Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

New Forms of Discrimination Based on the Aggregation of Data and their Effect on the Reliability and Fairness of Predictive Machine Learning Algorithms

This input paper was prepared by Professor James Maclaurin and Dr John Zerilli (University of Otago)

Suggested Citation

Maclaurin, J and Zerilli, J (2018). New Forms of Discrimination Based on the Aggregation of Data and their Effect on the Reliability and Fairness of Predictive Machine Learning Algorithms. Input paper for the Horizon Scanning Project “The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing” on behalf of the Australian Council of Learned Academies, www.acola.org.

New forms of discrimination based on the aggregation of data and their effect on the reliability and fairness of predictive machine learning algorithms.

by *James Maclaurin* and *John Zerilli*

The past decade has witnessed an unprecedented acceleration in both the sophistication and uptake of various algorithmic decision tools. From music and TV show recommendations, product and political advertising and opinion polling, to medical diagnostics, university admissions, job placement and financial services, the range of the potential application of these technologies is truly vast. In New Zealand this has led to widespread and rapid commercial adoption as well as use in government agencies including Corrections, Immigration, IRD, ACC [1, 2] and in a wide variety of contexts now understood as "Social Investment" [3, see esp. chs 7-13].

Champions of AI promote it as an upgrade to human reasoning: it is more accurate, and therefore more efficient, as well as more objective, and therefore more fair. This latter inference is problematic—that algorithmic decision-making tools are more objective because they are less biased than human decision-makers and that this objectivity enhances the justice and fairness of the decisions they make. Such assertions suggest that legal protection against unfair discrimination might not be relevant to 'objective' algorithmic decision-making, but recent research contradicts this inference. Human prejudice and algorithmic bias differ in character, but both are capable of generating unfair and discriminatory decisions. Tackling this problem will be particularly challenging owing to the contested nature of both fairness and discrimination.

To assess the risks of bias in automated decision-making, one must begin by looking at bias in human decision-making. Research into human decision-making has generated many important results over the past thirty years [4]. It is now understood that human prejudice is the result of various failures of reasoning [5]. For example, we often reason probabilistically from very small samples, and we regularly fail to update our beliefs in light of new information [6, 7]. At other times we abandon probabilistic reasoning altogether, relying instead on "generic" reasoning [8], judging that groups have particular characteristics irrespective of information about the frequency of those traits [9]. These generic judgements are pernicious as they are largely insensitive to evidence [10, 11]. For example, long-held beliefs about the criminality of ethnic minorities are not usually overcome by merely supplying evidence of the inaccuracy of such beliefs [12]. Moreover, emotions exert a powerful influence on human decision-making [13] and negative emotions like fear make us particularly prone to prejudice.

The accuracy of human decision-making is further decreased by a wide variety of psychological heuristics and biases. "Anchoring" and "framing" effects such as "availability" and "proximity" biases cause more recent or prominent events to exert disproportionate influence on human problem solving [4]. The tendency to see false correlations where none

exist is also well documented [14, 15]. The bias is at its strongest when a human subject is having to deal in small probabilities [4]. Furthermore, constraints imposed by short-term memory capacity limit our abilities at multi-factorial reasoning [16]. Because it is in the nature of complex decisions to present multiple relationships among many issues, our inability to assess these factors concurrently constitutes a significant limitation on our capacity to process complexity. The dangers of human bias are insidious, because "contemporary forms of prejudice are often difficult to detect and unknown to the prejudice holders" [17].

Discrimination and the law: Poor reasoning about group characteristics has caused historical injustice to marginalised groups. Partly for these reasons, and partly because it is unfair to further penalise disadvantaged groups [18], New Zealand's Human Rights Act 1993 protects citizens from grounds of reasoning including race, sex, and age in high risk circumstances such as employment and banking. This and various other pieces of legislation including the Bill of Rights Act, give New Zealand courts considerable power to remedy discrimination. In Australia, various federal laws also prohibit various discriminatory grounds of reasoning: the Racial Discrimination Act, the Sex Discrimination Act (protecting also gender, marital status and sexual orientation), the Age Discrimination Act and the Disability Discrimination Act. But it should be noted that prejudice and resulting discrimination also affect the operation and institutions of the law itself. Recent research suggests that the tendency to be unaware of one's own predilections is present even in those with regular experience of having to handle incriminating material in a sensitive and professional manner. In a recent review of psycho-legal literature comparing judicial and juror susceptibility to prejudicial publicity, the authors note that although "an overwhelming majority of judges and jurors do their utmost to bring an impartial mind to the matters before them…even the best of efforts may nonetheless be compromised" [19]. They write that "even accepting the possibility that judges do reason differently than jurors, the psycho-legal research suggests that this does not have a significant effect on the fact-finding role of a judge," and that "in relation to prejudicial publicity, judges and jurors are similarly affected."

So the problem of discrimination is widespread and complex, and up until now we have had legal protections that are generally accepted to be effective even though it is difficult to assess their actual efficacy on the accuracy and fairness of public decision-making. The use of such tools rests on the assumption that behaviours and experiences are universal and measurable. But even the use of standardized tools—or "structured professional judgments," as they are known—present a bias in how individuals are perceived, how behaviours are formulated, and how decisions are informed [25, 26]. It is in this context that algorithmic decision tools have been vigorously promoted [20-24].

Algorithmic bias: Amplifying these concerns, recent studies suggest that algorithmic decision tools may fail to live up to their promise of reducing harmful bias in decision-making [27-31]. Their probabilistic accuracy may in fact militate *against* fairness in most cases [32, 33].

It is useful here to distinguish intrinsic and extrinsic bias in decision-making systems. Intrinsic bias is built-in from scratch or results from inputs causing permanent change in the

system's structure and rules of operation. A Human Resources system designed by a male team to implement a set of rules that fail to accommodate the needs of female employees is intrinsically biased in its design. Ingrained unconscious prejudice in human reasoners that is effectively impervious to counter-evidence is also intrinsic. Extrinsic bias, on the other hand, derives from a system's inputs in a way that does not effect a permanent change in the system's internal structure and rules of operation. The output of such systems might be inaccurate or unfair but the system remains 'rational' in that new evidence is capable of correcting the fault.

What we usually call prejudice in humans is intrinsic bias [12, 17, 34] although of course misinformed humans can 'rationally' form inaccurate or unfair beliefs. Bias in algorithmic decision-makers can similarly be either intrinsic or extrinsic but differs in character from the corresponding human failings.

Intrinsic algorithmic bias can be the result of prejudiced developers or of ill-conceived software development (as in the Human Resources example above). Alternatively, intrinsic biases can arise from the inherent constraints imposed by the technology itself [35]. The way we represent data might have unexpected effects on the output of an algorithm, as when an algorithm that polls companies represented in an alphabetical list leads to increased business for those earlier in the alphabet [36]. Intrinsic algorithmic bias can also be the result of programming errors, as when poor design in a random number generator causes particular numbers to be favoured [36]. Some intrinsic bias is fundamentally historical, as when an algorithm is tied to rules that reflect current science, law or social attitudes.

The recent explosion in the use of artificial intelligence is largely driven by the development of algorithms that are not rule-based in the style of expert systems, but instead are capable of learning. Such "deep learning" networks can avoid intrinsic bias insofar as they can learn from their mistakes; but the cost of being able to learn is vulnerability to extrinsic bias. This has become a pressing issue in the development of ethical AI [35, 37].

In New Zealand, as elsewhere, algorithms designed to be accurate and fair routinely assess our creditworthiness, our desirability as employees, our reliability as tenants, and our value as customers. Extrinsic bias results from the fact that such apparently objective tools derive their power from historical data and hence actually aggregate decisions made by the very people whose potentially biased decision-making we are seeking to supplant [38]. Errors and biases latent in this "dirty data" tend to be reproduced in the outputs of machine learning tools [39, 40]. This is a significant problem, and one that is compounded by copyright and intellectual property laws, which presently limit the access users have to better quality training data [41]. Most extrinsic bias arises from the use of unrepresentative data sets. For instance, face recognition systems trained predominantly on Caucasian faces might reject the passport application photos of Asian persons, whose eyes appear closed [42]. Speech recognition systems, too, are known to make more mistakes decoding female voices than male ones [43]. Such situations arise from a failure to include members of diverse social groups in training data. The obvious solution is to diversify the training sets [44, 30] although there are both political and legal barriers in the way of this [41]. Moreover, the diversification of training

data presents a difficult technical problem. Demographic parity is achieved when a data set is equally representative of two groups (e.g. men and women), but where we are trying to be fair with regard to many different characteristics it is impossible to achieve demographic parity for all of them at once. Also, if the data available is strongly skewed in favour of a particular demographic group, discarding data in order to achieve demographic parity is likely to decrease the overall accuracy of the system [32].

Not all dirty data suffers from being unrepresentative. COMPAS is a commercial tool used in the criminal justice system which aids decisions about, amongst other things, parole. COMPAS scores, based on questionnaires filled out by prisoners, are predictive of risk of reoffending, but a recent study in the US shows a strong correlation between COMPAS score and race [45]. African Americans routinely have higher scores and so find it harder to get parole. The effects of historical injustice are writ large in such statistics. African Americans are likely to have lower incomes, to live in crime-ridden neighbourhoods, and to have diminished educational opportunities. This vicious circle is exacerbated by previous discriminatory patterns of policing [45, 29, 30]. This bias does not originate from unrepresentative data, which could be corrected by including more diverse ethnic groups in the training set. It stems from intrinsic human bias, with machines simply inheriting the bias from prevailing social conditions. So an algorithm that accurately predicts recidivism also unfairly penalises an already disadvantaged group. Moreover, because of these persistent correlations between race and disadvantage, modern AI, harnessing big data and machine learning, persistently detects race even when it receives no data specifically about this protected category [46].

Meanwhile work in data science shows that we can develop algorithms that are, in some sense, "fairer." The challenge, however, is that different notions of "fairness" are in conflict, meaning that it appears to be impossible to be "fairer" in every sense of that term [33, 20, 47]. Complicating the matter further, public safety and fairness also collide. As Corbett-Davies et al. [33] conclude after a rigorous statistical examination of the issue: "satisfying common definitions of fairness means one must in theory sacrifice some degree of public safety….Maximizing public safety requires detaining all individuals deemed sufficiently likely to commit a violent crime, regardless of race…There is…an inherent tension between minimizing expected violent crime and satisfying common notions of fairness."

We also note recent concern that algorithmic decision-making tools employing big data are effectively profiling New Zealanders in that they identify correlations between *apparently* unrelated and irrelevant characteristics and make predictions about behaviour at a group-level [48]. Thus, the individual is comprehended based on similarity to other people identified by the algorithm, rather than on their actual behaviour [49]. It is an open question whether an algorithm that employs a person's postcode to determine their creditworthiness is less pernicious than one that uses their skin colour to infer criminality. The line between static and dynamic variables is questionable: we might assume that things like postcode or employment status are within the individual's power to change, but in practice, they may very well not be.

REFERENCES

1. Chapman Tripp (2016) "Determining our Future: Artificial intelligence opportunities and challenges for New Zealand: A call to action." Auckland Institute of Directors, downloaded from https://www.iod.org.nz/Governance-Resources/Publications/Artificial-Intelligence (24/1/18).
2. Boyd, M. & Wilson, N. "Rapid developments in artificial intelligence: How might the New Zealand government respond?" *Policy Quarterly,* 13 (4), pp. 36-43.
3. Boston, J. & Gill, D. (2017) *Social investment: A New Zealand policy experiment*. Wellington: Bridget Williams.
4. Pomerol, J.-C. & Adam, F. (2008) "Understanding human decision making: A fundamental step towards effective intelligent decision support." In: *Intelligent decision making: An AI-based approach*, eds. G. Phillips-Wren, N. Ichalkaranje & L.C. Jain, pp. 41-76. Berlin: Springer.
5. Arpaly, N.. (2003) *Unprincipled virtue: An inquiry into moral agency*. Oxford University Press.
6. Fricker. M.. (2007) *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.
7. Gendler, T. (2011) "On the epistemic costs of implicit bias." *Philosophical Studies* 156(1): 33-63.
8. Begby, E. (2013) "The epistemology of prejudice." *Thought* 2(2): 90-99.
9. Leslie, S. (2017) "The original sin of cognition: Fear, prejudice, and generalization." *Journal of Philosophy* 114: 393-421.
10. Greenwald, A., & Banaji, M. (1995) "Implicit social cognition: Attitudes, self-esteem, and stereotypes." *Psychological Review* 102 (1): 4-27.
11. Saul, J. (2013) "Implicit bias, stereotype threat, and women in philosophy." In: *Women in philosophy: What needs to change?*, eds. K. Hutchison & F. Jenkins. Oxford University Press.
12. Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016) "A meta-analytical integration of over 40 years of research on diversity training evaluation." Available at: http://scholarship.sha.cornell.edu/ articles/974
13. Damasio, A.R. (1994) *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam's Sons.
14. Piattelli-Palmarini, M. (1995) *La r'eforme du jugement ou comment ne plus se tromper*. Paris: Odile Jacob.
15. Tversky, A. & Kahneman, D. (1974) "Judgment under uncertainty: Heuristics and biases." *Science* 185: 1124–1131.

16. Pohl, J. (2008) Cognitive Elements of Human Decision Making Jens. In: *Intelligent decision making: An AI-based approach*, eds. G. Phillips-Wren, N. Ichalkaranje & L.C. Jain, pp. 3-40. Berlin: Springer.
17. Plous, S. (2003) The psychology of prejudice, stereotyping, and discrimination. In: *Understanding prejudice and discrimination*, ed. S. Plous, pp. 3-48. New York: McGraw-Hill.
18. Khaitan, T. (2015) *A theory of discrimination law*. Oxford University Press.
19. McEwen, R., Eldridge, J. & Caruso, D. (2018) "Differential or deferential to media? The effect of prejudicial publicity on judge or jury." *International Journal of Evidence and Proof* 22(2): 124–143.
20. Hardt, M., Price, E. & Srebro, N. (2016) "Equality of opportunity in supervised learning." *30th Conference on Neural Information Processing Systems* (NIPS 2016). Available at: https://arxiv.org/pdf/1610.02413v1.pdf
21. Palk, G.R., Freeman, J.E. & Davey, J.D. (2008) "Australian forensic psychologists' perspectives on the utility of actuarial versus clinical assessment for predicting recidivism among sex offenders." *Proceedings 18th Conference of the European Association of Psychology and Law, Maastricht, The Netherlands*.
22. Craig, L.A. & Beech, A. (2009) Best practice in conducting actuarial risk assessments with adult sexual offenders. *Journal of Sexual Aggression* 15(2): 193-211.
23. Baird, J. & Stocks, R. (2013) Risk assessment and management: Forensic methods, human results. *Advances in Psychiatric Treatment* 19: 358–365.
24. Lawing, K., Childs, K.K., Frick, P.J. & Vincent, G. (2017) Use of Structured Professional Judgment by Probation Officers to Assess Risk for Recidivism in Adolescent Offenders. *Psychological Assessment* 29(6): 652-663.
25. Department of Corrections (as of March, 2015).
26. Tamatea, A. J. (2016) "Culture is our business: Issues and challenges for forensic and correctional psychologists." In: *ANZFSS 23rd International Symposium on the Forensic Sciences: Together InForming Justice*. Auckland, New Zealand.
27. O'Neil, C. (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, NY, USA.
28. Angwin, J. (2016) "Making Algorithms Accountable". *ProPublica*, downloaded from https:// www.propublica.org/article/making-algorithms-accountable (24/1/18).
29. Lum, K. & Isaac, W. (2016) To predict and serve? Bias in police-recorded data. *Significance* October 2016: 14-19.
30. Crawford, K. & Calo, R. (2016) There is a blind spot in AI research. *Nature* 538: 311-313.
31. Shapiro, A. (2017) Reform predictive policing. *Nature* 541: 458-460.

32. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. (2016) Algorithmic decision making and the cost of fairness. Proceedings of KDD'17. Available at: https://arxiv.org/pdf/1701.08230.pdf
33. Corbett-Davies, S., Pierson, E., Feller, A. & Goel, S. (2017) A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post* October 17, 2016.
34. Allport, G.W. (1954) *The nature of prejudice*. Cambridge, MA: Addison-Wesley.
35. Friedman, B. & Nissenbaum, H. (1996) Bias in computer systems. *ACM Transactions on Information Systems* 14(3): 330–347.
36. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016) The ethics of algorithms: Mapping the debate. *Big Data and Society* 16: 1-21.
37. Johnson, J.A. (2006) Technology and pragmatism: From value neutrality to value criticality. *SSRN Scholarly Paper, Rochester, NY: Social Science Research Network.* Available at: http://papers.ssrn.com/abstract=2154654
38. Citron, D. & Pasquale, F. (2014) "The Scored Society: Due Process for Automated Predictions", *Washington Law Review*, 89 (1), pp. 1-33.
39. Barocas, S. & Selbst, A.D. (2015) Big data's disparate impact. *SSRN Scholarly Paper, Rochester, NY: Social Science Research Network.* Available at: http://papers.ssrn.com/ abstract=2477899
40. Diakopoulos, N. (2015) Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3(3): 398–415.
41. Levendowski, A. (2017) How copyright law can fix artificial intelligence's implicit bias problem. *Washington Law Review* (forthcoming). Available at: https://ssrn.com/abstract=3024938
42. Griffiths, J. (2016) New Zealand passport robot thinks this Asian man's eyes are closed. *CNN.com* December 9, 2016.
43. Tatman, R. (2016) Google's speech recognition has a gender bias. *Making Noise and Hearing Things* July 12, 2016.
44. Klingele, C. (2016) The promises and perils of evidence-based corrections. *Notre Dame Law Review* 91(2): 537-584.
45. Larson, J., Mattu, S., Kirchner, L., and J Angwin "How we analyzed the COMPAS recidivism algorithm", *ProPublica*, 2016, downloaded from https://www.propublica.org/article/how-we- analyzed-the-compas-recidivism-algorithm.
46. Edwards, L. and Veale, M. (2017) "Slave to the algorithm? Why a 'Right to an Explanation is probably not the remedy you are looking for", *Duke Law and Technology Review* (16). pp. 18-84.

47. Kleinberg, J., Mullainathan, S. & Raghavan, M. (2017) Inherent trade-offs in the fair determination of risk scores. *8th Conference on Innovations in Theoretical Computer Science* (ITCS 2017). Available at: https://arxiv.org/pdf/1609.05807.pdf

48. Bonnett, G. (2018) "Immigration NZ using data system to predict likely troublemakers", Radio New Zealand (https://www.radionz.co.nz/news/national/354135/immigration-nz-using-data-system-to-predict-likely-troublemakers)

49. Newell, S. & Marabelli, M. (2015) "Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification.' *Journal of Strategic Information Systems* (forthcoming).