Trade-offs in Algorithmic Risk Assessment: an Australian Domestic Violence Case Study

Daniel McNamara, Timothy Graham, Ellen Broad, Cheng Soon Ong

31 August 2018

Actuarial methods have been part of criminal law and its enforcement in jurisdictions around the world for nearly a century.[1] 'Actuarial' methods employ probability theory to shape risk management tools designed to help humans make decisions about who to search, what geographical areas to police, eligibility for bail, eligibility for parole, the length of a criminal sentence and the kind of prison a convicted offender should be incarcerated in.[2] The criminal justice system can be said to have been employing algorithms and crunching 'big' data for decision-making long before these words became part of the popular lexicon surrounding automated decisions.

These days, a range of commercial and government providers are developing software that embed actuarial methods in code, using machine learning methods on large bodies of data and marketed under the umbrella of "artificial intelligence" (AI).[3] While the effects of using these kinds of probabilistic methods in criminal justice contexts – such as higher incarceration rates among certain racial groups and distorted future predictions — have been critiqued by legal and social science scholars for several years,[4] they've only recently become issues for the computer scientists and engineers developing these software solutions.

In-depth investigations of commercial criminal recidivism algorithms, like the COMPAS software developed by US-based company Equivant (formerly known as Northpointe), have become flash points in discussions of bias and prejudice in AI.[5] Within the computer science community, developing quantitative methods to potentially reduce bias and build fairer, more transparent decision-making systems is an increasingly important research area.[6] This chapter trials one quantitative approach to 'fairness', designed to reduce bias in the outputs of a pre-existing case study predicting domestic violence recidivism in the Australian context.

There is no one authoritative definition of fairness,[7] in computer science or in any other discipline. "Fairness" as a word carries significant cultural heritage.[8] John Rawls' famed "veil of ignorance" proposes an approach to fairness akin to an impartial observer, who does not know what status they will have in society (and how the definition of fairness agreed on) will affect them.[9] Other scholars have noted this abstract approach of fairness, when put into practice, does not reduce perceptions of unfair outcomes.[10] Previous explorations of varied definitions of fairness in disciplines as diverse as philosophy, law, neuroscience and information theory have concluded there is no single foundation on which to rest for the purposes of fair machine learning.[11]

To paraphrase the science fiction author Margaret Atwood: "Fair never means fairer for every-

[1] Bernard Harcourt. *Against Prediction: Profiling, Policing and Punishing in an Actuarial Age.* University of Chicago Press, 2006.

[2] Ibid.

[3] Richard Berk. *Criminal Justice Forecasts of Risk: A Machine Learning Approach.* Springer, 2012.

[4] Rice, Marnie and Harris, Grant. 'Violent Recidivism: Assessing Predictive Validity'. *Journal of Consulting and Clinical Psychology* 63 (1995).

[5] Julia Angwin et al. 'Machine Bias'. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[6] Arvind Nayaranan. *Tutorial: 21 Fairness Definitions and their Politics.* Conference on Fairness, Accountability and Transparency, https://www.youtube.com/watch?v=jIXIuYdnyyk. 2018.

[7] Shira Mitchell and Jackie Shadlen. 'Mirror Mirror: Reflections on Quantitative Fairness' (2018). https://speak-statistics-to-power.github.io/fairness/.

[8] Anna Wierzbicka. *English: Meaning and Culture.* Oxford University Press, 2006.

[9] John Rawls. *A Theory of Justice.* Harvard University Press, 1971.

[10] Stefan Trautmann and Gijs van de Kuilen. 'Process fairness, outcome fairness, and dynamic consistency: Experimental evidence for risk and ambiguity'. *Journal of Risk and Uncertainty* 53 (2016).

[11] Robert Williamson and Aditya Krishna Menon. 'The Cost of Fairness in Binary Classification'. *Proceedings of Machine Learning Research* 81 ().

one. It always means worse, for some".[12] This chapter does not assert its approach to fairness as the 'right' one. What is 'fair' is not a technical consideration, but a moral one.[13] We are interested in the insights that quantitative methods for fairness give human decision makers, allowing us to make explicit certain implicit trade-offs that have long been part of how humans make decisions. Efforts to quantify what is 'fair' allow us to measure the impact of these trade-offs.

Used effectively in a criminal justice context, these methods could help human decision makers make more transparent, informed decisions about a person's likelihood of recidivism. But they also speak to enduring challenges unpicking and rectifying bias in actuarial methods (and the AI systems that absorb these methods). Whatever definition of 'fairness' is employed, there are real world consequences. The impact of varying trade-offs in 'fair' decision making on victims and offenders should be handled with great caution in a domestic violence context.

1 Algorithmic Risk Assessment in an Australian Domestic Violence Context

In a 2016 paper, Australian researchers Robin Fitzgerald and Timothy Graham[14] evaluated the potential of existing administrative data drawn from the NSW Bureau of Crime Statistics and Research (BOCSAR) Re-offending Database (ROD) to predict domestic violence-related recidivism.[15] Being able to reliably and accurately assess which offenders, in which contexts, are likely to recommit domestic violence is a priority for law enforcement, victim support services and of course, for victims themselves.

Domestic violence (DV), also referred to as family violence or domestic abuse, is defined as a pattern of violence, intimidation or abuse between individuals in a current or former intimate relationship. A World Health Organization study found that within each of dozens of studies conducted around the world, between 10% and 69% of women reported having experienced physical abuse by an intimate partner, and between 5% and 52% reported having experienced sexual violence by an intimate partner.[16]

In Australia, one in six women and one in twenty men have experienced at least one instance of domestic violence since the age of 15.[17] On average, police in Australia respond to a domestic violence matter every two minutes.[18] These statistics emphasize the scale and the gendered nature of this issue. Indeed, aggregate prevalence rates further highlight the negative impact of DV and family violence more broadly. DV is one of the top ten risk factors contributing to disease burden among adult women,[19] and the economic costs of violence against women and children in Australia (including both domestic and non-domestic violence) are estimated at around $13.6 billion per year.[20] Existing statistics and surveys suggest that Indigenous communities face domestic violence issues at much greater rates than the rest of the population.[21]

1.1 The Evolution of Algorithmic Risk Assessments

Actuarial methods and probability theory have been employed to help humans make decisions in a criminal justice context for many years.[22] It's only recently that they've been embedded

[12] Margaret Atwood. *The Handmaid's Tale*. McClelland and Stewart, 1985.

[13] Nayaranan, *Tutorial: 21 Fairness Definitions and their Politics*.

[14] Robin Fitzgerald and Timothy Graham. 'Assessing the Risk of Domestic Violence Recidivism'. *Crime and Justice Bulletin* 189 (2016).

[15] NSW Bureau of Crime Statistics and Research. *Re-offending Statistics for NSW*. 2018.

[16] Etienne Krug et al. 'The World Report on Violence and Health'. *World Health Organization* (2002).

[17] Australian Bureau of Statistics. *Personal Safety Survey 2016*. 2017; Peta Cox. 'Violence Against Women in Australia: Additional Analysis of the Australian Bureau of Statistics' Personal Safety Survey'. *Horizons Research Report, Australia's National Research Organisation for Women's Safety* (2012).

[18] Clare Bulmer. 'Australian Police Deal with a Domestic Violence Matter Every Two Minutes', *ABC News*, 5 June 2015, http://www.abc.net.au/news/2015-05-29/domestic-violence-data/6503734.

[19] Australian Institute of Health and Welfare and Australia's National Research Organisation for Women's Safety. *Examination of the Health Outcomes of Intimate Partner Violence against Women: State of Knowledge Paper*. 2016; Australian Institute of Health and Welfare. *Family, Domestic and Sexual Violence in Australia*. 2018.

[20] Department of Social Services. *The Cost of Violence against Women and their Children. Report of the National Council to Reduce Violence against Women and their Children*. 2009.

[21] In NSW in 2016, 2.9% of the population were Indigenous (Australian Bureau of Statistics. *Census 2016*. 2017) while 65% of victims of family and domestic violence overall were Indigenous (Australian Bureau of Statistics. *Recorded Crime - Victims, Australia 2016*. 2017).

[22] Harcourt, *Against Prediction: Profiling, Policing and Punishing in an Actuarial Age*.

in software.[23] While these longstanding methods could be said to be 'algorithmic'[24] in nature – taking a rule-based approach to predictions – for the purposes of this chapter we use the term "algorithmic risk assessment" to refer to the more recent automated, software-driven systems. An example is the Public Safety Assessment,[25] which is used in the U.S. states of Kentucky, Arizona and New Jersey and several other U.S. counties.[26]

Algorithmic risk assessment systems have several potential advantages. They offer a mechanism to improve the accuracy of decisions made in the criminal justice system.[27] They are readily scalable, offering greater consistency than human judgment.[28] They offer increased transparency of decisions, assuming that the system's code, methodology and input data are accessible.[29] And they often have adjustable parameters (as in this work), which render trade-offs explicit in decision-making and allow them to be managed.

However, investigations of existing algorithmic risk assessment systems have demonstrated that these systems can – by choice – also be shrouded in secrecy, unnecessarily complex and/or reinforce existing bias.[30] It has been shown that COMPAS – which used over a hundred variables for predictions – performs no better than a logistic regression classifier using age and total number of previous convictions.[31] A controversial recent example of a risk assessment system in the Australian context is the Suspect Targeting Management Plan (STMP).[32] In the cases of both COMPAS[33] and STMP,[34] concerns have been raised that the systems are unfair, in the former case towards African-Americans and in the latter case towards Indigenous Australians.

1.2 Predicting Domestic Violence Recidivism using Administrative Data

A primary aim of any recidivism prediction is accuracy. That is, to accurately identify which offenders are most likely to recommit a crime and subsequently (1) adjust their access to bail or parole, or period of incarceration accordingly; and (2) understand the risk factors associated with recidivism in order to better target resources and programs aimed at crime prevention. But what is considered an 'accurate' prediction is complicated by risk-based, profiling approaches to policing that inevitably see certain populations overrepresented in data about past offenders, which is then used for making future predictions. Is a prediction based on this past data 'fair'? Answering this question depends on identifying and managing the trade-offs involved in the design of recidivism assessments.

Although domestic violence (DV) is a serious problem in Australia, to date there has been relatively little research on the risks associated with family violence and DV recidivism in the Australian context.[35] Recidivism in this paper refers to reoffending following index conviction. Broadly speaking, a 'recidivist' or 'reoffender' is an individual who is a repeat or chronic offender. In the context of DV recidivism, national and state-based agencies have begun to develop and implement computerized decision support systems (DSS) and risk assessment tools that draw on

[23] Sarah Desmarais and Jay Singh. 'Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States' (2013).

[24] Informally, an algorithm is simply a series of steps or operations undertaken to solve a problem or produce a particular outcome/output. For instance, in a rudimentary way a cake recipe can be thought of as an algorithm that, if the steps are followed precisely, produces a cake.

[25] Laura and John Arnold Foundation. *Public Safety Assessment: Risk Factors and Formula.* https://www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf. 2017.

[26] Laura and John Arnold Foundation. *Public Safety Assessment Frequently Asked Questions.* https://www.psapretrial.org/about/faqs.

[27] For example, a recent study using data from more than 750,000 pre-trial release decisions made by New York City judges found that, at the same jailing rate as the judges, an algorithm could reduce crime by 14.4-24.7%. Alternatively, without any increase in crime, an algorithm could reduce jail rates by 18.5-41.9%. (Jon Kleinberg et al. 'Human Decisions and Machine Predictions'. *The Quarterly Journal of Economics* 133.1 [2017])

[28] Ibid.

[29] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 'Interpretable Classification Models for Recidivism Prediction'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.3 (2017).

[30] Angwin et al., 'Machine Bias'.

[31] Julia Dressel and Hany Farid. 'The Accuracy, Fairness, and Limits of Predicting Recidivism'. *Science Advances* 4.1 (2018).

[32] NSW Police Force. 'NSW Police Force Corporate Plan 2016-18' (2016).

[33] Angwin et al., 'Machine Bias'.

[34] Vicki Sentas and Camilla Pandolfini. 'Policing Young People in NSW: A Study of the Suspect Targeting Management Plan'. *Youth Justice Coalition* (2017).

[35] Hayley Boxall, Lisa Rosevear, and Jason Payne. 'Identifying First Time Family Violence Perpetrators: The Usefulness and Utility of Categorisations Based on Police Offence Records'. *Trends and Issues in Crime and Criminal Justice* 487 (2015); Fitzgerald and Graham, 'Assessing the Risk of Domestic Violence Recidivism'.

standardized data (within and/or across agencies) to help understand the risk of DV recidivism for sub-groups within the population. There is increasing interest in evidence-based crime and social welfare governance that draw on data science and big data, perhaps due to a perception that these kinds of DSS and risk assessment tools are more efficient, objective and less costly than existing approaches.[36]

To be sure, the point of these DSS and risk assessment tools is to enhance, refine and better target programs and resources to prevent DV, rather than simply punishment and control. While computer-based DSS have been criticized in, for example, child welfare and protection,[37] recent studies suggest that DV-related risk assessment tools can be effective, particularly to assist under-resourced front-line agencies to make informed and speedy decisions about detention, bail and victim assistance.[38] A standard practice is to measure the accuracy of risk assessment tools using Receiver Operating Characteristic (ROC) curve analysis,[39] known as Area Under the Curve (AUC), and predictive risk assessment tools for DV recidivism have been shown to provide reasonably high levels of predictive performance, with AUC scores in the high 0.6 to low 0.7 range.[40]

1.3 Findings from Previous Studies

Fitzgerald and Graham[41] applied statistical methods to existing administrative data on NSW offenders who had recommitted domestic violence, to examine the kinds of factors – for example, socioeconomic status, history of past offences, Indigenous or non-Indigenous status – which were more predictive of future domestic violence offences. hey used logistic regression to examine the future risk of violent DV offending among a cohort of individuals convicted of any DV offence (regardless of whether it is violent or not) over a specific time period. They found that applying their models to unseen data achieved AUC of 0.69, indicating a reasonable level of predictive accuracy, on par with other risk assessment tools in other countries and contexts. A follow-up study explored using a decision tree induction approach on the same dataset.[42] Although these results show the potential for such models to be deployed to enhance targeted programs and resources for DV prevention, Fitzgerald and Graham also highlighted a significant problem that has yet to be addressed: in short, the authors found that their model was racially biased.

Fitzgerald and Graham argued that whilst DSS that incorporate logistic regression might offer a satisfactory tool for predicting the risk of domestic violence recidivism in the *overall population*, the efficacy is reduced for making predictions for particular sub-groups, particularly for individuals who identify as Indigenous. Indigenous status showed relatively large discrepancies in the test sample between the averages of the observed and predicted rates of violent DV reconviction. Indeed, Indigenous individuals were *more than twice as likely* to be predicted as reoffenders (29.4%) by the model compared to the observed rate (13.7%), whereas non-Indigenous individuals were less than *half as likely* to be predicted as reoffenders (2.3%) compared to the observed rate (6.1%).[43]

In other words, when it came to predicting DV recidivism for the Indigenous sub-group, Fitzgerald and Graham found that the model was biased on two fronts: over-predicting Indigenous reoffenders and under-predicting non-Indigenous reoffenders. If deployed as a risk assessment tool, this model could have serious negative consequences that may reinforce existing inequalities that have resulted from historical and contemporary injustices and oppression of Indigenous Australians.

[36] Philip Gillingham and Timothy Graham. 'Big Data in Social Welfare: The Development of a Critical Perspective on Social Work's Latest Electronic Turn'. *Australian Social Work* 70.2 (2017).

[37] Philip Gillingham. 'Risk Assessment in Child Protection: Problem Rather than Solution?' *Australian Social Work* 59.1 (2006).

[38] Ron Mason and Roberta Julian. 'Analysis of the Tasmania Police Risk Assessment Screening Tool (RAST), Final Report'. *Tasmanian Institute of Law Enforcement Studies, University of Tasmania* (2009); Jill Theresa Messing et al. 'The Lethality Screen: the Predictive Validity of an Intimate Partner Violence Risk Assessment for Use by First Responders'. *Journal of Interpersonal Violence* 32.2 (2017).

[39] Fawcett Tom. 'ROC Graphs: Notes and Practical Considerations for Researchers'. *HP Laboratories* (2004).

[40] Marnie Rice, Grant Harris, and Zoe Hilton. 'The Violence Risk Appraisal Guide and Sex Offender Risk Appraisal Guide for Violence Risk Assessment'. *Handbook of Violence Risk Assessment.* Routledge, 2010. AUC can be interpreted as the probability that a randomly selected reoffender will receive a higher risk score than a randomly selected non-reoffender. A random guess has expected AUC of 0.5 while the perfect prediction has AUC of 1.

[41] Fitzgerald and Graham, 'Assessing the Risk of Domestic Violence Recidivism'.

[42] Senuri Wijenayake, Timothy Graham, and Peter Christen. 'A Decision Tree Approach to Predicting Recidivism in Domestic Violence'. *ArXiv* (2018).

[43] Looking at the entire population the predicted (7.4%) and observed (7.6%) recidivism rates are relatively well-aligned. These large differences between predicted and observed recidivism rates only become visible looking separately at the Indigenous and non-Indigenous cohorts.

The output of the model not only reflects but also potentially *amplifies and reinforces* these inequalities. Indeed, the fact that Indigenous status (as an independent variable) appears at all in the dataset brings to light the politics of data collection and statistical forms of reasoning. The data provided through the BOCSAR Reoffending Database (ROD), and subsequently used in the study by Fitzgerald and Graham, reflects a 'practical politics' that involves negotiating and deciding what to render visible (and invisible) in an information system context.[44] This example shows that the issue of fairness in algorithmic decision-making is of utmost importance as we move towards computerized risk assessment tools in criminal justice and social welfare. At the same time, caution needs to be taken in how such fairness is defined and achieved.

2 Designing Fair Algorithmic Risk Assessments

The impact of an algorithmic risk assessment is determined by both its design and the context in which it is used. This context – which includes human judgment, policy settings and broader social trends – will remain an important determinant of outcomes in the justice system and elsewhere. No algorithm can rectify all of the past and present structural disadvantage faced by particular social groups. However, algorithmic risk assessments influence human decisions, which in turn determine the extent to which structural disadvantage is entrenched. Hence, algorithm design can play a part in making an overall system fairer – or indeed in reinforcing the unfairness of a system. Considerable research is underway to incorporate fairness into the design of algorithmic systems. This approach requires clear definitions of fairness, and modifications to algorithm design to accommodate these definitions.

2.1 Quantitative Definitions of Fairness

While defining fairness is a topic as old as human society, the advent of algorithmic predictions has necessitated the quantification of these definitions. We must be precise about what we mean if we are to embed fairness in computer code – a definition that seems simplistic or reductionist is still preferable to none at all. Therefore we necessarily consider a narrow subset of the possible meanings of 'fairness'. Quantitative definitions often describe fairness as avoiding discrimination on the basis of a particular kind of group membership, such as race or gender. Three types of definition have emerged, which we state informally:[45]

- **Parity**: Predictions should be similar for different groups
- **Independence**: Predictions should be independent of group membership
- **Causality**: Predictions should not be caused by group membership.

While each of these approaches has its advantages, our analysis focuses on definitions based on parity. A predictive model that achieves parity between groups is mathematically equivalent to one that is independent of group membership. However, (dis)parity may be measured on a continuous scale, unlike an all-or-nothing statement about independence. Unlike causality-based definitions,[46] parity measures can be computed using only an algorithm's outputs without the knowledge of its functional form, so that external auditing can be carried out without the cooperation of the algorithm's owner. Parity measures also do not require the selection of variables that are permitted to cause decisions (known as *resolving variables*[47]), which potentially could include proxies for group membership (e.g. 'redlining' where neighborhood is used a proxy for race). Finally, parity-based measures are arguably the simplest to understand for a lay audience, which is significant given the risk of excluding participants from non-quantitative backgrounds in debates about fairness.[48]

An important design choice is selecting a subset of the population to which these definitions are applied. We then ask for fair predictions – according to whichever definition we choose – only

[44] Geoffrey Bowker and Susan Star. 'How Things (Actor-Net)Work: Classification, Magic and the Ubiquity of Standards'. *Philosophia* 25.3 (1996).

[45] For further details, see Mitchell and Shadlen, 'Mirror Mirror: Reflections on Quantitative Fairness'.

[46] Matt Kusner et al. 'Counterfactual Fairness'. *Advances in Neural Information Processing Systems* (2017).

[47] Niki Kilbertus et al. 'Avoiding Discrimination through Causal Reasoning'. *Advances in Neural Information Processing Systems* (2017).

[48] Mitchell and Shadlen, 'Mirror Mirror: Reflections on Quantitative Fairness'.

within this subset, and permit differences in predictions between subsets. For example, in the recidivism context we might consider all individuals, or only those who reoffended, or only those who did not reoffend. If the subset consists of individuals who are similar according to some metric, we have a definition known in the quantitative fairness literature as *individual fairness*.[49]

Several mathematical results have shown that, for a particular set of fairness definitions, it is impossible for a predictive model to simultaneously satisfy all definitions in the set.[50] The COMPAS controversy showed this in practice: while *ProPublica*'s critique identified unfairness according to particular definitions,[51] COMPAS owner Equivant/Northpointe used different definitions to argue that the algorithm was not unfair.[52] Within a particular context, different definitions are aligned to the interests of particular stakeholders.[53] Furthermore, when predictions are also measured on their accuracy, the definitions of accuracy and fairness are in general not aligned.[54]

2.2 Defining Fairness in the Australian DV Recidivism Context

Parity-based definitions may be used to assess the fairness of a recidivism risk assessment model which generates a probability that an individual will reoffend. Given the issues associated with the context of DV in Australia, parity between Indigenous and non-Indigenous populations in the criminal justice system is of special interest. Consider the difference between Indigenous and non-Indigenous populations for each of the following:

- **Predicted reoffence rate**: the average probability of reoffence predicted by the model.
- **Predicted reoffence rate for non-reoffenders**: the average probability of reoffence predicted by the model, for those individuals who were not observed to reoffend.
- **Predicted reoffence rate for reoffenders**: the average probability of reoffence predicted by the model, for those individuals who were observed to reoffend.

Parity of predicted reoffence rates among non-reoffenders is referred to as *equality of opportunity*[55] in the quantitative fairness literature. If we also have parity of predicted reoffence rates among reoffenders, this is referred to as *equalized odds*[56] (also known as avoiding *disparate mistreatment*[57]). Enforcing these parity measures between Indigenous and non-Indigenous populations has some intuitive appeal, since it ensures that disagreements between the algorithm's predictions and the subsequently observed data do not disproportionately impact one racial group. However, these measures are sensitive to the way in which the reoffence data was collected. Profiling of particular populations, based on pre-existing risk assessments, can distort trends in reoffending. A feedback loop may be created, where this reoffence data in turn influences future risk assessments.[58]

Overall parity of predicted reoffence rate is referred to in the quantitative fairness literature

[49] Cynthia Dwork et al. 'Fairness Through Awareness'. *Innovations in Theoretical Computer Science Conference* (2012); Mitchell and Shadlen, 'Mirror Mirror: Reflections on Quantitative Fairness'.

[50] Alexandra Chouldechova. 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments'. *Big Data* 5.2 (2017); Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 'Inherent Trade-offs in the Fair Determination of Risk Scores'. *arXiv* (2016); Zachary Lipton, Alexandra Chouldechova, and Julian McAuley. 'Does Mitigating ML's Impact Disparity Require Treatment Disparity?' *arXiv* (2017); Geoff Pleiss et al. 'On Fairness and Calibration'. *Advances in Neural Information Processing Systems* (2017).

[51] Angwin et al., 'Machine Bias'.

[52] William Dieterich, Christina Mendoza, and Tim Brennan. 'COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity'. *Northpointe Inc.* (2016); Anthony Flores, Kristin Bechtel, and Christopher Lowenkamp. 'False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias'. *Federal Probation* 80 (2016).

[53] Nayaranan, *Tutorial: 21 Fairness Definitions and their Politics*.

[54] Sam Corbett-Davies et al. 'Algorithmic Decision Making and the Cost of Fairness'. *International Conference on Knowledge Discovery and Data Mining* (2017); Aditya Menon and Robert Williamson. 'The Cost of Fairness in Binary Classification'. *Conference on Fairness, Accountability and Transparency* (2018); Sam Corbett-Davies and Sharad Goel. 'The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning'. *arXiv* (2018).

[55] Moritz Hardt, Eric Price, and Nati Srebro. 'Equality of Opportunity in Supervised Learning'. *Advances in Neural Information Processing Systems* (2016).

[56] Ibid.

[57] Muhammad Bilal Zafar et al. 'Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment'. *International Conference on World Wide Web* (2017).

[58] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2017.

as *statistical parity*[59] or avoiding *disparate impact*.[60] We may not want overall parity of predicted reoffence rate if the observed rates of reoffence for Indigenous and non-Indigenous populations are different. However, overall parity has the advantage that it does not depend on the way that reoffence data was collected, which may systematically disadvantage one group.[61] Furthermore, an actual difference in reoffence rates may be the result of a complex historical process. In the case of Indigenous Australians this includes founding violence, structural violence and cultural breakdown, intergenerational trauma, disempowerment, and alcohol and drugs.[62] Legal decision-makers may wish to intervene in this process by reducing the discrepancy between incarceration rates for Indigenous and non-Indigenous populations.[63] To support this intervention, it may be appropriate for the design of a risk assessment system to incorporate greater parity in predicted reoffence rates. By contrast, other fairness definitions may be used to justify and perpetuate current rates of Indigenous incarceration.

A risk assessment model should also be accurate, subject to the previous caveat that reoffence data is likely to be imperfect and is possibly biased. While the AUC accuracy measure does not consider fairness with respect to group membership, it is related to fairness insofar as it measures the extent to which reoffenders are assessed as higher risk than non-reoffenders.

2.3 Techniques for Algorithmic Fairness

Recent work on quantitative fairness has, in addition to proposing fairness definitions, developed techniques to incorporate fairness into algorithm design.[64] One framework for organizing these fairness techniques divides them into three categories:

- **Pre-processing**: modify the data that the algorithm learns from[65]
- **In-processing**: modify the algorithm itself[66]
- **Post-processing**: modify the predictions produced by the algorithm.[67]

Pre-processing, the approach which we use in our analysis, has the advantage that it creates a *separation of concerns* between the data producer who controls the pre-processing and the data user who controls the algorithm. This means that fairness is guaranteed for any use of the pre-processed data, even if the data user is an *adversary* (i.e. they are deliberately unfair).[68] This has the potential to make regulation more practical to enforce.

Several pre-processing approaches have been proposed. To describe these, it is useful viewing a dataset as a sample from a probability distribution. The distribution jointly depends on a sensitive variable S, encoding an individual's group membership (e.g. their race), an input variable X, encoding other characteristics of the individual (e.g. their past criminal record), and a target variable Y, encoding something we wish to predict (e.g. whether or not the individual reoffended). The ultimate objective is to predict the target variable Y using the input variable X.

The result of the pre-processing is to produce a sample of a cleaned variable Z, which alters X so that it no longer contains information that can be used to infer S. This cleaned data can be used as an input to any subsequent algorithm instead of the original input data. In the following section we will use the concrete example of race as the sensitive variable S, past criminal record as the input variable X, reoffence as the target variable Y, and a cleaned version of past criminal record as Z. However, it is worth remembering that the approach works in general for other sets of variables.

[59] Dwork et al., 'Fairness Through Awareness'.

[60] Zafar et al., 'Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment'.

[61] Solon Barocas and Andrew Selbst. 'Big Data's Disparate Impact'. *California Law Review* 104 (2016).

[62] The Healing Foundation and White Ribbon Australia. 'Towards an Aboriginal and Torres Strait Islander Violence Prevention Framework for Men and Boys' (2017).

[63] As of 2017, the incarceration rate of Australia's Aboriginal and Torres Strait Islander population stood at 2434 per 100,000 people, versus 160 per 100,000 people for the non-Indigenous population. (Australian Bureau of Statistics. *Prisoners in Australia 2017*. 2017)

[64] For a review of this work in the context of recidivism prediction, see Richard Berk et al. 'Fairness in Criminal Justice Risk Assessments: the State of the Art'. *arXiv* (2017).

[65] e.g. Rich Zemel et al. 'Learning Fair Representations'. *International Conference on Machine Learning* (2013).

[66] e.g. Menon and Williamson, 'The Cost of Fairness in Binary Classification'.

[67] e.g. Hardt, Price, and Srebro, 'Equality of Opportunity in Supervised Learning'.

[68] Daniel McNamara, Cheng Soon Ong, and Bob Williamson. 'Provably Fair Representations'. *arXiv* (2017).

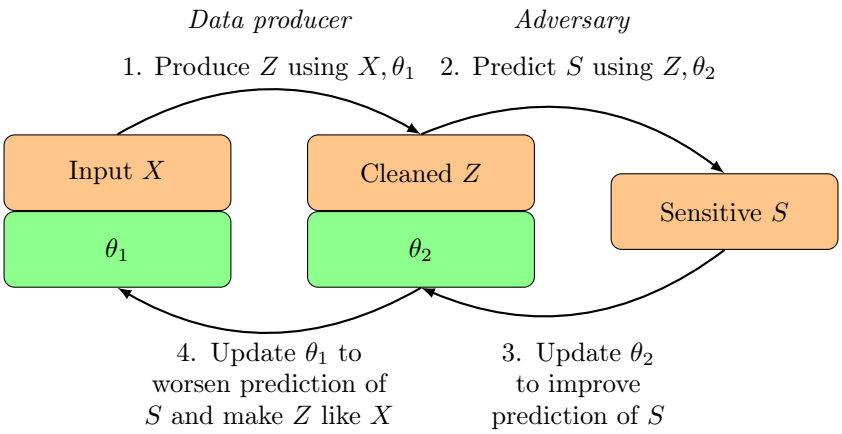

Figure 1: Learning fair representations with an adversary. In the text we use the example of X=criminal record, Z=the cleaned version of the criminal record record, S=race, Y=whether the person has reoffended. θ_1 and θ_2 are parameters of the learning algorithm.

One approach to pre-processing is to design the cleaned variable (Z) such that the distributions of Z conditioned on different values of race (S) are similar.[69] In addition to this requirement, the pre-processing procedure may optimize the independence of the cleaned variable (Z) and race (S).[70] Another pre-processing approach is to design the cleaned variable (Z) such that it is maximally informative about reoffence (Y), subject to a constraint that it is uninformative about race (S).[71]

2.4 Learning Fair Representations with an Adversary

We adopt a pre-processing approach[72] which involves learning a cleaned variable (Z) such that an adversary is unable to predict race (S) from it, while also trying to make the cleaned variable similar to the original input (X). In our case we assume that the data producer does not have access to whether the person has reoffended (Y),[73] which yields a simpler learning algorithm and is not affected by any bias in the way that we collect data on reoffences. We refer to this approach as *learning fair representations with an adversary*, since the pre-processing step can be seen as a modification to the representation of the data provided to the algorithm.

We introduce a parameter λ (lambda), a non-negative constant (once set, its value stays the same), to control the trade-off between the two objectives involved in the construction of the cleaned variable (Z). When λ is large, the algorithm focuses more on making the adversary unable to predict race (S). When λ approaches zero, the algorithm focuses more on making the original records and cleaned records similar. The algorithm does not provide any guidance as to how to select λ. Rather, this depends on a decision about the relative importance assigned to fairness and accuracy in the design of the algorithmic risk assessment. Such a decision is a social, political and regulatory one – the algorithm simply provides an implementation for whatever decision is made.

The learning steps of the algorithm are summarized in Figure 1.[74] The data producer learns a neural network parameterized by weights θ_1, which produces cleaned records from input records. The adversary learns a neural network parameterized by weights θ_2, which predicts race from the cleaned records. Observe that in this example we consider that records are cleaned if the adversary cannot use them to predict the sensitive variable, race (S). Four steps are repeated for each batch

[69] Michael Feldman et al. 'Certifying and Removing Disparate Impact'. *International Conference on Knowledge Discovery and Data Mining* (2015); James Johndrow and Kristian Lum. 'An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction'. *arXiv* (2017).

[70] Christos Louizos et al. 'The Variational Fair Autoencoder'. *International Conference on Learning Representations* (2016).

[71] AmirEmad Ghassami, Sajad Khodadadian, and Negar Kiyavash. 'Fairness in Supervised Learning: An Information Theoretic Approach'. *arXiv* (2018).

[72] This approach was proposed in Harrison Edwards and Amos Storkey. 'Censoring Representations with an Adversary'. *International Conference on Learning Representations* (2016).

[73] As in McNamara, Ong, and Williamson, 'Provably Fair Representations'.

[74] See McNamara, Ong and Williamson for further details. We also considered a variant of the adversary training objective proposed in David Madras et al. 'Learning Adversarially Fair and Transferable Representations'. *arXiv* (2018) but found it did not substantively change the results.

Table 1: Independent features in the BOCSAR dataset.

Feature	Description
Offender demographic characteristics	
Gender (**G**)	Whether the offender was recorded in ROD as male or female.
Age (**A**)	The age category of the offender at the index court finalization was derived from the date of birth of the offender and the date of finalization for the index court appearance.
Indigenous status (**IS**)	Recorded in ROD as 'Indigenous' if the offender had ever identified as being of Aboriginal or Torres Strait Islander descent, otherwise 'non-Indigenous'.
Disadvantage areas index (quartiles) (**DA**)	Measures disadvantage of an offender's residential postcode at the index offence. Based on the Socio-Economic Index for Areas (SEIFA) score (Australian Bureau of Statistics).
Index conviction characteristics	
Concurrent offences (**CO**)	Number of concurrent proven offences, including the principal offence, at the offender's index court appearance.
AVO breaches (**AB**)	Number of proven breach of Appended Violence Order (AVO) offences at the index court appearance.
Criminal history characteristics	
Prior juvenile or adult convictions (**PC**)	Number of Youth Justice Conferences or finalized court appearances with any proven offence(s) as a juvenile or adult prior to the index court appearance.
Prior serious violent offence conviction past 5 years (**P5**)	Number of Youth Justice Conferences or finalized court appearances in the 5 years prior to the reference court appearance with any proven homicide or serious assault.
Prior DV-related property damage offence conviction past 2 years (**P2**)	Number of Youth Justice Conferences or finalized court appearances in the 2 years prior to the reference court appearance with any proven DV property damage offence.
Prior bonds past 5 years (**PO**)	Number of finalized court appearances within 5 years of the reference court appearance at which given a bond.
Prior prison or custodial order (**PP**)	Number of previous finalized court appearances at which given a full-time prison sentence / custodial order.

of examples from the training data:

1. On receiving examples of X, the data producer passes them through a neural network with weights θ_1 to produce examples of Z
2. On receiving examples of Z, the adversary passes them through a neural network with weights θ_2 to predict the values of S
3. By comparing the true values of S to its predictions for these examples, the adversary updates θ_2 to improve its prediction of S in future
4. By comparing the true values of S to the adversary's predictions for these examples, the data producer updates θ_1 to worsen the adversary's prediction of S in future while also trying make Z similar to X. The trade-off between these two objectives is governed by the parameter λ.

Once learning is complete, for each individual the data producer passes their input record through a neural network with weights θ_1. This cleaned record is then provided to the data user, who uses it to make a prediction about whether the individual will reoffend.

3 Predicting DV Recidivism with the BOCSAR Dataset

We apply learning fair representations with an adversary to the prediction of DV recidivism in Australia with the BOCSAR ROD used in the study by Fitzgerald and Graham.[75] As a result, we achieve improved fairness compared to Fitzgerald and Graham's study on several measures. However, this case study also highlights the inevitable trade-offs involved. Our proposed approach allows us to reduce the disadvantage faced by Indigenous defendants incurred by using the original input data, but at the cost of predictive accuracy.

3.1 BOCSAR Dataset Experiments

The BOCSAR ROD contains 14776 examples and 11 categorical and ordinal input features for each example, as shown in Table 1. The input features are grouped to represent the offender,

[75] Fitzgerald and Graham, 'Assessing the Risk of Domestic Violence Recidivism'.

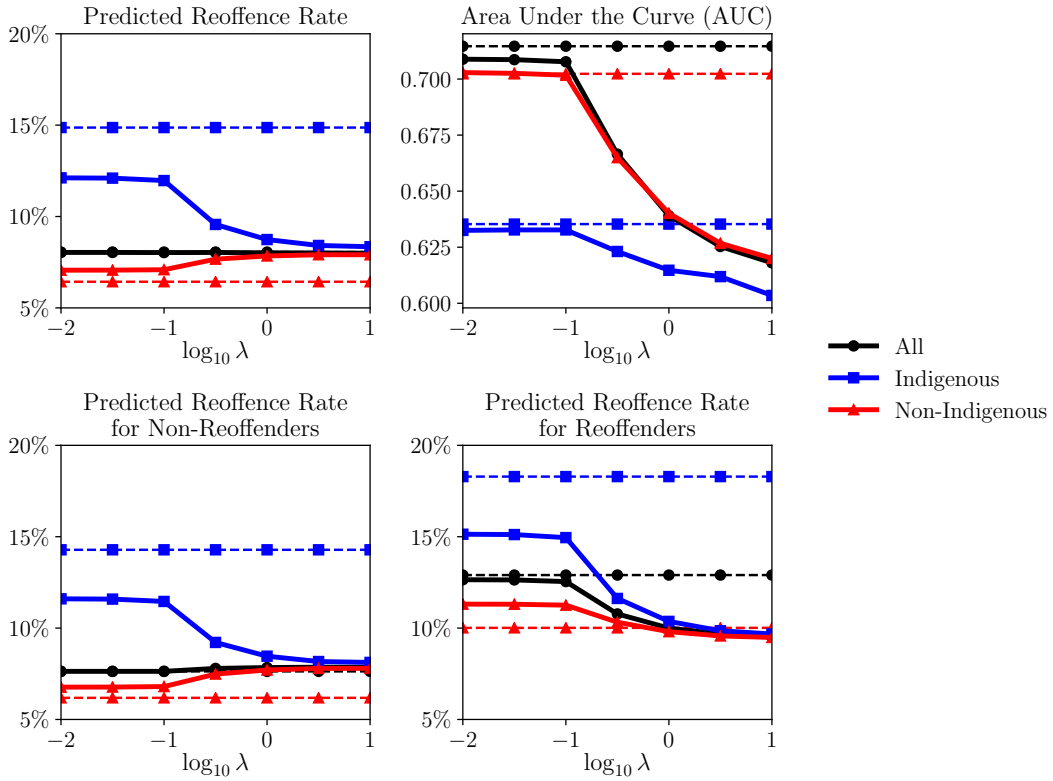

Figure 2: Results of applying pre-processing to the BOCSAR dataset. Baselines without pre-processing are shown as dashed lines. The y-axes show several fairness and accuracy measures of interest. The x-axes show the parameter λ used in pre-processing on a logarithmic scale.

index offence, and criminal history related characteristics of the offenders. The target variable is whether or not an individual re-committed a DV related offence within a duration of 24 months since the first court appearance finalization date. DV related offences include any physical, verbal, emotional, and/or psychological violence or intimidation between domestic partners. We use a random 50% sample for training and the remaining 50% for testing, as in some experiments in Fitzgerald and Graham.

Our baseline experiments use the original data, including the Indigenous status variable. We also tested the pre-processing method described in Section 2.4 for several values of the parameter λ. We predicted recidivism from the data using logistic regression as in Fitzgerald and Graham's study, which predicts the probability of reoffence for each individual. We computed the definitions of fairness and accuracy presented in Section 2.2, as shown in Figure 2. We computed each of these metrics for all individuals, for Indigenous individuals and for non-Indigenous individuals.

3.2 Discussion of the BOCSAR Dataset Results

We discuss our results by comparing the performance of the baseline method with our proposed pre-processing method. Using the original data, there are significant differences in the average predicted reoffence rates for Indigenous and non-Indigenous individuals. These predicted rates are closely related to the observed rates in the test set: for Indigenous 14.9% predicted vs 14.6% observed, and for non-Indigenous 6.4% predicted vs 6.5% observed. Our baseline does not display the severe overestimation of Indigenous reoffence observed in the Fitzgerald and Graham's model. Furthermore, the baseline test set AUC is 0.71 (slightly superior to the 0.69 previously reported by Fitzgerald and Graham), indicating that the model has some predictive accuracy.

However, there are still several potential issues with the baseline:

- variations in the way that reoffence data is collected among Indigenous and non-Indigenous populations may influence and be reinforced by predictions made by the model
- among observed non-reoffenders the average predicted reoffence rate is 14.3% for Indige-

nous vs 6.2% for non-Indigenous populations, indicating that a non-reoffending Indigenous individual is rated more than twice as risky as a non-reoffending non-Indigenous individual

- among observed reoffenders, the average predicted reoffence rate is 18.3% for Indigenous vs 10.0% for non-Indigenous populations, indicating that a reoffending non-Indigenous individual is rated only just over half as risky as a reoffending Indigenous individual[76]
- from a process perspective, it may be viewed as unfair that a person's Indigenous status is considered by the model.

Removing the Indigenous status column in the data is a possible step towards remediating these issues. It would address the final concern around fair process. However, our results show that the first three concerns stand even without the presence of this column. The solid lines on the left hand side of the plots, where λ approaches zero and the data is effectively left untouched except for the exclusion of the Indigenous status column, indicate that while the discrepancies between the Indigenous and non-Indigenous populations are not as acute as in the baseline case, they are still very much present. Information contained in the other columns still results in different outcomes for Indigenous and non-Indigenous populations, a phenomenon known as *redundant encoding*.[77]

Applying pre-processing with increasing values of λ, the above issues are addressed:

- the predicted reoffence rate for non-reoffenders is more similar for Indigenous and non-Indigenous populations (for $\lambda = 10$, 8.1% for Indigenous vs 7.8% for non-Indigenous)
- the predicted reoffence rate for reoffenders is more similar for Indigenous and non-Indigenous populations (for $\lambda = 10$, 9.7% for Indigenous vs 9.5% for non-Indigenous)
- the predicted reoffence rate overall is more similar for Indigenous and non-Indigenous populations (for $\lambda = 10$, 8.3% for Indigenous vs 7.9% for non-Indigenous).

There is a cost to pre-processing in terms of accurately predicting reoffence. The AUC drops to 0.62, so that the predictions are less accurate than the baseline (AUC 0.71), while still significantly more accurate than a random prediction (AUC 0.5).[78] Predicted reoffence rates are higher for non-reoffenders and lower for reoffenders than the baseline. This reduced accuracy is not surprising as the pre-processing removes information from the dataset. The decrease in predicted reoffence rates for reoffenders caused by the pre-processing is undesirable from the perspective of potential victims of domestic violence. Furthermore, this decrease is greater for Indigenous individuals, whose potential victims are more likely to also be Indigenous.

In summary, our approach improved on several measures of fairness compared to Fitzgerald and Graham's study. The naive approach of learning from the original input data results in a prediction that indicates that the average risk associated with Indigenous individuals is more than twice that of their non-Indigenous counterparts, even among non-reoffenders - while for a value of $\lambda = 10$ these risks are comparable. As discussed previously, this could not have been achieved simply by removing the Indigenous status column from the data. However, achieving comparable risks comes at the cost of overall predictive accuracy (AUC 0.71 to AUC 0.62). It is worth repeating that our approach does not prescribe a particular value of the trade-off parameter λ, but rather provides a quantitative tool to estimate the effect of this trade-off. We discuss further implications of fairness trade-offs in our conclusion.

4 Conclusion: Trade-offs in Algorithmic Risk Assessment

The Australian DV case study shows that without incorporating an explicit fairness criterion into algorithm design, individuals from one racial group may be marked higher risk than another, even when considering only reoffenders or only non-reoffenders. This is still true when race is simply

[76] It can be shown mathematically that if predicted reoffence rates are equal to observed reoffence rates for both Indigenous and non-Indigenous populations, and that the observed Indigenous and non-Indigenous reoffence rates are different from each other, and that the model is not perfectly accurate, then the predicted reoffence rate for non-reoffenders is different between Indigenous and non-Indigenous populations and/or the predicted reoffence rate for reoffenders is different between Indigenous and non-Indigenous populations.

[77] Dwork et al., 'Fairness Through Awareness'.

[78] It can be shown mathematically that given equal Indigenous and non-Indigenous predicted reoffence rates among reoffenders, among non-reoffenders and overall, the predicted reoffence rates for reoffenders and non-reoffenders must be equal (assuming that the observed Indigenous and non-Indigenous reoffence rates are unequal).

dropped from the input data: blindness is not enough. Incorporating a fairness criterion – such as via data pre-processing – yields more equal predicted reoffence rates for different racial groups: among reoffenders, among non-reoffenders and overall.

The case study also reveals an important trade-off involved in the design of algorithmic risk assessments. From the perspective of Indigenous defendants who in the baseline scenario were considered higher risk than non-Indigenous defendants, both among reoffenders and among non-reoffenders, this pre-processing makes the system fairer. The flipside is that non-Indigenous non-reoffenders are judged to be more risky. And all reoffenders – particularly Indigenous reoffenders – are judged to be less risky, which is not in the interests of potential victims.

The trade-off between the interests of different stakeholders is equally a part of human decision-making in the criminal justice system. The advantage of our approach is making this trade-off explicit and precisely controllable through a model parameter, which may be set according to whatever weighting is deemed appropriate by society. The approach we propose – involving an explicit trade-off between certain quantitative definitions of accuracy and fairness – also applies to other contexts where prediction algorithms are used to support decisions about individuals such as the provision of credit or insurance, and to other demographic groups besides racial groups.

There is a second trade-off involved here: between explicit and implicit explanations for decisions. Transparency allows individuals to better understand the social systems – including the criminal justice system – that make decisions about their lives. However, when the rationale for these decisions is laid bare, they may be less palatable than when they are opaque. Algorithms – with their stark rules implemented in code – have the effect of illuminating the myriad forms of inclusion and exclusion that invisibly form our social fabric. Perhaps the more profound trade-off is determining to what extent we are willing to shine that light.

References

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 'Machine Bias'. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Atwood, Margaret. *The Handmaid's Tale*. McClelland and Stewart, 1985.

Australian Bureau of Statistics. *Census 2016*. 2017.

— *Personal Safety Survey 2016*. 2017.

— *Prisoners in Australia 2017*. 2017.

— *Recorded Crime - Victims, Australia 2016*. 2017.

Australian Institute of Health and Welfare. *Family, Domestic and Sexual Violence in Australia*. 2018.

Australian Institute of Health and Welfare and Australia's National Research Organisation for Women's Safety. *Examination of the Health Outcomes of Intimate Partner Violence against Women: State of Knowledge Paper*. 2016.

Barocas, Solon and Andrew Selbst. 'Big Data's Disparate Impact'. *California Law Review* 104 (2016).

Berk, Richard. *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. Springer, 2012.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 'Fairness in Criminal Justice Risk Assessments: the State of the Art'. *arXiv* (2017).

Bowker, Geoffrey and Susan Star. 'How Things (Actor-Net)Work: Classification, Magic and the Ubiquity of Standards'. *Philosophia* 25.3 (1996).

Boxall, Hayley, Lisa Rosevear, and Jason Payne. 'Identifying First Time Family Violence Perpetrators: The Usefulness and Utility of Categorisations Based on Police Offence Records'. *Trends and Issues in Crime and Criminal Justice* 487 (2015).

Bulmer, Clare. 'Australian Police Deal with a Domestic Violence Matter Every Two Minutes', *ABC News*, 5 June 2015, http://www.abc.net.au/news/2015-05-29/domestic-violence-data/6503734.

Chouldechova, Alexandra. 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments'. *Big Data* 5.2 (2017).

Corbett-Davies, Sam and Sharad Goel. 'The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning'. *arXiv* (2018).

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 'Algorithmic Decision Making and the Cost of Fairness'. *International Conference on Knowledge Discovery and Data Mining* (2017).

Cox, Peta. 'Violence Against Women in Australia: Additional Analysis of the Australian Bureau of Statistics' Personal Safety Survey'. *Horizons Research Report, Australia's National Research Organisation for Women's Safety* (2012).

Department of Social Services. *The Cost of Violence against Women and their Children. Report of the National Council to Reduce Violence against Women and their Children.* 2009.

Desmarais, Sarah and Jay Singh. 'Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States' (2013).

Dieterich, William, Christina Mendoza, and Tim Brennan. 'COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity'. *Northpointe Inc.* (2016).

Dressel, Julia and Hany Farid. 'The Accuracy, Fairness, and Limits of Predicting Recidivism'. *Science Advances* 4.1 (2018).

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 'Fairness Through Awareness'. *Innovations in Theoretical Computer Science Conference* (2012).

Edwards, Harrison and Amos Storkey. 'Censoring Representations with an Adversary'. *International Conference on Learning Representations* (2016).

Fawcett Tom. 'ROC Graphs: Notes and Practical Considerations for Researchers'. *HP Laboratories* (2004).

Feldman, Michael, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 'Certifying and Removing Disparate Impact'. *International Conference on Knowledge Discovery and Data Mining* (2015).

Fitzgerald, Robin and Timothy Graham. 'Assessing the Risk of Domestic Violence Recidivism'. *Crime and Justice Bulletin* 189 (2016).

Flores, Anthony, Kristin Bechtel, and Christopher Lowenkamp. 'False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias'. *Federal Probation* 80 (2016).

Ghassami, AmirEmad, Sajad Khodadadian, and Negar Kiyavash. 'Fairness in Supervised Learning: An Information Theoretic Approach'. *arXiv* (2018).

Gillingham, Philip. 'Risk Assessment in Child Protection: Problem Rather than Solution?' *Australian Social Work* 59.1 (2006).

Gillingham, Philip and Timothy Graham. 'Big Data in Social Welfare: The Development of a Critical Perspective on Social Work's Latest Electronic Turn'. *Australian Social Work* 70.2 (2017).

Harcourt, Bernard. *Against Prediction: Profiling, Policing and Punishing in an Actuarial Age.* University of Chicago Press, 2006.

Hardt, Moritz, Eric Price, and Nati Srebro. 'Equality of Opportunity in Supervised Learning'. *Advances in Neural Information Processing Systems* (2016).

Johndrow, James and Kristian Lum. 'An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction'. *arXiv* (2017).

Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 'Avoiding Discrimination through Causal Reasoning'. *Advances in Neural Information Processing Systems* (2017).

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 'Inherent Trade-offs in the Fair Determination of Risk Scores'. *arXiv* (2016).

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 'Human Decisions and Machine Predictions'. *The Quarterly Journal of Economics* 133.1 (2017).

Krug, Etienne, Linda Dahlberg, James Mercy, Anthony Zwi, and Rafael Lozano. 'The World Report on Violence and Health'. *World Health Organization* (2002).

Kusner, Matt, Joshua Loftus, Chris Russell, and Ricardo Silva. 'Counterfactual Fairness'. *Advances in Neural Information Processing Systems* (2017).

Laura and John Arnold Foundation. *Public Safety Assessment Frequently Asked Questions.* https://www.psapretrial.org/about/faqs.

— *Public Safety Assessment: Risk Factors and Formula.* https://www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf. 2017.

Lipton, Zachary, Alexandra Chouldechova, and Julian McAuley. 'Does Mitigating ML's Impact Disparity Require Treatment Disparity?' *arXiv* (2017).

Louizos, Christos, Kevin Swersky, Yujia Li, Richard Zemel, and Max Welling. 'The Variational Fair Autoencoder'. *International Conference on Learning Representations* (2016).

Madras, David, Elliot Creager, Toniann Pitassi, and Richard Zemel. 'Learning Adversarially Fair and Transferable Representations'. *arXiv* (2018).

Mason, Ron and Roberta Julian. 'Analysis of the Tasmania Police Risk Assessment Screening Tool (RAST), Final Report'. *Tasmanian Institute of Law Enforcement Studies, University of Tasmania* (2009).

McNamara, Daniel, Cheng Soon Ong, and Bob Williamson. 'Provably Fair Representations'. *arXiv* (2017).

Menon, Aditya and Robert Williamson. 'The Cost of Fairness in Binary Classification'. *Conference on Fairness, Accountability and Transparency* (2018).

Messing, Jill Theresa, Jacquelyn Campbell, Janet Sullivan Wilson, Sheryll Brown, and Beverly Patchell. 'The Lethality Screen: the Predictive Validity of an Intimate Partner Violence Risk Assessment for Use by First Responders'. *Journal of Interpersonal Violence* 32.2 (2017).

Mitchell, Shira and Jackie Shadlen. 'Mirror Mirror: Reflections on Quantitative Fairness' (2018). https://speak-statistics-to-power.github.io/fairness/.

Nayaranan, Arvind. *Tutorial: 21 Fairness Definitions and their Politics*. Conference on Fairness, Accountability and Transparency, https://www.youtube.com/watch?v=jIXIuYdnyyk. 2018.

NSW Bureau of Crime Statistics and Research. *Re-offending Statistics for NSW*. 2018.

NSW Police Force. 'NSW Police Force Corporate Plan 2016-18' (2016).

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2017.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Weinberger. 'On Fairness and Calibration'. *Advances in Neural Information Processing Systems* (2017).

Rawls, John. *A Theory of Justice*. Harvard University Press, 1971.

Rice, Marnie, Grant Harris, and Zoe Hilton. 'The Violence Risk Appraisal Guide and Sex Offender Risk Appraisal Guide for Violence Risk Assessment'. *Handbook of Violence Risk Assessment*. Routledge, 2010.

Rice, Marnie and Harris, Grant. 'Violent Recidivism: Assessing Predictive Validity'. *Journal of Consulting and Clinical Psychology* 63 (1995).

Sentas, Vicki and Camilla Pandolfini. 'Policing Young People in NSW: A Study of the Suspect Targeting Management Plan'. *Youth Justice Coalition* (2017).

The Healing Foundation and White Ribbon Australia. 'Towards an Aboriginal and Torres Strait Islander Violence Prevention Framework for Men and Boys' (2017).

Trautmann, Stefan and Gijs van de Kuilen. 'Process fairness, outcome fairness, and dynamic consistency: Experimental evidence for risk and ambiguity'. *Journal of Risk and Uncertainty* 53 (2016).

Wierzbicka, Anna. *English: Meaning and Culture*. Oxford University Press, 2006.

Wijenayake, Senuri, Timothy Graham, and Peter Christen. 'A Decision Tree Approach to Predicting Recidivism in Domestic Violence'. *ArXiv* (2018).

Williamson, Robert and Aditya Krishna Menon. 'The Cost of Fairness in Binary Classification'. *Proceedings of Machine Learning Research* 81 ().

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna Gummadi. 'Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment'. *International Conference on World Wide Web* (2017).

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 'Learning Fair Representations'. *International Conference on Machine Learning* (2013).

Zeng, Jiaming, Berk Ustun, and Cynthia Rudin. 'Interpretable Classification Models for Recidivism Prediction'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.3 (2017).