Draft Ethical Report Concerning Predictive Modelling in the Allegheny Babies and Families Project

Michael Veale

Draft April 11, 2019

Contents

Introduction **2**

Short System Overview **4**

Proportionality of Targeting **5**

Proportionality of the Control Trial **7**

Data Minimisation **9**

Deployment and Function Creep **11**

Within the County 11

Beyond the County 13

Useful Transparency **15**

Holistic Evaluation and Maintenance **18**

Supporting Decision Support **21**

Introduction

Public bodies and agencies have been increasingly seeking to use new forms of data analysis to provide 'better' public services. These reforms have come in many guises: digital service transformations such as 'e-government 2.0' and the creation of 'integrated data infrastructures' (linked administrative datasets),[1] broadly aimed at 'improving the experience of the citizen', 'making government more efficient' and 'boosting business and the wider economy'.[2]

Some systems in this vein can be considered as attempting to *automate* administrative decisions—for example, streamlining the act of applying for services or documents which are relatively straightforward to provide. While there will likely always be edge-cases which require human oversight or intervention—hence rules around the world on human intervention in automated decisions[3]—in gen-

[1] Statistics New Zealand, "Integrated Data Infrastructure" (*Government of New Zealand*, 2016) ⟨https://perma.cc/9RXL-SV7P⟩ visited on October 4, 2018.

[2] John Manzoni, "Big data in government: the challenges and opportunities" (*GOVUK*, February 2017) ⟨https://perma.cc/GF7B-5A2R⟩ visited on October 4, 2018.

[3] See eg Lilian Edwards and Michael Veale, "Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?" (2018) 16(3) IEEE Security & Privacy 46 DOI: 10/gdz29v.

eral these operate best in uncontentious, well-defined domains.[4]

Other systems, such as those which are the topic of this report, attempt to *augment* traditional decision-making, such as administrative decision-making, with additional evidence which improves factors such as its efficacy or consistency. When considering individual decisions over citizens (rather than rule- and policy-making over phenomena of social concern), these systems usually become entwined in existing processes of clinical judgement. The role of clinical judgement—assessment by an individual such as a social worker, who determines a course of action—has many roles in a service delivery context, such as helping individuals navigate labyrithine processes, getting feedback on how as system works to improve it, ensuring unusual, mitigating or difficult to measure circumstances are substantially assessed and considered, or simply enabling face-to-face contact. Frontline workers providing clinical judgement also hold important tacit knowledge about complex policy problems which can be difficult to codify in rule-based systems, and the procedures they are faced with often have substantial issues and grey areas which leaves them having to make calls within the grey-zones of the rules, at times even acting as 'street-level ministers', heavily interpreting the rules in areas which are not clear.[5]

At the same time, more 'actuarial' judgement, which many will be familiar with through terms such as machine learning or predictive modelling, *can* bring strong benefits to a decision system. Statistical modelling is in general more 'accurate' than humans at challenging prediction tasks,[6] although accuracy is by far from the only relevant measure of performance of a system more broadly. Frontline individuals can exhibit undesirable biases or heuristics,[7] both in an 'irrational' psychological sense of mental shortcuts as well as in the broader societal senses of prejudice and discrimination. This is *not* to say that issues similar to this are not present in computing systems used for decision-support, as will be discussed, but that if a societal aim is to mitigate these issues and create a system which is both procedurally and substantively fair to the humans faced with it, neither clinical and actuarial systems should be idolised or romanticised, and opportunities for them to genuinely work together and reinforce each other's weaknesses should be sought.

This report forms one out of three perspectives from cross-disciplinary researchers looking at different aspects of the ethics of the risk-scoring system Allegheny County plans to deploy. My perspective in this third of the broader ethical review is as a researcher in technology policy surrounding the social impacts of data-driven predictive systems, particularly those in the public sector. This report therefore *does not* focus on issues such as the comparative efficacy of this mode of child and family protection compared to other evidence-based interventions, but particularly and more narrowly considers the

[4] Compare to the notion of 'structured problem' in Robert Hoppe, *The Governance of Problems: Puzzling, Powering and Participation* (Policy Press 2010).

[5] Michael Lipsky, *Street-level bureaucracy: Dilemmas of the individual in public services* (Russell Sage Foundation 2010).

[6] William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson, "Clinical versus mechanical prediction: a meta-analysis" (2000) 12(1) Psychological Assessment.

[7] Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases" (1974) 185(4157) Science 1124.

role of predictive modelling within the entire scheme.

Short System Overview

What follows is a brief lay summary of the system Allegheny County is proposing that this ethics report concerns.

Allegheny County (pop. 1.2 million) has an integrated Department of Human Services which provides an array of services aimed ensuring, among other things, the health, safety and well-being of children and families in early life. A new proposed programme, provisionally named, Allegheny Babies and Families (ABF), seeks to provide i) universal 'light touch' outreach to new mothers through a hospital-based community health worker (CHW); and ii) targeted outreach for 'priority families' determined through a validated statistical prevention model delivered in collaboration with research institutes including Auckland University of Technology in New Zealand. This second service is spurred by the recognition that 'the people who most need these services are not using them':[8] that prior to a crisis such as home removal (of a child), a significant proportion of families are not participating in the County's supportive services. Existing approaches have focussed on geographic outreach, seeing a major barrier as transportation to one of the County's 28 Family Support Centers, however geography is a far-from-perfect proxy for maltreatment, as poverty is not a direct cause of abuse, most people in poverty do not abuse or neglect children, and many abused or neglected children come from homes with moderate or high levels of income. The County's own research illustrated the socioeconomic span of cases of maltreatment, and reviews of the research point to other factors for understanding causes of child abuse, including untreated maternal depression or mental illness, substance abuse in caregivers, inexperienced/young mothers, and intimate partner violence. Consequently, the County wishes to create and valdidate a statistical 'prevention model' to target, initially, a subset of the 6,000 mothers who give birth anually at Magee Womens Hospital—the facility with the most births in the County. This model will be used to select families with the highest need for support, and reach out to them with an offer of voluntary services delivered through a range of partners and civil society organisations in a proactive manenr, as opposed to waiting for a crisis or for them to locate these resources. 240 families each year will be offered this targeted program service. As part of a commitment to evidence-based policy, Allegheny County wishes to rigorously measure the effectiveness of this programme, and as such, plans to integrate a randomised control trial into the above approach.

[8] Internal memo from Amy Malen on Allegheny Babies and Families.

Proportionality of Targeting

Historical analysis undertaken by Allegheny County has indicated that the services they provide to families are not being taken up by those who go on to experience a crisis such as a home removal. Prevention services triggered by referrals exist, in addition to a separate actuarial screening system, however approximately 65% of all child fatalities in the County, largely babies and toddlers, are never referred into the Department of Human Services before the fatal incident.

When considering the social impacts of a targeted solution, it is important to consider the alternatives. Some of these alternatives are things that the County could do today instead (or in addition to) the proposed course of action. Others are more ambitious, and would likely require a more significant amount of political will. All of them are important to consider however: even if the County is restricted in its actions, its attitude towards the trajectory of its ambitions concerning predictive technology is important, particularly given its vanguard role and consequent high profile in airing the opportunities and limitations of predictive technologies from operational experience.

Alternative approaches to targeting using machine learning. Machine learning and predictive systems are harmful when they are seen as a panacea for deep-rooted and complex problems. Particularly given their popularity, they risk obscuring or impeding thoughtful policy analysis and problem structuring, leading to solutions chasing problems rather than the other way around. In particular, while it is avoidable, the use of vendors' off-the-shelf models and systems can be problematic, as these systems can be insufficiently tailored for specific contexts, being sold in a similar form to achieve economies of scale to recoup development costs despite the highly varying nature of policy issues by region and sector, and can embody value-laden rules and decisions that should be subject to in-house development, notice and comment given their policy–like nature.[9]

Allegheny County has considered, and continues to invest in, alternative approaches seeking to rectify this challenge of non-uptake of services designed in part to reduce the risk of a crisis. These include placing service hubs in areas with high proportions of poverty and violence, where they can be strengthened by nearby civil society organisations. These, however, suffer from the challenge that using location and demographic as a proxy for service need is not wholly congruent with the data, which shows that demographic characteristics associated with cases of child maltreatment vary strongly across the County, with very noticeable differences in median income and race. Research findings instead support other factors, such as untreated maternal depression, substance abuse in caregivers, young mothers and intimate partner violence, which exist throughout the county.

[9] Danielle Keats Citron, "Technological due process" (2008) 85 Washington University Law Review 1249.

Recommendation 1 *Allegheny County should continue to vigorously look for and invest in other ways to reach out to families in need of support services that do not rely on algorithmic systems, to complement the strategies that do.*

Nature of the intervention. Interventions that are coercive or that could trigger coercive effect must be considered more carefully than those that do not. For example, the opening of an investigation into child welfare, even if only partially informed by an algorithm, involves an element of wielding the power of the state against a private citizen. The services offered by Allegheny County in this instance are *voluntary* in nature, and indeed are additional to awareness raising around universal services also planned as part of the same ABF scheme. This is positive from an ethical standpoint. However, there remain concerns around future uses of this score, as well as function creep (see below, in section Deployment and Function Creep). The County should ensure that all uses that area made of this score are supportive and voluntary in nature, rather than being used to trigger investigations.

There is already an argument that such a score, despite only being linked to voluntary services, might be indirectly linked to coercive powers, insofar as the increased proximity of the family to services provided by the County gives more scope for monitoring and for referral by a competent body, were that to be deemed appropriate. However, the same can be said of all protective services however delivered, and safeguards for appropriate referrals should be in place within those services (as they are already). Furthermore, the County should monitor the impact of these services on downstream referrals and increased data collection on these individuals over time to ensure that scores do not disproportionately result in increased surveillance.

Recommendation 2 *The County should pledge that this predictive system be only used to provide voluntary supportive services, rather than to start investigations or to directly inform coercive powers.*

Recommendation 3 *The County should monitor how scoring and service targeting affects the volume of data captured on these subset of individuals and on communities, and take appropriate measures to avoid these groups being disproportionately oversurveilled.*

Targeting as the only means to access services. The proposed plan uses the risk scoring model as the gatekeeper for the targeted service(s), with different services being offered in a graduated fashion at the highest level(s) of predicted risk to the child. A problem with this system is that the predictive system is the *only* gatekeeper in this context—there is no other proposed mechanism through which to access these advanced services. As a principle, to avoid or mitigate ethical issues resulting from unexpected failure modes or biases in the system presented (eg from those who only just moved into the County with no prior data record), and to both respect and safeguard families and communities who wish to avoid data collection, the predictive system should not be the *only* means of targeting

families for this scheme. This is not to say that an additional stream of eligibility needs to match the proposed mechanism in scope or scale: but an alternative means for screening-in families should exist. This might, for example, rely on the clinical judgement of relevant trained health workers. This is in line with the principle of *redundancy* in designing safety-critical systems, as well as in-line with the importance of qualitative methods and tacit knowledge in understanding and grappling with complex challenges.

Furthermore, the fact that certain services are only eligible for profiled individuals could present other problems. Knowing that individuals were receiving certain services would be sufficient information to reveal their minimum level of risk score on the predictive scale for reciept of that particular service. It might be that this doesn't present a problem, and that the way risk is perceived is not a negative one. There is a chance, however, that higher risk on the predictive scale brings some sort of stigma. Such stigma might be a reason for derision in the community, could foster and perpetuate exclusion, or even result in individuals refusing services for knowledge that it will effectively mean disclosing sensitive data. If it became known that the reasons for the highest scores were mostly related to issues such as health, substance abuse, or the like—highly private subjects for some—families choosing to receive these services may believe that they were disclosing this information to anyone who was aware that they were in receipt of them, and as a result might refuse these services.[10]

Recommendation 4 *Allegheny County should ensure that at least one other screening-in pathway for offering these targeted services exists. Such a pathway should have a primarily clinical element rather than solely relying on an actuarial model, and should aim to capture families whose cases and experiences might not be well quantified for procedural or substantive reasons.*

This report does not recommend the form that this alternative pathways should take, as it is not within the expertise of the author to advise on the details of such clinical processes. However, some considerations that align with analysis of the proposed evaluation methods below do point to some potential approaches (see section Proportionality of the Control Trial)

Proportionality of the Control Trial

One of the more controversial aspects of this scheme is the proposal to use a randomised control trial to assess the efficacy of the targeted services. The initial version of this proposal would see 240

[10] It should be noted that knowledge of the model might interact with the potential for stigma, and that equally transparency around models are encouraged in this report (see the section Useful Transparency). At the same time however, individuals often make 'folk theories' for systems regardless of whether they have knowledge of the innards or not, and therefore challenges such as the one hypothesised here are futile to avoid through secrecy. See e.g. Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik, "First I "Like" It, then I Hide It: Folk Theories of Social Feeds" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM 2016) DOI: 10.1145/2858036.2858494.

babies per year who the model indicated would meet the threshold for targeting of service offered the additional resources, in addition to the universal offering that is available to all mothers and families. 110 babies per year would be allocated to a control group, and would be treated the same as if they had not been selected for targeting (i.e. able to access core services).

This controvery is not unfamiliar to the medical field, and relates to the development of effective, potentially life-saving treatments, at the same time as assuring their efficacy in an evidence-based manner. The political and ethical challenges of observing or abstaining from an intervention when there is some reason to believe it could be effective is also not unfamiliar to the child protection field, with this being a reason for the shelving of a previous project in New Zealand which some of the researchers on this project had also worked on.[11]

It is important to ask whether there are more proportionate ways of achieving the aim of rigorously evaluating the impact of the targeted services, particularly in relation to the 110 babies a year who are determined to be high-need but who are not offered targeted services. This report will now reflect on some alterative options which may navigate these trade-offs with minimal impact on the rigour of a study, as evidence-based interventions area also of considerable ethical merit.

Option 1 Examine the opportunity of using data from hospitals elsewhere in the county which are not part of this scheme, treating the analysis as a *natural experiment*. Natural experiments are generally found where there are reasons that can be sufficiently classified as random which separate two populations that are otherwise the same. For example, postcode lotteries are a common form for studying natural experiments. Given that the focus of this project is on Magee Womens Hospital, it might be the case that suitable families who would be classified as high risk in a comparable way to those at Magee, but are within an institutitonal context not subject at all to the pilot trial, could be located. Their results could be compared after several years and any change examined. This approach might not be suitable however if the County has reason to believe that the populations admitted to hospitals other than Magee differ in important ways which are likely to act as confounders. It also suffers from the fact that it does not integrate with the other part of the scheme: the universal service offering, and therefore the comparison would be between families who are offered targeted services which include the universal service offering, and families who are offered the services in the manner the County currently operates at the other hospitals within its boundaries.

Option 1 might be ethically preferable to the proposed option given that no family in the other hospitals benefits from the ABF scheme (to my knowledge), and the infrastructure (e.g. the workers based in the hospital) to offer such a scheme is not present. It would have the drawback of limiting the options for families in these other hospitals to understand that their data were being processed for

[11] See Briefing from the Children's Action Plan to Anne Tolley, Minister for Social Development of New Zealand 'Vulnerable Children Predictive Modelling: design for Testing and Trialling' (7 November 2014).

research purposes, however this would be less important given that their data were not used for actionable decision-making at any point, and therefore could be defended as a public sector research use of data rather than data used for service delivery.

Option 2 Allow methods for families to opt-in to a separate assessment for targeted services regardless of whether they had been screened out or not. Concerns around the basis of excluding families from the heightened targeted services are discussed above in the section Proportionality of Targeting, and this is closely linked to Recommendation 4. This option aims to merge a method of tackling that with a method of tackling aspects of the trial issue above.

Given that both the universal and the targeted offerings are voluntary, the comparison group of most ethical salience is arguably the group of mothers who were determined to be of high need *and* offered and accepted targeted services, with the group of mothers who were determined to be of high need, not offered services due to the randomised control trial, and yet (hypothetically) *would* have accepted these targeted services had they been offered to her.

We might try to proxy this by making the assumption that the subset of mothers who would accept targeted services is within the subset who would accept (aspects of) the universal services. This might not be true, and it is worth looking at ways to validate this—there might be specific incentives in the targeted services that would attract mothers to choose them—but it may be a reasonable assumption to make. If this assumption is made, then the universal services become a potential venue for a clinical reassessment of whether targeted services (outside of the RCT) be offered. How this could be carried out is not for this report to suggest, and falls to those with different expertise. The impact of the magnitude of this reassessment might also require the power calculations in the original study plan to be recalculated to allow for the potential of 'drop-outs' from the control group. However, I believe this two-stage approach might enable trade-offs around proportionality to be made in a new way in relation to this project.

Data Minimisation

The data minimisation principle is a developed concept in law in different jurisdictions. One definition of it states that data should be 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.'[12]

In the case of this programme, a risk score is being developed, and individuals above or below that score can or cannot be targeted with services. At and above the targeting threshold, it may be useful to

[12] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1 (GDPR) 5(1)(c).

know the exact score, but **below the threshold, there appears to be minimum value in keeping or displaying it for *operational* purposes** (research purposes may differ, particularly concerning issues of model calibration).

As a result, it is recommended here to introduce a simple feature within the model software that takes the score calculated by the machine learning model deployed and looks to see if it it meets or exceeds a certain threshold. This threshold would, in the first instance, be the minimum score that allows eligibility for any service, which as this author understands it is currently 15. If it does, the score is retained in line with relevant standard operating procedures. If it does not, then a more generic message is retained, not revealing any details of the score, and the system only records that the user was not selected for targeting. The difference between these two conditions is due to the fact that targeting is planned *within* the services provided: i.e. those with a score 20 will be elgible for a programme of treatment that those with scores 15–19 will not. This also presents a light barrier against function creep (see below section), as miminising the results with respect to a certain purpose may make it more difficult to use downstream for other incompatible purposes.

Recommendation 5 *Scores should be redacted on production to 'Ineligible' if they fall below determined thresholds for service ineligibility, rather than unnecessarily keeping them in numeric form. An proportionate exemption to this could be determined if they were deemed necessary for validation or research, in which case these should be kept separately and redacted in the operational system.*

Once service delivery begins, it is unclear what role the score has. If no further use can be established, it is recommended that the score be purged entirely. In particular, this would also be useful because of the perception that such as score might be used as a 'black mark' on somebody's file—an argument which would be easier to refute if scores were deleted rather than retained for unclear or potential purposes.

Recommendation 6 *At the point at which there are no plans to make use of the value of a score through a standard operating procedure, the score should be purged.*

Data minimisation can also be considered in the sense of making the data coarser for any particular purpose. For example, if a score was desirable to retain to inform service providers that families who had previously refused or dropped out of schemes were a priority to re-engage, the question should be asked of what practical use retaining the difference between a 15 and a 20 would be on the re-engagement intervention. If the use was limited, then the score could be replaced with a marker that they were a priority to (re-)engage—a marker that could also be usefully triggered through other operating procedures.

Recommendation 7 *If there are justified cases for the score being retained for downstream service procision, tailoring or as flag for re-engagement, the score should be made as coarse as is possible and compatible with those purposes.*

Deployment and Function Creep

One underlying anxiety concerning predictive systems in the public sector is that by virtue of being created for one task, they establish an infrastructure consisting of many aspects—including data, technology, expertise and culture—which might expand beyond its original scope into areas its original democratic and societal mandate did not permit.

This section explores this from two starting points. The first is *within the County*, envisaging safeguards against creep in the current system by future officeholders who might want to deploy the system differently or share scores from the model more widely. The second starting point is *beyond the County*, considering the appetite for predictive systems in public bodies across the world more generally, and the pressure on a vanguard administration such as Allegheny County to provide software and models (if not data) to help them with their own challenges. Co-operation between public bodies on areas of technical difficulty is naturally welcome, however ethical questions do arise if the processes of the body to which the knowledge and software are being transferred to fall significantly below that of Allegheny County. Significant trust and reputation issues might follow given the provenance of the system.

Within the County

Allegheny County has, in previous work, given welcome consideration to delimiting of the role of the system in operating procedures within the welfare system. A previous system implemented by the County in 2016, the Allegheny Family Screning Tool (AFST),[13] provided call-screeners with ventiles (numbers 1–20 representing 5% percentiles of a probability distribution) estimating the risk that the child that is the subject of the call would be removed from home were they to be screened in. These inferred data were not shared beyond this stage of the process (e.g. to investigators downstream), which limits the role of these data as children who were screened-in manually or who were screened in on the recommendation of a high risk score are not distinguishable.

Some will be concerned that while that might be the policy today, it might not be robust to change in the future. Similarly, those who might have lost trust in a public service more generally might not trust assurances that this inferred data is deleted or not passed onto other actors in the system. From a technical standpoint, there are limitations to any formal guarantees that can be provided of this nature. Allegheny County could indeed delete the scores that were produced by this model, but the nature of the models used by the County to date have all been deterministic in nature, meaning that with a copy of the model at that time (which would be advisable to keep for purposes of auditing and

[13] See generally Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan, "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions" in *Conference on Fairness, Accountability and Transparency (FAT* 2018)* (2018).

accountability) and a copy of the data record in question (which is likely to persist and be expanded upon over time), such a score could be trivially recalculated.

The County may instead be able to provide some institutional and legal assurance of the path that they choose to go down. In many regimes around the world, data protection law or frameworks like it provide for the concept of *purpose limitation*.[14] While a similar regime may not be available in Pennsylvania law, similar constructs might be achieved through contracts or declarations which bind the County. For example, if such a predictive system were defined, some legally binding declaration could be made delimiting the purposes of this system in advance to a sufficiently narrow scope and set of actors. This agreement would then serve as a mechanism that could be used to hold future uses of this model to account—at least insofar as it would have to be actively and ideally publicly removed before the purposes of a score or a model could change. Organisations without such agreements have been heavily accused of function creep in other contexts, such as the use of data from the education sector in immigration enforcement in the United Kingdom.[15]

While it is not for this (non-US lawyer) to stipulate the legal form such an instrument would take, I can illustrate some of the characteristics that any suitable form should be tested against. It should be as binding as possible given its nature; it should allow for third-parties to challenge its enforcement; it should not be amendable without public declaration; and its core form should be public.

In the absence of an overarching framework governing the use of data and prediction in public services, the County should attempt to ensure that it remains proactively able to be held to account. Given the intentions of the current project this does not seem to place restrictions on what the system proposed intends to do, but instead acts as a guard against unwarranted function creep. In doing so, the County would also set an example for other public bodies and encourage them to do the same. This would be a collective benefit, as a single authority misusing scoring in a way that is seen as societally unacceptable is likely to tar trust across organisations seeking to use data and technology in proprortionate and safeguarded ways that are highly mindful of recent debates around the pitfalls of these practices.

Recommendation 8 *The County should examine options for creating a pledge in a binding and public manner to ensure that the intended purposes of the system do not expand without a clear and accountable process, and the safeguards do not diminish.*

This would sit alongside the existing initiatives for misuse in general that the County has in place, such as regular audits of the *Client View* portal for analysing the data of individuals.

[14] See eg GDPR, 5(1)(b). The notion of purpose and compatible purposes also arises in the California Consumer Privacy Act, 2018 Cal. Legis. Serv. Ch. 55 (A.B. 375).

[15] Damien Gayle, "Schools Census Used to Enforce Immigration Laws, Minister Says" (*The Guardian*, January 13, 2019) ⟨https://www.theguardian.com/politics/2019/jan/13/schools-census-used-for-immigration-enforcement-minister-says⟩ visited on April 10, 2019.

Given the significant powers that the County has, both in terms of being a safeguard for extremely vulnerable individuals as well as coercive powers of the state to e.g. remove children, it is noted that predictive modelling is by far from the most straight-forward method to do harm, were the organisation to find itself controlled by those with agendas that were not well-suited for addressing the difficult social challenges the Department for Human Services is faced with. That being said, computing systems can in many cases be seen as administrative rule-making systems,[16] and risk being used to disguise the true intent of policies that otherwise would be easier to uncover through tools such as FOIA. At the very least, they risk being *seen* as vehicles through which this could be done, even if they are not, and mechanisms of binding legal recourse are one way in which to build public trust, particularly in an environment where other agencies around the world may adopt less savoury uses of similar systems.

Beyond the County

As Allegheny County is one of the regions of the United States most advanced in this form of analysis, it is highly likely that other public organisations will seek to build on the experience the County has gained during building and deploying these systems. This itself brings ethical challenges.

The responsible downstream use of computationally-advanced statistical systems has been a subject of contention in a number of fields. General purpose technologies, like all similar tools, can be applied towards a wide variety of ends. It cannot be guaranteed that predictive tools such as those developed at Allegheny County will be redeployed elsewhere with comparable safeguards or ethical considerations. Not only might this harm individuals, families and communities, but the irresponsible deployment of technologies elsewhere might severely erode trust in more responsible deployments of similar technologies, as well as present serious reputational risk to the originators of these tools.

Some existing tools exist aiming to constrain the downstream use of software in society. This report does not seek to mandate any particular scheme, but recommends that Allegheny County consider feasible and effective safeguards within the legal regime and obligations they work under to

Licensing schemes. Licensing schemes attempt to legally constrain what can be done with code. While proprietary licensing is commonplace, if not the norm, in much commercial software, licensing has also been deployed to secure public value from developed code. Notable, the GNU General Public License (GPL) license enables software to be freely shared under the condition that any derivative products are also shared under the same license. This contrasts with permissive free software licenses, such as the MIT License, which places very few restrictions on downstream reuse.

[16] Citron (see n. 9); Michael Veale and Irina Brass, "Administration by Algorithm? Public Management meets Public Sector Machine Learning" in Karen Yeung and Martin Lodge (eds.), *Algorithmic Regulation* (Oxford University Press 2019).

Allegheny County, as emphasised elsewhere in this report (see the section Useful Transparency), should prioritise transparency measures aimed at empowered third parties. It should also ensure—using the limited leverage it has as the developer—that the deployment of similar systems elsewhere is accompanied by a similar minimal level of transparency.

One approach for this could be as follows. The County could release code associated with this project under licensing terms which allow open reuse, but which mandate the licensing of derived code under the same licenses as the original code was licensed under. In software licensing, this would be a sibling of *copyleft* licensing. For example, such a license could state that any organisation deploying such code towards a public function employ institutional methods to allow third parties to scrutinise its function, or more extremely, to release all code openly. *Copyleft* regimes have been criticised for extending these provisions to code which surrounds or interfaces with the code under license, insofar as they are distributed as an inextricable package.

It should be emphasised that this approach comes with particular limitations. If these downstream provisions are breached, the infringement by the party utilising the code is one of the *copyright* of the County. There exist an array of specific questions that follow from that, such as the enforceability of breaches in this contract by a third party.[17] There may be other legal approaches which could be used to this effect, and the County should consider exploring those.

Institutional programmes. Institutional programmes may also serve to support responsible dissemination of machine learning systems to other organisations. Allegheny County might wish to operate or accomodate employees from other public bodies being seconded in to learn more about how systems are deployed on the ground, and to pick up tacit knowledge. They may wish to establish high-quality documentation of both the technical deployment as well as the social processes involved across the organisation in relation to these systems.

Co-development of future systems. This is a more prospective arrangement that the County may wish to consider. In effect, this would result in consortia of two or more distinct governments facing comparable challenges joining their efforts and funding together to co-develop software systems, drawing on capacity that might be difficult for one alone. These consortia themselves would instigate knowledge sharing agreements for both technical and social processes surrounding the deployment of predictive and other systems, and further regions or governments wishing to benefit from the developed code and expertise would be able to join the consortium upon meeting certain determined membership conditions.

[17] See further Andrés Guadamuz, "Viral Contracts or Unenforceable Documents? Contractual Validity of Copyleft Licenses" (2004) 26(8) European Intellectual Property Review 331 ⟨https://ssrn.com/abstract=569101⟩.

Recommendation 9 *The County should produce and make public a strategy for ensuring that the system and their expertise is disseminated in the most responsible manner possible to other interested public sector bodies. This strategy may wish to consider legal instruments such as licensing schemes, and institutional arrangements such as secondment and collaboration agreements.*

Useful Transparency

Individuals should have a right to know why they were offered targeted services upon request. These 'algorithmic explanations' might promote perceptions of justice in the process as a whole,[18] and are commonly proposed,[19] and a version of them is recommended here.

Recommendation 10 *Allegheny County should ensure that easy-to-understand explanations of why an individual was offered targeted services are provided upon request. In documentation or during outreach, the opportunity to request this information should be actively presented.*

Furthermore, input data is important as a component of understanding systems—and in some cases, may be more important than providing access to the innards,[20] particularly when the type of variables that would be used in this kind of system might not be well-understood by citizens. Indeed, the use of data in this way would be a salient time to highlight to individuals that they can see copies of this data, and to explain the process through which that can be done. This is usually called a subject access request in laws around the world,[21] but in absence of it being present in a generic form in the United States, it would be recommended for Allegheny County to make particular effort to provide and publicise it in this case given the high-stakes nature of the subject matter.

Recommendation 11 *Allegheny County should ensure that individuals can request all information that relates to them used by the County in the creation of this score, and are actively informed about their ability to do this.*

[18] See Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt, "'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)* (ACM 2018) DOI: 10/cvcp.

[19] See e.g. their presence in international law in the Modernised Council of Europe Convention 108; in French administrative law, and in EU data protection law. See Mireille Hildebrandt, "The Dawn of a Critical Transparency Right for the Profiling Era" in Jacques Bus, Malcolm Crompton, Mireille Hildebrandt, and George Metakides (eds.), *Digital Enlightenment Yearbook* (IOS Press 2012) DOI: 10/cwpm; Edwards and Veale, "Enslaving the Algorithm" (see n. 3).

[20] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach, "Manipulating and Measuring Model Interpretability" [2018] arXiv:1802.07810 [cs] ⟨http://arxiv.org/abs/1802.07810⟩ visited on April 10, 2019.

[21] Jef Ausloos and Pierre Dewitte, "Shattering one-way mirrors—data subject access rights in practice" (2018) 8(1) International Data Privacy Law 4 DOI: 10/cwcf.

However, algorithmic explanations aimed at the individuals subject to algorithmic decisions should equally not be romanticised as a panacea or governance solution.[22] While explanations can be useful for a range of purposes, such as identifying errors or establishing trust, they face several limitations. Explanations burden decision-subjects with understanding and determining the nature of decisions that can affect them. Those who might need to challenge them most might be in vulnerable positions, and the least able to grapple with quite complex systems (and the bureaucracies or political economies behind them). This is especially the case in the situation in question, where women being approached for targeted services have just given birth, and can additionally be assumed to be in-need in variety of ways due to the fact they have been targeted by this system. Furthermore, many explanations can often be generated for each system, because of the non-linear nature of statistical systems. For those indivduals who were outliers in a dataset and might like an explanation most, it is not clear that the explanation will make sense or be of much use.[23]

It is therefore necessary to think of other forms of transparency that can augment individual explanations and reduce the burden on those affected by the decisions to oversee the system at the same time as experiencing it, and raising a young child. Primarily, in line with many movements around the world for open data and transparency on governmental systems, the County should aim to make the developed systems as open as possible. Allegheny County should consider publishing either full versions of its models, or parsimonious versions that estimate their core logics (eg a version of the more simple SLIM[24] model trained). Full versions of the model may not be possible or wise if such a model has sensitive variables that might leave it at risk of confidentiality attacks such as model inversion, which reveal a semblence of the data used to train it.[25] It is possible such publication could be unwise both reputationally and under laws such as HIPAA. However, many models, such as regression-based approaches, are generally resistant to model inversion attacks due to being destructive of data in the process of training for generalisable insights. Such systems might be published on a specific open data portal, or a code repository such as *GitHub*.

Recommendation 12 *Allegheny County should seek to publish version of the models it produces online, after ensuring they would not leak sensitive data.*

If a model cannot be released, Allegheny County should aim to release a 'model-centric explanation'[26]

[22] See Mike Ananny and Kate Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability" [2016] New Media & Society DOI: 10/gddxrg; Lilian Edwards and Michael Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not The Remedy You Are Looking For" (2017) 16 Duke L. & Tech. Rev. 18 DOI: 10/gdxthj.

[23] Edwards and Veale, "Slave to the Algorithm?" (see n. 22).

[24] Berk Ustun and Cynthia Rudin, "Supersparse Linear Integer Models for Optimized Medical Scoring Systems" (2016) 102(3) Machine Learning 349 DOI: 10/f8crhw.

[25] Michael Veale, Reuben Binns, and Lilian Edwards, "Algorithms That Remember: Model Inversion Attacks and Data Protection Law" (2018) 376 Phil. Trans. R. Soc. A 20180083 DOI: 10/gfc63m.

[26] Edwards and Veale, "Slave to the Algorithm?" (see n. 22).

or 'model card'[27] describing general features of the model. These would include summary statistics on model performance on various sub-groups and in various situations, a presentation of the core 'logics' extracted from the model (for example, random forest variable importance scores, despite their caveats,[28] can be useful in this situation). Such information should also include process-related elements such as data source, quality assurance, maintenance and re-training procedures, and the user interfaces it is deployed within. Ideally, such a model would also be presented in peer-reviewed work analysing it in context, as was undertaken with the AFST previously.[29]

Recommendation 13 *Allegheny County should publish and maintain 'model-centric' explanations online: metadata about the models including at least their main inputs, logics, and optimisation targets; performance in practice, includuing on sensitive and salient sub-groups; and practices and processes around model maintainance and use.*

The most promising approach for transparency is to aim it at empowered civil society, research or journalistic organisations, providing access to models. If models can be totally published, then these organsiations need only to be additionally access to the institutional and human infrastructures around them in the wider decision system. If they cannot be published, then other approaches might be possible. Sporadic physical arrangements for e.g. computational journalists to come and, through secure terminals and pre-installed software, examine data use, would be also be a possibility. Such arrangements are commonly used in statistical agencies for examining sensitive microdata, and inspiration on how that could be applied to models could be sought there—although there would be economic considerations.[30]

As the number of tools of this type grows within the County, it is worth the County considering what other scalable methods exist for oversight. As part of this report I have heard about the general strategy of openness towards both critics and those who want to understand more about the system by observing and studying it in practice. It would be worthwhile attempting to systematise efforts for internal review and oversight. Information days, either physically or online through webinars, might be useful in disseminating information more broadly. Other arrangments which are of a slower nature could be considered, such as joint-supervised PhD students with local universities, or projects for groups of postgraduates. Where research projects can lead to public, peer-reviewed outputs, this should be encouraged. While none of these presents a silver bullet for accountability and oversight, all of them contribute to general atmosphere of openness and collaboration that set tha bar for other agencies to meet.

[27] Margaret Mitchell et al., "Model Cards for Model Reporting" in *Proceedings of the 2nd ACM Conference on Fairness, Accountability and Transparency (ACM FAT* 2019)* (ACM 2019) ⟨https://arxiv.org/abs/1810.03993⟩.

[28] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution" (2007) 8(1) BMC Bioinformatics 1 DOI: 10.1186/1471-2105-8-25.

[29] Chouldechova, Benavides-Prado, Fialko, and Vaithianathan (see n. 13).

[30] See generally Felix Richie, "Access to sensitive data: Satisfying objectives rather than constraints" (2014) 30(3) Journal of Official Statistics.

Recommendation 14 *Allegheny County should invite researchers to examine and audit their systems and give them heightened access to software, data and front-line workers. A list of such collaborations should be published.*

Recommendation 15 *Allegheny County should seek to work with researchers to create and make publicly available peer-reviewed research on the models characteristics and their use in situ.*

Holistic Evaluation and Maintenance

Machine learning and decision support systems exist within a procedural and organisational context, which itself exists within a wider landscape of societal structures and challenges. In issues of welfare and social protection, this is especially salient.

Actuarial or machine learning systems used in high-stakes decision-making have been long accused of creating and perpetuating bias and discrimination in society.[31] As a result, a range of methods through which different fairness properties can be assessed and statistically assured have been developed.[32] It is likely useful to use such methods in this case, at least in an initial audit capacity, however the exact trade-offs will only become apparent through trying out many different types,[33] and there may be restrictions on the types of fairness issues that can be tackled based on limitations in measured or available sensitive data.[34] The County has prior experience of assessing deployed systems for bias and should draw upon the institutional routines and practices developed there.[35]

[31] See e.g. Oscar H Gandy, "Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems" (2010) 12(1) Ethics and Information Technology 29 DOI: 10/bzwqrx; Solon Barocas and Andrew D Selbst, "Big Data's Disparate Impact" (2016) 104 California Law Review 671 DOI: 10/gfgq9w.

[32] See e.g. Faisal Kamiran and Toon Calders, "Classifying without discriminating" in *2nd International Conference on Computer, Control and Communication, Karachi, Pakistan, 17–18 Feburary, 2009* (2009) DOI: 10/dtcfsn; Andrea Romei and Salvatore Ruggieri, "Discrimination data analysis: A multi-disciplinary bibliography" in Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (eds.), *Discrimination and Privacy in the Information Society* (Springer 2012); Sara Hajian and Josep Domingo-Ferrer, "Direct and indirect discrimination prevention methods" in Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (eds.), *Discrimination and privacy in the information society* (Springer 2012); Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, "Fairness through awareness" in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)* (2012) DOI: 10/fzd3f9; Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, "Certifying and removing disparate impact" in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015) DOI: 10/gfgrbk; Moritz Hardt, Eric Price, and Nati Srebro, "Equality of Opportunity in Supervised Learning" in DD Lee, M Sugiyama, UV Luxburg, I Guyon, and R Garnett (eds.), *Advances in Neural Information Processing Systems 29* (Curran Associates, Inc 2016); Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi, "Fairness beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment" in *Proceedings of the 26th International Conference on World Wide Web* (International World Wide Web Conferences Steering Committee 2017) DOI: 10/gfgq8r.

[33] On empirical comparisons in this area, see Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth, "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning" [2018] arXiv preprint ⟨http://arxiv.org/abs/1802.04422⟩.

[34] Michael Veale and Reuben Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data" (2017) 4(2) Big Data & Society DOI: 10/gdcfnz.

[35] Chouldechova, Benavides-Prado, Fialko, and Vaithianathan (see n. 13).

More broadly however, discrimination-aware data analysis comes with risks. Allegheny County has a range of policy levers they can use in order to understand and to tackle issues of poverty and inequality. This report calls for all issues of discrimination and bias—which involve both the under and the over targeting of certain (e.g. ethnic) groups—to be tackled in a holistic manner. While it is possible, and even tempting, to try to 'correct' for biases in a decision-support model, or even to consider introducing positive discrimination,[36] there is a risk that this both reifies the importance of the actuarial model in the broader process, falling into the trap of function creep,[37] as well as draws attention away from more substantive and sustainable policy mechanisms to address bias and discrimination, such as designing specific policy strategies to reach out or to consult with poorly represented groups. As has been previously noted in relation to child protection tools, 'fairness is a process property, not just a model property',[38] and this should be kept strongly in mind and at the heart of relevant intetventions.

Consequently, this report suggests that *while discrimination-aware data mining is useful, it must form part of a broader strategy*. The County should resist becoming a poster child for statistical fairness, and should seek instead to integrate it seemlessly as just one part of a broader strategy around how targeted interventions—whether that be actuarially or by other means such as geographic investment—interact with issues of concern such as discrimination and inequality. Many findings in actuarial fairness indicate the difficulties present when trying to reconcile different notions of formalised fairness, and that it can be mathematically impossible to have many types of fairness that might be desired at once.[39] This does not mean however that broader strategies cannot aim at fairness, particularly when the question is not a binary one (such as child removal) but a decision point with many potential interventions that are qualtitatively different from each other.

Recommendation 16 *The County should begin a programme of work on discrimination and inequality in information use and service provision more broadly which involves, but is not limited to, statistical understandings of fairness. Where deficits in modelling are discovered, particular attention should be paid to how different interventions might seek to mitigate these, rather than solely attempting to adjust the model to compensate.*

Recommendation 17 *The County's work in building the strategy in Recommendation 16 should be deeply participatory and open to comment.*

There are several further issues concerning modelling that are important whether personal data is involved or not, and these issues are relevant to an ethical analysis of this case.

[36] Sicco Verwer and Toon Calders, "Introducing positive discrimination in predictive models" in Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (eds.), *Discrimination and privacy in the information society* (Springer 2013).

[37] See the section Deployment and Function Creep.

[38] Chouldechova, Benavides-Prado, Fialko, and Vaithianathan (see n. 13) 13.

[39] See e.g. Alexandra Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments" (2017) 5(2) Big Data 153 DOI: 10/gdcdqf.

Firstly, statistical models in particular suffer from a risk of *concept drift*. Concept drift or dataset shift refers to changes in the conditional distributions of model inputs (e.g. DHS warehouse data) and outputs (e.g. a placement event).[40] For example, a model may have been built before some demographic, eocnomic or legal change which makes the population in question, and their social behaviour, significantly different to what it once was. It might also be that some individuals become more or less difficult to collect data on over time, and therefore end up under- or over-represented in a modelling process.

The consequences of concept drift can be severe. It might, for example, be that the model no longer represents the social phenomena on the ground as well as it could, or potentially could misrepresent it for certain demographic subgroups. In the Allegheny case, the nature and challenges of families with high scores or with crisis events may change over time. There are some tools that can help understand concept drift, and where technically feasible these should be considered by the modelling team.[41] These might include refreshing the model, dropping data over time, or weighting past data less heavily than recent data. Implementing these can be challenging, and they do not always cope well with certain types of drift that might be sudden or cyclical in nature. As reviews of the field have indicated, integrating separate, expert knowledge to detect when something has changed that the model has not considered is likely to be an important component of any concept drift detection strategy.[42]

Linking qualitative knowledge to understand drift in machine learning models is a challenge in public sector machine learning in general.[43] How might Allegheny County cope with it in an ongoing manner. Some lessons can already be taken from some of the preparatory work the County has done in interviewing '20/20' families to understand their needs. While instigating a regular programme of interviews *solely* for the purpose of better understanding a predictive model might be disproportionate, this report instead recommends an integrated evidence-gathering strategy across the County which allows interviews and other data (such as focus groups with specialist social workers) to be fed regularly into modelling and maintenance teams. For example, risk scores might become part of an interview sampling strategy for more general evidence gathering across the County, enabling these interviews to play a useful and importantly a specific role in many different functions of the Department of Human Services.

[40] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence, *Dataset shift in machine learning* (The MIT Press 2009).

[41] J Gama, Indre Žliobaitė, A Bifet, M Pechenizkiy, and A Bouchachia, "A survey on concept drift adaptation" (2013) 1(1) ACM Comput. Surv. DOI: 10/gd893p.

[42] See eg ibid. 30.

[43] See generally Michael Veale, Max Van Kleek, and Reuben Binns, "Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)* (ACM 2018) DOI: 10/ct4s 5.

Recommendation 18 *Maintenance of the model over time should have a strong qualitative dimension, seeking feedback from different sources of on-the-ground knowledge, such as front-line workers, to understand how the model performs on different groups, and how the phenomena being modelled might be changing over time.*

Supporting Decision Support

Much has been written about how decision support systems might end up determining decisions rather than just providing one of several pieces of relevant evidence. Broadly this has been characterised as an issue of *automation bias*, where individuals either over- or underrely on computerised information compared to a rational model of how much it should be relied upon.[44]

In this case, automation bias presents a concern where case-workers are manually involved in reviewing a case where the score is visible. Automation bias might also work in more complex ways, such as in interaction with features such as perceived stereotypes.

Allegheny County must actively work to mitigate and pre-empt automation bias in predictive systems, as such biases may serve to undermine the aims of the tools being deployed.

In the case of the Allegheny Family Screening Tool, researchers and practitioners affiliated to the County's project analysed the level of over- or under-reliance, in terms of how often a child the system had flagged as a 'mandatory' (i.e. highly recommended) screen-in was manually overriden by supervisors. It is noted that some organisational operating procedures are already in place concerning this particular tool: for example, that only supervisors, rather than front-line screeners, are allowed to 'screen-out' scores that are in the top 15% of risk.[45] Despite this barrier, and while there *is* an obserable drop in screen-outs over the 17–18 ventile threshold that represents this level of risk, a considerable proportion of overrides are visible.[46]

Training and investment in tools beyond predictive scores In Allegheny County, the data that is used to calculate the score is available to decision-makers through their *Client View* system. This platform also contains a variety of documents, plans and assessments which decision-makers can examine. It is important that investment, training and testing in this system is maintained, in order to present decision-makers with feasible informational alternatives to the score provided. While the proposed programme involves the drawing up of lists to give to programme providers rather than decision-support in a clinical setting (as the AFST does), this recommendation is still important for

[44] Linda J Skitka, Kathleen L Mosier, and Mark Burdick, "Does automation bias decision-making?" (1999) 51 International Journal of Human-Computer Studies 991 DOI: 10/bg5rb7; Jaap J Dijkstra, "User agreement with incorrect expert system advice" (1999) 18(6) Behaviour & Information Technology 399 DOI: 10/fsnqm9.

[45] Chouldechova, Benavides-Prado, Fialko, and Vaithianathan (see n. 13) footnote 6.

[46] ibid. 12.

those who are drawing up those lists or otherwise involved in monitoring and evaluating ths scores. As a general principle around this kind of use of predictive system, they must be deployed with consideration for user testing, improvement, training and maintenance of existing, complementary systems.

Workload management in the presence of risk scoring. Where individuals are working alongside systems using this score in their role, and where the score is supposed to make their jobs faster or more efficient, care should be taken that there remains substantive time to oversee each individual case as appropriate. The County should be mindful of the risk that where resources are constrained, there is a greater tendency to want to rubber-stamp machine outputs due to limited capacity to provide meaningful oversight.

Organisational structures to support disagreement and monitor reliance. Wherever risk scoring meets organisational process, and the operating procedures permit some override, it needs to be ensured that staff involved have the confidence and support to 'disagree with the machine'. In the ABF case, there do appear to be some cases where discretion is required. The DHS programme staff aggregating families into groups to tier services and to provide lists to the service providers may be empowered to include families who were not scored sufficiently in these lists. The community health workers reaching out to mothers in the hospital may need discretion to be able to offer targeted services where they see real need. Those monitoring and evaluating the system in the DHS may need confidence to spot errors and to raise them with senior staff or external modellers.

Operating procedures may rightly discourage disagreement with the system where there is strong evidence to support that judgement. In the AFST system, disagreement was discouraged through defaulting those calls with a screening score of over 18 to being 'screened-in', with the only mechanism of override available being an appeal to the supervisor on duty at the time. Despite this, a significant number of overrides were recorded, which in many ways is a strong sign of critical independence and discussion of the models rather than a rubber-stamping of decisions.

Allegheny County should establish appropriate monitoring measures for disagreement or over- or under-reliance, and follow these up with qualitative studies. Such studies should be carefully designed not to appear to blame those disagreeing with the system, but to ensure they feel that their feedback is crucial to making a better system in future, and identifying the failure modes they saw using the knowledge they have gained from the job. Supervisors in roles where scores are being used by those they manage should receive training aimed at helping those under them feed-back on the systems being used, and ensure that they encourage those under them to maintain a critical view of all systems they work with. This should especially be emphasised for new starters—it is possible that the levels of disagreement in the AFST system will decrease over time as the number of screeners who

have never known work without this system decreases. Keeping on top of these phenomena will be a continuous effort, but can be built into existing training schemes to reduce burden.

Recommendation 19 *The County should develop methods to monitor oversight, critical examination and the use of these scores by all those who have access to them or are tasked with approaching families for voluntary targeted services.*

Recommendation 20 *The County should ensure that investment decisions do not create undesirable over-reliance on scores, such as through under-investment in alternative sources of information, or through reduction of time available for each task due to efficiencies of automated systems.*

References

Ananny M and Crawford K, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability" [2016] New Media & Society DOI: 10/gddxrg.

Ausloos J and Dewitte P, "Shattering one-way mirrors—data subject access rights in practice" (2018) 8(1) International Data Privacy Law 4 DOI: 10/cwcf.

Barocas S and Selbst AD, "Big Data's Disparate Impact" (2016) 104 California Law Review 671 DOI: 10/gfgq9w.

Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, and Shadbolt N, "'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)* (ACM 2018) DOI: 10/cvcp.

Chouldechova A, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments" (2017) 5(2) Big Data 153 DOI: 10/gdcdqf.

Chouldechova A, Benavides-Prado D, Fialko O, and Vaithianathan R, "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions" in *Conference on Fairness, Accountability and Transparency (FAT* 2018)* (2018).

Citron DK, "Technological due process" (2008) 85 Washington University Law Review 1249.

Dijkstra JJ, "User agreement with incorrect expert system advice" (1999) 18(6) Behaviour & Information Technology 399 DOI: 10/fsnqm9.

Dwork C, Hardt M, Pitassi T, Reingold O, and Zemel R, "Fairness through awareness" in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)* (2012) DOI: 10/fzd3f9.

Edwards L and Veale M, "Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not The Remedy You Are Looking For" (2017) 16 Duke L. & Tech. Rev. 18 DOI: 10/gdxthj.

– "Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?" (2018) 16(3) IEEE Security & Privacy 46 DOI: 10/gdz29v.

Eslami M, Karahalios K, Sandvig C, Vaccaro K, Rickman A, Hamilton K, and Kirlik A, "First I "Like" It, then I Hide It: Folk Theories of Social Feeds" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM 2016) DOI: 10.1145/2858036.2858494.

Feldman M, Friedler SA, Moeller J, Scheidegger C, and Venkatasubramanian S, "Certifying and removing disparate impact" in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015) DOI: 10/gfgrbk.

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, and Roth D, "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning" [2018] arXiv preprint ⟨http://arxiv.org/abs/1802.04422⟩.

Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, and Bouchachia A, "A survey on concept drift adaptation" (2013) 1(1) ACM Comput. Surv. DOI: 10/gd893p.

Gandy OH, "Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems" (2010) 12(1) Ethics and Information Technology 29 DOI: 10/bzwqrx.

Gayle D, "Schools Census Used to Enforce Immigration Laws, Minister Says" (*The Guardian*, January 13, 2019) ⟨https://www.theguardian.com/politics/2019/jan/13/schools-census-used-for-immigration-enforcement-minister-says⟩ visited on April 10, 2019.

Grove WM, Zald DH, Lebow BS, Snitz BE, and Nelson C, "Clinical versus mechanical prediction: a meta-analysis" (2000) 12(1) Psychological Assessment.

Guadamuz A, "Viral Contracts or Unenforceable Documents? Contractual Validity of Copyleft Licenses" (2004) 26(8) European Intellectual Property Review 331 ⟨https://ssrn.com/abstract=569101⟩.

Hajian S and Domingo-Ferrer J, "Direct and indirect discrimination prevention methods" in B Custers, T Calders, B Schermer, and T Zarsky (eds.), *Discrimination and privacy in the information society* (Springer 2012).

Hardt M, Price E, and Srebro N, "Equality of Opportunity in Supervised Learning" in DD Lee, M Sugiyama, UV Luxburg, I Guyon, and R Garnett (eds.), *Advances in Neural Information Processing Systems 29* (Curran Associates, Inc 2016).

Hildebrandt M, "The Dawn of a Critical Transparency Right for the Profiling Era" in J Bus, M Crompton, M Hildebrandt, and G Metakides (eds.), *Digital Enlightenment Yearbook* (IOS Press 2012) DOI: 10/cwpm.

Hoppe R, *The Governance of Problems: Puzzling, Powering and Participation* (Policy Press 2010).

Kamiran F and Calders T, "Classifying without discriminating" in *2nd International Conference on Computer, Control and Communication, Karachi, Pakistan, 17–18 Feburary, 2009* (2009) DOI: 10/dtcfsn.

Lipsky M, *Street-level bureaucracy: Dilemmas of the individual in public services* (Russell Sage Foundation 2010).

Manzoni J, "Big data in government: the challenges and opportunities" (*GOVUK*, February 2017) ⟨https://perma.cc/GF7B-5A2R⟩ visited on October 4, 2018.

Mitchell M et al., "Model Cards for Model Reporting" in *Proceedings of the 2nd ACM Conference on Fairness, Accountability and Transparency (ACM FAT* 2019)* (ACM 2019) ⟨https://arxiv.org/abs/1810.03993⟩.

Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, and Wallach H, "Manipulating and Measuring Model Interpretability" [2018] arXiv:1802.07810 [cs] ⟨http://arxiv.org/abs/1802.07810⟩ visited on April 10, 2019.

Quiñonero-Candela J, Sugiyama M, Schwaighofer A, and Lawrence ND, *Dataset shift in machine learning* (The MIT Press 2009).

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1.

Richie F, "Access to sensitive data: Satisfying objectives rather than constraints" (2014) 30(3) Journal of Official Statistics.

Romei A and Ruggieri S, "Discrimination data analysis: A multi-disciplinary bibliography" in B Custers, T Calders, B Schermer, and T Zarsky (eds.), *Discrimination and Privacy in the Information Society* (Springer 2012).

Skitka LJ, Mosier KL, and Burdick M, "Does automation bias decision-making?" (1999) 51 International Journal of Human-Computer Studies 991 DOI: 10/bg5rb7.

Statistics New Zealand, "Integrated Data Infrastructure" (*Government of New Zealand*, 2016) ⟨https://perma.cc/9RXL-SV7P⟩ visited on October 4, 2018.

Strobl C, Boulesteix A.-L, Zeileis A, and Hothorn T, "Bias in random forest variable importance measures: Illustrations, sources and a solution" (2007) 8(1) BMC Bioinformatics 1 DOI: 10.1186/1471-2105-8-25.

Tversky A and Kahneman D, "Judgment under Uncertainty: Heuristics and Biases" (1974) 185(4157) Science 1124.

Ustun B and Rudin C, "Supersparse Linear Integer Models for Optimized Medical Scoring Systems" (2016) 102(3) Machine Learning 349 DOI: 10/f8crhw.

Veale M and Binns R, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data" (2017) 4(2) Big Data & Society DOI: 10/gdcfnz.

Veale M, Binns R, and Edwards L, "Algorithms That Remember: Model Inversion Attacks and Data Protection Law" (2018) 376 Phil. Trans. R. Soc. A 20180083 DOI: 10/gfc63m.

Veale M and Brass I, "Administration by Algorithm? Public Management meets Public Sector Machine Learning" in K Yeung and M Lodge (eds.), *Algorithmic Regulation* (Oxford University Press 2019).

Veale M, Van Kleek M, and Binns R, "Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)* (ACM 2018) DOI: 10/ct4s.

668 Verwer S and Calders T, "Introducing positive discrimination in predictive models" in B Custers,
669 T Calders, B Schermer, and T Zarsky (eds.), *Discrimination and privacy in the information society*
670 (Springer 2013).
671 Zafar MB, Valera I, Gomez Rodriguez M, and Gummadi KP, "Fairness beyond Disparate Treatment & Dis-
672 parate Impact: Learning Classification without Disparate Mistreatment" in *Proceedings of the 26th*
673 *International Conference on World Wide Web* (International World Wide Web Conferences Steering
674 Committee 2017) DOI: 10/gfgq8r.