Yee-Fui Ng and Stephen Gray***

DISADVANTAGE AND THE AUTOMATED DECISION

Abstract

Automated decision-making is becoming increasingly prevalent in Australia. Familiar examples include MyGov for tax and social security benefits, and the use of SmartGates when arriving in Australia. Yet vulnerable populations have been detrimentally affected by the Australian government's use of automated processes. This is illustrated by the Robodebt debacle, where errors of methodology of government decision-making resulted in incorrect or inflated debt calculations for over 470,000 individuals. This article focusses on the impact of automated decision-making on vulnerable individuals. It will closely examine automated decision-making in the context of social security, focussing on the Robodebt and ParentsNext programs, as well as the recent incursions of automated decision-making into income management for vulnerable youth and changes proposed under the National Disability Insurance Scheme. Further, this article will also consider the related issues raised by automated decision-making in the context of criminal justice, from the investigatory stage — including facial recognition technology and police databases to identify offenders — through to sentencing. Finally, this article will develop guiding principles to protect the rights of vulnerable populations. These include safeguards at all stages of the automated decision-making process, including the design, implementation and evaluation of new technologies.

I Introduction

Like many societies worldwide, Australia is inching into territory once occupied only by science fiction writers — what Simon Chesterman called the 'robot century'.[1] From our romantic lives[2] to our shopping habits, our biometrics

* BCom, LLB (Hons) (Melb); PhD (Monash); Associate Professor, Faculty of Law, Monash University; Acting Director, Australian Centre for Justice Innovation.

** BA, LLB (Hons) (Monash); LLM (Melb); PhD (Monash); Senior Lecturer, Faculty of Law, Monash University; Deputy Director, LLB, Monash University.

1 Simon Chesterman, 'The Robot Century' (Blog Post, 25 January 2021) <https://simonchesterman.com/2021/01/25/the-robot-century/> ('The Robot Century').

2 David Tuffley, 'Love in the Time of Algorithms: Would You Let Artificial Intelligence Choose Your Partner?', *The Conversation* (online, 18 January 2021) <https://theconversation.com/love-in-the-time-of-algorithms-would-you-let-your-artificial-intelligence-choose-your-partner-152817>.

and our criminal records, electronic information about our personal characteristics, life histories and preferences is captured by governments or private corporations.[3] It may then be used for an increasing range of purposes, and to make increasingly significant decisions, of which the average citizen remains mostly unaware.[4] Some of these purposes are clearly benign and beneficial, bringing economic benefit and augmenting human potential.[5] Others raise issues that impact on fundamental rule of law values, including transparency and consistency, and equality before the law.[6] There is a question as to whether it is possible to design guiding principles which will enable us to disentangle the various complex and competing considerations at stake, in deciding whether the benefits of a particular form of decision-making outweighs its dangers. The difficulties of such a task are obvious — but without any attempt, we forfeit the chance to walk with open eyes into a new era of automated decision-making, and instead stumble 'zombie-like' into it,[7] blind.

While other literature in this field has focussed on the effects of automated decision-making on rule of law or public law values,[8] or on discrete aspects of decision-making such as social security law,[9] we consider the particular implications of this fast-developing field for one group of citizens: the vulnerable, or those 'able to be cast as outsiders rather than as rights-bearing citizens'.[10] Vulnerable persons are simultaneously most likely to be adversely affected by automated decision-making, due partly to the superficially neutral but practically discriminatory assumptions built

3 Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (Hachette, 2019) 11.

4 Ibid. According to Shoshana Zuboff, ordinary life is now so deeply immersed and saturated in the machinery of surveillance capitalism, and our dependency on it so total, that it 'produces a psychic numbing that inures us to the realities of being tracked, parsed, mined, and modified'.

5 See 'Singapore's Approach to AI Governance', *Personal Data Protection Commission Singapore* (Web Page, 21 June 2022) <https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework>.

6 Monika Zalnieriute, Lyria Bennett Moses and George Williams, 'The Rule of Law and Automation of Government Decision-Making' (2019) 82(3) *Modern Law Review* 425, 427.

7 Philip Alston, *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, 74th sess, Agenda Item 70(b), UN Doc A/74/493 (11 October 2019) 21 [77] ('*Report of the Special Rapporteur*').

8 See, eg: Zalnieriute, Bennett Moses and Williams (n 6); Will Bateman, 'Algorithmic Decision-Making and Legality: Public Law Dimensions' (2020) 94(7) *Australian Law Journal* 520.

9 Terry Carney, 'Artificial Intelligence in Welfare: Striking the Vulnerability Balance?' (2020) 46(2) *Monash University Law Review* 23 ('Artificial Intelligence in Welfare'); Asher Wright and Yee-Fui Ng, 'Services Australia's Single Touch Payroll Program: The Enduring Legacy of Robodebt, or a Fundamentally Different System?' (2022) 33(2) *Public Law Review* 127.

10 Carney, 'Artificial Intelligence in Welfare' (n 9) 25.

into machine algorithms, and also — due in part to lack of literacy or IT skills — are least likely to be able to challenge such decisions.

We argue that automated decision-making needs to take particular account of the interests of the vulnerable by paying close attention to the guiding principles of empowerment, harm minimisation and transparency. We contend that guiding principles need to be built into any analysis of whether to adopt a new form of automated decision-making, or to continue its use. This requires more than a simple checklist, but rather a substantial consideration of the real and potential effects of the proposal upon the vulnerable.

Our analysis proceeds in four stages. First, Part II considers definitional issues, and in particular, what automated decision-making actually is, and what forms of it are relevant for the purposes of this article. This Part also briefly outlines accepted understandings of vulnerability. Part III then evaluates automated decision-making in the context of social security, focussing on the government programs Robodebt and ParentsNext, as well as the recent incursions of automated decision-making into income management for vulnerable youth and changes proposed under the National Disability Insurance Scheme ('NDIS'). Part IV considers the issues raised by automated decision-making in the context of criminal justice, from the investigatory stage, including facial recognition technology and police databases to identify potential offenders, through to sentencing. Finally, Part V proposes key elements we suggest are necessary to protect the vulnerable in the context of automated decisions.

II Definitions

A *Automated Decision-Making*

Automated decision-making involves the operation of artificial intelligence ('AI'). 'AI is a compendious and fluid term' that can be classified depending on the type of model or process used.[11] Although there are narrower definitions,[12] this article takes an expansive view of AI to encompass a wide-ranging constellation of technologies, from the more basic expert systems that merely automate decision-making, to the more sophisticated forms of machine learning systems that can make predictions or decisions using machine or human-based inputs.[13] Expert systems are

11 Terry Carney, 'Automation in Social Security: Implications for Merits Review?' (2020) 55(3) *Australian Journal of Social Issues* 260, 261 ('Automation in Social Security').

12 For example, the OECD definition of 'AI system' confines it to machine learning systems. See OECD, *Recommendation of the Council on Artificial Intelligence* (22 May 2019) art I: '[a]n AI system is a machine-based system that can … make predictions, recommendations, or decisions'.

13 This broader approach is consistent with that taken by Australian law reform bodies and scholars. See, eg: Australian Human Rights Commission, *Human Rights and Technology* (Final Report, 1 March 2021) 37 ('*AHRC Final Report*'); Carney, 'Automation in Social Security' (n 11).

rule-based deterministic systems that follow 'a series of pre-programmed rules written by humans', while predictive machine learning systems deploy 'rules that are inferred by the system from historic data'.[14] Expert systems such as Robodebt use pre-programmed rules to reach a decision, such as that a person is eligible for a benefit.[15] Governments have used various forms of such systems since the 1980s.[16] Indeed some decisions on benefits or fines are rule-based, and leave no discretion to the decision-maker, making them more suitable for automation. On the other hand, an example of a predictive system that determines whether a person is likely to be a recidivist, or commit a further crime in the future, is the United States' sentencing tool called the Correctional Offender Management Profiling for Alternative Sanctions ('COMPAS').[17] This type of decision is more problematic from a rule of law point of view, and doubts have arisen in the context of both criminal sentencing and the calculation of tax benefits, about whether a decision made without human input constitutes a legal 'decision' at all.[18]

As Tania Sourdin has proposed, one way of classifying types (or taxonomies) of decisions made using AI involves three main categories: (1) the 'supportive' (helping and advising people involved in the system); (2) the 'replacement' (replacing functions previously carried out by humans); and (3) the 'disruptive' (a completely different form of decision-making).[19] Following Sourdin's taxonomy, Monika Zalnieriute, Lyria Bennett Moses and George Williams propose a spectrum moving from AI-assisted decisions, 'where automation plays a supporting role' to AI-made decisions, where decisions 'are made entirely by machines'.[20]

Forms of AI such as facial recognition technology are types of AI-assisted or 'supportive' technology, as a surveillance program or police database may be programmed to identify individuals regarded as suspect, or people fitting a profile considered suspicious.[21] This then supports the human 'decision' to arrest, search or charge. At the other end of the spectrum, an AI system may identify relevant

14 *AHRC Final Report* (n 13) 37.

15 Yee-Fui Ng et al, 'Revitalising Public Law in a Technological Era: Rights, Transparency and Administrative Justice' (2020) 43(3) *University of New South Wales Law Journal* 1041, 1068.

16 Nigel Stobbs, Dan Hunter and Mirko Bagaric, 'Can Sentencing Be Enhanced by the Use of Artificial Intelligence?' (2017) 41(5) *Criminal Law Journal* 261, 270.

17 Zalnieriute, Bennett Moses and Williams (n 6) 437.

18 Yee-Fui Ng and Maria O'Sullivan, 'Deliberation and Automation: When Is a Decision a Decision?' (2019) 26(1) *Australian Journal of Administrative Law* 21, 26–31; Ng et al (n 15) 1058.

19 Tania Sourdin, 'Judge v Robot? Artificial Intelligence and Judicial Decision Making' (2018) 41(4) *University of New South Wales Law Journal* 1114, 1117 ('Judge v Robot?').

20 Zalnieriute, Bennett Moses and Williams (n 6) 432.

21 Jake Goldenfein, 'Australian Police Are Using the Clearview AI Facial Recognition System With No Accountability', *The Conversation* (online, 4 March 2020) <https://theconversation.com/australian-police-are-using-the-clearview-ai-facial-recognition-system-with-no-accountability-132667>.

information based on predetermined criteria, such as the likelihood of compliance with a payment plan, 'and then make a decision based upon that information without engaging a human decision-maker', such as whether an applicant qualifies for a welfare benefit.[22]

B *Vulnerability*

Various definitions or taxonomies of the 'vulnerable' have also been proposed. One strand of literature contends that vulnerability is a universal human condition, with all humans being prone to chronic and episodic dependency,[23] and 'governments therefore have a responsibility to respond affirmatively to that vulnerability by ensuring that all people have equal access to the societal institutions that distribute resources'.[24] Another strand argues that some people or groups should be seen as especially vulnerable, and should therefore be entitled to special protection of the state.[25] Both strands of literature, however, can and do advocate for state action to protect vulnerable populations.

'Vulnerability is typically associated with victimhood, deprivation, dependency, or pathology.'[26] Other indicators of vulnerability include: (1) age; (2) low income; (3) unemployment; (4) having 'long-term disabilities'; (5) having 'a lower educational attainment'; (6) being a 'rural dweller …'; and (7) belonging to an 'ethnic minority …'.[27] Vulnerability has been characterised as a 'multidimensional concept', which may be focussed on both the changeable, individual characteristics of people, as well as on external factors.[28] For instance, people may be 'inherently' vulnerable (for example if they are mentally ill), or vulnerable in a 'situational' context (suffering from poverty), or in a 'pathogenic' manner (as a result of discriminatory or defective policies or laws)[29] — although, once again, these categories overlap. There are two overlapping senses of vulnerability: (1) consent-based; and (2) fairness-based.[30] Consent-based vulnerability will arise where people have

22 Zalnieriute, Bennett Moses and Williams (n 6) 432.

23 Martha Albertson Fineman, 'The Vulnerable Subject: Anchoring Equality in the Human Condition' (2008) 20(1) *Yale Journal of Law and Feminism* 1, 9; Jonathan Herring, *Vulnerable Adults and the Law* (Oxford University Press, 2016) 7–15.

24 Nina A Kohn, 'Vulnerability Theory and the Role of Government' (2014) 26(1) *Yale Journal of Law and Feminism* 1, 3.

25 See generally Jennifer Collins, 'The Contours of Vulnerability' in Julie Wallbank and Jonathan Herring (eds), *Vulnerabilities, Care and Family Law* (Routledge, 2015) 22.

26 Fineman (n 23) 8.

27 Tania Sourdin and Mirella Atherton, 'Vulnerability and Dispute Resolution in the Banking and Finance Sector' (2019) 9(1) *Social Business* 69, 73.

28 Ibid.

29 Wendy Rogers, Catriona Mackenzie and Susan Dodds, 'Why Bioethics Needs a Concept of Vulnerability' (2012) 5(2) *International Journal of Feminist Approaches to Bioethics* 11, 23–5.

30 Philip J Nickel, 'Vulnerable Populations in Research: The Case of the Seriously Ill' (2006) 27(3) *Theoretical Medicine and Bioethics* 245, 247–9.

diminished autonomy and are less able to safeguard their own interests (ie unable to fully consent).[31] Fairness-based vulnerability will arise among those people who are powerless in society and thus unable to protect their basic rights, therefore requiring special protections (ie to ensure fairness).[32]

Another related concept is that of disadvantage. Disadvantage is traditionally associated with poverty, but also has links to broader notions of deprivation (exclusion from the minimum acceptable standard of living in a society), low capabilities and functioning, and social exclusion from participation and social connections.[33] Disadvantage can be a result of problems linked to vulnerability, 'such as unemployment, discrimination, poor skills, low incomes, poor housing, high crime, bad health and family breakdown', which are mutually reinforcing and may have a compounding effect.[34]

We acknowledge the complexities and nuances of the literature. Our purpose is to consider the impacts of automated government decision-making on these populations.

III Automated Decision-Making in Social Security

The Australian government has been enthusiastic about the adoption of digital technologies in the sphere of social security. The Department of Human Services has proclaimed that it is 'continuing to transform its services by moving towards digital service delivery'.[35]

The historical design of Australian social security with its 'rigid eligibility categories and tight arithmetic logic' is conducive to the expert system form of automated decision-making, meaning that 'Australia has led the world' in the adoption of information technology in social security.[36] The first generation of rule-based AI

---

31 Ibid.

32 Ibid.

33 Rosalie McLachlan, Geoff Gilfillan and Jenny Gordon, 'Deep and Persistent Disadvantage in Australia' (Staff Working Paper, Productivity Commission, July 2013) 5; Amartya Sen, 'A Sociological Approach to the Measurement of Poverty: A Reply to Professor Peter Townsend' (1985) 37(4) *Oxford Economic Papers* 669, 670.

34 Social Exclusion Unit, *Breaking the Cycle: Taking Stock of Progress and Priorities for the Future* (Report, Office of the Deputy Prime Minister (UK), September 2004) 3 [3].

35 Department of Human Services, Submission No 66 to Senate Community Affairs References Committee, Parliament of Australia, *Design, Scope, Cost-Benefit Analysis, Contracts Awarded and Implementation Associated with the Better Management of the Social Welfare System Initiative* (2017) 1.

36 Terry Carney, 'Social Security Law: What Does the Politics of "Conditional Welfare" Mean for Review and Client Representation?' (2011) 46(3) *Australian Journal of Social Issues* 233, 236 ('Social Security Law').

involving the coding of simple legal rules with deductive reasoning steps has been increasingly widely adopted in the Australian social security context.[37] By contrast, machine learning technologies have not yet made any incursions into the Australian welfare system, compared to their use in other comparable liberal democracies.[38] The government has made legislative changes to enable the use of automated decision-making in social security determinations, including by inserting deeming provisions that deem decisions made by a computer to be made by the Departmental Secretary.[39]

This article will examine four aspects of automation of social security law: (1) the Robodebt saga; (2) the ParentsNext program; (3) income management for vulnerable youth; and (4) under the NDIS — and analyse these within the context of broader trends of social welfare policy in Australia.

A *Automated Debt Recovery: The 'Robodebt' Saga*

The catastrophic implementation of Centrelink's online compliance initiative (known by the derogatory moniker 'Robodebt') is the most visible recent example of automated decision-making by the Australian government. It involved a 'data matching method' for issuing and pursuing social security overpayment debts, extrapolating from the Australian Taxation Office's ('ATO') data 'the total amount and period over which employment income was earned and applying … that average to every separate fortnightly rate calculation period'.[40] From 2016, the automated system

> *automatically* issued letters to targeted welfare recipients asserting that they owe a debt for every case where they could not disprove the possible over-payment, effectively shifting the onus of proof from the department to the individual.[41]

---

37 See Carney, 'Artificial Intelligence in Welfare' (n 9).

38 For example, AI risk assessment tools that identify child welfare cases with a high probability of serious child injury or death are utilised in several US states: Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Press, 2018) 127.

39 *Social Security (Administration) Act 1999* (Cth) s 6A. The *Social Services and Other Legislation Amendment (Omnibus) Act 2020* (Cth) sch 1 makes a minor amendment to ensure that offences in the Act are against the Human Services Department, rather than an individual agency officer, to facilitate online service delivery. For discussion on deeming provisions, see Ng and O'Sullivan (n 18).

40 Ng et al (n 15) 1068.

41 Ibid (emphasis in original). See also 'Acceptable Evidence for Verifying Income when Investigating Debts 107-02040020', *Services Australia* (Web Page) <https://operational.servicesaustralia.gov.au/public/Pages/debts/107-02040020-01.html>.

These errors resulted in inaccurate debt calculations for more than 470,000 welfare recipients.[42] As Yee-Fui Ng argues:

> these large-scale incorrect calculations have reduced public trust in computer-supported government decision-making and led to grave repercussions for vulnerable low-socioeconomic debtors, including individuals experiencing severe mental health issues …[43]

The Robodebt debacle thus highlights the issues that arise if the design of AI technology is deficient, with poor use of data points in its income averaging calculations. It also illustrates the issues of legality in the irrational methodology of decision-making and accountability where Centrelink sought to shift the onus of proof, as well as the lack of transparency which led to recipients being unaware that the decision was automated and assuming that it had been checked and was accurate.[44] These issues were ventilated by a parliamentary committee[45] and the Commonwealth Ombudsman,[46] which raised various issues of procedural fairness and transparency.

In 2019, the basis for raising debts under the Robodebt program was held to be unlawful. In a test case by a debtor subject to a Robodebt, the Davies J of the Federal Court in consent orders in *Amato v Commonwealth* (*'Amato'*)[47] held that automated decisions made on the basis of income averaging alone (the Robodebt method) were irrational and therefore unlawful.[48] In *Prygodicz v Commonwealth [No 2]*

---

42 'Income Compliance Program Refunds', *Services Australia* (Web Page, 22 August 2022) <https://www.servicesaustralia.gov.au/individuals/subjects/information-about-refunds-income-compliance-program> ('Income Compliance Program Refunds').

43 Yee-Fui Ng, 'The Rise of Automated Decision-Making in the Administrative State: Are Kerr's Institutions Still "Fit for Purpose"?', *Australian Public Law* (Forum Post, 20 August 2021) <https://www.auspublaw.org/blog/2021/08/the-rise-of-automated-decision-making-in-the-administrative-state-are-kerrs-institutions-still-fit-for-purpose>.

44 Ng et al (n 15) 1068–70; Wright and Ng, 'Services Australia's Single Touch Payroll Program: The Enduring Legacy of Robodebt, or a Fundamentally Different System?' (n 9) 142; Paul Henman, 'Of Algorithms, Apps and Advice: Digital Social Policy and Service Delivery' (2019) 12(1) *Journal of Asian Public Policy* 71, 76–7.

45 Senate Community Affairs References Committee, Parliament of Australia, *Design, Scope, Cost-Benefit Analysis, Contracts Awarded and Implementation Associated with the Better Management of the Social Welfare System Initiative* (Report, June 2017) ix, 32–4, 107.

46 Commonwealth Ombudsman, *Centrelink's Automated Debt Raising and Recovery System: A Report about the Department of Human Services' Online Compliance Intervention System for Debt Raising and Recovery* (Investigation Report No 2, April 2017); Commonwealth Ombudsman, *Centrelink's Automated Debt Raising and Recovery System* (Implementation Report No 1, April 2019).

47 Order of Davies J in *Amato v Commonwealth* (Federal Court of Australia, VID611/2019, 27 November 2019).

48 Ibid 6 [9].

('*Prygodicz*'),[49] the Commonwealth agreed to settle a class action brought on behalf of 600,000 persons affected by Robodebt for $112 million, without any admission of liability.[50] The settlement agreement was approved by the Federal Court in June 2021.[51] The primary claims in *Prygodicz* were based on unjust enrichment and negligence.[52] However, as noted by Murphy J: 'In the course of the proceedings the Commonwealth admitted that it did not have a proper legal basis to raise, demand or recover asserted debts which were based on income averaging from ATO data.'[53]

In approving the class action settlement, Murphy J found that the applicant would have had good prospects of proving the debts issued were ultra vires and unlawfully imposed if the matter had proceeded to trial.[54] Further, Murphy J commented that due to the 'asymmetry in resources, capacity and information that existed between' the vulnerable social security recipients and the Commonwealth, the Commonwealth should have ensured that it had a proper basis to raise, demand and recover social security debts.[55] Accordingly, Murphy J declared that a decision that the applicant owed a debt under s 1223 of the *Social Security Act 1991* (Cth) was not validly made where the Robodebt income averaging method was utilised as the sole methodology.[56]

Following these cases, the Government agreed that from July 2020, it would refund all repayments of debts imposed using income averaging information from the ATO,[57] thereby refunding 470,000 unlawful debts.[58] In addition, Centrelink announced that it will no longer raise debts solely on the basis of the erroneous income averaging utilised in Robodebt, although the income averaging method is still utilised as one of the data points in raising debts.[59]

---

49 (2021) 173 ALD 277.

50 Ibid 281 [8].

51 Ibid 282 [14].

52 Ibid 279–80 [3].

53 Ibid 280 [4].

54 Ibid 311–12 [145].

55 Ibid 280–1 [7].

56 Ibid Annexure C.

57 'Income Compliance Program Refunds' (n 42).

58 'Justice for Hundreds of Thousands of People Affected by Robo-Debt', *Victoria Legal Aid* (Web Page, 29 May 2020) <https://www.legalaid.vic.gov.au/justice-hundreds-thousands-people-affected-robo-debt>.

59 'Media Hub: Online Income Compliance Programme Update', *Services Australia* (Web Page, 19 November 2019) <https://mediahub.servicesaustralia.gov.au/media/online-income-compliance-programme-update/>.

As Murphy J stated during the *Prygodicz* proceedings, the Robodebt debacle has been 'a very sorry chapter in Australian public administration'.[60] It represented a monumental failure of AI design and implementation, with incorrect debts levied on hundreds of thousands of alleged debtors. These governmental errors in automated decision-making have led to stress, anxiety, stigma, and even suicides within the vulnerable populations of alleged debtors.[61] Despite this, there is nothing in the *Amato* and *Prygodicz* decisions that would preclude the future use of automated decision-making in a welfare context, as long as decisions are made utilising a more accurate methodology.

B *ParentsNext Program and the Targeted Compliance Framework*

The ParentsNext program is a pre-employment program for young, at-risk, sole parents who receive parenting payments.[62] The program requires its participants to complete activities in a participation plan determined by a ParentsNext provider, with the aim of reducing welfare dependency, increasing female employment, and helping to close the gap for Indigenous employment.[63] Half of the households that receive Parenting Payment live in poverty, with an over-representation of single mothers in this cohort being at risk of financial stress.[64]

The ParentsNext program has been imposed in a punitive fashion through a targeted compliance framework involving automated sanctions that applied to more than 75,000 welfare recipients of parenting payments.[65] These automated sanctions include payment suspensions that are automatically triggered by a person receiving demerit points, leaving some families without money for daily essentials.[66] Approximately one in five participants in the program were subject to an automated sanction.[67] An intensive version of the ParentsNext program specifically targets regions with high concentrations of Aboriginal and Torres Strait Islander populations.[68] The Senate

60 Luke Henriques-Gomes, 'Robodebt Responsible for $1.5bn Unlawful Debts in "Very Sorry Chapter", Court Hears', *The Guardian* (online, 7 May 2021) <https://www.theguardian.com/australia-news/2021/may/07/robodebt-responsible-for-15bn-unlawful-debts-in-very-sorry-chapter-court-hears>.

61 Tom Stayner, 'Mothers Whose Sons Took Their Lives after Robodebts Detail Anguish in Heartbreaking Letters', *SBS News* (online, 18 August 2020) <https://www.sbs.com.au/news/mothers-whose-sons-took-their-lives-after-robodebts-detail-anguish-in-heartbreaking-letters>.

62 Carney, 'Artificial Intelligence in Welfare' (n 9) 37.

63 Senate Community Affairs References Committee, Parliament of Australia, *ParentsNext, Including Its Trial and Subsequent Broader Rollout* (Report, March 2019) 1.

64 Ibid 1 [1.5].

65 Ibid 2–3.

66 Ibid 55.

67 Ibid 13 [1.52].

68 Ibid 29 [2.54].

Community Affairs References Committee found that the program was causing 'anxiety, distress and harm' for many parents, including women escaping violence.[69]

The automation of sanctions in the ParentsNext program has created significant issues for those who were targeted by the scheme. The design of the sanctions system requires onerous and continuous digital reporting by recipients and generates demerit points for failing to self-report within the same day, creating an intense burden on the vulnerable sole parents on these programs.[70] As opposed to reporting to a human, who is able to discuss and review a recipient's individualised circumstances, the ParentsNext program automatically issued 'dubious demerit points unable to be [immediately] reviewed or corrected until they crystallised into a formal sanction'.[71] In addition, the targeted sanctions system resulted in Jobactive Employment Service providers (the private-sector providers tasked with operational responsibility for the ParentsNext system) taking the responsibility of compliance away from skilled caseworkers who assessed individual needs, and instead giving it to clerical staff who were only trained to apply rigid rules and were unskilled in applying discretion.[72] Due to this faulty system, Centrelink had to withdraw 50,000 warning strikes or potential suspensions.[73] Therefore, the implementation of automated sanctions prioritised the efficiency of cost-cutting, at the expense of the accuracy of decisions and protection of vulnerable single parents.

C *Income Management of Vulnerable Youth*

Income management is a controversial welfare policy that involves mandatory quarantine of a proportion of a person's welfare payments and imposing prohibitions on how those funds could be used (such as bans on purchasing drugs or alcohol). This has the stated aim of reducing violence or harmful behaviour and encouraging socially responsible behaviour.[74] Income management in Australia was implemented in three waves: the first wave targeted Indigenous welfare recipients as part of the 2007 Northern Territory Emergency Response; the second wave involved specific categories that 'automatically applied to welfare recipients residing in government targeted geographical locations'; and the third wave introduced the cashless debit card.[75] As of June 2020, approximately 37,000 Australians had been placed on

69 Ibid 71 [4.1].

70 Carney, 'Artificial Intelligence in Welfare' (n 9) 38.

71 Ibid 40.

72 Ibid.

73 Ibid.

74 Shelley Bielefeld, 'Cashless Welfare Transfers for "Vulnerable" Welfare Recipients: Law, Ethics and Vulnerability' (2018) 26(1) *Feminist Legal Studies* 1, 2.

75 Ibid 2–3. See also Stephen Gray, 'The Healthy Welfare Card: Indigenous Empowerment or "Remote Control"?' in Claire Spivakovsky, Kate Seear and Adrian Carter (eds), *Critical Perspectives on Coercive Interventions: Law, Medicine and Society* (Routledge, 2018) 135.

income management programs, mostly on a compulsory basis.[76] In August 2022, the incoming Labor Government introduced legislation to wind down the cashless debit card, but left the longer-standing 'Basics Card' in place.[77]

Despite the increasing use of income management, the evidence of its effectiveness in achieving positive outcomes is unclear.[78] In fact, some studies have shown that income management is not an effective means for reducing alcohol and drug abuse, does not achieve its objectives, and has had 'a negative effect on the autonomy, wellbeing, and financial management capacity of many of those forced onto the program'.[79]

Although much of income management is not yet automated, automated decision-making has been applied to vulnerable welfare recipients. In particular, it has been applied to disengaged youth living at home and young people receiving payments such as Special Benefit or Youth Allowance due to it being unreasonable for them to live at home.[80] This may reflect a possibility of further automation in the future.

The automation of income management for vulnerable youth was introduced in 2013, where the individualised social worker referrals 'were superseded by Centrelink data-mining software which automatically set in motion the *application* of income management to anyone meeting the "trigger payment" conditions'.[81] This automation has been criticised by the Commonwealth Ombudsman, who identified failures in the automated decision-making process. The Ombudsman found that the automation of decisions to extend vulnerable youth welfare payments beyond 12 months fettered the discretion of the decision-maker (which was an automated

---

76 Philip Mendes et al, 'Is Conditional Welfare an Effective Means for Reducing Alcohol and Drug Abuse? An Exploration of Compulsory Income Management across Four Australian Trial Sites' (2021) 56(2) *Australian Journal of Political Science* 153, 154.

77 See Social Security (Administration) Amendment (Repeal of Cashless Debit Card and Other Measures) Bill 2022 (Cth). For discussion, see Elise Klein, 'Has Labor Learnt from the Failure of the Cashless Debit Card?', *The Conversation* (online, 4 August 2022) <https://theconversation.com/has-labor-learnt-from-the-failure-of-the-cashless-debit-card-188065>.

78 Philip Mendes, Jacinta Waugh and Catherine Flynn, 'Income Management in Australia: A Critical Examination of the Evidence' (2014) 23(4) *International Journal of Social Welfare* 362, 370.

79 Philip Mendes et al (n 76) 166. See J Rob Bray et al, *Evaluating New Income Management in the Northern Territory* (Final Evaluation Report, September 2014) xxii.

80 Commonwealth Ombudsman, *Administration of Income Management for 'Vulnerable Youth': Department of Human Services* (Report No 1, February 2016) 6 [1.3], 12 [3.10]–[3.11] ('Ombudsman Report').

81 Terry Carney, 'Vulnerability: False Hope for Vulnerable Social Security Clients?' (2018) 41(3) *University of New South Wales Law Journal* 783, 813 (emphasis in original) ('Vulnerability').

system) by removing any capacity for discretion.[82] The Department of Human Services agreed to remove this aspect of automated decision-making.[83]

Further, the Ombudsman found that the automated decision-making process failed to give sufficient consideration to mandatory subjective legislative criteria that can only be made by a qualified decision-maker based on individual circumstances.[84] The Ombudsman found that this failure may 'exacerbate mental health issues or housing instability'.[85] However, Centrelink rejected the Ombudsman's recommendations to reinstate 'personal assessment and reasoned decision-making', claiming that the 'existing procedures and staff training would ensure that all criteria were fully considered',[86] meaning that this form of automated decision-making still exists.

D *The National Disability Insurance Scheme*

The NDIS provides funding of $22 billion annually to approximately 500,000 Australians with permanent and significant disability, including 'intellectual, physical, sensory, cognitive and psychosocial disability'.[87] It provides eligible participants with individualised resource packages 'under a personal plan geared to the needs of their particular disability', with the aim of maximising participant control.[88] The requirement to 'tailor … entitlements to the specific needs, living circumstances and preferences of the person' would imply social caseworker facilitation, rather than 'administrative routinisation … or digitisation and automation'.[89] However, the scheme has been criticised as undermining this vision and 'instead applying bureaucratic, standardised administrative logics', due to the imperatives of cost-cutting and meeting completion targets.[90]

In part, these shortcomings arise from the 'insurance' approach enshrined at the heart of the NDIS. According to Bruce Bonyhady, a foundational advocate of the NDIS, an insurance approach requires that 'expenditure is factored in over the life of an individual, and Scheme sustainability is measured by calculating the total future

82 Ombudsman Report (n 80) 18 [3.42].

83 Ibid 35.

84 The subjective criteria included whether being subject to income management might place a person's mental, physical or emotional wellbeing at risk: see ibid 13 [3.13]–[3.14].

85 Ombudsman Report (n 80) 14 [3.21].

86 Carney, 'Vulnerability' (n 81) 813. See also ibid 54–5.

87 'What is the NDIS?', *NDIS* (Web Page, 14 September 2021) <https://www.ndis.gov.au/understanding/what-ndis>.

88 Terry Carney et al, 'National Disability Insurance Scheme Plan Decision-Making: Or When Tailor-Made Case Planning Met Taylorism and the Algorithms' (2019) 42(3) *Melbourne University Law Review* 780, 782.

89 Ibid 783.

90 Ibid 780.

costs of all those who are insured'.[91] In theory, an insurance approach minimises overall costs by investing in short-term capacity-building for affected individuals, resulting in their long-term improvement.[92] As a result, in calculating whether proposed supports are 'reasonable and necessary' under the scheme, the decision-maker must consider, amongst other matters: 'whether there is evidence that the support will substantially improve the life stage outcomes for, and be of longterm benefit to, the participant';[93] and 'whether funding or provision of the support is likely to reduce the cost of the funding of supports for the participant in the long term'.[94] In practice, this means that funding may be denied to people with deteriorating or terminal conditions who are assessed as not likely to 'get better' as a result of funding. For example, in 2018 a scientist with a deteriorating motor neurone disability was denied funding for wheelchair and home modifications because he was assessed as having a poor life expectancy.[95]

Partially because of this approach, the digitisation of disability entitlements has proceeded apace. The government is introducing a disability app for NDIS participants to claim expenses,[96] and has trialled blockchain technology for NDIS participants' budgets.[97]

More controversially, the National Disability Insurance Agency was developing a mechanism dubbed 'robo-planning'. This mechanism would involve independent assessors determining whether people with disabilities are eligible for funding and the amount they receive based on a computer algorithm. This algorithm would then recommend a 'personalised budget' before a disabled person sees a human to plan their support program.[98] This went beyond the recommendation of the Auditor-

91 Bruce Bonyhady, 'Reducing the Inequality of Luck: Keynote Address at the 2015 Australasian Society for Intellectual Disability National Conference' (2016) 3(2) *Research and Practice in Intellectual and Developmental Disabilities* 115, 116–17.

92 Ibid 117.

93 *National Disability Insurance Scheme (Supports for Participants) Rules 2013* (Cth) r 3.1(b).

94 Ibid r 3.1(c).

95 Melinda James and Gavin Coote, 'Concerns about "Overloaded" NDIS Following Leading Scientist's Fight for Special Wheelchair', *ABC News* (online, 10 March 2018) <http://www.abc.net.au/news/2018-03-10/concerns-about-motor-neurone-disease-ndis/9533350>.

96 Stilgherrian, 'NDIS Sidesteps Blockchain in Government Kitchen Sink Debt-Chasing App', *ZDNet* (online, 12 April 2021) <https://www.zdnet.com/article/ndis-gets-a-government-app-with-blockchain-but-no-ethics/>.

97 'Blockchain Case Study: Commonwealth Bank and the NDIS', *Digital Transformation Agency* (Web Page) <https://www.dta.gov.au/help-and-advice/technology/blockchain/do-you-need-blockchain/blockchain-case-studies/blockchain-case-study-commonwealth-bank-and-ndis>.

98 Stilgherrian (n 96).

General to implement data matching to combat fraudulent claims,[99] as it would incorporate AI within the decision-making processes, rather than using the data merely as a measure to detect fraudulent claims.

Bonyhady has raised concerns that the government was 'removing the individualised nature of the NDIS and replacing it with an algorithm which will see many individuals receiving less support'.[100] Following discussions with the state and territory disability Ministers, the federal government decided to scrap the independent assessor process and work on a new model of assessment.[101]

The proposed robo-planning algorithm would have introduced an expert system into an area which ordinarily relies on tailor-made responses and would be the first example of automated decision-making intruding into the disability sector in determining benefits. This is alongside the potential for further data-mining technologies to be used to detect fraudulent claims — taking a step closer towards the Robodebt 'minefield'. It is a positive step that the federal government has decided not to go down this path. However, close attention needs to be paid to the new model that will be developed in its stead.

E *Implications of Automation in Social Services*

The use of automation in welfare sits alongside a broader trend towards punitive tendencies in social welfare policy in Australia. Social welfare policy is predicated on two main interpretations of social disadvantage: the individualistic or behaviouristic interpretation and the structuralist approach.[102] The individualistic or behaviouristic approach attributes a person's poverty and unemployment to personal characteristics, such as incompetence, immorality or laziness.[103] By contrast, the structuralist approach, which is based on social democratic philosophy, requires guarantees of social rights, including income security 'outside the operations of the labour

99 Australian National Audit Office, *National Disability Insurance Scheme Fraud Control Program* (Auditor-General Report No 50, 2019) 11, 41–2.

100 Bruce Bonyhady, 'An Analysis of the NDIA's Proposed Approach to Independent Assessments: A Response to the National Disability Insurance Agency (NDIA) Consultation' (Consultation Paper, Melbourne Disability Institute, February 2021). See also Denham Sadler, '"Robo-Planning" Will "Blow-Up" NDIS: Key Architect', *InnovationAus* (online, 26 April 2021) <https://www.innovationaus.com/robo-planning-will-blow-up-ndis-key-architect/>.

101 Nas Campanella, Leonie Thorne and Celina Edmonds, 'NDIS Minister Says Independent Assessments Model Is "Dead", in Win for Disability Advocates', *ABC News* (online, 9 July 2021) <https://www.abc.net.au/news/2021-07-09/ndis-disability-independent-assessments-model-dead-after-meeting/100277324>; 'Independent Assessments', *Parliament of Australia* (Web Page) <https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/National_Disability_Insurance_Scheme/IndependentAssessments>.

102 Carney, 'Automation in Social Security' (n 11).

103 Philip Mendes, 'Compulsory Income Management: A Critical Examination of the Emergence of Conditional Welfare in Australia' (2013) 66(4) *Australian Social Work* 495, 495.

market'.[104] The structuralist approach recognises 'that blaming the poor for their plight reflects a lack of compassion and is unlikely to improve their prospects'.[105]

In Australia, there was a fundamental shift in social security policy during the conservative Howard Government from a structural to an individualistic approach to social disadvantage.[106] The Howard Government strongly promoted a contractual or conditional mode of welfare involving 'mutual obligations', where social security payments provided to persons unemployed and above the working age involved return responsibilities for the recipient, such as actively looking for work.[107] This, coupled with the politics of 'welfare blame' (based on the individualistic approach that blames individuals for their social disadvantage), led to punitive sanctions for breaches of compliance requirements.[108]

Alongside the move towards a quid pro quo system for welfare benefits, the Australian social security system has also moved 'towards greater paternalism and away from honouring the Henderson Poverty Report's endorsement in the 1970s of autonomy in social security'.[109] In this vein, social security payments are increasingly imposed subject to 'conditional welfare' restrictions — that is, restrictions on the permissible expenditure of social security payments for certain categories of recipients, such as income management for vulnerable people.[110]

The increasingly widespread use of automation in governmental welfare programs points to the broader issue of the 'digital welfare state'.[111] The focus on punishing welfare non-compliance via automated systems reflects the State's individualistic/behaviouristic approach to disadvantage, to the detriment of vulnerable participants. Former United Nations Special Rapporteur on extreme poverty and human rights, Philip Alston, warned about the risks of the digital welfare state, where 'digital data and technologies … are used to automate, predict, identify, surveil, detect, target and punish'.[112] In a targeted welfare system, a heavy-handed approach that adopts automated systems, rather than personalised processes with appropriate oversight,

104 Ibid 496.

105 Ibid.

106 Ibid 495.

107 Peter Yeend, 'Mutual Obligation/Work for the Dole', *Parliament of Australia* (E-Brief, 15 June 2004) <https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/Publications_Archive/archive/dole>.

108 Terry Carney and Gaby Ramia, 'Welfare Support and "Sanctions for Non Compliance" in a Recessionary Work Labour Market: Post-Neoliberalism or Not?' (2010) 2(1) *International Journal of Social Issues* 277.

109 Carney, 'Vulnerability' (n 81) 784–5. See also Commission of Inquiry into Poverty, Parliament of Australia, *Poverty in Australia* (Parliamentary Paper No 210, April 1975) 304.

110 Carney, 'Social Security Law' (n 36) 235.

111 *Report of the Special Rapporteur* (n 7) 3[3].

112 Ibid.

adversely affects vulnerable people the most as these groups are most likely to be in the cohort of social security recipients.

Two main issues arise out of automated decision-making in welfare. The first is the enactment of harsh rules with very little or no discretion. The second is the use of computerised systems to implement said rules, which further exacerbates the harshness and reduces the possibility of accountability. It is more difficult to review automated decisions due to the opacity of AI.[113] Although the use of AI itself is not to blame, it makes it practically and politically easier to impose these systems on vulnerable populations. At a broader level, the implementation of automated systems as a cost-cutting measure — coupled with the individualist interpretation of social disadvantage — pushes social security mechanisms towards a more punitive approach and away from an individualised and tailored approach that accounts for the personal circumstances of vulnerable social security recipients.[114]

IV Automated Decision-Making in the Criminal Justice System

While the use of AI in the criminal justice system is relatively new, it raises a pertinent concern to criminologists. The use of AI may reinvigorate the debate on whether people choose to commit crime, or whether criminals are born not made — and to the extent that they are 'born', whether it ought to be possible to predict criminal propensity, whether by DNA profiling, crime data analysis or indeed by the discredited measurement of skulls.[115] More recent discussions have proceeded in the awareness of the disturbing risk that allowing courts to take genetic or related factors into account might 'diminish any notion of personal responsibility and enhance a fatalistic attitude'.[116] Nevertheless, research raising dangerous ethical possibilities continues to be pursued, at least in some parts of the world. For example, in China, research based on still facial images of 1,856 real persons — with half of these persons convicted of criminal offences — suggested that physical features such as lip curvature, eye inner corner distance and nose-mouth angle could be used to predict criminality.[117]

It is important to understand at the outset of this discussion that the criminal justice system impacts disproportionately on the vulnerable and disadvantaged.

113 Ng et al (n 15) 1041, 1052–4, 1073.

114 Simone Jane Casey, 'Towards Digital Dole Parole: A Review of Digital Self-Service Initiatives in Australian Employment Services' (2002) 57(1) *Australian Journal of Social Issues* 111.

115 Clive Coleman and Clive Norris, *Introducing Criminology* (Willan Publishing, 2000) 24.

116 Justice Michael Kirby, 'The Future of Criminal Law' (1999) 23(5) *Criminal Law Journal* 263, 271.

117 Aleš Završnik, 'Criminal Justice, Artificial Intelligence Systems, and Human Rights' (2020) 20(5) *ERA Forum* 571.

In Australia, the most egregious example is the impact on Indigenous Australians. The Royal Commission into Aboriginal Deaths in Custody report of 1991[118] highlighted the problem of Indigenous Australians being incarcerated at a rate many times higher than non-Indigenous Australians, with the rate of over-incarceration increasing in more recent years.[119] Indigenous Australians are over-represented at even higher rates in juvenile detention, making up approximately 96% of juveniles in detention in the Northern Territory.[120] One of the main causes of over-incarceration is the operation of police discretion. The Australian Law Reform Commission cited evidence in its 2017 report that 'Aboriginal and Torres Strait Islander young people are more likely to be arrested than their non-Indigenous counterparts even after other factors such as the offence, offending history and background factors are taken into account'.[121]

It is also frequently argued that Indigenous Australians are treated disproportionately harshly in sentencing — or, more subtly, that factors such as any prior record, together with the refusal of courts to take routine account of systemic deprivation and historical disadvantage in sentencing Indigenous Australians, have disproportionate and discriminatory impact on this vulnerable and disadvantaged group.[122] It is important, therefore, that forms of AI that are being introduced into the criminal justice system do not perpetuate or accentuate this disadvantage.

Forms of AI exist or are proposed at various stages of the criminal justice system, from the use of surveillance technology to identify possible criminal offending, to the use of automated profiling systems in police investigation, as well as its application to elements of the trial process itself, and finally to its controversial recent application, in sentencing. This article argues that these AI systems do indeed operate in a disproportionate and potentially discriminatory way against vulnerable people, including particularly Indigenous Australians. Special attention therefore needs to be paid to guiding principles in the development of this form of decision-making, to ensure that the rights and interests of vulnerable people are protected.

---

118 *Royal Commission into Aboriginal Deaths in Custody* (Final Report, 15 April 1991).

119 Australian Bureau of Statistics, 'Prisoners in Australia, 2016' (Catalogue No 4517.0, 8 December 2016); Australian Law Reform Commission, *Pathways to Justice: An Inquiry into the Incarceration Rate of Aboriginal and Torres Strait Islander Peoples* (Final Report No 133, 22 December 2017) 21 ('*ALRC Final Report*').

120 Stephen Gray, 'Scoring the Intervention: Fail Grades on Closing the Gap, Human Rights' (2016) 8(23) *Indigenous Law Bulletin* 10, 13.

121 *ALRC Final Report* (n 119) 453.

122 Thalia Anthony, Lorana Bartels and Anthony Hopkins, 'Lessons Lost in Sentencing: Welding Individualised Justice to Indigenous Justice' (2015) 39(1) *Melbourne University Law Review* 47, 75.

A *Police Investigations: Facial Recognition Technology*

Perhaps the best-known recent debate about the application of AI to the criminal justice process concerns facial recognition technology ('FRT'). This technology analyses an individual's geometric facial features, comparing an algorithm created from the captured image with existing data derived from a driver's licence, social media account, or police database.[123] It is increasingly being trialled or deployed in various contexts around Australia, including at airports,[124] banks and shopping centres,[125] as well as by private companies such as 7-Eleven Australia, which reportedly uses it for 'customer feedback'.[126] Additionally, Australian police agencies have reportedly used a private facial recognition service called 'Clearview AI', which looks for a match with an uploaded image of a person's face, through searching its database of several billion images collected from the web.[127] Police agencies initially denied they were using the service until a list of Clearview AI's customers was stolen and distributed online, showing both federal and state police.[128] No regulator exists to scrutinise or test the reliability or suitability of private technologies such as this, with the only testing apparently having been done in the United States by the company itself.[129] In late 2021, however, the Australian Information Commissioner, Angelene Falk, issued a determination that Clearview AI had breached the *Privacy Act 1988* (Cth), ordering the company to cease collecting facial images and biometric templates, and to destroy those it had

123 Joe Purshouse and Liz Campbell, 'Privacy, Crime Control and Police Use of Automated Facial Recognition Technology' (2019) 43(3) *Criminal Law Review* 188, 188.

124 Niamh Kinchin, 'AI Facial Analysis is Scientifically Questionable: Should We Be Using It for Border Control?', *The Conversation* (online, 24 February 2021) <https://theconversation.com/ai-facial-analysis-is-scientifically-questionable-should-we-be-using-it-for-border-control-155474>.

125 Liz Campbell, 'Why Regulating Facial Recognition Technology Is So Problematic: And Necessary', *The Conversation* (online, 26 November 2018) <https://theconversation.com/why-regulating-facial-recognition-technology-is-so-problematic-and-necessary-107284>.

126 Rick Sarre, 'Facial Recognition Technology Is Expanding Rapidly across Australia: Are Our Laws Keeping Pace?', *The Conversation* (online, 10 July 2020) <https://theconversation.com/facial-recognition-technology-is-expanding-rapidly-across-australia-are-our-laws-keeping-pace-141357>.

127 Jake Goldenfein, 'Australian Police Are Using the Clearview AI Facial Recognition System with No Accountability', *The Conversation* (online, 4 March 2020) <https://theconversation.com/australian-police-are-using-the-clearview-ai-facial-recognition-system-with-no-accountability-132667>.

128 Ibid. See also Ryan Mac, Caroline Haskins and Logan McDonald, 'Clearview's Facial Recognition App Has Been Used by the Justice Department, ICE, Macy's, Walmart and the NBA', *Buzzfeed News* (online, 28 February 2020) <https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement>.

129 Ibid.

already collected.[130] It is unclear whether this has yet occurred or whether police have ceased using the service.

Following a Council of Australian Governments ('COAG') agreement in 2017,[131] the federal government embarked on a process designed to legalise the collection and sharing of facial images and other identity information among government agencies Australia-wide. As Sarah Moulds points out, it also legalised some sharing with private organisations.[132] The package of legislation, known as the 'identity matching laws', aimed to set up a national 'hub' for the sharing of such information under the scrutiny of the Department of Home Affairs.[133] Its aims included identifying missing individuals, including in times of disaster or emergency, as well as combating identity crime and promoting community safety.[134] The Identity-Matching Services Bill 2019 (Cth) was subjected to scrutiny by the Parliamentary Joint Committee on Intelligence and Security, which produced a report in 2019 recommending significant changes to the proposed laws, including greater protection for individual privacy and human rights.[135]

These concerns are well-founded. They mirror the views expressed by the Human Rights Law Centre, which argued that the scheme could be used to identify any Australian, regardless of whether they were suspected of a crime. The Australian Human Rights Commission ('AHRC') has argued that the facial-matching software used could discriminate against particular racial or gender groups.[136] While it appears that the proposed legislation has now lapsed, in July 2021 the federal government announced an Intergovernmental Agreement on data sharing, which commits '[a]ll jurisdictions … to share data across jurisdictions as a default position, where it can be done securely, safely, lawfully and ethically'.[137]

In the United Kingdom, facial recognition technology has been used by police for some years, reportedly to monitor public spaces on a 'trial' basis. A system known

130 'Clearview AI Breached Australians' Privacy', *Office of the Australian Information Commissioner* (Web Page, 3 November 2021) <https://www.oaic.gov.au/updates/news-and-media/clearview-ai-breached-australians-privacy>.

131 Sarah Moulds, 'Who's Watching the "Eyes"? Parliamentary Scrutiny of National Identity Matching Laws' (2020) 45(4) *Alternative Law Journal* 266, 267.

132 Ibid; Department of Parliamentary Services (Cth), *Bills Digest* (Digest No 21 of 2019–20, 26 August 2019) 2, 3.

133 See: Identity-Matching Services Bill 2019 (Cth) cl 19; Explanatory Memorandum, Identity-Matching Services Bill 2019 (Cth) 2. See also Australian Passports Amendment (Identity-Matching Services) Bill 2019 (Cth).

134 Moulds (n 131) 267.

135 Parliamentary Joint Committee on Intelligence and Security, Parliament of Australia, *Advisory Report on the Identity-Matching Services Bill 2019 and the Australian Passports Amendment (Identity-Matching Services) Bill 2019* (Report, October 2019).

136 Moulds (n 131) 268.

137 National Cabinet, *Intergovernmental Agreement of Data Sharing between Commonwealth and State and Territory Governments* (9 July 2021) 1.

as 'Neoface' was first used by Leicestershire Police in 2014 — and again by other police forces at various carnivals and music festivals over the next four years — but no results of the 'trial' have been published.[138] The use of this technology was challenged on a human rights basis in *Wood v Commissioner of Police for the Metropolis*,[139] with the United Kingdom Court of Appeal finding that the police surveillance of a campaigner against the arms trade was in breach of art 8 of the *European Convention on Human Rights* ('*ECHR*').[140] A European Commission expert body recommended the banning or curbing of such technologies in June 2019.[141] In July 2019, the United Kingdom Information Commissioner's Office began an investigation into the use of FRT by law enforcement. However, despite the privacy and human rights concerns, the United Kingdom resisted pressure to enact statutory rules governing the deployment of FRT.[142]

In August 2020, a significant decision from the United Kingdom Court of Appeal changed the terms of this debate. *Bridges v CC South Wales Police*[143] concerned the lawfulness of the use of live automated facial recognition technology by the South Wales Police Force. The trial involved the use of a system called 'AFR Locate' (in fact, a development of Neoface),[144] which utilised surveillance cameras to capture digital images of members of the public. These were then compared to digital images of people on a police watchlist.[145] The watchlist included people with outstanding warrants, people who had escaped from custody, people suspected of crimes, missing persons, or people simply of interest to the police — including individuals regarded as vulnerable.[146] When the software identified a possible match, a police officer (or 'system operator') would compare the digital images to determine if a match had in fact been made.[147] The general public would be alerted to use of AFR

138 Purshouse and Campbell (n 123) 190.

139 [2009] EWCA Civ 414.

140 *Convention for the Protection of Human Rights and Fundamental Freedoms*, opened for signature 4 November 1950, 213 UNTS 221 (entered into force 3 September 1953) ('*ECHR*').

141 Elizabeth Denham, 'Live Facial Recognition Technology: Data Protection Law Applies', *Wired Gov* (Blog Post, 9 July 2019) <https://www.wired-gov.net/wg/news.nsf/articles/Blog+Live+facial+recognition+technology+data+protection+law+applies+10072019091000?open>; High-Level Expert Group on Artificial Intelligence, 'Policy and Investment Recommendations for Trustworthy AI' (26 June 2019) 20 [12.4] <https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence>.

142 Purshouse and Campbell (n 123) 203.

143 [2020] EWCA Civ 1058 ('*Bridges*').

144 Ibid [10].

145 Ibid [8].

146 Ibid [13].

147 Ibid [15].

technology at events where it was used. Images would be deleted from the system after (at most) 31 days.[148]

Bridges challenged the use of the technology on the basis that it contravened arts 8, 10 and 11 of the *ECHR*, as well as European data protection law, and s 149 of the *Equality Act 2010* (UK).[149] The Court of Appeal accepted that South Wales Police's use of the technology was an interference with Bridges' rights under art 8(1) of the *ECHR*, and was not 'in accordance with the law' for the purpose of art 8(2) of the *ECHR*. The Court's conclusion that there was an insufficient legal framework for the use of the technology was on the basis that too much discretion was left to individual police officers, and that '[i]t is not clear who can be placed on the watchlist nor is it clear that there are any criteria for determining where AFR can be deployed'.[150] It also considered that the use of the technology was in breach of public sector equality duties, in that the police had not done everything reasonable to be satisfied that the software used did not have a racial or gender bias.[151]

Thus, it is clear that police use of FRT in investigating crime or identifying suspects raises a significant set of privacy and human rights issues. In 2021, the AHRC Final Report on Human Rights and Technology recommended a moratorium on the use of FRT until legislation can be passed regulating its use and expressly protecting human rights.[152] Therefore, FRT needs to be carefully deployed in high stakes situations that impact upon a person's fundamental rights of life, liberty or property, such as in criminal investigations, in order to avoid wrongful arrests and detention.[153] Accordingly, FRT should only be deployed when the accuracy of the technology is confirmed for its intended purpose, and when there are strong legislative guidelines regulating its use, as well as the ability for individual appeals over errors from the use of this technology.

B *Police Powers and the Automated Decision*

As is well-known, police are entrusted with broad powers in the course of their overall duty to uphold and enforce the law. Their powers include the power to stop and question individuals suspected of involvement in crime, powers of entry, search and seizure, the power to take steps to prevent the commission or continuation of an offence, and the power of arrest. In making such decisions, police in recent years have turned increasingly to 'tools' which promise to identify those considered at increased risk of criminal behaviour, hence theoretically saving police time and

148 Ibid [18]–[19].

149 Ibid [32], [52].

150 Ibid [91]. See also Paul Schwartfeger, 'Automating Bias?' (2020) (August) *New Law Journal* 15.

151 *Bridges* (n143) [201].

152 *AHRC Final Report* (n 119) 115–16, 120.

153 Daniel E Ho et al, 'Evaluating Facial Recognition Technology: A Protocol for Performance Assessment in New Domains' (2021) 98(4) *Denver Law Review* 753, 754–6.

scarce resources.[154] On occasions, these 'tools' have included forms of automated decision-making.

One example of such a tool is the 'Suspect Target Management Plan' ('STMP') developed by New South Wales Police in order to 'identify, assess and target people "suspected of being recidivist offenders or responsible for emerging crime problems"'.[155] First developed in 1999, with a later iteration in 2005, it uses a 'quantitative risk assessment tool, designed to identify individuals' risk of re-offending'.[156] Once placed on the plan, an individual will be targeted by New South Wales Police officers, including by attending the person's 'house on a regular basis, and using police powers of stop and search, and move on directions, whenever police encounter the individual'.[157] A suspicion exists that the plan involves 'the use of particular algorithms, or risk assessment tools, to calculate a person's risk of offending or re-offending'.[158] However, it is impossible to know the extent to which automated decision-making forms part of the STMP, as the policy and operational arrangements surrounding the STMP are not made publicly available.

There is evidence that the 'STMP disproportionately targets young people, particularly Aboriginal and Torres Strait Islander people', and that it 'has the effect of increasing vulnerable young people's contact with the criminal justice system'.[159] The plan lacks a statutory basis, with the result that the STMP 'may be inadvertently diminishing police understanding of the lawful use of [their] powers'.[160] It is true that criticism of the police for targeting Indigenous people is not new. It was a major focus of the 1991 Royal Commission into Aboriginal Deaths in Custody, as noted above, and so may be said to pre-date the use of forms of automated decision-making. However, the introduction of forms of automated decision-making into the process by which police target, arrest, and charge people represents a disturbing new element. There is no transparency in either the algorithms used to identify people as targets, or in the very existence of automation as part of the decision-making process. The secrecy surrounding the STMP 'poses significant risks to effective and fair policing', and 'may in particular intensify the conditions that escalate conflict between young Aboriginal and Torres Strait Islander peoples and police'.[161]

---

154 For example, as Vicki Sentas and Camilla Pandolfini point out, the Suspect Target Management Plan in New South Wales 'seeks to effectively target command resources' to address crime problems: Vicki Sentas and Camilla Pandolfini, *Policing Young People in NSW: A Study of the Suspect Targeting Management Plan* (Report, Youth Justice Coalition New South Wales, 2017) 5 [1.4] quoting New South Wales Ombudsman, *The Consorting Law: Report on the Operation of Part 3A, Division 7 of the Crimes Act 1900* (Report, April 2016) 118.

155 Ibid 1, 5.

156 Ibid 5.

157 Ibid.

158 Ibid 6.

159 Ibid 1.

160 Ibid 1.

161 Ibid 29.

Another example of police use of predictive risk assessment tools is a Queensland Police Service trial 'using artificial intelligence (AI) to determine the future risk posed by known domestic violence perpetrators'.[162] Relying on an algorithm developed using Queensland Police Service administrative data, the tool is designed to identify people regarded as 'high risk', and ensure police visit them at home before domestic violence or other crimes are committed. As with the STMP in New South Wales, there is a 'lack of transparency in the specific kinds of data analysed'.[163] More disturbingly, there is a concern that the use of AI 'could reinforce existing biases in the criminal justice system', creating 'an endless feedback loop between police and those members of the public who have the most contact with police', that is to say, Aboriginal and Torres Strait Islander people.[164] In other words, the use of forms of AI does not merely reinforce existing biases. It potentially aggravates them, because the system identifies and then further targets particular vulnerable individuals, increasing the likelihood of escalating conflict and ultimately more serious consequences.

C *Sentencing and the Automated Decision*

So far in Australia, there is little evidence that AI has become directly involved in the criminal sentencing process, with academic discussion mainly concerned with predictions about possible impacts of AI on the judicial function in the future.[165] In the United States, however, automated decision-making has become far more directly and practically enmeshed with criminal sentencing. The decision-making tool, COMPAS, is used in various United States jurisdictions to predict which convicted offenders pose the highest risk of re-offending.[166] COMPAS gives an individual a 'score' or risk assessment using an algorithm which processes or interprets the individual's personal characteristics and history, including criminal history, education, employment, age and substance abuse history, as well as criminal associates, pro-criminal attitudes and an 'antisocial personality'.[167]

However, it is not known how, precisely, the COMPAS risk assessment tool performs this task. This is because the tool itself, and the algorithm on which it is

---

162 Heather Douglas and Robin Fitzgerald, 'QLD Police Will Use AI To "Predict" Domestic Violence before It Happens: Beware the Unintended Consequences', *The Conversation* (online, 17 September 2021) <https://theconversation.com/qld-police-will-use-ai-to-predict-domestic-violence-before-it-happens-beware-the-unintended-consequences-167976>.

163 Ibid.

164 Ibid.

165 See, eg, Sourdin, 'Judge v Robot?' (n 19) 1115–16.

166 Zalnieriute, Bennett Moses and Williams (n 6) 437.

167 Hannah Bloch-Wehba, 'Access to Algorithms' (2020) 88(4) *Fordham Law Review* 1265, 1288–9.

based, remains a trade secret.[168] Nevertheless, in *State of Wisconsin v Loomis*,[169] the Supreme Court of Wisconsin approved the use of such tools, providing the judge did not fully delegate their decision-making function, and still considered the defendant's arguments on the question of future re-offending.[170]

Zalnieriute, Bennett Moses and Williams argue that the use of risk assessment tools such as COMPAS violates the principle of equality before the law. This is not because such tools explicitly use race or other impermissible characteristics in the calculation of risk, but rather because of 'the fact that automation can infer rules from historical patterns and correlations' in order to produce 'racially or otherwise biased assessments'.[171] An investigation by ProPublica in 2016 found that African-Americans were more likely than Caucasians to be given a false positive score. This result appears to flow, not from the explicit use of race as a variable, but from other information which itself may be the product of racial biases, such as numbers of Facebook 'likes', or the number of times a defendant has been stopped and questioned by police.[172] This form of decision-making is the product of a system that lacks transparency, denies defendants the opportunity to participate in the findings and processes of the court, and over which humans exercise insufficient supervisory control.[173]

In contrast, Nigel Stobbs, Dan Hunter and Mirko Bagaric argue that a broader use of automated decision-making could actually lead to improvements in current sentencing practice, which they consider to be unduly reliant on the inconsistent or even capricious decisions generated by human judges.[174] Such decisions, they argue, are particularly lacking in objectivity when they are the product of the intuitive or instinctive synthesis method of sentencing, which 'neither requires (nor permits) judges to set out with any particularity the weight (in mathematical or proportional terms) accorded to any particular consideration'.[175] Such sentences lack transparency, even to the point of amounting to arbitrary detention, and can lead to unpredictability and numerical inconsistency in sentencing, again perhaps to the point of eroding public confidence in the administration of justice.[176] They argue further that there is no evidence that increased discretion leads to greater fairness,

168 Zalnieriute, Bennett Moses and Williams (n 6) 437.

169 881 NW 2d 749 (Wis, 2016).

170 Ibid [56].

171 Zalnieriute, Bennett Moses and Williams (n 6) 452.

172 Ibid, citing Michal Kosinski, David Stillwell and Thore Graepel, 'Private Traits and Attributes Are Predictable from Digital Records of Human Behavior' (2013) 110(15) *Proceedings of the National Academy of Sciences* 5802.

173 Zalnieriute, Bennett Moses and Williams (n 6) 453.

174 Stobbs, Hunter and Bagaric (n 16) 265. See also Tania Sourdin, *Judges, Technology and Artificial Intelligence* (Edward Elgar, 2021) 72, 78–9 ('*Judges, Technology and Artificial Intelligence*').

175 Stobbs, Hunter and Bagaric (n 16) 265.

176 Ibid 266–7.

and that there is in fact evidence that discretion leads to sentencing decisions being impacted by subconscious bias, such as bias against Indigenous defendants or in favour of those with an attractive appearance.[177] This argument is supported by Aleš Završnik, who considered that human decision-making is 'often flawed', with 'stereotypical arguments and prohibited criteria such as race, sexual preference or ethnic origin often creeping … into judgments'.[178]

The difficulty with an automated sentencing process determined by a computer algorithm, as Stobbs, Hunter and Bagaric acknowledge, is that while the various considerations to be factored into sentencing are well-known, the precise weight to be accorded to each of those in an individual case is not easily discernible. Stobbs, Hunter and Bagaric assert that this considerable difficulty may be overcome by systematic research involving reading a large number of sentencing decisions in each jurisdiction, and then breaking them down in order to ascertain the precise weight accorded to various aggravating or mitigating circumstances.[179]

Thus, the form of automated decision-making proposed by Stobbs, Hunter and Bagaric goes considerably further than the COMPAS tool currently in use in the United States. In fact, the authors consider the risk assessment tool to measure the chance of future offending could be an 'additional feature' incorporated into the overall automated sentencing process.[180] The authors acknowledge the limited studies into the effectiveness of such tools, and the significant reservations regarding their accuracy.[181] However, with considerable confidence, the authors assert the possibility of crime predictive tools being 'far more accurate than unstructured judicial observations, so long as they are adapted to the local population in which they are to be used'.[182]

This confidence is arguably misplaced. Studies on the accuracy of COMPAS have produced mixed results.[183] For instance, Julia Dresel and Hany Farid found that COMPAS was only as accurate as an online poll of 400 random people without criminal or legal training,[184] while Zhiyuan Lin et al found that the COMPAS algorithm could perform better than a human when feedback on whether the person

177 Ibid 268.

178 Aleš Završnik, 'Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings' (2021) 18(5) *European Journal of Criminology* 623, 633.

179 Stobbs, Hunter and Bagaric (n 16) 271–2.

180 Ibid 272–3.

181 In China, a 'social credit' system has become more inextricably linked to judicial decision-making, with social credit data increasingly likely to be used 'to determine both credibility and outcomes': Sourdin, *Judges, Technology and Artificial Intelligence* (n 174) 75, 228.

182 Ibid 274.

183 Ibid 73–4.

184 Julia Dressel and Hany Farid, 'The Accuracy, Fairness, and Limits of Predicting Recidivism' (2018) 4(1) *Science Advances* eaao5580:1–5, 2.

had in fact reoffended was removed, as such data may not be available in real life.[185] At any rate, risk assessment tools already in use have been shown to operate in a discriminatory or unequal way against minority groups, despite the absence of any explicitly discriminatory criteria in the algorithm.[186] It is difficult to see how the algorithm itself could be made 'free of the discrimination that permeates the present sentencing regime',[187] given that the algorithm incorporates the same variables and criteria. It may also incorporate criteria likely to operate in an indirectly discriminatory way. For example, it may include a criterion that residents from a particular area have been convicted of crimes more than residents from another area, without acknowledging that those residents are more often from a marginalised group.[188]

Moreover, the algorithms appear to be written without the involvement of legally trained individuals, and thus the translation of rules into code may not reflect the correct interpretation of complex legislation, statutory presumptions, and case law.[189] The same criticism of 'intuitive sentencing' regimes administered by human judges may be made to an even greater degree of automated sentencing regimes, given the confidentiality of the algorithms on which they are based. Judges are public figures, subject to considerable scrutiny, unlike the anonymous programmers who have produced an algorithm.[190] An algorithm needs to be kept up-to-date to reflect changes in law and policy, which is something unlikely to occur in a program developed by a private company and sold to a decision-making authority.[191]

It is true that it is theoretically possible that 'all integers of the algorithm should be known to the court', as well as to the parties and the wider community.[192] However, this is unlikely in practice given that the algorithms currently in use, and the information on which they are based, have been developed by private companies for private profit. The assertion that an algorithm would somehow produce an 'objective' assessment of the multiple varying sentencing factors is reminiscent of the notion that a utilitarian calculus or an economic theory of law might produce mathematically precise and justifiable decisions. It is easy to assert in theory, but

185 Zhiyuan 'Jerry' Lin et al, 'The Limits of Human Predictions of Recidivism' (2020) 6(7) *Science Advances* eaaz0625:1–8, 1.

186 Zalnieriute, Bennett Moses and Williams (n 6) 452.

187 Stobbs, Hunter and Bagaric (n 16) 274.

188 Richard Berk and Jordan Hyatt, 'Machine Learning Forecasts of Risk to Inform Sentencing Decisions' (2015) 27(4) *Federal Sentencing Reporter* 222, 225.

189 Justice Melissa Perry, 'iDecide: Administrative Decision-Making in the Digital World' (2017) 91(1) *Australian Law Journal* 29.

190 For commentary on the scrutiny of judges: see, eg, Chief Justice Gerard Brennan, 'The Role of the Judge' (Speech, National Judicial Orientation Programme, 13 October 1996) 1 [5].

191 Deirdre K Mulligan and Kenneth A Bamberger, 'Procurement as Policy: Administrative Process for Machine Learning' (2019) 34(3) *Berkeley Technology Law Journal* 773, 817.

192 Stobbs, Hunter and Bagaric (n 16) 274.

seemingly impossible to show how it would work in the human world of judicial decision-making.

V Guiding Principles for Ensuring Automated Decision-Making Protects Vulnerable Populations

The use of AI in government and judicial decision-making can 'improve efficiency, certainty, predictability and consistency'.[193] However, as seen from the discussion of social security and criminal justice above, when used as a decision-making tool, AI has some key differences from human decision-making that must be considered. The first is the issue of transparency: due to the 'black box' nature of the system itself, as well as due to proprietary interests in the AI, which can protect the inner workings as trade secrets, the reasoning behind the decision is not always discoverable.[194] The second consideration relates to issue of algorithmic bias, where the training of machine learning programs has the possibility of ingraining existing biases in the AI (or even creating new ones).[195] Third, the issues of privacy and data protection must be considered due to the new challenges AI systems present.[196]

It is clear that automated government decision-making has advanced at a rapid pace in recent decades. The new public management ('NPM') movement that has swept through many Western democracies worldwide from the late 1970s has been responsible for an augmentation in AI technologies towards 'digital era governance'.[197] NPM involves inserting business-like principles into government, leading to a focus on performance and consequently the measuring, monitoring and auditing of agency outcomes.[198] This has translated into administrators being focussed on

193 Ng et al (n 15) 1049.

194 Andrew D Selbst and Solon Barocas, 'The Intuitive Appeal of Explainable Machines' (2018) 87(3) *Fordham Law Review* 1085, 1094.

195 Australian Human Rights Commission, *Using Artificial Intelligence To Make Decisions: Addressing the Problem of Algorithmic Bias* (Technical Paper, November 2020) ('*Using Artificial Intelligence to Make Decisions*'); UK Centre for Data Ethics and Innovation, *Review into Bias in Algorithmic Decision-Making* (November 2020) 119.

196 Sourdin summarises these issues as fairness, transparency and explainability, responsibility and accountability; robustness and reliability; privacy and trust; and safety and security: *Judges, Technology and Artificial Intelligence* (n 174) 237. In comparison, Paul Henman identifies issues of accuracy, bias and discrimination; legality, due process and administrative justice; responsibility, accountability, transparency and explainability; and power, compliance and control: Paul Henman, 'Improving Public Services Using Artificial Intelligence: Possibilities, Pitfalls, Governance' (2020) 42(4) *Asia Pacific Journal of Public Administration* 209.

197 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) 121, 122.

198 Paul Henman and Michael Adler, 'Information Technology and the Governance of Social Security' (2003) 23(2) *Critical Social Policy* 139, 148.

case numbers and targets, rather than the impacts of AI on vulnerable populations. Digitised administration may also 'be used to install new political or ideological agendas, such as job shedding, replacement of skilled with un/semi-skilled staff, enhanced managerial control of workers, and increasing surveillance and control of citizens'.[199] As a result, new technologies in government have been deployed in ways that sometimes detrimentally affect vulnerable populations. Targeted governmental intervention for vulnerable populations in Australia has been problematic. The focus on vulnerability is 'susceptible to abuse by powerful interests intent on increasing coercive, surveillance, discipline and disentitlement for those designated as "vulnerable"'.[200] It can expose vulnerable populations to invasive and paternalistic interventions.[201] This can be seen in the area of social security, where coercive paternalistic interventions that have proved to be deeply detrimental to welfare recipients have been imposed in the name of protecting the vulnerable.[202]

The use of AI in the welfare state has supported the 'informatisation' of organisations, including the surveillance of claimants through data matching procedures to identify welfare fraud and over-payments, as well as the monitoring and measuring of departmental staff rates of processing cases.[203] As Paul Henman argues: 'algorithms constitute a particular, predictive way of thinking about the practice of government, a practical way of governing the future in the present, and the present in the future'.[204] Nevertheless, the use of AI for vulnerable populations could be a positive force, as it may enhance customer service through the creation of a 'one stop shop' in welfare service provision by bringing together disparate information and organisations.[205] Expert systems can also improve customer service by 'helping claimants and advocates … better understand the reasons for an administrative decision', ensuring that 'claimants receive their full entitlement to benefit', and enhancing the accountability of welfare organisations.[206]

---

199 Paul W Fay Henman, 'Administrative Justice in a Digital World: Challenges and Solutions' in Marc Hertogh et al (eds), *The Oxford Handbook of Administrative Justice* (Oxford University Press, 2021) 459, 473.

200 Bielefeld (n 74) 1–2.

201 Herring (n 23) 35.

202 See, eg, Andrew Forrest, *The Forrest Review: Creating Parity* (2014) 103, where it is argued that there should be 'a cashless welfare card system, not just for vulnerable first Australians, but for vulnerable people across Australia'. See also Explanatory Memorandum, Social Security Legislation Amendment (Debit Card Trial) Bill 2015 (Cth) 3, which states the restriction on welfare recipients using the cashless debit card on alcohol, gambling and illegal drugs 'is to ensure that vulnerable people are protected from abuse of these substances, and associated harm and violence'.

203 Henman and Adler (n 198) 147–8.

204 Paul Henman, 'Governing by Algorithms and Algorithmic Governmentality: Towards Machinic Judgement' in Marc Schuilenburg and Rik Peeters (eds), *The Algorithmic Society: Technology, Power and Knowledge* (Routledge, 2021) 19, 23.

205 Henman and Adler (n 198) 148.

206 Ibid 149.

This promise has not been realised, however, as digital welfare systems have 'tended to reinforce the knowledge barriers' between the department and the claimant.[207] Paul Henman and Michael Adler's survey of 13 OECD countries (including Australia) has found that the computerisation of social security has 'decreased the simplicity of social security policy … decreased the number of staff … and local offices … and decreased the personal contact between claimants and staff'.[208] They thus concluded that 'computer technologies are more likely to be used to control rather than to empower staff and claimants'.[209]

An even longer-standing bias lies against people involved with the criminal justice system, which has always operated disproportionately against the vulnerable and disadvantaged, including Aboriginal and Torres Strait Islander people. The increased use of forms of AI at various stages of the criminal justice system carries a strong risk that these existing biases will be aggravated. Police use of automated 'tools' to identify and target particular individuals seems likely to further stigmatise and criminalise those people, while the use of algorithmic tools in sentencing also carries the risk of aggravating existing biases in the system. The punitive nature of both the criminal justice and welfare systems points towards 'a long tradition of linking crime and poverty (and more recently, welfare) in the discourses and practices of the state', thus leading to a 'convergence' of the welfare and criminal justice systems.[210]

Given these deep-seated troubling issues, in order to protect vulnerable populations, we argue that three main guiding principles are required to provide particular protection to disadvantaged populations: the principles of empowerment, harm minimisation, and transparency. Although these broad normative principles are desirable for government policy-making more generally, in the area of automated decision-making these are particularly salient, and we will highlight the particular implications these principles have for these systems.

First, the principle of empowerment requires that automated decision-making promotes the autonomy and capabilities of vulnerable people. Second, the principle of harm minimisation suggests a protective approach should be adopted in the design and implementation of AI systems, including non-discrimination and reducing algorithmic bias. Third, the principle of transparency requires that AI systems provide meaningful information about their design and the basis for decisions, to enable the public to challenge and seek legal redress for harmful AI decisions.

---

207 Ibid.

208 Ibid 157.

209 Ibid 159.

210 Zoe Staines et al, 'Governing Poverty: Compulsory Income Management and Crime in Australia' (2021) 29(4) *Critical Criminology* 745, 747.

D *The Principle of Empowerment*

According to an 'ethics of vulnerability' framework, the aim of governmental interventions is to support and foster autonomy and promote capabilities, rather than increase powerlessness and 'at worst … compound vulnerability or create new forms of pathogenic vulnerability'.[211] Promotion of empowerment could be achieved by 'increasing self-esteem, skill levels [and] financial management' in order to open up options for people rather than constraining them.[212] AI systems properly deployed are good at optimisation and are able to improve coordinative processes between government authorities and vulnerable populations, as well as enhance government service provision to individuals.

In order to promote the dignity and empowerment of people affected by AI systems, the design phase of new technologies should incorporate elements of co-design and consultation with those affected to ensure their needs are taken into account, with multiple feedback loops and adaptation based on the feedback.[213] Further, there should be meaningful community engagement with those affected before, during and after the technology is deployed, and vulnerable populations should be provided with the ability to assess and potentially reject the use of AI systems.[214]

Beyond consulting with affected populations, in order to allow broader public policy participation in the design of AI systems, there needs to be 'input and oversight by stakeholders with both substantive and technological capacity at multiple points over the design and implementation timeline'.[215] Thus, stakeholders with policy and technical expertise, as well as affected populations, would be able to collaboratively provide input about technical choices that have broader policy and political implications. This should be combined with regular periodic reviews by these communities of expertise and lived experience to ensure that AI systems continue to meet broader public policy goals.[216]

E *The Principle of Harm Minimisation*

Where AI systems make errors based on poor design or faulty data, these mistakes are compounded over hundreds of thousands of decisions. The broad-scale and inflexible implementation of deficient technologies has the potential to cause great

---

211 Catriona Mackenzie, 'The Importance of Relational Autonomy and Capabilities for an Ethics of Vulnerability' in Catriona Mackenzie, Wendy Rogers and Susan Dodds, *Vulnerability: New Essays in Ethics and Feminist Philosophy* (Oxford University Press, 2013) 33, 40.

212 Henman and Adler (n 198) 150.

213 Carney, 'Artificial Intelligence in Welfare' (n 9).

214 Meredith Whittaker et al, *AI Now Report 2018* (Report, AI Now Institute, December 2018) 22.

215 Deirdre K Mulligan and Kenneth A Bamberger, 'Saving Governance-By-Design' (2018) 106(3) *California Law Review* 697, 772.

216 Ibid.

harm to vulnerable populations. Accordingly, the principle of harm minimisation is a protective approach that seeks to ensure that the design of AI systems is non-discriminatory and free from bias. This principle also considers the impact on vulnerable populations through an AI impact assessment, and ensures a careful approach to rolling out and auditing new technologies.

Despite the promises of efficiency and cost-effectiveness, machine learning algorithms can be trained on datasets that contain human bias,[217] thus resulting in predictions that are tainted with unfair discrimination.[218] For instance, a United States study has shown that facial recognition technologies generate a disproportionate number of false positives up to 10 of 100 times more among African and Asian faces than for Caucasians.[219] As academics at New York University noted, '[g]iven the deep and historical racial biases in the criminal justice system, most law enforcement databases are unlikely to be "appropriately representative"'.[220]

Accordingly, the AHRC has made recommendations to combat algorithmic bias.[221] First, they suggested that the collection and utilisation of more appropriate data to train the machine learning programs will improve accuracy.[222] In particular, more data should be acquired on under-represented minority groups.[223]

Second, the AHRC recommended that data is pre-processed in order to mask protected attributes.[224] This may reduce the risk of discrimination.[225] However, even when racial data is not used as an input, the creation of proxies for race from certain data points is still a fundamental issue when seeking to remove algorithmic bias in this manner.[226]

---

217 Australian Human Rights Commission and World Economic Forum, 'Artificial Intelligence: Governance and Leadership' (White Paper, January 2019) 9 ('Artificial Intelligence: Governance and Leadership').

218 Ibid.

219 Natasha Singer and Cade Metz, 'Many Facial-Recognition Systems Are Biased, Says US Study', *New York Times* (online, 19 December 2019) <https://www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html>.

220 Whittaker et al (n 214) 16.

221 'Using Artificial Intelligence to Make Decisions' (n 195) 22–30.

222 Ibid 24–5.

223 Ibid 25.

224 Ibid.

225 Ibid.

226 Issie Lapowsky, 'Crime-Predicting Algorithms May Not Fare Much Better than Untrained Humans', *WIRED* (online, 17 January 2018) <https://www.wired.com/story/crime-predicting-algorithms-may-not-outperform-untrained-humans>.

The principle of harm minimisation also suggests there should be an AI impact assessment for both rule-based and machine learning systems[227] that assesses the benefits, risks and safety of AI from a legal, technical and ethical perspective, as well as consultation with stakeholders affected by the use of those technologies. These assessments would prompt agencies to consider the political consequences of algorithmic design and implementation.

To further minimise harm, in the implementation phase, new technologies should be rolled out in a careful manner, with a human 'in the loop' to provide and maintain oversight at key phases.[228] This should include a testing process and incremental rollout of the technology. The technology should be piloted on a contained sample prior to implementation to ensure that it meets its design criteria. Further, there should be evaluation of the effectiveness and efficiency of the technology once it is deployed. Government agencies should monitor outcomes through regular, periodic audits of a sample of automated decisions to check that the automated technology is working consistently with its design criteria, checking especially for error rates, bias, and unanticipated effects on individuals. There should be periodic independent reviews on the effectiveness, performance, accuracy, and security of automated decisions.

F *The Principle of Transparency*

Government transparency is a democratic ideal based on the concept 'that an informed citizenry is better able to participate in government; thus providing an obligation on government to provide public disclosure of information'.[229]

A major challenge associated with automated decision-making is its opacity. AI-made decisions may be inscrutable due to the complexity and sophistication of the technology, which involves rules that are so numerous, intricate and interdependent that they defy practical inspection.[230] The inherent difficulty in understanding an algorithm is classified as a 'technical black box'.[231]

---

227 See Alessandro Mantelero, 'AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment' (2018) 34(4) *Computer Law and Security Review* 754.

228 Richard M Re and Alicia Solow-Niederman, 'Developing Artificially Intelligent Justice' (2019) 22(2) *Stanford Technology Law Review* 242, 282.

229 Chris Draffen and Yee-Fui Ng, 'Foreign Agent Registration Schemes in Australia and the United States: The Scope, Risks and Limitations of Transparency' (2020) 43(4) *University of New South Wales Law Journal* 1101, 1106–7, citing Daniel J Metcalfe, 'The History of Government Transparency' in Padideh Ala'i and Robert G Vaughn (eds), *Research Handbook on Transparency* (Edward Elgar, 2014) 247, 249.

230 Selbst and Barocas (n 194) 1094.

231 Han-Wei Liu, Ching-Fu Lin and Yu-Jie Chen, 'Beyond *State v Loomis*: Artificial Intelligence, Government Algorithmization and Accountability' (2019) 27(2) *International Journal of Law and Information Technology* 122, 135.

Achieving algorithmic transparency would require clarity on the 'fact, extent and operation of automation in decision-making'.[232] The lack of awareness by individuals subject to Robodebt notices that the decisions were automated meant that they were less likely to question the determinations issued. In this vein, art 15 of the *General Data Protection Regulation* requires data controllers to provide data subjects with information about the existence of automated decision-making, 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject'.[233] However, commentators have debated the scope of this 'right to explanation', including what might constitute meaningful information to satisfy this disclosure requirement.[234]

There are several types of information that may provide meaningful transparency in the context of AI systems. The first relates to whether the explanation is about individualised reasons for the specific decision, for example 'the weighting of features, machine-defined case-specific decision rules, information about reference or profile groups'.[235] Alternatively, this relates to the general operation of the AI system (or 'system functionality') — that is, 'the logic, significance, envisaged consequences, and general functionality' of a system.[236] These categories of information may overlap, particularly in machine learning systems, where machine-defined case-specific rules form part of the model.[237] It is also important to consider the temporal aspect of an explanation, that is, whether the explanation should be provided — 'before or after automated decision-making'.[238]

A fulsome explanation that would satisfy the dictates of transparency would require all elements of disclosure, that is, both individualised information about the particular decision and generalised information about the operation of the AI system. Further, where machine learning is utilised, there should be

> accurate documentation of the decision logic, including the principles behind the machine learning model, training and testing processes; and a statement of

---

232 Ng et al (n 15) 1052. See also Toby Walsh et al, *Closer to the Machine: Technical, Social, and Legal Aspects of AI* (Report, Office of the Victorian Information Commissioner, August 2019) 50.

233 *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 L 119/1, art 15(1)(h).

234 Andrew D Selbst and Julia Powles, 'Meaningful Information and the Right to Explanation' (2017) 7(4) *International Data Privacy Law* 233, 239.

235 Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7(2) *International Data Privacy Law* 76, 78.

236 Ibid.

237 Selbst and Powles (n 234).

238 Ibid 238.

> reasons should be logged for all predictions or decisions at the point in time that they are made.[239]

Providing individualised reasons for automated decisions fulfils dignitary goals of recognising the personhood and autonomy of those affected by automated decisions, particularly vulnerable people.[240] In the context of vulnerable populations, there should be a user friendly non-technical explanation for the reasons for the decision that is 'comprehensible by a lay person'.[241] Individualised reasons also fulfil a justificatory purpose, as these provide the necessary information and evidence to enable affected individuals to challenge government decisions, which will in turn ensure that public sector decisions are made based on 'legally acceptable reasoning and are legitimized by acceptable process or oversight'.[242]

On the other hand, systemic or aggregate transparency about the operation of AI systems fulfils a different goal, which is an instrumental one of ensuring that automated decisions are accurate, rational and non-discriminatory.[243] The ability to scrutinise any faults in AI systems will enable more accountable government decision-making and lead to corrections of algorithmic design.[244] The provision of both individualised and generalised disclosure for automated decisions for both technical and non-technical audiences will thus enable government decision-making to be subject to scrutiny by a wider range of stakeholders.[245]

In temporal terms, at the very outset of the development of technology, there needs to be formal openness about the code itself, as well as publicity about the political nature of the questions resolved by design choices, which can often by obscured by the translation of rules into code.[246] Following the automated decision, there should also be *ex post* disclosure to affected individuals that the decision has been automated, including advising them of their potential avenues of challenge. This is particularly important for vulnerable populations, who may not otherwise understand how they may challenge an automated decision.

In short, there is a need for continued vigilance in evaluating the impact of new technologies, particularly on vulnerable populations. Our guiding principles provide a basis for safeguarding the rights of vulnerable populations through *empowerment*

---

239 Ng et al (n 15) 1075.

240 Margot E Kaminski, 'Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability' (2019) 92(1) *Southern California Law Review* 1529, 1534.

241 *AHRC Final Report* (n 119) 62.

242 Kaminski (n 240) 1534.

243 Ibid.

244 Ibid 1578.

245 Danielle Keats Citron, 'Technological Due Process' (2008) 85(6) *Washington University Law Review* 1249, 1284.

246 Mulligan and Bamberger (n 191) 772.

by incorporating their needs and feedback throughout the AI process, *harm minimisation* by ensuring that AI decisions are non-discriminatory and free of algorithmic bias, implemented carefully and regularly audited, and *transparency*, which allows people to understand the basis of AI decisions and enables legal challenge of harmful AI decisions.

VI Conclusion

It is undeniable that whatever our concerns and fears, the AI horse has left its computer-generated stable and bolted into the centre of our lives. Increasingly, all manner of decisions are made by algorithms, which reach conclusions — or at the very least proposals — which are rarely questioned in practice.[247] Given the reality of the 'robot century', it is redundant to question the overall benefits AI may bring — its potential to improve predictability, consistency and address the well-known (and easily, in this context, forgotten) shortcomings in human decision-making, in addition to its obvious benefits in efficiency. Thus, it is important to recognise the abundance of opportunities presented by technological developments to streamline and enhance the efficiency and consistency of government decision-making and service delivery. Technology has significantly contributed to the nature and practice of the welfare state,[248] as well as the surveillance and punishment of crime.

However, as this article has shown, the use of automated decision-making in social security and criminal justice, with its 'technocratic predictive logic',[249] risks perpetuating, intensifying and institutionalising discrimination and bias by adopting blanket rules over sections of the population that are particularly vulnerable. As Zoe Staines et al explain:

> [W]hile the social security system increasingly recoils and punishes, seeking to push its dependents into waged labor, the criminal justice system intensifies its focus on the poor and unruly and simultaneously ignores the infractions of the powerful.[250]

It is true that many of the problems arising from the use of AI in decision-making have been recognised by governments in recent times. The problem of algorithmic bias, for example, has been carefully outlined in a 2020 publication by the AHRC, which more recently again reported on human rights and technology.[251] Chesterman

---

247 Simon Chesterman, 'Artificial Intelligence and the Problem of Autonomy' (2020) 1(2) *Notre Dame Journal on Emerging Technologies* 210, 239–40.

248 Henman and Adler (n 198) 139.

249 Paul Henman, 'Targeted! Population Segmentation, Electronic Surveillance and Governing the Unemployed in Australia' (2004) 19(2) *International Sociology* 173, 173.

250 Staines et al (n 210) 755.

251 *Using Artificial Intelligence to Make Decisions* (n 195); *AHRC Final Report* (n 119).

goes further, arguing that the 'past few years have seen a proliferation of guides, frameworks and principles'.[252] These include soft norms developed by Singapore, Australia and New Zealand, texts by the European Union, the G7, and the OECD, as well as a set of principles known as the Rome Call for AI Ethics, endorsed by the Pope.[253]

Given all the soft law-generating activity, one might expect greater legal recognition of the necessity to ensure that decisions made by AI are consistent with broad principles such as transparency, accountability, non-discrimination and privacy, not to mention the existing law. It is arguable that, in fact, there are some signs the opposite may occur; for example, in a concerning decision in *Pintarich v Deputy Commissioner of Taxation*,[254] the Full Federal Court raised a fundamental question of whether decisions made by automated systems are decisions at all, and hence within the scope of judicial review.[255]

In this context, it is important to ensure not just that government pays lip service to AI frameworks, but that it pays real respect to them in daily practice. It is easy to be distracted by the superficial glamour and cost-saving potential of a newly developed algorithm, particularly when slickly presented by its proponents, and when the career advancement of senior public servants hinges on its swift implementation. The vulnerable, in such a game, are easy targets. For this reason, our article has focussed particularly on the impact of automated decision-making on the vulnerable, arguing that given the government's significant coercive powers, it is imperative to ensure that new technologies protect individual rights. This article has proposed a guiding framework on issues that government agencies should consider in the design, implementation and evaluation of new technologies to protect vulnerable populations. These safeguards will ensure that the design and implementation of AI in government promote empowerment of vulnerable populations, minimise harm, and are transparent, thus enabling legal redress.

A consistent, considered approach to AI will enable the criminal justice system and the Commonwealth government to 'reap the benefits of new technologies while minimising its attendant risks, as well as protect the individual rights and freedoms that are fundamental to our democracy'.[256] We have allowed the Trojan Horse of AI within our gates; it is important that the horse be tamed and put to use, rather than trample on the rights of the vulnerable, potentially running amok.

252 Chesterman, 'The Robot Century' (n 1).

253 Ibid.

254 (2018) 262 FCR 41.

255 Ibid. See also Ng and O'Sullivan (n 18) 27.

256 Ng et al (n 15) 1077.