community of family members, social workers, and CSC services employees in the US by Brown, Chouldechova, Putnam-Hornstein, Tobin, and Vaithianathan. When discussing the way that outcomes of predictive risk models were defined, participants in this study expressed worries that 'this framing focuses attention on predicting a negative outcome ('failure') based on negative inputs that capture 'deficits' instead of 'strengths.' There is concern that such approaches risk anchoring workers to a disproportionately negative view of the situation, which may in turn drive negative actions (Brown et al., 2019, p. 9). Moving beyond such an emphasis on negative possibilities, an approach to ML modelling that also seeks to optimise for positive and more holistically defined outcomes for children and families would not only carry forward the significant progress being made in reconceptualising outcomes frameworks for effective relation-based social care, it would also be more consistent with the ethical purposes that were articulated in the previous part of this study.

Steps in the direction of more strengths-based approaches have recently been made by researchers in CSC, who stress the importance of identifying 'protective factors' in children's lives that are indicative of their resilience to the harmful long-term effects of trauma and adversity on their health and wellbeing. Studies, in this area, have stressed that a child's resilience depends on numerous factors such as a safe and nurturing familial, educational, and community environment as well as on the promotion of mechanisms to support effective parental coping and family functioning (Bethell, Newacheck, Hawes & Halfon, 2014; Schofield, Lee & Merrick, 2013; Fraser et al., 2014; Academy Health, 2014). Banyard, Hamby and Grych (2017) have incorporated research in the psychology of resilience, thriving, and character strength into a platform for 'protective interventions.' They built a 'Resilience Portfolio model' based on strengths-building resources where 'protective factors or strengths are drawn from across the ecological framework of individuals, families, communities, and society, [and] resources are defined as protective factors that are outside of the person that support positive functioning including social supports and positive community factors like collective efficacy' (Banyard, Hamby & Grych, 2017, p. 90). Building off of such a 'poly-strengths' perspective, Walsh, Joyce, Maloney, and Vaithianathan (2020) examine a longitudinal cohort of integrated administrative data 'to identify potential protective factors as a first step in designing programs for families identifies that would enable frontline workers to take a strengths-based approach' to using ML in CSC (Walsh, Joyce, Maloney & Vaithianathan, 2020, p. 1).

Rethinking problem formulation and outcome definition, in this light, would involve asking a wider set of initial questions about how algorithmic decision-assistance systems may better inform relation-driven, strengths-based, and user-centred social care practices that aim to make a positive difference in the lives of children in need and their families. In their outcomes framework, 'How do we know if children's social care services make a difference?', La Valle, Hart, Holmes, and Pinto pose three questions intended to orient thinking about how to assess that positive outcomes for children are actually being achieved:

- Are children in need safe where they live, both at home and in their community?
- Have they been supported by CSC services to be healthy and happy, that is to achieve developmental, physical, cognitive, social and emotional milestones?
- Have they been supported by CSC services to make progress in education and to have positive educational experiences? (La Valle et al., 2019, p. 9)

These questions helpfully move beyond the conventional focus on immediate safety risk and harm prevention to consider outcomes that involve: the achievement of stability and permanence in the lives of affected children and their families; the progress they are making in their behavioural, emotional, and social development; the support they are getting to steward mental health; and the educational experiences that will allow them to make progress in obtaining the cognitive and practical know-how to advance and flourish (La Valle et al., 2019, p. 10). The recalibration of problem formulation and outcome definition in designing ethical ML systems in CSC should take this range of concerns into account from the start. This would involve a holistic approach to thinking about the qualitative scope of outcome definition. It would also entail a 'long arc' view of risks (Vaithianathan et al., 2017) and benefits that moves beyond the priority of predicting immediate harm so that algorithmic decision support provides a field of vision for frontline social workers that is oriented to longer-term impacts and results.

However, given the existing state of play in CSC data recording and collection practices as well as the scarcity of useable data resources of relevance, it could be difficult to link such a range of concerns to measurable target variables, which capture the integrative character of the positive outcomes of interest. Unlike in healthcare where the specific diagnosis, treatments, and outcomes can be empirically observed and recorded, in CSC there are inherent difficulties both in measuring positive outcomes and in gauging the causative factors that contributed to them (Gillingham, 2016, p. 1053 discussing Billings, Dixon, Mijanovic & Wennberg, 2006 and Parton, 1998). It is often difficult to link the CSC services provided with a particular long-term outcome for a child and family without considering the vast array of potential interceding factors that could have contributed to that particular outcome. It is also difficult to encapsulate the impact of complex CSC processes, which occur in very context-specific environments and over potentially extended periods of time.

Still, the use of such kinds of positive target variables may be made possible, in part, by the employment of longitudinal data, which can capture the experiences of children in need and their families over time and assist researchers to assess the relative impacts of different types of interventions. At present, there is,

largely, a lack of data that would afford access to these sorts of path-specific outcomes, with much of the national administrative data for children's social care focused on risk-centred, quantitative outputs rather than qualitative, positive outcomes (La Valle et al., 2019). However, as the ultimate goal of CSC services is to provide effective support to those in need across the spectrum of outcomes, researchers who wish to develop ethically justifiable ML applications in CSC should explore ways to improve the landscape of data resources, so that a more holistic approach to writing ML programs for CSC can be supported by high-quality data (Abiteboul et al., 2017). Active efforts have to be made to create a better data landscape that is more amenable to picking up patterns indicative of positive outcomes that foster the wellbeing and flourishing of children in need and their families.

Pre-processing

Supervised ML models rely on large datasets of labelled examples to identify correlations between different features and the target variable. Decisions made during the pre-processing stage involve curating and cleaning the data, preparing it for training, as well as finding ways to handle missing or inaccurate data points. Pre-processing decisions also involve crafting the space of input variables that will be the basis for modelling the distribution of concern. This includes selecting the most relevant features, trimming down the feature space by reducing the number of attributes, aggregating or binning input variables to reduce a model's dimensionality, and transforming variables to meet the model's predictive or classificatory needs. All decisions about curating, annotating, ordering, cleaning, and otherwise preparing the data for processing can have impacts on the model's performance as well as its interpretability, and all of them may involve human biases and subjective choices.

Pre-processing is also the stage when many elements of algorithmic bias can be redressed, through a variety of techniques, such as suppressing proxy variables that are correlated with sensitive attributes, changing the labels of some objects to counter discrimination, reweighing different records, or resampling (Kamiran & Calders, 2012, p. 2-3). Researchers should explore the latest technical methods to detect and mitigate biases that may be lurking in the dataset (Berk, Heidari, Jabbari, Kearns, & Roth, 2017, p. 25-27; d'Alessandro, O'Neil, & LaGatta, 2017).

Due to their far-reaching impacts, decisions made at the pre-processing stage should be taken carefully and in collaboration with domain experts. The overall composition of the dataset should be closely examined for potential discriminatory influences, data sufficiency, as well as the distribution of measurement inaccuracies and missing data across different groups of the population. The impact of pre-processing decisions on the performance of the ML model and on different groups of the population should be accounted for and any inequitable impacts across different groups of the population should be avoided. Decisions made about data curation and cleaning during this pre-processing phase—and across the ML production pipeline—should be recorded along with the reasons and justifications for making them (Gebru et al., 2018; Holland, Hosny, Newman, Joseph, & Chmielinsku, 2018; Leslie, 2019).

Annotation and labelling

Extreme care should be taken to ensure that processes of annotating and labelling the data are undertaken in a bias-mitigating and explanation-enabling manner. Choices made about how to categorise and classify features and how to add meta-data can be crucial for understanding, justifying, and explaining the results of a trained model downstream. When data is being curated for ML applications in CSC, appropriate efforts should be made to attach rich contextual information and ample meta-data, so that processing results make optimal sense to users when they query the rationale behind a given output. These results should reflect reasonable expectations about the determinants an outcome that can be tied back to the relevant factors and signals contained in the feature space from where statistical inferences have been drawn. For instance, when unstructured data are transformed into structured attributes and used in a model's feature space, they should be annotated to include the reasoning behind their inclusion, i.e. how they might serve as factors to support evidence-based reasoning about the causal influences contributing to a given algorithmic result (ICO & Turing, 2019).

Feature determination and engineering

Human decision-making enters into the feature determination and engineering stage when the attributes that will serve as input variables for the model are chosen, sifted, and organised. At this stage, researchers identify groupings of variables that hold more predictive power and 'trim away' features which are not correlated with the model's target variable, so that the processing is more efficient and the feature space is as reasonably sparse as possible. Limiting the totality of attributes to input features that are influential and correlated to the target variable without excluding any data of relevance to the accurate mapping of the underlying distribution facilitates the creation of a more transparent and interpretable ML model. The more features that are used to classify or predict an output, the more complex and high-

dimensional a model and its operations will be and the less potentially explainable.

One of the priorities that should drive decisions made about feature determination is that the selected attributes should be pertinently related to the target variable in a way that is easily understandable and rationally conveyable. When choosing the features that will serve as input variables, judgements are made about what sorts of information may or may not be relevant or rationally required to yield a reasonable, accurate, unbiased classification or prediction. ML project teams should therefore consider the ethical permissibility of using features with predictive power but no reasonable causal connection to the production of the outcome of interest.

Consider again the fact that a parent's previous interaction with the foster care system or a child's previous contact with children's services were found to have predictive value for future child abuse or neglect in the US context (Glaberson, 2019, p. 344, discussing Williams & Monroe, 2017, DePanfilis & Zuravin, 1999, Fluke, et al., 2008, Sledjeski et al., 2008). Using these variables in an ML model in CSC may contribute to its performance insofar as they are correlated to the outcome of safety risk. However, whether or not it is reasonable and fair to consider a parent's own earlier contact with foster care or a child's need-based interaction with children's services as variables that should factor into a decision about the specific circumstance of child and family is much more contentious. Extreme care should be taken when weighing the inclusion of variables like these, which may be correlated with an outcome of interest in complex ways but are clearly not attributable to decision subjects as causal factors behind the paths of action that may produce a predictive target. Unreflectively relying on such signals of adversity, poverty, and social disadvantage as useable correlates of harmful outcomes may function to reinforce a family's marginalisation, its stigmatization, and its unfair treatment while also failing to give due regard to the agency and particular life context of the parent.

The feature engineering jobs of aggregating, extracting, or decomposing attributes from datasets may also introduce human appraisals that have biasing effects. Decisions made about combining variables into wider categories or even eliminating them altogether may introduce bias into a trained system by de-emphasising important characteristics that should have remained discernible in the data. For example, if a dataset used for predictive analytics in CSC includes information about past contact with local police forces but combines types/degrees of offenses or excludes variables indicating neighbourhoods of arrest activity, it may obscure signals of discrimination (originating, for instance, in over-policing or over-surveillance) that are crucial for fairness-aware modelling. This would prevent bias-mitigating efforts to differentiate between excessive police involvement in communities and individual patterns of behaviour that may be predictive with regard to child safety.

For this reason, both discrimination awareness and cognisance of the reasonableness of included attributes and grouped features should play a large role at this stage of the AI model-building workflow as should domain knowledge and policy expertise. ML project teams should proceed aware that choices made about grouping or separating, and including or excluding features, as well as more general judgements about the comprehensiveness or coarseness of the total set of features, may have significant consequences for vulnerable or protected groups.

Model building

When ML project teams move to the model scoping, selection, and training/testing stages of system development, they are faced with a new set of design challenges that are directly related to the ethical permissibility and justifiability of the resulting models. At this construction phase of the ML lifecycle, project teams should focus on:

- Performance and accuracy - how their models can be built to be optimally accurate and to meet performance criteria suitable for the system's risks and objectives,
- Fairness and bias mitigation - how they can be deliberatively designed in a fairness-aware manner, and
- Interpretability - how they can employ appropriately interpretable algorithmic techniques so that designers can ensure that the systems they build are not discriminating and are safe, accurate, and reliable.

We will explore each of these areas in turn below.

Performance and accuracy

The performance of an ML model is often the dimension of its design that is considered to be most important to optimise. When it comes to predictive analytics, the accuracy of ML models' predictions is critically linked to their usefulness for informing and supporting the judgment of social workers (Church

& Fairchild, 2017, p. 70). However, focusing on the general accuracy of an ML system (i.e. its error rates or the proportion of correct predictions) without closely examining the types of errors being made is not a sufficient way to ensure that its outcomes are ethically justifiable. A tool that correctly identifies all cases of child abuse, but also marks safe situations as being high risk (false positives) will not only undermine user trust, it may inflict a great deal of trauma on inaccurately targeted children and families (Munro, 2019, p. 3; Glaberson, 2019, p. 340-341 discussing Coleman, 2005, p. 436-437, 441), and it will place erroneous burdens on the already resource-restricted field of CSC. From the other side of the error spectrum, a model that has high rates of false negatives (i.e. fails to detect situation where a child is at risk) will lead CSC services to miss identifying occasions where intervention may be necessary.

It is crucial to keep in mind that, since perfect accuracy is not possible in ML systems (which are inherently probabilistic), choices have to be made by project teams and designers about how error types are prioritised and distributed. While these choices frequently involve adjusting the mathematical machinery of the model in order to constrain the allocation of error types, they also have an inherent 'normative valence' (Lehr & Ohm, 2017, p. 698). Determining which types of errors are more important to limit than others entails making value-based decisions about which sorts of possible harms are more tolerable than others (Glaberson, 2019, p. 338-343). On the one hand, when the safety of affected children is emphasised, the costs of false negatives in the predictive outcomes of risk models may be stressed. On the other, where invasions of family privacy and autonomy, the trauma of unwarranted child removal, and the destabilisation of family functioning by needless intervention are matters of concern, the costs of false positives may be emphasised. Discussions about how to balance these contending values and prioritise errors types when building an ML model should be inclusive and actively involve practitioners, families, and other affected stakeholders.

As a general rule, information about how these decisions are taken and how they are operationalised in the cost-tuning of an ML model should be documented and made available to implementers in an understandable and accessible way. The human decision-maker, who is supported by a model's results should have a solid working knowledge of how it may go wrong and at what rates it is designed to do so. This will, in part, involve training users to be able to evaluate, work with, and clearly communication an ML system's performance and accuracy measurements (such as its sensitivity, specificity, and precision). Enabling users/implementers to gain a working understanding of a model's overall performance and of how precise, sensitive, or specific it is in a context-sensitive and use case-based manner will allow them to utilise its results more effectively, to understand the limitations of its performance in a more critically informed and reflective manner, and to better grasp the trade-offs that have been made between error types. In addition, integrating confidence intervals, which indicate ranges of certainty for specific instances, into this provision of performance information will better enable them to weigh the degree of uncertainty that is at play in the statistical outputs they are considering.

While ensuring the accuracy of an ML model must remain a design priority, maintaining an active awareness of the limitations and pitfalls of validating the performance of a trained system should be treated as equally important. Performance metrics for a specific ML system are conventionally obtained when a trained model is tested and validated on the 'unseen' portion of its training dataset—often referred to as a test set or validation sample. This 'holdout method' helps ML system designers to combat endogenous issues such as overfitting (i.e. when a model maps onto the underlying distribution of its training data so well that it does not generalise when applying this mapping function to new, out-of-sample data). However, it does not ensure performance as it relates to exogenous factors such as poorly chosen or misdefined target variables, or concept drift (viz., real-world changes in the underlying societal phenomenon that is being modelled). To mitigate these exogenous factors, continual performance monitoring and external validation should be put into place.

Fairness and bias

An aspect of model design that is closely related to performance and accuracy is outcome fairness, which has to do with placing formal constraints on the allocation of errors and outcomes as they are distributed among subgroups of a population. Outcome fairness is a specific way of prioritising or calibrating errors between these groups or calibrating the likelihood of specific outcomes vis-à-vis the features of individuals and groups.

To be able to assess this formal/distributive fairness of a model, researchers must identify an appropriate fairness benchmark – a formal fairness definition. Such a fairness benchmark can serve as a yardstick against which to optimise a model's allocation of errors and outcomes and can also serve as a last step through which to mitigate biases that may have

entered the model through data and design decisions. The fairness definition chosen should be made clear and explicit to affected stakeholders in advance in order to ensure the justifiability of the model and the appropriate transparency and publicity of its fairness position.

There are many ways in which fairness can be defined in the context of ML. Researchers, working together with domain experts and stakeholders, should choose a definition that is appropriate to a given ML system's impacts on the individuals and communities it affects, to the domain in which it will be deployed, and to the specific context of its use case. This choice should also consider the data on which the model is trained and tested and the feasibility of taking the technical steps necessary to incorporate fairness criteria into the tool during the pre-processing, modelling, or post-processing, given the kind or kinds of algorithmic technique(s) being used. Other issues and challenges that need to be considered in making this decision include (Leslie, 2019, p. 18):

- The incompatibility of different fairness definitions or unavoidable trade-offs between them
- The limitation of these kinds of formal fairness definitions to the distributive or allocative consequences in the use of the model
- The need for data about the protected characteristics or other sensitive features of interest in order to operationalise the formal definition of fairness reached (As we have discussed above, such data is frequently not available in a detailed, consistent, and accurately recorded manner in the field of CSC services. And, even when available, issues of privacy and data protection may arise.)

Generally, fairness definitions focus on either fairness between groups or between individuals. Here is a table of some of the more common approaches:

Table 6. An overview of some formal definitions of outcome fairness

Some Formalisable Definitions of Outcome Fairness (Adapted from: Leslie, 2019)	
Type of Fairness	**Definition**
Demographic/ Statistical Parity **Group Fairness**	An outcome is fair if each group in the selected set receives benefit in equal or similar proportions, i.e. if there is no correlation between a sensitive or protected attribute and the allocative result. This approach is intended to prevent disparate impact, which occurs when the outcome of an algorithmic process disproportionately harms members of disadvantaged or protected groups (Dwork, Hardt, Pitassi, Reinhold & Zemel, 2012; Zemel, Wu, Swersky, Pitassi & Dwork, 2013).
True Positive Rate Parity **Group Fairness**	An outcome is fair if the 'true positive' rates of an algorithmic prediction or classification are equal across groups. This approach is intended to align the goals of bias mitigation and accuracy by ensuring that the accuracy of the model is equivalent between relevant population subgroups. This method is also referred to as 'equal opportunity' fairness because it aims to secure equalised odds of an advantageous outcome for qualified individuals in a given population regardless of the protected or disadvantaged groups of which they are members (Hardt, Price & Srebro, 2016).
False Positive Rate Parity **Group Fairness**	An outcome is fair if it does not disparately mistreat people belonging to a given social group by misclassifying them at a higher rate than the members of a second social group, for this would place the members of the first group at an unfair disadvantage. This approach is motivated by the position that sensitive groups and advantaged groups should have similar error rates in outcomes of algorith-mic decisions (Zafar, Valera, Rodriguez & Gummadi, 2017; Chouldechova, 2017).

Positive Predictive Value Parity **Group Fairness**	An outcome is fair if the rates of positive predictive value (the fraction of correctly predicted positive cases out of all predicted positive cases) are equal across sensitive and advantaged groups. Outcome fairness is defined here in terms of a parity of precision, where the probability of members from different groups actually having the quality they are predicted to have is the same across groups (Kleinberg, Mullainathan & Raghavan, 2016; Chouldechova, 2017).
Individual Fairness **Individual Fairness**	An outcome is fair if it treats individuals with similar relevant qualifications similarly. This approach relies on the establishment of a similarity metric that shows the degree to which pairs of individuals are alike with regard to a specific task (Dwork et al., 2012).
Counterfactual Fairness **Individual Fairness**	An outcome is fair if an automated decision made about an individual belonging to a sensitive group would have been the same were that individual a member of a different group in a closest possible alternative (or counterfactual) world. Like the individual fairness approach, this method of defining fairness focuses on the specific circumstances of an affected decision subject, but, by using the tools of contrastive explanation, it moves beyond individual fairness insofar as it brings out the causal influences behind the algorithmic output. It also presents the possibility of offering the subject of an automated decision knowledge of what factors, if changed, could have influenced a different outcome. This could provide them with actionable recourse to change an unfavourable decision (Kusner, Loftus, Russel & Silva, 2017; Ustun, Spangher & Liu, 2019).

While the fairness definitions contained in the above table are widely in use at the moment, it is important to keep in mind that other fairness benchmarks and other approaches to bias mitigating model calibration are being developed all the time. The appropriateness of applying one or another of these fairness definitions will depend on the particular use case, on the data available to the ML project team, and on the beliefs about equity and allocative justice that are held by policy-makers and affected stakeholders. It should also be noted, in this connection, that the plurality of available fairness definitions demands that democratic processes of inclusive deliberation and conversation be undertaken to establish the legitimacy of the fairness characterisations which are ultimately operationalised in the trained model. Unavoidable indeterminacy with regard to value judgements about how to properly define fairness implies that the concept itself must be viewed in an open and dynamic way. Processes of reaching reciprocal understandings of fairness benchmarks therefore call for the continuous participation of affected stakeholders in making consensus-based determinations about how to delimit such definitions in the particular contexts that impact them.

In the CSC context, ML project teams should engage with domain and policy experts, as well as with individuals who may be impacted by ML tools to discuss and to identify appropriate approaches to fairness for their specific applications. Research into how people perceive different fairness definitions is as yet a developing area, though useful work is already being done (Grgić-Hlača, Zafar, Gummadi, & Weller, 2016; Binns et al., 2018; Green & Chen, 2019; Saxena et al., 2019). In the context of loan decision-making, Saxena et al. surveyed people online to see what they thought about three fairness definitions and found that participants preferred a calibrated concept of fairness that selects individuals in proportion to their merit, followed by one wherein similar people are treated similarly, finally, followed by one that prioritises the best candidate for a loan regardless of their protected characteristics (Saxena et al., 2019). This suggests that decisions involving socially sensitive features require forms of model calibration that redress inequitable patterns, so that the merits and circumstance of individual instances can be understood clearly, fairly, and knowledgeably. Whether or not the outcomes of this research generalise to other fields is a question that warrants further domain-situated exploration.

Be that as it may, the widespread acknowledgement of the plurality of fairness definitions does suggest that it is misguided to limit these definitions to formalisable criteria of error and outcome distribution. Thus Grgić-Hlača et al. argue that, in addition to formal notions of distributive or outcome-based

fairness definitions, considerations of 'the fairness of the process of decision making' should be integrated into deliberations about the equity of ML decision-support' (Grgić-Hlača et al., 2016, p. 1). For them, what matters in procedural fairness, in particular, is 'which input features are used in the decision process and how including or excluding the features would affect outcomes' (Grgić-Hlača, Zafar, Gummadi, & Weller, 2018, p. 1). Thinking in a criminal justice context, they emphasise considerations of:

1. **Feature volitionality:** Does the feature represent the result of volitional (i.e., voluntarily chosen) decisions made by the individual (e.g., number of prior offenses); or rather is it the result of circumstances beyond their control (e.g., age or race) (Beahrs, 1991)?

2. **Feature reliability:** How reliably can a feature be assessed (e.g., in credit assessments, opinions towards bankruptcy may be harder to reliably assess than number of prior bankruptcies) (Trankell, 1972)?

3. **Feature privacy:** Does use of the feature give rise to a violation of the individual's privacy (GDPR 2016)?

4. **Feature relevance:** Is the feature causally related or not to the decision outcomes (Kilbertus et al. 2017; Kusner et al. 2017)? (List taken directly from Grgić-Hlača et al., 2018)

This list of normative feature characteristics (volitionality, reliability, privacy, and relevance) points us toward the significance of considering how the actual architectural components of an ML model contribute to the rationale behind its results. The architectural components of the logic underlying a model's output—elements such as the relative importance of features, feature interactions, significant inferences, etc.—should be treated as essential factors in considerations about the fairness and potential biases of an ML system. The role that input attributes play both in relation to the target variable and in relation to each other provides the evidentiary basis for explaining and ethically justifying the results of any model. If we cannot understand and interpret the rationale behind a result based upon a clear view of how the components of a model work together to produce it, we will not be able to ensure that the inferences and correlations that are contributing to the generation of a particular outcome are fair, equitable, and reasonable. Possibilities of lurking proxies for discrimination and biased correlations buried deep within opaque, high-dimensional feature spaces and model architectures make the interpretability of an ML model a critical partner in safeguarding bias mitigation and the overall fairness of the system itself.

Interpretability

The priority of ensuring the **interpretability** of ML tools intended for use in the domain of CSC should play a central role in the model building process. The transparency of the trained system (or set of systems) should be considered from the start, because the establishment of the equity, safety, and reliability of the model will significantly hinge on its degree of interpretability (i.e. its intelligibility to human reasoning) and explainability (i.e. the conveyability of the logic behind its results). The use of an opaque or 'black box' model in predictive analytics that directly impact individuals and families can present insurmountable challenges for a project team in terms of:

- the verification and validation of the system's own operational integrity,

- its usability as easily accessible support for the evidence-based reasoning of implementers/ users, and

- the confirmation that potential discriminatory patterns and inferences lurking in its fit to the underlying distribution have been mitigated or excluded.

These issues related to the safe and ethical functioning of a predictive ML model are magnified in high-impact and safety-critical domains such as CSC, for system errors, unreliable performance, and lurking biases may have life and death consequences.

The demand for **transparent explanation** is of equal importance in ML design and in CSC practice. Providing clear and accurate information is essential to decision-making. This imperative of transparency is integrated into ethical frameworks in social work like the BASW Code of Ethics through ideas such as giving people the full information they need to make informed choices and allowing people to access information about themselves. Similarly, in the responsible implementation of ML systems, the provision of clear explanations of algorithmic outputs is a key and non-substitutable component of offering affected decision recipients justifications for decisions reached (Wachter, Mittelstadt & Floridi, 2017). The explainability of algorithmic results provides access to the logic behind them and can help prevent errors, assist implementers and affected individuals

to ascertain the appropriateness of criteria used, and ultimately increase trust in such tools. (Doshi-Velez & Krotz, 2017, p. 2; Vaithianathan, 2017, p. 5). Moreover, explainable ML systems can help human decision-makers determine whether influential factors justify their interferences about the personal life of a particular child or family (Church & Fairchild, 2017, p. 77).

The advantages of using interpretable ML systems in CSC derive from the capacity of certain algorithmic techniques like decision trees, rule lists, and regression-based analysis and its extensions to track reasonable expectations and to replicate humanly accessible logical inferences in a plain and understandable way. In particular, methods like logistic regression, regularised regression (LASSO), and generalised linear models are both **linear** and **monotonic**. Combined with a reasonably **sparse feature space**, algorithms like these allow for ML model building that yields optimally interpretable systems.

Traits of regression-based models that allow for optimal explainability and transparency (ICO & Turing, 2019)

- **Linearity:** Any change in the value of the predictor variable is directly reflected in a change in the value of the response variable at a constant rate. The interpretable prediction yielded by the model can therefore be directly inferred from the relative significance of the parameter/weights of the predictor variable and have high inferential clarity and strength.
- **Monotonicity:** When the value of the predictor changes in a given direction, the value of the response variable changes consistently either in the same or opposite direction. The prediction yielded by the model can therefore be directly inferred. This monotonicity dimension is a highly desirable interpretability condition of predictive models in many heavily regulated sectors, because it incorporates reasonable expectations about the consistent application of sector-specific selection constraints into automated decision-making systems.
- **Sparsity/Non-Complexity:** The number of features (dimensionality) and feature interactions is low enough and the model of the underlying distribution is accessible enough to enable a clear understanding of the function of each part of the model in relation to its outcome.

The importance of interpretability in CSC derives both from the domain context and the use case. In general, high-impact, safety-critical, or other potentially sensitive environments heighten demands for the thoroughgoing accountability and transparency of ML systems (Rudin, 2019). Specifically, when the use of algorithmic decision-support critically and directly affects individuals and families, implementers (and the individual and families themselves) must be able to understand the logic behind the risk scores or predictions generated. Insofar as these results have direct bearing on individual-impacting decisions, their interpretability will be crucial for protecting the rational autonomy, individual dignity, and due-process rights of affected parties. For the use of an ML system in CSC to be ethically justifiable, implementers must be able to ***meaningfully incorporate the outputs of these systems into their own rational deliberations as evidence-based decision support. Interpretability is crucial, in this respect, because an algorithmically generated risk score is only as meaningful to reasoning human decision-makers as the clarity and accessibility of the rationale behind it***. In cases where an ML system is assisting decisions that directly impact children and families, unthinking reliance on a high or low risk score, without reflective engagement and understanding of the logic underlying it (as well as an active weighing of that logic against the concrete circumstances of the lives impacted), signals a rational deficit in the implementation, which does significant moral harm to those subjected to the tangible coercion of the result.

Despite this sense that optimal interpretability should remain an indispensable prerequisite in individual-impacting ML systems, there remains active and ongoing discussion in the field of CSC about the 'trade-off' or balancing between performance/accuracy and explainability. Because the curviness and non-linearity of more complex systems like artificial neural nets and support vector machines are able to better fit complicated patterns in underlying data distributions, they tend to produce predictions with higher degrees of accuracy than more rigid, but interpretable, rule-based or linear models but at the cost of a significant degree of transparency. The mathematical complexity of these systems means that their logical innerworkings exceed human-scale understanding (Burell, 2016). Similarly, since ensemble methods like random forests or adaptive boosting combine multiple interpretable base algorithms in order to improve overall system performance, they tend to increase the accuracy of trained models at the expense of a significant degree of interpretability. In this case, the aggregation of multiple base learner models through voting, sequential reweighing, or averaging obscures explanatory access to the relative

influence of individual features on the aggregated model's specific outputs.

Although concerns about sacrificing transparency for performance in algorithmic decision-support are well-established (Roberts, O'Brien, & Pecora, 2018; Brauneis & Goodman, 2018; Church & Fairchild, 2017; Munoz, Smith, & Patil, 2016; Nash, 2017; among others), some researchers have explored possible ways to incorporate 'black box' algorithms into predictive risk modelling in CSC. Jolley (2012), for example, uses neural network algorithms to predict the risk of recurrent child maltreatment. In order to address interpretability issues, she employs a 'post-hoc' method of building a decision tree on the outputs of the neural net so that a partial account of predictor-response relationships can be accessed (following Farragi, Leblanc, & Crowley, 2001). While Jolley's experiment does yield improved performance on predicting recurrent maltreatment from existing administrative data, her model—even with the limited information generated by the supplementary explanation technique employed—preserves the opacity of the system. It fails to provide a level of interpretability appropriate for understanding the actual influence of feature variables and their interactions on the outputs. This model, in turn, also presents insufficient assurances about architectural aspects of fairness and bias mitigation.

More recently, working with administrative CSC data from Broward County, Florida, Schwartz, Nowakowski-Sims, Ramos-Hernandez, and York combine decision tree algorithms with ensemble methods to show that the deployment of the latter techniques produces performance improvements in comparison with the use of logistic regression (Schwartz et al., 2017 building off of Watkins et al., 2013). While these researchers deliberately selected decision trees for reason of the 'added advantage of their [transparency] and interpretability', the study leaves unclear how the incorporation of ensemble methods affects the interpretability and explainability of their system. It also does not clarify how this introduction of a degree of opacity ultimately impacts the implementation and delivery component at the level of the evidence-based reasoning of the user.

In a real-world example, Chouldechova et al. (2018) also exploit ensemble methods (random forest and XGBoost) in their rebuild of a predictive risk model for the Allegheny Tool. While these researchers make mention of the widespread reluctance '...due to interpretability concerns...to adopt a more complex model despite evidence of improved prediction accuracy', they use the more complex systems anyway. They do not provide any substantive justification for why it may be ethically permissible to do so, aside from an allusion to the 'limited utility' of reaching an understanding of the correlational character of ML decision-support systems, more generally (Chouldechova et al., p. 11, 2018).

If anything, these research examples show that effective redress of the problem of interpretability in ML applications in CSC is still at an inchoate stage in its development. Several tools for supplementary explanation of 'black box' ML systems such as LIME and SHAP have been recently introduced for use as open source software, but significant issues with the certainty and reliability of the explanations they enable have also been discovered and underscored (Molnar, 2019; Alvarez-Melis & Jaakkola, 2018; Mittelstadt, Russell, & Wachter, 2018; Leslie, 2019). This is not to say that actionable ways of interpreting and explaining complex and as yet opaque ML systems will not be discovered in the near future. Much valuable research is underway into the technical dimensions of supplementary explanation facilities and the practical dimensions of the responsible implementation of interpretable ML (ICO & Turing, 2019; Royal Society, 2019; for reviews of the technical components of explainable AI, see Adadi & Berrada, 2018; Došilović, Brčić, & Hlupić, 2018; Eisenstadt & Althoff, 2018; Pedreschi et al., 2018; Gilpin et al., 2019; Mittelstadt et al., 2019).

It should be emphasized, however, that until sound methods for enabling the interpretability and explainability of 'black box' ML decision-support systems have been widely tested, validated, and confirmed, the use of these systems should be checked by careful considerations of the aforementioned risks of bias, injustice, and injury to individual dignity. The use of appropriately interpretable ML systems must remain a priority in CSC to protect individuals and families from potentially harmful but simultaneously unexplainable outcomes. That being said, it is widely accepted that, with a combination of solid domain knowledge, context awareness, and good data science, building optimally accurate and intrinsically interpretable ML models may be possible (Rudin, 2019; Rudin & Ustun, 2018; Kim, Khanna, & Koyejo, 2016; Lou, Caruana, & Gehrke, 2012). Where properly handled data resources produce well-structured, meaningful representations and domain expertise is diligently incorporated into model architectures, maximally effective interpretable techniques may often, in fact, be preferable to opaque systems. Careful data pre-processing and iterative model development can, in these cases, hone the accuracy of such interpretable systems in ways that may make the advantages gained by the combination of their

performance and transparency outweigh the benefits of more semantically non-transparent approaches.

Nonetheless, when considering the role of interpretability in the range of possible data scientific tools that may be deployed in the field of CSC (beyond those used for predictive risk modelling), specifics of the use case and the context of application matter greatly. While optimal interpretability should remain an indispensable prerequisite in individual-impacting ML systems, other applications that classify or predict at organisational- or population-levels may constructively draw upon more complex (and less intrinsically interpretable) algorithmic techniques with beneficial macro-level effects. As a recent example, the Data Science for Social Good Summer Fellowship, hosted jointly by the University of Warwick and The Alan Turing Institute, partnered with Ofsted to build an ML tool for the 'data driven prioritisation of independent fostering agencies in England and Wales' (Brundyn et al., 2019). In building this system, data scientists drew upon ensemble methods (like random forest algorithms) to create an ML model that flags independent fostering agencies in need of targeted inspection and support from Ofsted, thereby directing limited governmental resources where they could most usefully be applied. While less interpretable at the local, instance level, this trained model provides users access to the relative importance of features in the operation of the model as a whole—in this case features such as historical and last inspections and proximate surveys. The usefulness of such an ML tool therefore derives both from the efficiency gains in the provision of critical services to children in need and the insights facilitated by a global understanding of the trained system itself.

Implementation

One significant dimension of model production that we have not yet discussed is the aspect of implementation. Where ML models are developed in the context of CSC with a view to being used in practice, strategies for their responsible implementation have to enter into the design and deployment lifecycle early on and the training of users and implementers has to be prioritised. Cognitive biases can influence whether and how users and affected individuals interact with ML models. This can affect and prejudice the way in which tools are handled and how their results are interpreted. It also consequently leads to their potential overuse, underuse, or misuse. Appropriate user training and the provision of clear information to intended implementers and decision subjects will play a fundamental role in countering such possible biases. This is equally valid where ML models are intended to be used for research purposes or in practice, for individual-level assessments or for population-level analysis.

To begin with, automation bias can mean that humans rely on automated decision-making systems to the point where they neglect contradictory advice from non-automated sources, even if this overreliance leads to erroneous outcomes (Mosier & Skitka, 1996). It can affect both experts and non-experts, individuals and teams. This can be due to a perception of automated systems as being superior to or more computationally powerful than humans (Dzindolet, Pierce, Beck & Dawe, 2002; Lee & See, 2004) and ultimately may lead to automation complacency – the tendency for people not to carry out sufficient oversight of automated decision-making and implicitly and unjustifiably trust it. Automation complacency is most likely to exist when individuals use highly reliable systems or systems where errors are not immediately perceivable and their attention is consumed by a number of competing tasks, resulting in insufficient oversight (Parasuraman & Manzey, 2010, p. 385, 387-388, 403) – situations that could arise if ML models are intended to be used to alleviate high demands on social workers in practice.

Potential risk aversion may also lead to a misuse of ML models that is similar to overreliance. An ethnographical study in 2007-2008 Australia found that practitioners used decision tools as a way to ensure accountability and consistency within their organisations rather than to support decision-making (Gillingham, 2009). Risk averseness may also make social workers more likely to investigate or to keep cases open for longer than necessary where ML models raise risk alerts.

A former social worker, speaking to one of our researchers, as part of the roundtable, shared their thoughts on the potential impacts of risk assessments informed by ML models. They pointed out that, where unwarranted investigations take place or cases are kept open longer than necessary as a result of a high-risk score, the relationship between a social worker and a family can ultimately be greatly upset. Opening investigations and failing to substantiate them or keeping cases open longer than necessary may also undermine the trust of families in the CSC system in general and dissuade them from reaching out when in need of support. Depending on the problem formulation chosen, ML models in CSC can provide predictions going months or years into the future. This highlights the need for clarity among social workers, policy-makers, citizens, and ML developers about how best to design the interface with and use of ML tools.

On the opposite side of automation bias is algorithmic aversion which may lead individuals to underuse algorithmic advice to their detriment. People trust themselves in areas where they are knowledgeable and tend to underuse outside advice, whether algorithmic or human (Logg, Minson & Moore, 2018). There is also aversion to the idea of algorithms or autonomous machines making decisions with a moral impact, such as with legal or medical decisions, even where such decisions have positive outcomes (Bigman & Gray, 2018).

Beyond aversion, the trust individuals place in algorithmic decision-making in general can depend on the tasks to be completed (Lee, 2018). Automated decision-making is trusted for tasks involving mainly mechanical skills, provided the algorithm's reliability and lack of bias is ensured, but individuals find algorithmic decisions less trust-worthy and less justifiable for more complex tasks due to algorithmic limitations in considering outliers, exceptions and unquantifiable variables. The implications of algorithms being more distrusted in the highly complicated and multivariate CSC context are clear.

Indeed, a public servant speaking at the stakeholder roundtable highlighted that one of their take-aways from exploring ML tools in CSC practice was that social workers did not use the risk scoring features of the tools they had access to. This is not to say that the tool was completely unused. Discussions with social workers and audits of cases, where the tool was deployed, showed that what social workers found most helpful was the access to additional information, such as school attendance of children, as well as its visualisation of such information which allowed them to have conversations about the information with the families concerned. A company developing analytics tools at the roundtable also shared that, from their experience, an initial exhilaration about having all information in a central location is common among local authorities. The ease of information access was also a reason for another public body at the roundtable to pivot a project originally intended to provide general level strategic overview of the functioning of the CSC system towards instead providing social workers with information and supporting their practice.

Overall, to ensure the benefits of algorithmic support for complex decision-making, forethought should go into how to design the interface of users and affected decision subjects with ML models, how to train these users, what procedures and practices to incorporate into implementation processes, and all in all, how to ensure the models are actually supporting humans. Since ML research into CSC may require many resources, sufficient forethought should be put into creating tools that will be deployed appropriately by their intended users and to supplement the employment of such tools with sufficient training regimes.

User interface

User interface and set-up can play a key role in ensuring that the benefits of ML tools are reaped without diminishing the role of human oversight. To effectively support, rather than replace human judgment, automation should be incorporated in a way that provides information integration to support skilful human processes rather than recommending specific actions (Crocoll & Coury, 1990; Rovira, McGarry & Parasuraman, 2007; Sarter & Schroeder, 2001).

This is not only ethically preferable but also unavoidable in areas such as CSC, given the complexity and nuance outlined earlier in this report. For example, even where an ML tool identifies a heightened risk score for a child, how this information is used, alongside social work practice is determined by the actions and decisions of those working in CSC. These risks need to be considered within the particular context of the child and family and their expressed needs and desires. Human input will be necessary to appropriately adapt solutions to the circumstances of the particular family given the difficulties involved in identifying proportionate intervention to support families and to avoid children receiving a service that is insufficient or, conversely, too intensive for their level of need (Forrester, 2017).

To support human decision-making effectively, the user interface should be transparent about the rationale behind a tool's outputs and about its performance limitations and levels of uncertainty. It should include information about the factors that led to its output for a particular child or family and provide optimal support for evidence-based judgments—whether in the form of accessible and plain language explication of the statistical generalisations or through visualisations. Frontline CSC practitioners should be trained and clearly instructed as to how to treat such information knowledgeably and appropriately. They should also be trained to utilise the results of predictive analytics with an active knowledge of the difference between correlation and causation. The fact that we can know and show what the influential factors for a particular prediction are does not imply direct knowledge into what the causes of a predicted risk are. ML models work on the basis of correlation rather than causation. Be that as it may, weighing these factors as they relate to the particular circumstances of affected children and families is the most crucial element of

the delivery of algorithmic decision-support, for these factors provide part of the inferential underpinnings for the social worker's well-considered and evidence-anchored judgement of the situation of interest.

The transparency of other elements of the ML tool may also be helpful for users. The interface should ideally also make clear the system's compliance with a particular chosen fairness standard and the model's confidence level about this particular output, (Leslie, 2019, p. 21-22; McGuirl & Sarter, 2006) visually and interactively highlighting uncertainty levels at each stage of a machine learning pipeline (Sacha, Senaratne, Kwon, Ellis & Keim, 2016).

The way algorithmic outputs are presented should also be carefully considered and planned so as to counter a range of biases to which individuals could be subject. These include automation bias, confirmation bias (Shafir, 1993; Klauer, Musch & Naumer, 2000; Taber & Lodge, 2006), and information-framing bias (Tversky, Kahneman, 1981), each of which may influence the way people interpret algorithmic outputs. A well-designed user interface can not only improve trust and usability, but also enhance the accuracy of operators correctly rejecting wrong suggestions, counteracting automation complacency. The amount of information provided to users should be relevant, accurate, clear, and manageable (from a cognitive perspective) in order to increase user investment and trust in such tools (Chen et al., 2014; Lee & See, 2004; Mercado et al., 2016).

Training and procedures

Educating CSC managers and practitioners to employ evidence-based reasoning is key to safeguarding and improving the quality of CSC. Diligent training regimes are consequently a necessary dimension of the responsible implementation of any potential ML systems, because these tools are intended to augment such a capacity for empirically-informed judgement. This is a crucial component of the delivery side of ML innovation in CSC not least because of the need to secure and maintain the agency and autonomy of frontline practitioners. The rise of a culture of risk management and auditing protocols is already viewed as having increased the formalisation of the profession and the potential diminution of the role of independent professional judgment (Broadhurst, Hall, Wastell, White & Pithouse, 2010). Social workers have expressed concern that their flexibility and freedom to respond to the dynamic needs of children and families has been curbed by standardisation and 'central prescription' of working practices (Munro, 2011, p. 6-7). For this reason, training provided should ensure they are confident and able to use ML models with appropriate autonomy to utilise their expertise in tandem with the supplementary information provided by statistical analysis. Training and professional development opportunities should be cultivated and delivered with a view to allowing these users to more effectively exercise their professional judgment in using ML outputs.

One of the central motivations behind this incorporation of appropriately intensive training into the responsible implementation of ML systems in CSC is the desire to prevent risk assessment tools from deskilling social workers and ultimately hindering the development of their expertise (Gillingham, 2010). Both in the work environment and in the procedures established for accountable workplace practices, such tools should not be viewed as an alternative to the exercise of professional know-how. Rather, they should be put in their place, on a practical level, as just one service-assisting affordance, that provides decision-makers with an additional ability to consider patterns in the available data that would have otherwise gone unnoticed.

Putting ML systems in their proper place in the children's social care environment will also involve training implementers to understand how human judgments and values have been integrated into the models they are using. Implementers should have a working awareness of the original motivations and purposes behind the choice to use an ML system, so that they can assess how to align the system's results in a particular context with bigger picture organisational and care-giving goals as well as weigh those goals against family-defined ends and preferences. They should correspondingly grasp the reasoning behind the tool's problem formulation and target variable in order to know what the model is intended to predict or score and to comprehend the possible limitations thereof. Without clear definitions and context-setting, users will find it difficult to know what is being measured or detected (Garrison, 2012, p. 26).

Users of statistical outputs should additionally be trained in understanding and analysing statistical outputs and their limitations. Most people, especially untrained ones, are liable to fallacies and misjudgements when drawing conclusions and insights on the basis of statistical information. Individuals may exhibit overconfidence in future predictions on the basis of historically consistent data (illusion of validity) (Kahneman & Tversky, 1973; Tversky & Kahneman, 1974). They may see importance in random streaks and clusters of data (clustering illusion) (Gilovich, 1991). They may neglect to pay sufficient attention to general trends (base rate

fallacy), (Bar-Hillel, 1980) to limitations in insights that are due to smaller dataset sizes (extension neglect) (Kahneman & Tversky, 1999), and to probabilities of events materialising (neglect of probability) (Bar-Hillel, 1980). To make it clear that ML models are meant to supplement human judgments and to counter such potential biases, user training should include use case examples and exercises to illustrate how statistical evidence can and should be weighed in practice, as well as how it can be misjudged (Leslie, 2019, p. 22). Among other things, intended users should be preventively exposed to failures and errors of the model (Bahner, Hüper & Manzey, 2008).

Finally, users of ML models should be made accountable for their responses and for overall performance, thus incentivising them to be more active in overseeing automation (Mosier, Skitka, Heers & Burdick, 1998). In the context of CSC, the monitoring of automation should be proportionate to the reliability of the automation, (Moray, 2003) as well as to the potential impacts that automation-supported decisions can have on children's and family's lives.

V. WHAT IS TO BE DONE?

RECOMMENDATIONS FOR STEERING DIRECTION OF THE USE OF MACHINE LEARNING IN CHILDREN'S SOCIAL CARE

Recommendations

In this final section we outline some preliminary recommendations for steering the present direction of the use of ML in CSC, both in its application to practical, real-world problems and as a medium for research insight and discovery. We also include some provisional suggestions for optimising the capacity of future data scientific research and innovation in CSC to produce tangible societal benefits and advance individual, familial, and public wellbeing.

1. **Mandate the responsible design and use of ML models in CSC at the national level:** In our Family engagement workshop, family members stressed the need for nationally mandated public standards to guide the ethical design and deployment of ML in CSC. Such standards should protect affected stakeholders against the misuse of data in social care settings. They should also provide LA's with guidelines for designing, procuring, and implementing ML models in CSC fairly, ethically, and responsibly.

2. **Connect practitioners and data scientists across local authorities to improve ML innovation and to advance shared insights in applied data science through openness and communication:** The current siloing of research and innovation practices is stifling progress in the public sector use of ML and data science in CSC. Research into mechanisms that may help foster network-building and insight-sharing among practitioners and researchers is much needed.

3. **Institutionalise inclusive and consent-based practices for designing, procuring, and implementing ML models:** The participants of our family engagement workshop emphasised the importance of consent and deliberative involvement at all points across the ML lifecycle. Local authorities should actively pursue the creation of engagement processes, which will optimise the consent-based involvement of affected stakeholders from start to finish of any ML innovation project. This may include the integration of citizens' juries, citizens' assemblies, distributed dialogue, and other means of deliberative democratic participation into ML design, procurement, and implementation workflows.

4. **Fund, initiate, and undertake active research programmes in system, organisation, and participant readiness:** There is a need for much empirical research to be done to identify and better understand the barriers and enablers to effective integration of responsible ML innovation into CSC settings. Use-case based studies, comparative analyses of innovation interventions in care setting, and experimentally-anchored examinations of the factors and contexts of system, organisation, and participant (SOP) readiness will empower organisations and practitioners to move beyond trial-and-error implementation processes and toward reflective and intentional innovation intervention. Insights from this programme of research will better enable the deliberate and ecology-aware development and implementation of ethically-designed ML models. It will also allow for more effectual resource allocation and more targeted capacity building.

5. **Understand the use of data in CSC better so that recognition of its potential benefits and limitations can more effectively guide ML innovation practices:** Responsible data science in CSC must better understand how data has and is being used in the field. This will involve landscape scoping and empirical investigation of which local authorities have been using data, for what purposes, and to what degrees of success or failure. This knowledge will help manage wider expectations about what is possible in the use of ML in CSC and help to connect stakeholders in leveraging experience to understand which projects and undertakings to commission or decommission.

6. **Use data insights to describe, diagnose and analyse the root causes of the need for CSC, experiment to address them:** The focus of much of applied data science and data scientific research in CSC should be recalibrated, so that its resources can be directed at understanding, diagnosing, and addressing the root causes behind the deeper social-structural problems and dynamics that are generating expanding needs for CSC services. Existing local authority

and community-based data can be used to pursue descriptive and diagnostic insights at the institutional, organisational, and socioeconomic-structural levels so that underlying causal influences and social factors can be identified and better understood. Policies and interventions can then be designed based on real world determinants of the need for CSC, and rigorously and responsibly evaluated.

7. **Focus on individual- and family-advancing outcomes, strengths-based approaches, and community-guided prospect modelling:** Research is needed to explore how positive (individual- and family-advancing) outcomes can be integrated into data analytics in CSC. Part of developing such prospect assessment models would involve inclusive, family- and community-integrating processes of objective setting, problem formulation, and outcome definition as well as multi-stakeholder and interdisciplinary approaches to model planning and implementation. Through these processes of co-creation, the analytics would come to better reflect the best interests of the communities to which they apply. Exploring the possibilities of strengths-based, prospective approaches would also involve creating a better data landscape capable of capturing how children and families experience CSC, as well as patterns indicative of positive outcomes that foster the wellbeing and flourishing of children in need and their families. At the same time, those working toward cultivating this data landscape would have to safeguard the interests of affected data-subjects—in particular, those most vulnerable to over-collection and the potential harms of data misuse—by working through privacy-preserving and consent-based programming. This starting point in an improved data landscape would call upon data scientists to develop novel approaches to these analytics that enable holistic considerations of developmental, physical, cognitive, social and emotional needs of affected individuals.

8. **Improve data quality and understanding through professional development and training:** Data collection, analysis, and use of ML models should be built into social care and social work training. The importance of accurate, impartial data collection should be emphasised in social care training by all those who may contribute to data collection and analysis (e.g. administrators, foster carers, residential home workers, social workers, technical experts). The analysis of data and its limitations and use of ML models should be integrated into social worker training to ensure understanding of and the responsible use of ML in CSC that follows the follows the ethical values of:

- **Respect** the dignity of individual persons, empower them, and value the uniqueness of their aspirations, cultures, contexts, and life plans
- **Connect** with each other sincerely, openly, and inclusively, and prioritise trust, solidarity, and interpersonal collaboration
- **Care** for the wellbeing of each and all, and serve others with empathy, selflessness, and compassion
- **Protect** the priorities of social justice and the public interest by ensuring equity, recognising diversity, and challenging discrimination and oppression

BIBLIOGRAPHY

Aarons, G. A., & Palinkas, L. A. (2007). Implementation of evidence-based practice in child welfare: Service provider perspectives. *Mental Health Services Research, 34*, 411–419. DOI: 10.1007/s10488-007-0121-3.

Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(1), 4–23. DOI: 10.1007/s10488-010-0327-7.

Aarons, G. A., Fettes, D. L., Sommerfeld, D. H., & Palinkas, L. A. (2012). Mixed methods for implementation research: Application to evidence-based practice implementation and staff turnover in community-based organizations providing child welfare services. *Child Maltreatment, 17*(1), 67–79. DOI: 10.1177/1077559511426908.

Aarons, G.A., Fettes, D.L., Hurlburt, M.S., Palinkas, L.A., Gunderson, L., Willging, C.E., & Chaffin, M.J. (2014). Collaboration, negotiation, and coalescence for interagency-collaborative teams to scale-up evidence-based practice. *Journal of Clinical Child and Adolescent Psychology*, *43*(6), 915–928. DOI: 10.1080/15374416.2013.876642.

Abiteboul, S., Arenas, M., Barceló, P., Bienvenu, M., Calvanese, D., David, C.,.... Yi, K. (2017). Research directions for principles of data management (Dagstuhl Perspectives Workshop 16151). *Dagstuhl Manifestos* (pp. 1-28). Schloss Dagstuhl: Dagstuhl Publishing. Retrieved from https://arxiv.org/pdf/1701.09007.pdf.

Academy Health (2014). *Promoting early and lifelong health: The challenge of adverse childhood experiences (ACEs) and the promise of resilience*. AcademyHealth. Retrieved from https://www.academyhealth.org/about/programs/adverse-childhood-experiences-aces.

Access Now & Amnesty International (2018). *The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems*. Access now. Retrieved from https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/.

ACM US Public Policy Council (2017). *Statement on Algorithmic Transparency and Accountability*. Association for Computing Machinery (USACM). Retrieved from https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

Adadi, A. & Berrada, M., (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138-52160. DOI: 10.1109/ACCESS.2018.2870052.

All Party Parliamentary Group for Children (2017). *No good options: Report of the inquiry into children's social care in England*. Retrieved from https://www.ncb.org.uk/sites/default/files/uploads/No%20Good%20Options%20Report%20final.pdf

Alvarez-Melis, D. & Jaakkola, T.S. (2018). On the robustness of interpretability methods. In *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)* (pp.66 – 71). Stockholm, Sweden. Retrieved from https://arxiv.org/pdf/1806.08049.pdf.

Amoore, L. (2009). Algorithmic war: everyday geographies of the war on terror. *Antipode: A Radical Journal of Geography, 41*(1), 49-69. DOI: 10.1111/j.1467-8330.2008.00655.x.

Amoore, L. & Raley, R. (2016). Securing with algorithms; Knowledge: decision, sovereignty. *Security Dialogue, 48*(1), 3 – 10. DOI: 10.1177/0967010616680753.

Ash, J. (1997). Organizational factors that influence information technology diffusion in academic health sciences centers. *Journal of the American Medical Informatics Association, 4(*2), 102–111. DOI: 10.1136/jamia.1997.0040102.

Association of Directors of Children's Services (2018). *Safeguarding pressures phase 6. Research report.* Manchester: The Association of Directors of Children's Services Ltd (ADCS). Retrieved from https://adcs.org.uk/safeguarding/article/safeguarding-pressures-phase-6.

Australian Association of Social Workers (2010). *AASW Code of Ethics*. Retrieved from https://www.aasw.asn.au/document/item/1201.

Bahner, J.E., Hüper, A.D. & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies, 66* (9), 688-699. DOI: 10.1016/j.ijhcs.2008.06.001.

Banks, S.J. (2011). Ethics in an age of austerity: social work and the evolving new public management. *Journal of Social Intervention: Theory and Practice, 20*(2), 5-23. Retrieved from http://dro.dur.ac.uk/9101/1/9101.pdf?DDD34+dss4ae.

Banyard, V., Hamby, S., & Grych, J. (2017). Health effects of adverse childhood events: Identifying promising protective factors at the intersection of mental and physical wellbeing. *Child Abuse and Neglect, 65*, 88–98. DOI:10.1016/j.chiabu.2017.01.011.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*. *44* (3), 211–233. DOI: 10.1016/0001-6918(80)90046-3.

Barocas, S. & Selbst, A.D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671-732. DOI: 10.15779/Z38BG31.

Beahrs, J. O. (1991). Volition, deception, and the evolution of justice. *Bulletin of the American Academy of Psychiatry & the Law*, *19*(1), 81-93.

Becan, J. E., Bartkowski, J. P., Knight, D. K., Wiley, T. R. A., DiClemente, R., Ducharme, L. ,..... Aarons, G.A. (2018). A model for rigorously applying the exploration, preparation, implementation, sustainment (EPIS) framework in the design and measurement of a large scale collaborative multi-site study. *Health & Justice, 6*(1), 9. DOI: 10.1186/s40352-018-0068-3.

Beer, D. (2017). The social power of algorithms. *Information, Communication & Society, 1,* 1-13. DOI: 10.1080/1369118X.2016.1216147.

Beninger, K., Newton, S., Digby, A., Clay, D. & Collins, B. (2017). *Newcastle City Council's Family insights programme*. Department for Education. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/624837/Newcastle_City_Council_s_Family_Insights_Programme.pdf.

Berk, R., Heidari, H., Jabbari, S., Kearns, M. & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. Retrieved from https://arxiv.org/pdf/1703.09207.pdf.

Bethell, C. D., Newacheck, P., Hawes, E., & Halfon, N. (2014). Adverse childhood experiences: Assessing the impact on health and school engagement and the mitigating role of resilience. *Health Affairs*, *33*(12), 2106–2115. DOI: 10.1377/hlthaff.2014.0914.

Biestek, F.P. (1957). *The casework relationship.* George Allen & Unwin.

Bigman, Y.E. & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21-34. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010027718302087.

Billings, J., Dixon, J., Mijanovic, T. & Wennberg, D. (2006). Case findings for patients at risk of readmission to hospital: Development of algorithm to identify high risk patients. *British Medical Journal*, *12*, 327-333. DOI: 10.1136/bmj.38870.657917.AE.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 377. ACM. Retrieved from https://arxiv.org/pdf/1801.10408.pdf.

Boden, M., Bryson, J., & Caldwell, D. (2009). *Principles of robotics: Regulating robots in the real world*. UKRI EPSRC. Retrieved from: https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.

Bogost, I. (2015). The cathedral of computation. *The Atlantic, 15*. Retrieved from https://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/.

Bowyer, S., Gillson, D., Holmes, L., Preston, O., & Trivedi, H. (2018). *Edge of care cost calculator: Change project report.* Devon: Research in Practice.

Brandon, M., Belderson, P. Warren, C. Gardner, R. Dodsworth, J., & Black, J. (2008). *Analysing child deaths and serious harm through abuse and neglect: What can we learn?.* Nottingham: DCSF Publishing.

Brandon, M. Bailey, S. Belderson, P. Gardner, R. Sidebotham, P. Dodsworth, J., & Black, J. (2009). *Understanding serious case reviews and their impact.* Nottingham: DCSF Publishing.

Brandon, M. Sidebotham, P., Bailey, S. Belderson, P. Hawley, C. Ellis, C., & Megson, M. (2012). *New learning from serious case reviews: a two-year report for 2009-2011.* University of East Anglia & University of Warwick / Department of Education.

Bratton, B.H. (2016). *The stack: On software and sovereignty.* MIT Press.

Brauneis, R. & Goodman, E.P. (2018). Algorithmic transparency for the smart city. *Yale Journal of Law & Technology, 20*(103), 103-176. Retrieved from https://www.yjolt.org/sites/default/files/20_yale_j._l._tech._103.pdf.

Broadhurst, K., Hall, C. Wastell, D., White, S., & Pithouse, A. (2010). Risk, instrumentalism and the humane project in social work: Identifying the *informal* logics of risk management in children's statutory services. *British Journal of Social Work*, *40*, 1046-1064. DOI: 10.1093/bjsw/bcq011.

Bromfield, L.M. & Higgins, D. (2004). The limitations of using statutory child protection data for research into child maltreatment. *Australian Social Work, 57*(1), 19-30. DOI: 10.1111/j.0312-407X.2003.00110.x.

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A. & Vaithianathan, R. (2019). Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-

making in child welfare services. In *CHI 2019: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* Paper No. 41. New York: ACM. DOI: 10.1145/3290605.3300271.

Brown, R. & Ward, H. (2013). *Decision-making within a child's timeframe: An overview of current research evidence for family justice professionals concerning child development and the impact of maltreatment*. Childhood Wellbeing Research Centre. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/200471/Decision-making_within_a_child_s_timeframe.pdf

Brownlee, J. (2015). *8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset*. Retrieved from https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

Brundyn, A., Budd, J., Hernandez, V., Joseph, J., Sidgwich, J., & de Unanue, A.(2019). *Data-driven Prioritisation of Independent Fostering Agency Inspections*. Retrieved from http://www.dssgfellowship.org/project/foster-agencies-inspection/.

Bryant, B., Parish, N., & Rea, S. (2016). *Action research into improvement in local children's services: Final research report, Spring 2016*. London: ISOS Partnership, Local Government Association.

Bryson, J. J. (2017). The meaning of the EPSRC principles of robotics. *Connection Science, 29*(2), 130-136. Retrieved from https://joanna-bryson.blogspot.com/2016/03/the-meaning-of-epsrc-principles-of.html.

Burell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1). DOI: 10.1177/2053951715622512.

Bywaters, P. (2015). Inequalities in child welfare: Towards a new policy, research and action agenda. *British Journal of Social Work, 45*(1), 6-23.

Bywaters, P., Brady, G., Sparks, T. & Bos, E. (2014). Inequalities in child welfare intervention rates: the intersection of deprivation and identity. *Child & Family Social Work, 21*(4). DOI: 10.1111/cfs.12161.

Bywaters, P., Brady, G., Sparks, T., Bos, E., Bunting, L., Daniel, B., . . . Scourfield, J. (2015). Exploring inequities in child welfare and child protection services: Explaining the 'inverse intervention law'. *Children and Youth Services Review*, *57*, 98-105. DOI: 10.1016/j.childyouth.2015.07.017.

Bywaters, P., Brady, G., Bunting, L., Daniel, B., Featherstone, B., Jones, C., ... Webb, C. (2017). Inequalities in English child protection practice under austerity: A universal challenge?. *Child & Family Social Work*, *23*, 53-61. DOI: 10.1111/cfs.12383.

Cabot, R.C., (1973). *Social service and the art of healing*. National Association of Social Workers.

Canadian Association of Social Workers (2005). *Code of ethics*. Retrieved from https://www.casw-acts.ca/sites/default/files/attachements/casw_code_of_ethics.pdf.

Capatosto, K. (2017). *Foretelling the future: A critical perspective on the use of predictive analytics in child welfare*. Kirwan Institute Research Report, Columbus, Ohio, US: Kirwan Institute. Retrieved from http://kirwaninstitute.osu.edu/wp-content/uploads/2017/05/ki-predictive-analytics.pdf.

Cardon, D. (2016). Deconstructing the algorithm: four types of digital information calculations. In R. Seyfert & J. Roberge (Eds.), *Algorithmic Cultures: Essays on meaning, performance and new technologies* (pp. 95-110). New York: Routledge.

Care Crisis Review (2018). *Care Crisis Review: Options for Change*. London: Family Rights Group. Retrieved from https://www.frg.org.uk/images/Care_Crisis/CCR-FINAL.pdf.

Chawla, N.V. (2010). Data mining for imbalanced datasets: An overview. In Maimon, O. & Rokach, L. (Eds.), *Data Mining and Knowledge Discovery Handbook* (2nd ed.) (pp. 853 – 867). London: Springer.

Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A. & Barnes, M. (2014). Situation awareness–based agent transparency (No. ARL-TR-6905). *Aberdeen Proving Ground*, MD: U.S. Army Research Laboratory.

Children's Workforce Development Council (2009). *The common assessment framework for children and young people: A guide for practitioners*. Retrieved from https://webarchive.nationalarchives.gov.uk/20130102192341/https://www.education.gov.uk/publications/eOrderingDownload/CAF-Practitioner-Guide.pdf.

Church, C.E. & Fairchild, A.J. (2017). In Search of a Silver Bullet: Child Welfare's Embrace of Predictive Analytics. *Juvenile & Family Court Journal, 68*(1), 67-81. DOI: 10.1111/jfcj.12086.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data, 5*(2), 153-163. DOI: 10.1089/big.2016.0047.

Chouldechova, A., Putnam-Hornstein, E., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research*, *81*, 1-15. Retrieved from http://proceedings.mlr.press/v81/chouldechova18a.

Coleman, D.L. (2005). Storming the castle to save the children: The ironic costs of a child welfare exception to the fourth amendment. *William and Mary Law Review*, *47*, 413-540.

Cowls, J. & Floridi, L. (2018). *Prolegomena to a White Paper on an Ethical Framework for a Good AI Society*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3198732.

Cowls, J. & Floridi, L. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). DOI: 10.1162/99608f92.8cd550d1.

Crampton, D. (2007). Research review: Family group decisionmaking: A promising practice in need of more programme theory and research. *Child and Family Social Work, 12*, 202–209. DOI: 10.1111/j.1365-2206.2006.00442.x.

Crocoll, W. M. & Coury, B. G. (1990). Status or recommendation: selecting the type of information for decision aiding. *In Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1525–1528). DOI: 10.1177/154193129003401922.

Cross, T.P. & Casanueva, C. (2009). Caseworker judgments and substantiation. *Child Maltreatment*, *14*(1), 38-52. DOI: 10.1177/1077559508318400.

Cuccaro-Alamin, S., Foust, R., Vaithianathan, R., & Putnam-Hornstein, E. (2017). Risk assessment and decision making in child protective services: Predictive risk modeling in context. *CYSR Children and Youth Services Review*, *79*, 291-298. DOI: 10.1016/j.childyouth.2017.06.027.

Curtis, C.M. & Denby, R.W. (2011). African American children in the child welfare system: Requiem or reform. *Journal of Public Child Welfare, 5*(1), 111-137. DOI: 10.1080/15548732.2011.542731.

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data, 5*(2), 120-134. DOI: 10.1089/big.2016.0048.

Damanpour, F. (1991). Organizational innovation: A meta-analysis of effects of determinants and moderators. *Academy of Management Journal, 34*(3), 555–590.

Dawes, R.M., Faust, D. & Meehl, P.E. (1989). Clinical vs actuarial judgement. *Science*, *243*(4899), 1668-1674.

Degli Esposti, M., Humphreys, D. K., Jenkins, B. M., Gasparrini, A., Pooley, S., Eisner, M., & Bowes, L. (2019). Long-term trends in child maltreatment in England and Wales, 1858-2016: An observational, time-series analysis. *Lancet Public Health, 4*(3), e148-e158.

Dencik, L, Hintz, A., Redden, J. & Warne, H. (2018). *Data Scores as Governance: Investigating uses of citizen scoring in public services*. Cardiff, UK: Data Justice Lab, Cardiff University, Open Society Foundations. Retrieved from https://datajustice.files.wordpress.com/2018/12/data-scores-as-governance-project-report2.pdf.

DePanfilis, D. & Zuravin, S. J. (1999). Predicting Child Maltreatment Recurrences During Treatment. *Child Abuse & Neglect*, *23*(8), 729–743.

Department for Digital, Culture, Media & Sport (2018). *Data Ethics Framework*. GOV.UK. Retrieved from https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework.

Department for Education (2018). *Children in Need of help and protection. Preliminary longitudinal analysis.* London: Department for Education. Retrieved from https://www.gov.uk/government/publications/children-in-need-of-help-and-protection-data-and-analysis.

Department for Education (2019). *Children looked after by local authorities in England. Guide to the SSDA903 collection 1 April 2018 to 31 March 2019 - Version 1.3.* London: Department for Education. Retrieved from https://www.gov.uk/government/publications/children-looked-after-return-2018-to-2019-guide.

Department for Education (2019a). *Statistics: Looked-after children*. Retrieved from https://www.gov.uk/government/collections/statistics-looked-after-children

Dingwall, R., Eekelaar, J., & Murray, T. (2014). *The protection of children* (2nd Ed.). New Orleans: Quid Pro Books.

Doshi-Velez, F & Krotz, M. (2017). Accountability of AI Under the Law: The Role of Explanation. Retrieved from https://arxiv.org/pdf/1711.01134.pdf.

Došilović, F.K., Brčić, M., Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, 2018(pp. 210-215). DOI: 10.23919/MIPRO.2018.8400040.

Doyle, M. & Dolan, M. (2002). Violence risk assessment: combining actuarial and clinical information to structure clinical judgements for the formulation and management of risk. *Journal of Psychiatric and Mental Health Nursing, 9,* 649-657.

Dunleavy, P. & Hood, C. (1994). From old public administration to new public management. *Public Money and Management, 14*, 9-16. DOI: 10.1080/09540969409387823

Dwork, C., Hardt, M., Pitassi, T., Reinbold. O. & Zemel, R. (2012). Fairness through awareness. In *ITCS '12 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference,* (pp. 214-226). New York: ACM. DOI: 10.1145/2090236.2090255.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*, 79–94.

Edmondson, A. C. (2004). Learning from failure in health

care: Frequent opportunities, pervasive barriers. *Quality and Safety in Health Care, 13(*suppl 2), ii3–ii9.

Ebert, L., Amaya-Jackson, L., Markiewicz, J. M., Kisiel, C., & Fairbank, J. A. (2012). Use of the breakthrough series collaborative to support broad and sustained use of evidence-based trauma treatment for children in community practice settings. *Administration and Policy in Mental Health, 39*(3), 187–199. DOI: 10.1007/s10488-011-0347-y.

Eisenstadt, V. & Althoff, K.D. (2018). *A Preliminary Survey of Explanation Facilities of AI-Based Design Support Approaches and Tools*. Retrieved from https://www.dfki.de/fileadmin/user_upload/import/9983_LWDA_2018_paper_59.pdf

English, D.J., Bangdiwala, S.I. & Runyan, D.K. (2005). The dimensions of maltreatment: Introduction. *Child Abuse and Neglect*, *29*(5), 441-460. DOI: 10.1016/j.chiabu.2003.09.023.

EPSRC (2011). *Principles of robotics*. Retrieved from: https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.

Eubanks, V. (2018). *Automating Inequality*. St Martin's Press.

EU High-Level Expert Group on AI (2019). *Ethics Guidelines for Trustworthy AI*. European Commission. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Faraggi, D., LeBlanc, M., & Crowley, J. (2001). Understanding neural networks using regression trees: An application to multiple myeloma survival data. *Statistics in Medicine, 20*, 2965-2976.

FAT/ML (2016). *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*. Fairness, Accountability, and Transparency in Machine Learning. Retrieved from https://www.fatml.org/resources/principles-for-accountable-algorithms.

Featherstone, B. (2001). Where to for feminist social work?. *Critical Social Work, 2*(1). Retrieved from: https://ojs.uwindsor.ca/index.php/csw/article/download/5619/4592?inline=1.

Featherstone, B., Morris, K., & White, S. (2014). A Marriage Made in Hell: Early Intervention Meets Child Protection. *The British Journal of Social Work*, *44*(7). 1735-1749. DOI: /10.1093/bjsw/bct052.

Ferlie, E. B. & Shortell, S. M. (2001). Improving the quality of health care in the United Kingdom and the United States: A framework for change. *Milbank Quarterly, 79*(2), 281–315.

Finch, T.L., Rapley, T., Girling, M., Mair, F.S., Murray, E., Treweek, S..... May, C.R. (2013). Improving the normalization of complex interventions: Measure development based on normalization process theory (NoMAD): study protocol. *Implementation Science, 8*(1), 43. Retrieved from https://implementationscience.biomedcentral.com/articles/10.1186/1748-5908-8-43.

Fixsen, D.L., Naoom, S.F., Blas., K.A., Friedman, R.M., Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute.

Floridi, L. & Lord Clement-Jones, T. (2019). The five principles key to any ethical framework for AI. *New Statesman.* Retrieved from https://tech.newstatesman.com/policy/ai-ethics-framework.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V.,.... Vayena, E. (2018). AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*(4), 689-707. DOI: 10.1007/s11023-018-9482-5.

Fluke, J. D., Shusterman, G. R., Hollinshead, D. M., & Yuan, Y.T. (2008). Longitudinal analysis of repeated child abuse reporting and victimization: Multistate analysis of associated factors. *Child Maltreatment, 13*(1), 76–88.

Forrester, D. (2017). Outcomes in children's social care. *Journal of Children's Services, 12*(2-3), 144-157. DOI: 10.1108/JCS-08-2017-0036.

Fraser, J. G., Griffin, J. L., Barto, B. L., Lo, C., Wenz-Gross, M., Spinazzola, J., ... & Bartlett, J. D. (2014). Implementation of a workforce initiative to build trauma-informed child welfare practice and services: Findings from the Massachusetts Child Trauma Project. *Children and Youth Services Review, 44,* 233-242. DOI: 10.1016/j.childyouth.2014.06.016.

Future of Life Institute (2017). *Asilomar AI Principles.* Future of Life Institute. Retrieved from https://futureoflife.org/ai-principles/.

Gambrill, E. & Shlonsky, A. (2000). Risk assessment in context. *Children and Youth Services Review, 22*(11-12), 813-837. DOI: 10.1016/S0190-7409(00)00123-7.

Garrison, M. (2012). Taking the risks out of child protection risk analysis. *Journal of Law and Policy*, *21*(1), 5-35. Retrieved from https://brooklynworks.brooklaw.edu/jlp/vol21/iss1/2/.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. Retrieved from https://arxiv.org/abs/1803.09010

General Social Care Council (2010). *Codes of Practice for Social Care Workers.* Retrieved from: https://www.scie.org.uk/workforce/files/CodesofPracticeforSocialCareWorkers.pdf?res=true

Ghate, D. (2016). From programs to systems: Deploying implementation science and practice for sustained real world effectiveness in services for children and families. *Journal of Clinical Child & Adolescent Psychology, 45*(6), 812-826. DOI: 10.1080/15374416.2015.1077449

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P.J. Boczkowski & K. A. Foot (Eds.), *Media Technologies: Essays on Communication, Materiality, and Society* (pp. 167-194). London: MIT Press Scholarship.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillingham, P. (2009). *The use of assessment tools in child protection: An ethnomethodological study* (Doctoral dissertation). The University of Melbourne. Retrieved from https://minerva-access.unimelb.edu.au/bitstream/handle/11343/35144/114684_Final%20Thesis%20%282%29%20Philip%20Gillingham.pdf?sequence=1&isAllowed=y.

Gillingham, P. (2010). 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*(4), 412-421. DOI: 10.1111/j.1365-2206.2011.00756.x.

Gillingham, P. (2016). Predictive risk modelling to prevent child maltreatment and other adverse outcomes for service users: Inside the 'black box' of machine learning. *British Journal of Social Work*, *46*(4), 1044-1058. DOI: 10.1093/bjsw/bcv031.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: The Free Press.

Glaberson, S. (2019). Coding Over the Cracks: Predictive Analytics and Child Protection. *Fordham Urban Law Journal,* 46(2), 307-363. Retrieved from: https://ir.lawnet.fordham.edu/ulj/vol46/iss2/3/.

Gleacher, A. A., Nadeem, E., Moy, A. J., Whited, A. L., Albano, A. M., Radigan, M., ... & Eaton Hoagwood, K. (2011). Statewide CBT training for clinicians and supervisors treating youth: the New York State evidence based treatment dissemination center. *Journal of Emotional and Behavioral Disorders, 19*(3), 182-192. DOI: 10.1177/1063426610367793.

Glipin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining Explanations: An Overview of Interpretability of Machine Learning. Retrieved from https://arxiv.org/pdf/1806.00069.pdf.

Glisson, C. & Schoenwald, S. K. (2005). The ARC organizational and community intervention strategy for implementing evidence-based children's mental health treatments. *Mental health services research, 7*(4), 243-259.

Gray, M. (2010). Moral sources and emergent ethical theories in social work. *The British Journal of Social Work, 40*(6), 1794-1811.

Green, B. & Chen, Y. (2019). Disparate Interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *FAT* '19, Proceedings of the Conference on Fairness, Accountability, and Transparency,* (pp. 90-99). New York: ACM. DOI: 10.1145/3287560.3287563.

Grgić-Hlača, N., Zafar, M.B., Gummadi, K.P., & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS '16 Proceedings of Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems*. Barcelona, Spain. Retrieved from https://people.mpisws.org/~gummadi/papers/process_fairness.pdf.

Grgić-Hlača, N., Zafar, M.B., Gummadi, K.P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Thirty Second AAAI Conference on Artificial Intelligence*. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16523/15670.

Griffith TL, Zammuto RF, Aiman-Smith L. (1999) Why new technologies fail: overcoming the invisibility of implementation. *Ind Manage, 41*, 29–34.

Grol, R. P., Bosch, M. C., Hulscher, M. E., Eccles, M. P., & Wensing, M. (2007). Planning and studying improvement in patient care: the use of theoretical perspectives. *The Milbank Quarterly, 85*(1), 93-138.

Hamilton, G. (1951). *Theory and practice of social casework*. Columbia University Press.

Hardt, M., Price E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *NIPS '16 Proceedings of the 30th Conference on Neural Information Processing Systems*, (pp. 3323-3331). Barcelona Spain.

Hart, S.D. (1998). The role of psychopathy in assessing risk for violence: conceptual and methodological issues. *Legal and Criminological Psychology, 3*. 121-137.

Hay, T. & Jones, L. (1994). Societal interventions to prevent child abuse and neglect. *Child Welfare, 73*(5), 379-403.

Healy, K. (2001). Reinventing critical social work: Challenges from practice, context and postmodernism. *Critical Social Work* 2(1). Retrieved from https://ojs.uwindsor.ca/index.php/csw/article/download/5618/4591?inline=1

Herrenkohl, R.C. (2005). The definition of child maltreatment: From case study to construct. *Child Abuse and Neglect, 29*(5), 413-424.

HM Government (2018). *Working together to safeguard children: A guide to inter-agency working to safeguard and promote the welfare of children*. Retrieved from https://

assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/779401/Working_Together_to_Safeguard-Children.pdf.

Hoagwood, K.E., Olin, S.S., Kerker, B.D., Kratochwill, T.R., Crowe, M., & Saka, N. (2007). Empirically based school interventions targeted at academic and mental health functioning. *Journal of Emotional and Behavioral Disorders, 15*, 66-92.

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. *arXiv:1805.03677*.

Holmes, L., McDermid, S., Padley, M., & Soper, J. (2010). *Exploration of the costs and impact of the Common Assessment Framework*. Research Report DFE-RR210, Department of Education. Retrieved from https://www.gov.uk/government/publications/exploration-of-the-costs-and-impact-of-the-common-assessment-framework

Holmes, L. & McDermid, S. (2012). *Understanding costs and outcomes in child welfare services: A comprehensive costing approach to managing your resources*. London: Jessica Kingsley Publishers.

House of Commons (2018). *Local Government Association briefing Westminster Hall debate on findings of the Care Crisis Review*. Retrieved from https://www.local.gov.uk/parliament/briefings-and-responses/westminster-hall-debate-findings-care-crisis-review-house

House of Commons (2019). *Children's social care in England*. Retrieved from https://researchbriefings.parliament.uk/ResearchBriefing/Summary/CDP-2018-0284#fullreport.

House of Lords, Select Committee on Artificial Intelligence (2019). *AI in the UK: ready, willing and able?*. House of Lords, Report of Session 2017-19, HL Paper 100. Retrieved from https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

Howe, D. (1992). Child abuse and the bureaucratisation of social work. *The Sociological Review, 40*(3), 491-508.

Hugman, R. (2005). Exploring the paradox of teaching ethics for social work practice. *Social Work Education: The International Journal,* (5), 535-545.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*. Institute of Electrical and Electronics Engineers (IEEE). Retrieved from https://ethicsinaction.ieee.org/.

Ife, J. (1997). *Rethinking social work: Towards critical practice*. Longman.

Information Commissioner's Office & The Alan Turing Institute (2019). *Explaining decisions made with AI: Draft guidance for consultation*. Retrieved from https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-and-the-turing-consultation-on-explaining-ai-decisions-guidance/

International Federation of Social Workers (2018). *Global social work statement of ethical principles*. Retrieved from https://www.ifsw.org/global-social-work-statement-of-ethical-principles/.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R*. London: Springer. Retrieved from http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Learning, 1,* 389 – 399.

Jolley, J.M. (2012). *Applying neural network models to predict recurrent maltreatment in child welfare cases with static and dynamic risk factors* (Doctoral dissertation). University of Washington in St. Louis.

Joseph, M. V. (1989). Social work ethics: Historical and contemporary perspectives. *Social Thought, 15* (3-4), 4-17.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. DOI: 10.1037/h0034747.

Kahneman, D. & Tversky, A. (1999). Evaluation by moments: Past and future. In D. Kahneman and A. Tversky (Eds.) *Choices, Values and Frames*, New York: Cambridge University Press and the Russell Sage Foundation. Retrieved from http://www.vwl.tuwien.ac.at/hanappi/TEI/momentsfull.pdf.

Kamiran, F. & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*, 1-33. DOI 10.1007/s10115-011-0463-8.

Katz, R.V., Green, B.L., Kressin, N.R., Kegeles, S.S., Wang, M.Q., James, S.A., Russell, S.L., Claudio, C., & McCallum, J.M. (2008). The legacy of the Tuskegee Syphilis Study: Assessing its impact on willingness to participate in biomedical studies. *Journal of Health Care for the Poor and Underserved, 19*(4), 1168-1180.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Scholkopf, B. (2017). Avoiding discrimination through causal reasoning. In *NIPS '17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, (pp. 656-666). USA: Curran Associates Inc.

Kim, B., Khanna, R., & Koyejo, O.O. (2016). Examples are not enough, learn to criticize! Criticism for Interpretability.

In *Proceedings from the 30th Conference on Neural Information Processing Systems (NIPS 2016).* Barcelona, Spain. Retrieved from https://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability.pdf.

Klauer, K.C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, *107* (4), 852–84. DOI: 10.1037/0033-295X.107.4.852.

Klein, K. J. & Knight, A. P. (2005). Innovation implementation: Overcoming the challenge. *Current Directions in Psychological Science, 14*(5), 243–246.

Klein, K. J. & Sorra, J. S. (1996). The challenge of innovation implementation. *Academy of Management Review, 21*(4), 1055–1081.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. Retrieved from https://arxiv.org/abs/1609.05807

Klingele, C. (2016). The promises and perils of evidence-based corrections. *Notre Dame Law Review*, *91*(2), 537-584. Retrieved from https://scholarship.law.nd.edu/ndlr/vol91/iss2/2/

Koepke, J.L. & Robinson, D.G. (2018). Danger ahead: Risk assessment and the future of bail reform. *Washington Law Review, 93*, 1725 – 1807. Retrieved from https://digitalcommons.law.uw.edu/wlr/vol93/iss4/4/

Kohl, P.L., Jonson-Reid, M. & Drake, B. (2009). Time to leave substantiation behind: Findings from a national probability study. *Child Maltreatment*, *14*(1), 17-26. DOI: 10.1177/1077559508326030.

Kuhse, H. & Singer, P. (1998). *A companion to bioethics*. (2nd ed.) Wiley-Blackwell.

Kusner, M.J., Loftus, J., Russel, C., & Silva, R. (2017). Counterfactual fairness. In *NIPS '27 Proceedings of the 31st International Conference on Neural Information Processing Systems*, (pp. 4069-4079). Long Beach, California, USA. Retrieved from https://papers.nips.cc/paper/6995-counterfactual-fairness.pdf.

La Valle, I., Hart, D., Holmes, L., & Pinto, V.S. (2019). *How do we know if children's social care services make a difference? Development of an outcomes framework*. UK: Nuffield Foundation, Rees Centre, University of Oxford. Retrieved from http://www.education.ox.ac.uk/wp-content/uploads/2019/07/CSCS-Outcomes-Framework-July-2019.pdf.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 275-284). Retrieved from https://cs.stanford.edu/~jure/pubs/contraction-kdd17.pdf.

Lang, J. M., Franks, R. P., Epstein, C., Stover, C., & Oliver, J. A. (2015). Statewide dissemination of an evidence-based practice using Breakthrough Series Collaboratives. *Children and Youth Services Review*, 55, 201–209.

Latonero, M. (2018). *Governing artificial intelligence: Upholding human rights & dignity*. Data & Society. Retrieved from https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

Lee, J. D. & See, J. (2004). Trust in automation and technology: Designing for appropriate reliance. *Human Factors, 46*, 50–80.

Lee, M.K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 1-16. DOI: 10.1177/2053951718756684.

Leeman, J., Calancie, L., Hartman, M. A., Escoffery, C. T., Herrmann, A. K., Tague, L. E., ... & Samuel-Hodge, C. (2015). What strategies are used to build practitioners' capacity to implement community-based interventions and are they effective?: a systematic review. *Implementation Science*, *10*(1), 80.

Lehr, D. & Ohm, P. (2017). Playing with the data: What legal scholars should learn about machine learning. *University of California, Davis, 51*, 653 – 717.

Leschied, A.W., Chiodo, D., Whitehead, P.C., Hurley, D., & Marshall, L. (2003). The empirical basis of risk assessment in child welfare: The accuracy of risk assessment and clinical judgement. *Child Welfare, 82*(5), 527-540.

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. London, UK: The Alan Turing Institute. Retrieved from https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

Levi, A. (2008). The ethics of nursing student international clinical experiences. *Journal of Obstetric, Gynecologic & Neonatal Nursing, 38*(1), 94-99.

Local Government Association (2017). *LGA budget submission: Autumn 2017*. Retrieved from https://www.local.gov.uk/parliament/briefings-and-responses/lga-autumn-budget-submission-2017

Logg, J., Minson, J. & Moore, D.A. (2018). Algorithm appreciation: People prefer algorithmic to human judgment. *Harvard Business School NOM Unit Working Paper No. 17-086*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2941774.

Loewenberg, F.M. & Dolgoff, R. (1982). *Ethical Decisions for Social Work Practice*. F.E. Peacock Publishers.

Lonne, B., Harris, M., Featherstone, B., & Gray, M. (2016). *Working ethically in child protection*. Taylor and Francis Inc.

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 150-158). New York: ACM. DOI: 10.1145/2339530.2339556.

Luna, F. & Macklin, R. (1998). Research involving human beings. In H. Kuhse & P. Singer (Eds.), *A Companion to Bioethics* (2nd ed.) (pp. 457-468). Wiley-Blackwell.

Mancini, J. A., & Marek, L. I. (2004). Sustaining community-based programs for families: conceptualization and measurement. *Family Relations*, *53*(4), 339-347.

Mapp, S.C. (2008). *Human rights and social justice in a global perspective: An introduction to international social work*. Oxford University Press.

Margetts, H. & Dorobantu, C. (2019). Rethink government with AI. *Nature*. Retrieved from https://www.nature.com/articles/d41586-019-01099-5.

Marion-Young, I. (1990). *Justice and the politics of difference*. Princeton University Press.

May, C. (2006). Mobilising modern facts: health technology assessment and the politics of evidence. *Sociology of health & illness*, *28*(5), 513-532.

May, C. & Finch, T. (2009) Implementation, embedding, and integration: An outline of Normalization Process Theory. *Sociology, 43*, 535-554.

May, C. R., Mair, F. S., Dowrick, C. F., & Finch, T. L. (2007). Process evaluation for complex interventions in primary care: understanding trials using the normalization process model. *BMC Family Practice*, *8*(1), 42.

May, C. R., Mair, F., Finch, T., MacFarlane, A., Dowrick, C., Treweek, S., ... Murray, E. (2009). Development of a theory of implementation and integration: Normalization Process Theory. *Implementation Science, 4*(1), 29.

McGuirl, J. M. & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, *48*, 656–665. Retrieved from https://bib.dbvis.de/uploadedFiles/uncertainty_trust.pdf.

McDermid, S. (2008). The nature and availability of child level data on children in need for use by Children's Services practitioners and managers. *Research Policy and Planning, 26*(3), 183-192.

Meagher, G. & Parton, N. (2004). Modernising social work and the ethics of care. *Social Work & Society: International Online Journal, 2*(1), 10-27.

Mendel, P., Meredith, L. S., Schoenbaum, M., Sherbourne, C. D., & Wells, K. B. (2008). Interventions in organizational and community context: a framework for building evidence on dissemination and implementation in health services research. *Administration and Policy in Mental Health and Mental Health Services Research*, *35*(1-2), 21-37.

Mercado, J.E., Rupp, M.A., Chen, J.Y.C., Barnes, M.J., Barer, D., & Procci, K. (2016). Intelligent agent transparency in human-agent teaming for multi-UxV management. *Human Factors*, *58* (3), 401-415.

Metz, A. & Albers, B. (2014). What does it take? How federal initiatives can support the implementation of evidence-based programs to improve outcomes for adolescents. *Journal of Adolescent Health*, *54*(3), S92-S96.

Ministry of Housing, Communities and Local Government (2019). *National evaluation of the Troubled Families Programme 2015 – 2020: Family outcomes – national and local datasets, Part 4*. London: Ministry of Housing, Communities and Local Government. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/786891/National_evaluation_of_the_Troubled_Families_Programme_2015_to_2020_family_outcomes___national_and_local_datasets_part_4.pdf

Mittelstadt, B., Russel, C., & Wachter, S. (2019). Explaining explanations in AI. In *FAT* '19 Proceedings of the Conference on Fairness, Accountability, and Transparency*, (pp. 279 – 288). New York: ACM. DOI: 10.1145/3287560.3287574.

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable.* Retrieved from https://christophm.github.io/interpretable-ml-book/.

Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics*, *31*, 175–178.

Mosier, K. L. & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and application*, (pp. 201–220). Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates, Inc.

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision-making and performance in hightech cockpits. *International Journal of Aviation Psychology*, *8*, 47–63.

Moullin, J. C., Dickson, K. S., Stadnick, N. A., Rabin, B., & Aarons, G. A. (2019). Systematic review of the exploration, preparation, implementation, sustainment (EPIS) framework. *Implementation Science*, *14*(1), 1.

Muńoz, C., Smith, M., & Patil, D.J. (2016). *Big data: A*

report on algorithmic systems, opportunity, and civil rights. Executive Office of the President. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

Munro, E. (1999). Common errors of reasoning in child protection work. *Child Abuse and Neglect, 23*(8), 745-758. Retrieved from http://eprints.lse.ac.uk/358/.

Munro, E. (2011). *The Munro review of child protection: Final report – A child-centred system*. Department for Education. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/175391/Munro-Review.pdf

Munro, E. (2019). Predictive analytics in child protection. *Knowledge for Use – K4U, CHESS Working Paper*. Retrieved from https://www.researchgate.net/publication/332528200_Predictive_analytics_in_child_protection.

Murphy, K. (2012). *Machine learning: A probabilistic perspective*. London: The MIT Press.

Murray, E., Treweek, S., Pope, C., MacFarlane, A., Ballini, L., Dowrick, C., ... & Ong, B. N. (2010). Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. *BMC medicine*, *8*(1), 63.

Murray, L. K., Dorsey, S., Skavenski, S., Kasoma, M., Imasiku, M., Bolton, P., ... Cohen, J.A. (2013). Identification, modification, and implementation of an evidence-based psychotherapy for children in a low-income country: The use of TF-CBT in Zambia. *International Journal of Mental Health Systems, 7*(1), 24.

Murray, L. K., Skavenski, S., Michalopoulos, L. M., Bolton, P. A., Bass, J. K., Familiar, I., ... & Cohen, J. (2014). Counselor and client perspectives of trauma-focused cognitive behavioral therapy for children in Zambia: A qualitative study. *Journal of Clinical Child & Adolescent Psychology*, *43*(6), 902-914.

Nadeem, E. & Ringle, V. A. (2016). De-adoption of an evidence-based trauma intervention in schools: A retrospective report from an urban school district. *School Mental Health, 8*(1), 132–143. DOI: 10.1007/s12310-016-9179-y

Nadeem, E., Jaycox, L. H., Kataoka, S. H., Langley, A. K., & Stein, B. D. (2011). Going to scale: Experiences implementing a schoolbased trauma intervention. *School Psychology Review, 40*(4), 549–568.

Nash, M. (2017). *Examination of using structured decision-making and predictive analytics in assessing safety and risk in child welfare*. Los Angeles, CA: Los Angeles County Office of Child Protection. Retrieved from http://file.lacounty.gov/SDSInter/bos/bc/1023048_05.04.17OCPReportonRiskAssessmentTools_SDMandPredictiveAnalytics_.pdf.

National Association of Social Workers (2013). *NASW standards for social work practice in child welfare*. Retrieved from https://www.socialworkers.org/LinkClick.aspx?fileticket=zV1G_96nWoI%3D&portalid=0

National Association of Social Workers (2017). *NASW code of ethics*. Retrieved from https://www.socialworkers.org/About/Ethics/Code-of-Ethics/Code-of-Ethics-English.

Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implementation science*, *10*(1), 53.

Northern Ireland Social Care Council (2003). *The national occupational standards for social work.* Retrieved from https://www.basw.co.uk/resources/national-occupational-standards-social-work

NSPCC (2018). *Issues to consider when looking at child abuse statistics*. Retrieved from https://learning.nspcc.org.uk/media/1310/issues-consider-looking-child-abuse-statistics.pdf.

O'Grady, N. (2015). A politics of redeployment: Malleable technologies and the localisation of anticipatory calculation. In L. Amoore 7 V. Piotukh (Eds.), *Algorithmic Life*, (pp. 86-100). Oxford: Routledge.

O'Neil, C. (2016). *Weapons of math destruction; How big data increases inequality and threatens democracy*. USA: Crown.

Office for National Statistics (2019). *User guide to crime statistics for England and Wales*. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/methodologies/userguidetocrimestatisticsforenglandandwales

Orme, J. (2008). Feminist social work. In M. Gray and S.A. Webb (Eds), *Social work theories and methods* (pp. 65–75). London: Sage.

Otway, O. (1996). Social work with children and families: From child welfare to child protection. In N. Parton (Ed.) *Social Theory, Social Change and Social Work* (pp. 152 - 171). Oxford: Routledge.

Parasuraman, R. & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52* (3), 381-410.

Parton, N. (1998). Risk, advanced liberalism and child welfare: The need to rediscover uncertainty and ambiguity. *British Journal of Social Work*, *28*(1), 5-27. DOI: /10.1093/oxfordjournals.bjsw.a011317.

Parton, N. (2003). Rethinking professional practice: The contributions of social constructionism and the feminist 'ethics of care'. *The British Journal of Social Work, 33*(1),

1-16.

Pedreschi, D., Gianotti, F., Guidotti, R., Monreale, A., Pappalardo, L., Ruggieri, S., & Turini, F. (2018). Open the black box: Data-driven explanation of black box decision systems. Retrieved from https://arxiv.org/abs/1806.09936

Pelton, L.H. (1989). *For reasons of poverty: A critical analysis of the public child welfare system in the United States.* New York: Praeger Publishers.

Powell, B. J., Patel, S. V., Haley, A. D., Haines, E. R., Knocke, K. E., Chandler, S., ... & Aarons, G. A. (2019). Determinants of implementing evidence-based trauma-focused

interventions for children and youth: A systematic review. *Administration and Policy in Mental Health and Mental Health Services Research*, 1-15. DOI: 10.1007/s10488-019-01003-3

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., ... & Hensley, M. (2011). Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*(2), 65-76.

Reamer, F.G. (1985). The emergence of bioethics in social work. *Health & Social Work,* 10, 271-281.

Reamer, F.G. (2014). The evolution of social work ethics: Bearing witness. *Advances in Social Work, 15*(1), 163-181.

Reder, P., Duncan, S. & Gray, M. (1993). *Beyond blame: Child abuse tragedies revisited*. London: Routledge.

Reverby, S.M. (2009). *Examining Tuskegee: The infamous syphilis study and its legacy*. The University of North Carolina Press.

Rhodes, M.L. (1986). *Ethical dilemmas in social work practice*. Routledge & Kegan Paul.

Rizzuto, T. E., & Reeves, J. (2007). A multidisciplinary meta-analysis of human barriers to technology implementation. *Consulting Psychology Journal: Practice and Research*, *59*(3), 226.

Roberts, Y.H., O'Brien, K., & Pecora, P.J. (2018). *Safe, strong, supportive: Considerations for implementing predictive analytics in child welfare*. Casey Family Programs. Retrieved from https://caseyfamilypro-wpengine.netdna-ssl.com/media/Considerations-for-Applying-Predictive-Analytics-in-Child-Welfare.pdf.

Rose, S.J., & Meezan, W. (1996). Variations in perceptions of child neglect. *Child Welfare, 75*(2), 139-160.

Rovira, E., McGarry, K. & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*, 76–87.

Rudin, C. & Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *INFORMS Journal on Applied Analytics, 48*(5), 399-486, C3. DOI: 10.1287/inte.2018.0957.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Learning* (1), 206-215. DOI: 10.1038/s42256-019-0048-x.

Ruscio, J. (1998). Information integration in child welfare cases: An introduction to statistical decision making. *Child Welfare, 3*(2), 143-156.

Sabalauskas, K. L., Ortolani, C. L., & McCall, M. J. (2014). Moving from pathology to possibility: Integrating strengths-based interventions in child welfare provision. *Child care in practice*, *20*(1), 120-134.

Sacha, D., Senaratne, H., Kwon, B.C., Ellis, G. & Keim, D.A. (2016). The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, *22* (1), 240-249.

Sarter, N. B. & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, *43*, 573–583.

Saxena, N.A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C., & Liu, Y. (2019). How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *AIES '19 Proceedings of the 2019 AAAI / ACM Conference on AI, Ethics, and Society,* (pp. 99-106). New York: ACM. Retrieved from https://econcs.seas.harvard.edu/files/econcs/files/saxena_ai19.pdf.

Schofield T.J., Lee R.D., & Merrick M.T. (2013). Safe, stable, nurturing relationships as a moderator of intergenerational continuity of child maltreatment: A meta-analysis. *Journal of Adolescent Health, 53*(4 Suppl). S32–8. DOI: 10.1016/j.jadohealth.2013.05.004

Schwartz, I.M., York, P., Nowakowski-Sims, E., & Ramos-Hernandez, A. (2017). Predictive and prescriptive analytics, machine learning and child welfare risk assessment: The Broward County experience. *Children and Youth Services Review 81*(C), 309-302.

Sebba, J., Luke, N., McNeish, D., & Rees, A. (2017). *Children's Social Care Innovation Programme: Final evaluation report*. Department for Education. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/659110/Children_s_Social_Care_Innovation_Programme_-_Final_evaluation_report.pdf

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory and Cognition*, *21*(4), 546–556. Retrieved from https://link.springer.com/content/pdf/10.3758%2FBF03197186.pdf.

Sigel, B. A., Kramer, T. L., Conners-Burrow, N. A., Church, J. K., Worley, K. B., & Mitrani, N. A. (2013). Statewide dissemination of trauma-focused cognitive-behavioral therapy (TF-CBT). *Children and Youth Services Review, 35*(6), 1023–1029. DOI: 10.1016/j.child youth.2013.03.012.

Simon, H. (1997). *Administrative behaviour: A study of decision-making processes in administrative organisations.* New York: The Free Press.

Skegg, A. (2005). Brief note: Human rights and social work: A western imposition or empowerment to the people?. *International Social Work, 48*(5), 667 – 672.

Sledjeski, E. M., Dierker, L. C., Brigham, R., & Breslin, E. (2008). The use of risk assessment to predict recurrent maltreatment: A classification and regression tree analysis (CART). *Prevention Science*, *9*(1), 28–37. DOI: 10.1007/s11121-007-0079-0.

Stevenson, O. (2007) *Neglected Children and their Families* (2nd ed.). Oxford: Blackwell Publishing.

Stewart, A. & Thompson, C. (2004). *Comparative evaluation of child protection assessment tools*. Queensland: Griffith University.

Stone, B. (1998). Child neglect: Practitioners' perspectives. *Child Abuse Review*, *7*(2), 87–96.

Straus, M.A. & Kantor, G.K. (2005). Definition and measurement of neglectful behavior: Some principles and guidelines. *Child abuse & neglect, 29*(1), 19-29. DOI: 10.1016/j.chiabu.2004.08.005.

Strifler, L., Cardoso, R., McGowan, J., Cogo, E., Nincic, V., Khan, P. A., ... & Treister, V. (2018). Scoping review identifies significant number of knowledge translation theories, models, and frameworks with limited use. *Journal of clinical epidemiology*, *100*, 92-102.

Striphas, T. (2015). Algorithmic culture. *European Journal of Cultural Studies, 18*(4-5), 395 – 412. DOI: 10.1177/1367549415577392.

Surden, H. (2014). Machine learning and law. *Washington Law Review, 89*, 87-115.

Tabak, R. G., Khoong, E. C., Chambers, D. A., & Brownson, R. C. (2012). Bridging research and practice: models for dissemination and implementation research. *American Journal of Preventive Medicine*, *43*(3), 337-350.

Taber, C. S. & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50* (3), 755–69.

Taylor, C. & White, S. (2001). Knowledge, truth and reflexivity: The problem of judgement in social work. *Journal of Social Work, 1*(1), 37-59.

The Allegheny County Department of Human Services (2017). *Data brief: Racial disproportionality in Allegheny County's child welfare system*. Allegheny County. Retrieved from https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/10/ACDHS_CYF-Race-Disproportionality-Brief_100217-final.pdf

The Royal Society (2017). *Machine learning: the power and promise of computers that learn by example*. Report, The Royal Society. Retrieved from https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

The Royal Society (2019). *Explainable AI: The basics* (Policy Briefing). Retrieved from https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

The Policy, Ethics, and Human Rights Committee (2012). *The code of ethics for social work: Statement of principles*. British Association of Social Workers (BASW). Retrieved from http://cdn.basw.co.uk/upload/basw_112315-7.pdf.

Trankell, A. (1972). *Reliability of evidence: Methods for analyzing and assessing witness statements*. Oxford: Beckmans.

Tronto, J. (1993). *Moral boundaries: A political argument for an ethic of care.* London: Routledge.

Tupper, A., Broad, R., Emanuel, N., Hollingsworth, A., Hume, S., Larkin, C., Ter Meer, J., & Sanders, M. (2016). *Decision-making in children's social care: Quantitative data analysis*. Behavioural Insights Team, Department for Education. Retrieved from https://www.bi.team/wp-content/uploads/2016/07/Social_Worker_Decision_Making__BIT.pdf.

Tupper, A., Broad, R., Emanuel, N., Hollingsworth, A., Hume, S., Larkin, C., Ter Meer, J., & Sanders, M. (2017). *Decision-making in children's social care – Quantitative data analysis.* Behavioural Insights Team, Department for Education. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/623394/Social_Worker_Decision_Making_v7.12.pdf.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185* (4157), 1124–1131. DOI: 10.1126/science.185.4157.1124. Retrieved from https://science.sciencemag.org/content/185/4157/1124.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211* (4481), 453–58. Retrieved from https://science.sciencemag.org/content/211/4481/453

Université de Montréal (2017). *Montreal declaration for a responsible development of artificial intelligence*. Retrieved from https://www.montrealdeclaration-responsibleai.com/.

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *FAT* '19 Proceedings of the Conference on Fairness, Accountability, and Transparency*, (pp. 10-19). New York: ACM.

Vaithianathan, R., Putnam-Hornstein, E., Jiang, N., Nand, P., & Maloney, T. (2017). *Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation*. New Zealand: AUT University, Centre for Social Data Analytics.

Vaithianathan, R. (2017). *Five lessons for implementing predictive analytics in child welfare*. The Chronicle of Social Change. Retrieved from https://chronicleofsocialchange.org/opinion/five-lessons-implementing-predictive-analytics-child-welfare/27870

Van Dijk, J. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Veale, M. & Brass, I. (2019). Administration by algorithm? Public management meets public sector machine learning. In K. Yeung & M. Lodge (Eds.), *Algorithmic Regulation* (pp. 121-149). Oxford: Oxford University Press.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76-99. DOI: 10.1093/idpl/ipx005.

Waegeman, W., Dembczyński, K., & Hüllermeier, E. (2018). Multi-target prediction: A unifying view on problems and methods. Retrieved from https://arxiv.org/abs/1809.02352

Walsh, M. C., Joyce, S., Maloney, T., & Vaithianathan, R. (2020). Exploring the protective factors of children and families identified at highest risk of adverse childhood

experiences by a predictive risk model: An analysis of the growing up in New Zealand cohort. *Children and Youth Services Review*, *108*. DOI: /10.1016/j.childyouth.2019.104556

Ward, H., Holmes, L., & Soper, J. (2008). *Costs and consequences of placing children in care.* London: Jessica Kingsley Publishers.

Watkins, S., Jonsson-Funk M., Brookhart, M.A., Rosenberg, S.A., O'Shea T.M., & Daniels, J. (2013). An empirical comparison of tree-based methods for propensity score estimation. *Health Services Research, 48*(5), 1798-1817. DOI: 10.1111/1475-6773.

Weiner, B. J. (2009). A theory of organizational readiness for change. *Implementation Science, 19*(4), 67.

Welbourne, P. (2002). Culture, children's rights and child protection. *Child Abuse Review, 11*(6), 345-358. DOI: 10.1002/car.772.

Wenocur, K., Parkinson-Sidorski, M., & Snyder, S. (2016). Provision of child trauma services in emergency family housing (practice note). *Families in Society-the Journal of Contemporary Social Services*, *97*(3), 253–258.

What Works for Children's Social Care (2018). *Outcomes framework*. Retrieved from https://whatworks-csc.org.uk/research/outcomes-framework-for-research/.

White, A. & Walsh, P. (2006). *An issues paper: Risk assessment in child welfare*. NSW Department of Community Services. Retrieved from http://www.community.nsw.gov.au/__data/assets/pdf_file/0005/321647/research_riskassessment.pdf

Williams, L. & Monroe, K. (2017). *Using predictive analytics to identify high risk child welfare cases*. Ohio Department of Job and Family Services, Ohio Casa, Celebrate Kids! Conference, September 20, 2017.

Wilkinson, R. & Pickett, K. (2009). *The spirit level: Why more equal societies almost always do better.* London: Allen Lane.

Yoo, J., Brooks, D., & Patti, R. (2007). Organizational constructs as predictors of effectiveness in child welfare interventions. *Child Welfare, 86*(1), 53–78.

Zafar, M.B., Valera, I., Rodriguez, M.G., & Gummadi, K.P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW '17 Proceedings of the 26th International Conference on World Wide Web*, (pp. 1171-1180). Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Zahra, S. A. & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. *Academy of management review*, *27*(2), 185-203.

Zarsky, T. Z. (2017). An analytic challenge: Discrimination theory in the age of predictive analytics. *ISJLP*, *14*, 11.

Zemel, R., Wu, Y, Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, Proceedings of Machine Learning Research 28*(3), 325-333.

Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking Artificial Intelligence Principles. Retrieved from https://arxiv.org/abs/1812.04814.

Ziewitz, M. (2016). Governing algorithms: Myth, mess, and methods. *Science, Technology, & Human Values, 41*(1), 3-16. DOI: 10.1177/0162243915608948

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology, 30*(1), 75-89. DOI: 10.1057/jit.2015.5.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* New York: Public Affairs.

What Works *for* Children's Social Care

wwccsc@nesta.org.uk
@whatworksCSC
whatworks-csc.org.uk

Alliance for Useful Evidence | nesta | scie | CASCADE | Cardiff University Prifysgol Caerdydd | What Works Network