ARTIFICIAL INTELLIGENCE IN WELFARE: STRIKING THE VULNERABILITY BALANCE?

TERRY CARNEY*

Artificial intelligence in public administration is both inevitable and potentially quite beneficial. Its assistive form offers access, efficiency and convenience; while the gains potentially are even larger in its augmentive 'machine learning' form. Offsetting risks include disadvantaging technology poor clients, and poor design which fails adequately to reflect social welfare principles or provide adequate accountability and redress for errors; a risk heightened for machine learning. This paper reviews some of the different forms and settings for AI in social security and argues that the Australian experience to date has been very mixed due to poor or rushed AI designs, poor understanding of client characteristics, and inadequate understanding of dynamics within contracted-out government services settings.

I INTRODUCTION

Australia's social security system historically is highly categorical (many separate payments), strongly needs-based (means testing), and draws many fine policy distinctions between recipients.[1] Huge numbers of decisions are made by its administration (called Centrelink) with very low levels of staff overheads. Like tax, its administration involves application of a vast and complex body of hard law,[2] along with reams of soft-law policy guidelines about how to apply the law.

Since discretion was largely eliminated from social security law over 40 years ago,[3] its rule-based form makes it a prime case for use of first-generation artificial

* Emeritus Professor of Law, The University of Sydney (Eastern Avenue, University of Sydney, NSW 2006, AUSTRALIA; fax: +61 2 9351 0200; email: terry.carney@sydney.edu.au); Visiting Research Professor, University of Technology Sydney; Associate Investigator, ARC Centre of Excellence for Automated Decision-Making and Society <https://www.admscentre.org.au/>.

[1] Peter Whiteford, 'The Australian Tax-Transfer System: Architecture and Outcomes' (2010) 86(275) *Economic Record* 528; Terry Carney, 'Conditional Welfare: New Wine, Old Wine or Just the Same Old Bottles?' in Peter Saunders (ed), *Revisiting Henderson: Poverty, Social Security and Basic Income* (Melbourne University Press, 2019) 100.

[2] Vast even in its 'potted' or heavily digested form: Westlaw AU, *The Laws of Australia* (online at 25 February 2020) 22 Insurance and Income Security, '22.3 Social Security'.

[3] See generally Terry Carney, *Social Security Law and Policy* (Federation Press, 2006) ch 2.

intelligence ('AI'). This essentially involves computerisation of data and its application in the administration of simple legal rules amenable to coding as deductive reasoning steps (expert systems 'automation').[4] More advanced AI developments using sophisticated techniques such as data-mining and 'machine learning' systems to mimic or assist in making more complex decisions involving a discretionary element might prove more problematic, but little space for such complexity remains in social security. More sophisticated AI deployed in more challenging public administration spaces is a topic for a different paper. The hinge for this article is that AI in Australian social security ought to be the paradigm case of being largely a positive experience, with few drawbacks. With such a thicket of 'rules' already laid out in legislation or the voluminous policy guides,[5] it could be anticipated that translating such policies into decision rules would be more straightforward than in other areas of public administration.

This article reviews examples of the different forms and settings for AI in social security in Australia. The review of arguably this easiest or most straightforward branch of Australian public administration finds that the AI experience to date has been very mixed. This result is attributed to poor or rushed AI designs, poor understanding of client characteristics and personal vulnerabilities, and inadequate understanding of the dynamics within contracted-out social security services settings. It is argued that it is not sensible to generalise from gravely bungled AI episodes to insist on retention of human rather than automated or even machine learning decision-making. Both human and AI systems have potential advantages for social security clients and both have potential weaknesses. Neither is intrinsically the greater or lesser threat to rule of law values.[6] But as with all new technologies, adjustments need to be made. An important question it will be revealed, is determining the benchmark values or principles against which to make or to assess those adjustments.

While there is no lack of evaluative principles to channel and shape AI deployment in welfare,[7] it will be argued that the United Nations ('UN') Special Rapporteur on Extreme Poverty and Human Rights is correct in contending that the primary

[4] AI progression, from its initial to its more complex forms, is commonly simplified as successive 'waves' for the purposes of discussion of legal and social implications: see, eg, the summary of such typologies in Part II of L Thorne McCarty, 'Finding the Right Balance in Artificial Intelligence and Law' in Woodrow Barfield and Ugo Pagallo (eds), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar, 2018) 55, 57–65.

[5] See, eg, Department of Social Services (Cth), 'Social Security Guide', *Guides to Social Policy Law* (Web Page, 21 March 2016) <http://guides.dss.gov.au/guide-social-security-law> ('Social Security Guide').

[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.

[7] Joe Tomlinson, *Justice in the Digital State: Assessing the Next Revolution in Administrative Justice* (Policy Press, 2019) 12. Tomlinson states that '[t]he difficulty is not … in suggesting concepts that may be relevant to the digitalisation of administrative justice, but in making sense of what to do with all the concepts that are often thrown around'. See also 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, 1045–8.

goal should be countering the risk of a 'digital welfare dystopia'. In his October 2019 report to the UN General Assembly Philip Alston warned against

> stumbling, zombie-like, into a digital welfare dystopia … in which *unrestricted data-matching* is used to expose and punish the slightest irregularities in the record of welfare beneficiaries (while assiduously avoiding such measures in relation to the well-off); *evermore refined surveillance options* enable around-the-clock monitoring of beneficiaries; conditions are imposed on recipients that undermine individual autonomy and choice in relation to sexual and reproductive choices and choices in relation to food, alcohol, drugs and much else; and *highly punitive sanctions are able to be imposed on those who step out of line*.[8]

It will be suggested here that it is vastly more difficult to avoid that dystopic form of digital welfare state than in other spheres of public life, precisely because welfare recipients are so vulnerable and so readily able to be cast as 'outsiders' rather than as rights-bearing citizens. Particular consideration will be paid to the aspects of unreasonable use of data-matching and the operation of sanctioning in Australian social security.

The analysis begins by sketching some of the main forms of AI, the impacts supposedly typically associated with each, and its deployment in Australian social security and wider government administration. Contrary to previous understandings it is concluded that AI is potentially disruptive of administrative governance and legal values in *all* its forms, not only in its sophisticated machine learning guise (Part II). Part III(A) then takes two Australian welfare examples — automated debt raising (or robodebt) and young at-risk sole parents (ParentsNext). The first case study highlights the dystopian welfare risk consequent on misuse of data-matching and other AI design failures of robodebt. The second case study (ParentsNext) profiles the way AI facilitated structural outsourcing of the welfare compliance regime, risking the welfare dystopia of excessive sanctioning and generation of overlays of 'pathogenic' vulnerability as a product of state action. Part III(B) considers the vulnerability challenge for social security clients of rendering accountable opaque AI processes and decisions, along with legal and extra-legal avenues for accommodating wider implications for AI in social security and public administration. A short conclusion summarises the lines of argument and assesses the risk of a welfare dystopia.

[8] Philip Alston, *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc A/74/493 (11 October 2019) 21–2 [77] (emphasis added) ('*Report of the Special Rapporteur on Extreme Poverty and Human Rights*').

II WHAT "IS" AI IN SOCIAL SECURITY AND GOVERNMENT ADMINISTRATION?

Artificial intelligence is a very broad term: 'AI can include machine learning, natural language processing, expert systems, vision, speech, planning and robotics.'[9] The recent Australian government consultation paper on AI and ethics likewise adopted a broad definition of AI as '[a] collection of interrelated technologies used to solve problems autonomously and perform tasks to achieve defined objectives without explicit guidance from a human being'.[10]

While space does not permit going into detail about the mathematical and engineering differences between various forms of AI which might be deployed in public administration, some differences are worth noting. First wave expert systems essentially involve labour intensive encoding into computer language of the legal or administrative rule in question. The process or steps for applying those encoded rules can then be expressed as an algorithm. So-called second wave applications, such as face recognition, instead rely on statistical learning (and probability) often using neural networks (algorithms, loosely modelled on the human brain, which recognise patterns). Machine learning is a particular form of this, involving parsing large data sets to detect patterns, commonly 'training' on one half of the data, with ongoing refining occurring on the remainder (and then progressive adaptation to fresh data).[11] To date all of these fall short of *replicating* complex human reasoning such as that in legal work or adjudication.[12] Any third wave capable of greater sophistication remains speculative.[13]

What AI is *not* also is quite important. So, a taxonomy is required which explains differences between the various types of AI that do qualify.

A *An AI End-User Taxonomy*

AI is not simply the 'mining' of data points (already done manually in the mid-

[9] Tania Sourdin, 'Judge v Robot? Artificial Intelligence and Judicial Decision-Making' (2018) 41(4) *University of New South Wales Law Journal* 1114, 1116, citing Michael Mills, 'Artificial Intelligence in Law: The State of Play 2016 (Part 1)' *Legal Executive Institute* (online, 23 February 2016) <http://legalexecutiveinstitute.com/artificial-intelligence-in-law-the-state-of-play-2016-part-1/>.

[10] D Dawson et al, 'Artificial Intelligence: Australia's Ethics Framework' (Discussion Paper, Data61 CSIRO, 2019) 14 <https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf>.

[11] This ability to appear to 'escape' from the originally encoded rules (strictly speaking merely adapt them) is in other contexts termed a 'deep learning' capability.

[12] McCarty (n 4); Frank Pasquale and Glyn Cashwell, 'Prediction, Persuasion, and the Jurisprudence of Behaviourism' (2018) 68(Supp No 1) *University of Toronto Law Journal* 63; Sourdin (n 9). McCarty provides an accessible treatment of the importance and difficulty in constructing appropriate computer 'languages' capable of capturing this level of linguistic complexity (legalXML and ruleML being two popular non-profit open-standards choices): McCarty (n 4).

[13] McCarty (n 4) 57.

19th century to set shipping lanes) and nor is it the use of algorithmic decision-trees or other mathematical processes as such.[14] One helpful classification of the functional impacts of AI from a legal or administrative perspective postulates three main types: the 'supportive' (*aids* to human decisions), the 'replacement' (automation replacing previously human decision-making), and the 'disruptive' (where AI results in different *forms* of administration and justice).[15] However like all heuristics, lived experience proves to be more nuanced and complicated, as now discussed.

1 *Digitisation and Expert System Automation*

In many popular usages, AI is simply synonymous with 'digitisation' of public administration. Australia digitised social security records and information quite early.[16] AI to assist in making or to automate decision-making was facilitated by enacting provisions designed to equate electronic decisions with those made by or with human intervention, first by giving effect to electronic application of a rule-based decision. This was followed by the current provision deeming automated decisions, made in accordance with programs authorised and controlled by the Secretary, to be decisions of the Secretary.[17]

For most purposes a decision is a decision,[18] whether made by human hand or

[14] Rebecca Williams, 'Rethinking Deference for Algorithmic Decision-Making' (Research Paper No 7/2019, Faculty of Law, University of Oxford, 31 August 2018) 3–4.

[15] Sourdin (n 9) 1117, citing Tania Sourdin, 'Justice and Technological Innovation' (2015) 25(2) *Journal of Judicial Administration* 96. Disruption is used in its more descriptive sense of a significant change, rather than the more teleological character of displacing old markets with new, and by implication superior, forms: 'Disruptive Innovation', *Christensen Institute* (Web Page) <https://www.christenseninstitute.org/disruptive-innovations/>.

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

[17] *Social Security (Administration) Act 1999* (Cth) s 6A. For a nice summary: see 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, 30–1. This has been the position since 2001, though in 1999 general authority was given to make or record a decision by computer, and from 1989 to that date, simply to 'record' it by computer: Will Bateman, 'Automatic Public Law' (Conference Paper, Public Law Weekend, Centre for International and Public Law, 3 November 2018). Equivalent provisions of the *Social Security Act 1991* (Cth) ('*Social Security Act*') covering automatic rate adjustment or cancellation decisions existed earlier, such as s 75A:

> 75A If:
>
> (a) a person is receiving an age pension on the basis of data in a computer; and
>
> (b) the pension is automatically terminated or the pension rate is automatically reduced by the operation of a provision of this Act; and
>
> (c) the automatic termination or reduction is given effect to by the operation of a computer program approved by the Secretary stopping payment or reducing the rate of payment of the pension;
>
> there is taken to be a decision by the Secretary that the automatic termination or rate reduction provision applies to the person's pension.

[18] For example a decision made by AI is no more problematic for conduct of merits review than one with human involvement: Carney, 'Automation in Social Security: Implications for Merits Review?' (n 16).

an automated computer process,[19] but as later discussed, serious doubts arise about whether an AI decision lacking human input constitutes a 'decision' for the purpose of judicial review.[20] While such AI otherwise is generally unproblematic, it does give rise to some additional issues, including the intelligibility of on-screen information or the screen dumps provided to bodies such as merits review tribunals,[21] or the harvesting of dubious information from social media postings as a basis for investigation or sanctioning of clients, and the probity of such information in the age of 'false information'.[22]

Automation which to varying degrees *displaces* human decision-makers by a replacement 'expert system' also is not recent.[23] Such automation is best suited to closed rule-based decision-making where subjective judgment (political, professional or otherwise) is not engaged.[24] Its main contributions lie in saving time (efficiency) and greater reliability of decision-making (eg good arithmetic in complex social security overpayment debt calculations). Thus the United Kingdom ('UK') government promotes deployment of AI as 'empowering', a way of 'help[ing] achieve government goals of economy, efficiency, and effectiveness while at the same time promoting the good governance values of transparency, accountability, and participation'.[25]

Embedding *replacement* AI within legacy systems constructed around human decision-making without system redesign, can be problematic.[26] Centrelink's Online Compliance Initiative ('OCI' or 'robodebt') is a classic example of the pitfalls of seeking to build greater ambitions than the existing system design,

[19] An example of the latter is cancellation of disability support pension on loss of portability eligibility due to living outside Australia beyond the allowed period: *Re Kampf and Secretary, Department of Families, Housing, Community Services and Indigenous Affairs* [2013] AATA 189. Other more subsidiary links between computerised calculation and the decision, are to be found in the scoring and application of the Carer Allowance eligibility measure, the Child Disability Assessment Tool: *Re Secretary, Department of Family and Community Services and Davies* [2001] AATA 101. An example of adverse consequences when auto-generated cancellation notices issued due to incorrect coding of correctly reported information are not queried within three months of issue (meaning no back-payment on correction) is *Re Estate of Thomas Biggin and Secretary, Department of Family and Community Services* [2000] AATA 125.

[20] Ng and Maria O'Sullivan (n 17) 27–31.

[21] Carney, 'Automation in Social Security: Implications for Merits Review?' (n 16) 262–3.

[22] Lyndal Sleep and Kieran Tranter, 'Social Media in Social Security Decision-Making in Australia: An Archive of Truth?' (2018) 22(4) *Media and Arts Law Review* 442.

[23] Zalnieriute, Bennett Moses and George Williams (n 6) 432–3.

[24] Thus it is ill-suited to individual tailoring of services in a casework arena, such as the National Disability Insurance Scheme ('NDIS'): Terry Carney et al, 'National Disability Insurance Scheme Plan Decision-Making: Or When Tailor-Made Caseplanning Met Taylorism & the Algorithms?' (2019) 42(3) *Melbourne University Law Review* 780.

[25] Carol Harlow and Richard Rawlings, 'Proceduralism and Automation: Challenges to the Values of Administrative Law' in Elizabeth Fisher, Jeff King and Alison L Young (eds), *The Foundations and Future of Public Law* (Oxford University Press, 2020) 275, 292, citing John Morison, 'Modernising Government and the E-Government Revolution: Technologies of Government and Technologies of Democracy' in Nicholas Bamforth and Peter Leyland (eds), *Public Law in a Multi-Layered Constitution* (Hart Publishing, 2003) 157.

[26] This can even delay adoption, leading to a 'slow and surprising creep' of uptake: 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, 123.

computer hardware and data quality permitted.[27] The OCI system was simply incapable of converting Australian Tax Office ('ATO') information about half-yearly or annual earnings into the fortnight-by-fortnight figures social security law insists on as the basis for raising a valid debt.

Assistive and replacement AI can however also have a *transformative* (ie arguably a 'disruptive') impact. With OCI robodebt that disruptive impact was that formerly accurate and legally sound debts completely lost both attributes (at a considerable human cost to affected individuals and a very substantial cost to government revenue when ruled illegal). Another, but more subtle, example of such administrative disruption from a supposedly assistive/replacement AI expert system was the introduction in the mid-1990s of the jobseeker classification instrument ('JSCI'). The JSCI allocates people to one of three differently remunerated streams for the purpose of determining the level of service to be provided to jobseekers and the amount of the government payment to employment providers under their government contracts.

As Mark Considine and colleagues found:

> The automation of interactions with jobseekers is evident in our own survey data. As [that data shows], the work performed by frontline staff has become increasingly computerized. The percentage of frontline staff who agree with the statement 'Our computer tells me what steps to take with clients/jobseekers and when to take them' soared from 17 per cent in 1998 to 47 per cent and then 50 per cent in 2008 and 2012.[28]

As others evocatively characterise it, this involved a shift from the former 'street-level bureaucracy' — as clients held real face-to-face conversations about their placement needs — to instead become a form of 'screen-level bureaucracy', where an officer sits at a computer and engages with their screen.[29] In the result 'less effort is exerted in getting to know jobseekers and fewer inputs from jobseekers

[27] For a recent overview, see Terry Carney, 'Bringing Robo-Debts before the Law: Why It's Time to Right a Legal Wrong' (2019) 58 (August) *Law Society of New South Wales Journal* 68. On 19 November 2019, the day after conceding the Federal Court test case challenge in the case of *Amato v Commonwealth* (Federal Court of Australia, VID611/2019, commenced 6 June 2019) <https://www.comcourts.gov.au/file/FEDERAL/P/VID611/2019/order_list>, the government announced that debts would no longer be raised as formerly on the basis solely of Australian Tax Office data match calculations of 'average fortnightly income', but instead would require further proof. Although the Minister has spoken of obtaining 'additional proof points', the law requires that Centrelink prove income for each and every fortnight across any debt period: Carney, 'Bringing Robo-Debts before the Law: Why It's Time to Right a Legal Wrong' (n 27).

[28] Mark Considine, Phuc Nguyen and Siobhan O'Sullivan, 'New Public Management and the Rule of Economic Incentives: Australian Welfare-to-Work from Job Market Signalling Perspective' (2018) 20(8) *Public Management Review* 1186, 1199. In part this less accommodating view by frontline staff in employment agencies is attributable to the 'pathologising' of welfare receipt as a consequence of conditionality of welfare: Michael McGann, Phuc Nguyen and Mark Considine, 'Welfare Conditionality and Blaming the Unemployed' (2020) 52(3) *Administration and Society* 466.

[29] Considine, Nguyen and Siobhan O'Sullivan (n 28) 1199, citing Catherine McDonald, Greg Marston and Amma Buckley, 'Risk Technology in Australia: The Role of the Job Seeker Classification Instrument in Employment Services' (2003) 23(4) *Critical Social Policy* 498, 508.

are considered for the purpose of service tailoring'.[30] Just as with the National Disability Insurance Scheme ('NDIS'),[31] there are always policy implications whenever expert systems intrude into areas calling for personalisation or tailor-making of responses.[32]

E-governance based around expert systems which more efficiently process digitised data records, however, is of quite a different order to the machine learning systems next discussed.

2 *Machine-Learning*

Machine learning, as already foreshadowed, involves writing algorithms which interrogate data sets and automate decision-making in part or in whole, through their ability to distil and apply underlying patterns in the data.[33] As Rebecca Williams explains:

> Algorithms can now either be trained, or learn by themselves to see patterns in big data. In 'predictive', or 'supervised' data mining, a dataset is divided into two. One half is used as training data (usually consisting of a collection of annotated objects or individuals) and the machines is [sic] taught to decide, on the basis of this training data, whether a new example falls into the relevant category or not. The second half of the data set is then used to determine whether or not the algorithm has learned to make the relevant distinctions correctly. With 'descriptive' or 'unsupervised' data mining, algorithms determine for themselves any commonalities they find between data objects in a particular dataset.[34]

Machine learning is being deployed in all spheres of private sector and government administration, including incorporation within government regulatory regimes.[35] This is the subset of AI that Veale and Brass have chosen in non-technical terms to call 'augmentive' decision-making. For instance,

> [t]he nature of the analytic capacity that algorithmic augmentation systems are supposed to improve, particularly in the context of linked administrative data combined with additional data sources, is that it is possible to 'mine' data for insights public professionals alone would miss. In areas such as tax fraud

[30] Considine, Nguyen and Siobhan O'Sullivan (n 28) 1199.

[31] See above n 24.

[32] See, for example, the discussion of risk scoring under Austria's instrument to determine levels of job provider support, and other international examples in *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc A/74/493 (n 8) 10–11 [27], 18 [63].

[33] Veale and Brass (n 26).

[34] Rebecca Williams (n 14) 4, citing Bart W Schermer, 'The Limits of Privacy in Automated Profiling and Data Mining' (2011) 27(1) *Computer Law and Security Review* 45, 46.

[35] Susan C Morse, 'Government-to-Robot Enforcement' [2019] (5) *University of Illinois Law Review* 1497.

> detection, ambitions do not stay at replicating existing levels of success with reduced staff cost, but to do 'better than humans'.[36]

Centrelink's 'risk profiling' tools are more in this vein,[37] but as yet machine learning has not been deployed in Australia to make substantive decisions about eligibility for social security.

B *AI in 'Services Australia'*

The body now known as Services Australia and its predecessors have had carriage of AI within government over much of the recent past.[38] Its record of administration of AI has however been a chequered one.

1 *AI and the Digital Transformation Project*

Although digitisation was already relatively well advanced in certain parts of the federal bureaucracy such as the Department of Social Services ('DSS'), it was anticipated to accelerate following the mid-2015 establishment of the Digital Transformation Office under then Communications Minister Malcolm Turnbull. But this proved to be premature. Its high-profile head Paul Shetler left in October 2016 as the office was relegated to an 'agency' and lost other prize recruits.[39] It then passed through several ministerial hands before landing with then Department of Human Services ('DHS') Minister Keenan between December 2017 and the May 2019 federal election.[40] Perhaps as a consequence, the record of achievement in digital transformation has been poor. Soon after the 2017 downgrade of the office for instance, Shetler publicly excoriated robodebt as a cataclysmic IT failure that

[36] Veale and Brass (n 26) 126, quoting Cas Milner and Bjarne Berg, *Tax Analytics: Artificial Intelligence and Machine Learning* (Research Report, PwC Advanced Tax Analytics & Innovation, 2017) 15. For discussion of how the UK envisages 'linked administrative data', see: Nigel Shadbolt et al, 'Linked Open Government Data: Lessons from Data.gov.uk' (2012) 27(3) *Institute of Electrical and Electronics Engineers Intelligent Systems* 16.

[37] Scarlet Wilcock, 'Policing Welfare: Risk, Gender and Criminality' (2016) 5(1) *International Journal for Crime, Justice and Social Democracy* 113, 120–1. For an outline of current initiatives, including the more proactive 'real time risk profiling', see Department of Human Services (Cth), *Annual Report 2017–18* (Report, 2018) 126–40 <https://www.servicesaustralia.gov.au/sites/default/files/2018/10/8802-1810-annual-report-web-2017-2018.pdf>. See generally, Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Press, 2017).

[38] In something of an irony given concerns about lack of human decision-making in the since successfully challenged Online Compliance Initiative ('robodebt'), in June 2019 the re-elected Morrison Government removed 'Human' from the title of the former Department of Human Services, the portfolio then called Services Australia, before absorbing it as an agency within the Department of Social Security under machinery of government changes announced by the Prime Minister on Thursday, December 5, 2019: Scott Morrison, Prime Minister, 'New Structure of Government Departments' (Media Release, Commonwealth, 5 December 2019) <https://www.pm.gov.au/media/new-structure-government-departments>.

[39] 'Home Page', *Digital Transformation Agency* (Web Page) <https://www.dta.gov.au/>.

[40] Paul Smith, 'Government's "Mind-Boggling" Digital Transformation Policy Steps Out of a Time Warp', *The Australian Financial Review* (online, 26 November 2018) <https://www.afr.com/technology/governments-mindboggling-digital-transformation-policy-steps-out-of-a-time-warp-20181123-h188yu>. Responsibility for digital transformation has however been placed with Stuart Robert, the Government Services Minister (and Minister for the NDIS).

would not be tolerated in the private sector.[41]

The AI responsibility of the now 'agency' of Services Australia is one for which little documentary guidance can be found, beyond the then DHS's 13-page *Technology Plan 2016–20*. Its principal stated aim is for government services to be as accessible as online banking or shopping,[42] but the document reads more as aspirational public relations than specific plan.[43] One of the few concrete proposals was for development of 'virtual assistants', but the rollout of 'Nadia', the prototype virtual assistant in the NDIS, was soon aborted.[44] The plan does reference the important new tool of 'co-design' in the development of AI, but in the form expressed in that document it falls well short of the 'agile design' ideal for optimally delivered digital transformation projects,[45] where top-down public service design is supposed to be replaced by ground-up engagement with the needs of end-users as established through 'prototyping, testing and research'.[46]

Rather worryingly for vulnerable citizens, co-design in the DHS Plan instead is expressed as including '[o]nline self-help communities … naturally formed by customers, providers and others who share a common interest, with online forums and crowd-sourced assistance'.[47] Here co-design has become code for a partial outsourcing of government responsibilities to facilitate citizen access to programs and services. Such drift is a serious and well recognised risk for co-design, as Tomlinson has warned. Even at its best, co-design conversations are often somewhat artificial and unduly constrained, since they always remain at the behest of government willingness adequately to fund the process and to listen to feedback and are always about 'how' rather than 'whether' to digitise.[48]

2 *AI in Centrelink*

Centrelink itself is over four years into a welfare payment infrastructure transformation program. Improving operation of the student Youth Allowance payment ('YA') was identified as an early priority, but the results have been disappointing, despite considerable optimism expressed in annual reports of the

[41] Christopher Knaus, 'Centrelink Crisis "Cataclysmic" Says PM's Former Head of Digital Transformation', *The Guardian* (online, 6 January 2017) <https://www.theguardian.com/australia-news/2017/jan/06/centrelink-crisis-cataclysmic-turnbull-former-head-digital-transformation>.

[42] Department of Human Services (Cth), *Technology Plan 2016–20* (Plan, 2017) 7 <https://www.humanservices.gov.au/sites/default/files/2017/03/13297-1703-technology-plan-summary.pdf> ('*Technology Plan 2016–20*').

[43] In 13 pages devoting significant space to graphics of light bulbs, there are but four pages containing substantive text: *Technology Plan 2016–20* (n 42).

[44] Justin Hendry, 'NDIS' Great Bot Hope Nadia Takes More Time Off for Stress Leave', *iTnews* (online, 10 December 2018) <https://www.itnews.com.au/news/ndis-great-bot-hope-nadia-takes-more-time-off-for-stress-leave-516592>.

[45] Tomlinson (n 7) 73.

[46] Ibid 75, 76 respectively.

[47] *Technology Plan 2016–20* (n 41) 9.

[48] Tomlinson (n 7) 77.

DHS.

The 2017–18 annual report for example touted progress made with the YA system to provide 'faster and more consistent decisions' freeing staff to 'support customers with complex needs and circumstances'.[49] It cited various supposed achievements, including reducing median claim processing times to three weeks, reduction in the number of claims questions by 'almost 70 per cent, from 117 to 37', and pre-population of later claims with answers from initial claims.[50] It claimed that: 'The program has successfully developed the capability to progressively automate the processing of student claims, meaning that some students will find out in near real time whether they will receive a payment'; however, the median processing time for new claims was still three weeks compared to a featured initial YA claim that had taken four months.[51] This shows how distant remains the goal of processing in 'real time', or of reaching the performance standards of online banking or shopping.

One reason for slow progress and patchy outcomes of AI in Centrelink appears to be a lack of concrete benchmarks, with a 2019 Audit Office follow-up report on call centre operations finding the only performance target for that program to be that of boosting take-up by 5% annually.[52] One of the few success stories so far, if only because of Reserve Bank collaboration, is immediate bank transfers of emergency payments, such as disaster relief.[53]

Centrelink has also managed to develop a number of smartphone app interfaces for digital communication of information about payments, including provision of downloadable letters, advice of future appointments and as a means of uploading any required documents or other information. But these too have proved to be controversial, shifting the geography of governance of clients into a 'virtual' space,[54] and posing challenges for protection of basic rights.[55]

[49] Department of Human Services (Cth), *Annual Report 2017–18* (n 36) 142.

[50] Ibid.

[51] Ibid 143, 142 respectively.

[52] Denham Sadler, 'Audit Puts Heat on DHS Digital Shift', *InnovationAus* (online, 26 February 2019) <https://www.innovationaus.com/audit-puts-heat-on-dhs-digital-shift>, discussing Australian National Audit Office, Department of Human Services (Cth), *Management of Smart Centres' Centrelink Telephone Services: Follow-Up* (Auditor-General Report No 28, 21 February 2019) <https://www.anao.gov.au/sites/default/files/Auditor-General_Report_2018-2019_28.pdf>.

[53] Dylan Bushell-Embling, 'Centrelink Adopts Real-Time Urgent Welfare Payments', *GovTech Review* (Web Page, 6 March 2019) <https://www.govtechreview.com.au/content/gov-digital/news/centrelink-adopts-real-time-urgent-welfare-payments-866902798>.

[54] Lyndal Sleep and Kieran Tranter, 'The Visiocracy of the Social Security Mobile App in Australia' (2017) 30(3) *International Journal for the Semiotics of Law* 495, especially at 506.

[55] Paul Henman, 'Of Algorithms, Apps and Advice: Digital Social Policy and Service Delivery' (2019) 12(1) *Journal of Asian Public Policy* 71, 75–8.

III WHAT ARE THE VULNERABILITY CHALLENGES OF AI IN SOCIAL SECURITY?

A major concern regarding the deployment and design of AI in social security is the impact on the vulnerable. Vulnerability is generally accepted to be a universal feature of the human condition, waxing and waning over the life-course and in response to external events.[56] Three analytically distinct if overlapping forms of vulnerability have been postulated:[57] those 'inherent' to the person; those which are 'situational'; and those which are 'pathogenic' (exacerbated by, or manufactured by, defective social policies).[58] Situational vulnerabilities are 'context-specific'. They are located in and amplified by 'the personal, social, political, economic, or environmental situation of a person or social group', and can be short or long-term.[59] Inherent and situational vulnerabilities can be latent or 'occurrent' (ie actualised by external circumstances).

While social security clients encounter manifold inherent vulnerabilities (such as mental illness) along with situational vulnerabilities (generational or locational poverty)[60] it is the additional harms that are a *product* of state action ('pathogenic vulnerability') that is of particular interest in this part of the paper. This section opens by elaborating the robodebt experience before drawing out the more complex implications of a case study of the administration of ParentsNext.

A *Two Case Studies of AI-Induced Vulnerability in Social Security*

1 *Robodebt: The Measure that Derailed Steady Expert System and AI Development?*

The most notorious contemporary application of AI in social security has been the OCI (robodebt) initiative, struck down as an illegal and invalid automation algorithm by the Federal Court in November 2019, more than three years after it

[56] Jonathan Herring, *Vulnerable Adults and the Law* (Oxford University Press, 2016) ch 2.

[57] 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.

[58] Recently Mianna Lotz has explored situations of so-termed 'discretionary' (volitional assumption of) vulnerability and conceptualisation of cognate notions of 'resilience': Mianna Lotz, 'Vulnerability and Resilience: A Critical Nexus' (2016) 37(1) *Theoretical Medicine and Bioethics* 45.

[59] Rogers, Mackenzie and Dodds (n 57) 24.

[60] See Terry Carney, 'Vulnerability: False Hope for Vulnerable Social Security Clients?' (2018) 41(3) *University of New South Wales Law Journal* 783.

began.[61] This pre-election boosted savings measure[62] severely tarnished the brand for AI in social security administration, for several reasons.

Robodebt built on existing data matching exchanges of simple earnings information between the DSS and the ATO. In place of the past practice of investigating and proving any debt amounts, robodebt assumed that there was a debt whenever the *average* fortnightly earnings calculated from ATO data did not agree with information previously reported to DSS for what frequently were fluctuating casual fortnightly earnings. It did so without adjusting for the incommensurate concepts (averages were projected over 26 weeks when actual earnings figures for each and every fortnight were required to establish any debt); it flouted the legal obligation to prove social security debts (unlawfully requiring people to *disprove* the supposed 'debt');[63] and breached model litigant[64] and other ethical principles in order to avoid public rulings of invalidity. Thus, no merits review appeal was ever made to the publicly accessible General Division of the Administrative Appeals Tribunal ('AAT') against rulings of invalidity of robodebt by the lower tier Social Services and Child Support Division of the AAT (which is not public).[65]

Consequently a massive impost of what proved to be false or inflated debts continued to be imposed on vulnerable people.[66] Between its July 2016 commencement and March 2019, 500,281 robodebts were raised, valued at $1.25 billion, of which

[61] This is not an isolated example: see, eg, Jane Millar and Peter Whiteford, 'Policy Choices and Automation: How Benefits Systems Can Create Unjust Debts', *Austaxpolicy: Tax and Transfer Policy Blog* (Blog Post, 21 February 2020) <https://www.austaxpolicy.com/policy-choices-and-automation-how-benefits-systems-can-create-unjust-debts/>; Terry Carney, 'Robodebt Failed Its Day in Court, What Now?', *The Conversation* (online, 28 November 2019) <http://theconversation.com/robodebt-failed-its-day-in-court-what-now-127984>.

[62] Peter Martin, 'Extortion is No Way to Fix the Budget', *The Sydney Morning Herald* (online, 11 April 2018) <https://www.smh.com.au/politics/federal/extortion-is-no-way-to-fix-the-budget-20180411-p4z8x2.html>.

[63] Terry Carney, 'The New Digital Future for Welfare: Debts Without Legal Proofs or Moral Authority?' [2018] (1) *University of New South Wales Law Journal Forum* 1. See also Peter Hanks, 'Administrative Law and Welfare Rights: A 40-Year Story from *Green v Daniels* to "Robot Debt Recovery"' [2017] (89) *Australian Institute of Administrative Law Forum* 1. For a remarkably uncritical analysis from the Ombudsman's Office of its equally anodyne report on OCI: Amie Meers et al, 'Lessons Learnt about Digital Transformation and Public Administration: Centrelink's Online Compliance Intervention' (Conference Paper, Commonwealth Ombudsman, Australian Institute of Administrative Law National Administrative Law Conference, 20–1 July 2017) <https://www.ombudsman.gov.au/__data/assets/pdf_file/0024/48813/AIAL-OCI-Speech-and-Paper.pdf>. Similar disregard for the principle of legality has been found in other digital welfare initiatives internationally: *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc A/74/493 (n 8) 14–15 [42]–[43].

[64] Continuing to defend raising of those debts on that basis at internal (Authorised Review Officer) and external (Child Support Division of the AAT) review was a clear breach of the Commonwealth's 'model litigant' policy: for an outline see, Eugene Wheelahan, 'Model Litigant Obligations: What Are They and How Are They Enforced?' (Seminar Paper, Federal Court Ethics Seminar Series, 15 March 2016) <http://www.fedcourt.gov.au/digital-law-library/seminars/ethics-seminar-series/20160315-eugene-wheelahan>.

[65] For a brief overview: Terry Carney, 'Robo-Debt Illegality: A Failure of Rule of Law Protections?', *Australian Public Law* (Blog Post, 30 April 2018) <https://auspublaw.org/2018/04/robo-debt-illegality/>.

[66] Terry Carney, 'Robo-Debt Illegality: The Seven Veils of Failed Guarantees of the Rule of Law?' (2019) 44(1) *Alternative Law Journal* 4.

57,386 were reduced, 24,788 waived in part, and 31,160 fully waived.[67] Although amounts *recovered* are not fully known, the net revenue returns were very small; and across all debt categories just 59% were subject to repayment arrangements.[68] Prior to its invalidation by the Federal Court (by consent no less), the adjustments to robodebt had been inconsequential.[69]

In place of the government's own estimated $721,000 substantial economic cost of repaying (with interest) all past robodebts raised by the same illegal and false averaging methodology (increased to $1.2 billion after settling a class action in November 2020), the government should have taken the time to properly engage and design the system at the outset,[70] or in the three years during which its illegality was on notice. This would have avoided the massive infliction of pathogenic vulnerability the robodebt system generated for between a quarter and half a million social security clients already experiencing an elevated incidence of inherent and situational vulnerabilities. Only after the scheme was struck down did the government fast-track a better designed scheme to match Centrelink clients' fortnightly reporting of their earnings against ATO 'one-touch' fortnightly employer reports of payments in close to real time.[71] It introduced in February 2020 and enacted — with delayed effect from December 2020 — legislation to simplify the definition of 'income' so that Centrelink and ATO data would for the first time use the conceptually equivalent definitions of actual 'receipt' and actual 'payment' of earnings.[72]

[67] Luke Henriques-Gomes, 'Centrelink Still Issuing Incorrect Robodebts to Meet Targets, Staff Claim', *The Guardian* (online, 29 May 2019) <https://www.theguardian.com/australia-news/2019/may/29/centrelink-still-issuing-incorrect-robodebts-to-meet-targets-staff-claim>, citing Department of Human Services (Cth), Answer to Question on Notice to Senate Community Affairs Legislation Committee, Parliament of Australia, *Online Compliance Intervention: South Australia* (5 April 2019).

[68] Department of Human Services (Cth), *Annual Report 2017–18* (n 36) 177.

[69] It was unprofessional that the Commonwealth Ombudsman's Office did not address the fundamental issues of legality or reliability, largely accepting the sufficiency of 'improvements' or implementation of earlier (and frankly inadequate) recommendations: Commonwealth Ombudsman, *Centrelink's Automated Debt Raising and Recovery System* (Implementation Report No 1, April 2019) <http://www.ombudsman.gov.au/__data/assets/pdf_file/0025/98314/April-2019-Centrelinks-Automated-Debt-Raising-and-Recovery-System.pdf>.

[70] Although the focus was not on social security recipients but instead on fraud by employers in failing to make their contributions towards contributory social security in Austria, good design is possible: Johannes Himmelbauer et al, 'Towards a Data-Driven Approach for Fraud Detection in the Social Insurance Field: A Case Study in Upper Austria' in Andrea Kő et al (eds), *Electronic Government and the Information Systems Perspective: 8th International Conference* (Springer, 2019) 70.

[71] The 2019–20 Federal Budget foreshadowed a July 2020 roll out of an improved interface between Centrelink and the ATO (later delayed until 14 December 2020 due to COVID), where earned income would be reported to Centrelink as it 'is *received* during the fortnight' rather than, as previously, estimate any income either 'earned, derived or received' (which caught monies before they were paid). Client reports of earnings then could be matched against ATO single touch payroll ('STP') fortnightly data of payments made to those 'recipients with employers utilising STP': Commonwealth, 'Budget 2019–20: Budget Measures' (Budget Paper No 2, Parliament of Australia, 2 April 2019) 158 (emphasis added); *Social Security Act* (n 17) s 8. The ATO strongly encouraged small employers to adopt one touch software: see 'Software Solutions for Single Touch Payroll', *Australian Taxation Office* (Web Page, 2 July 2020) <https://www.ato.gov.au/business/single-touch-payroll/in-detail/low-cost-single-touch-payroll-solutions/>, but its full adoption among hospitality and other small scale employers seems unlikely before 2022 on present indications.

[72] *Social Services and Other Legislation Amendment (Simplifying Income Reporting and Other Measures) Act 2020* (Cth).

Robodebt, however, was a straightforward or simple case study compared to the multi-faceted interactions between AI, governance and other contextual characteristics found to be in play with the program for young sole parents.

2 *ParentsNext: Synergies between AI, Governance Modes and Other Contextual Factors?*

The administrative arrangements for program delivery, or its 'governance form', is a crucial feature in shaping the issues and outcomes from adoption of AI within public administration.

Veale and Brass astutely observe that take-up of the more sophisticated 'augmentative' form of AI has coincided with the two or so decade prominence of the new public management ('NPM') governance modality, developed as part of the shift to neoliberal forms of government where functions are contracted out to private sector providers. As they write:

> In many ways, this logic continues the more quantified approach to risk and action found in the wide array of managerialist tools and practices associated with New Public Management. These have long had an algorithmic flavour, including performance measures and indicators, targets, and audits. … Particularly in areas where professional judgement plays a key role in service delivery, such as social work, augmentation tools monitor and structure work to render individuals countable and accountable in new ways, taking organizations to new and more extreme bureaucratic heights of predictability, calculability, and control.[73]

Australia's ParentsNext program, with its double pincher movement intersection with client suspension of payments and other breach penalties, illustrates some of the social and legal consequences of this model.

ParentsNext is an 'investment welfare' program for young mothers to address potential risks of long-term welfare dependence identified in the McClure Report.[74] Piloted in 2016 it was extended nationally from July 2018 (except for remote regions). ParentsNext targets recipients of Parenting Payment on that payment for more than six months without receiving any income from a job and with a child under six years. Its introduction coincided with a new personal 'compliance' framework for all working age social security payment recipients. Called the 'Targeted' Compliance Framework ('TCF'), this reform is designed to reduce excessive and undiscriminating sanctioning (loss or reduction of payments)

[73] Veale and Brass (n 26) 126–7.

[74] Reference Group on Welfare Reform, Department of Social Services (Cth), *A New System for Better Employment and Social Outcomes* (Final Report, February 2015) <https://www.dss.gov.au/sites/default/files/documents/02_2015/dss001_14_final_report_access_2.pdf> ('McClure Report ').

for failure to meet obligations or activities associated with a payment (mutual obligations),[75] by instead fostering client compliance mainly by suspending and then restoring payments (with back pay) on resumption of compliance. Actual rate reductions or non-payment periods are reserved for those few 'wilfully' doing the wrong thing. Both ParentsNext and TCF also strongly embraced digital (eg smartphone) engagement as the principal way of reporting compliance (to be notified 'on the day') and for communicating compliance status (a 'traffic light' system for alerting people to their 'at risk' or actual breach status). Various features have attracted concern, including from the UN Special Rapporteur.[76]

Consistent with neoliberal NPM theory,[77] not only the service (activation of people on welfare) but also compliance monitoring and sanctioning have now effectively been fully contracted-out to the private sector (Jobactive Employment Services providers). This form of delegation of operational responsibilities for welfare programs is now the norm across working-age payment administration, including for ParentsNext.[78] To enable this, the legal authority to delegate powers to an 'officer' has been expanded to include: 'a person engaged (whether as an employee or otherwise) by … an organisation that performs services for the Commonwealth'.[79] Earlier contracting out of service provision to the private sector had already altered the qualities of the information emanating from those agencies to ground eligibility and compliance decisions. The operating 'culture' of private-for-profit providers is unlike that of the public service, being more 'enterprising' and flexible.[80] Prior to such extensive operational devolution personal compliance information at least was assessed by an officer of the public service, trained in and expected to understand the administrative law protections around lawful exercise of such powers. Protections which included due process and jurisdictional constraints (doctrines such as reasonableness and relevant considerations) and application of public administration precepts such as like-treatment of like-cases.

Under the latest iteration of ParentsNext two new features heighten concern: the automation of information flows; and generation of adverse consequences as the default setting. First, the recipient of the payment bears greater responsibility for self-reporting compliance on the day. Second, any absence of a positive

[75] Compliance sanctions rose from 1.47 million in 2014–15 to 2.17 million in 2016–17: Sally Whyte, 'Job Seekers Penalised Millions of Times by Private Job Services', *The Sydney Morning Herald* (online, 2 November 2018) <https://www.smh.com.au/politics/federal/job-seekers-penalised-millions-of-times-by-private-job-services-20181101-p50dee.html>.

[76] *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc A/74/493 (n 8) 12 [31].

[77] Considine, Nguyen and Siobhan O'Sullivan (n 28).

[78] Compliance reporting of information or participation in activities is strongly encouraged to be made by way of a smart phone 'app'. This is conducted under the auspices of the job provider rather than by interacting with a public servant employed by Centrelink, but the information feeds directly into a Centrelink computer system so decision-making effectively takes place in a real time 'virtual' space.

[79] *Social Security (Administration) Act 1999* (Cth) s 234(7)(c).

[80] McGann, Nguyen and Considine (n 28); Considine, Nguyen and Siobhan O'Sullivan (n 28).

report automatically records as a 'demerit point' breach. This is the default unless the provider, when prompted, finds that there was a reasonable excuse for non-compliance and enters that opinion into the system to dissolve that demerit point. These demerit points are of considerable practical importance. Once three demerits accrue within a six month period, the person moves into a 'warning' zone and the provider must conduct a review of their capability to comply with their existing obligations (and change them if not).[81] On reaching five demerits within six months, a second capability review is conducted *within* the department; if found capable, the person moves to the 'penalty zone' where subsequent unexcused breaches successively attract a 50% rate reduction of a fortnightly payment followed by 100% on the next occasion,[82] and then loss (cancellation) of payment for four weeks.[83] The new division of responsibility between private sector job service providers and the public service under ParentsNext essentially 'automates' the demerit point stage of compliance processing, removing the human 'in the loop' role formerly played by in-house officers of the department. Rectitude and other public administration values now are postponed until the concluding stages of sanctioning, such as the conduct of capability reviews.

These features of digital engagement and surgical targeting of compliance sanctions were strongly endorsed by the Employment Services Expert Advisory Panel Report of December 2018.[84] However implementation of TCF was not immediately accompanied by the Expert Panel's crucial main recommendation to radically reform funding incentives for Jobactive employment services providers so that they reward working with those *most* in need of services (due to vulnerability or complex needs). The initial rollout of TCF retained long-standing payment structures from the 1990s which in practice instead favoured concentrating on the 'low hanging fruit' of working with *easy to place* or even 'self-placing' clients.[85] The resultant lack of spare agency funds — in combination with administrative

[81] Department of Social Services (Cth), Social Security Guide (n 5) [3.11] (Mutual Obligation Requirements); *Social Security Act* (n 17) ss 607B (Newstart), 544B(7) (Youth Allowance). From March 2020 Newstart was renamed Jobseeker allowance.

[82] *Social Security (Administration) Act 1999* (Cth) ss 42AF(2)(c), 42AN(3) respectively.

[83] Ibid ss 42AF(2)(d), 42AP(5).

[84] Employment Services Expert Advisory Panel, Department of Jobs and Small Business (Cth), *I Want to Work: Employment Services 2020* (Report, 2018) <https://docs.jobs.gov.au/system/files/doc/other/final_-_i_want_to_work.pdf>.

[85] The perverse features encouraging 'parking' challenging clients to concentrate on 'creaming' of the easiest to service, or 'churning' of repeat clients who generate up-front payments was first identified in Mark Considine, *Enterprising States: The Public Management of Welfare-to-Work* (Cambridge University Press, 2001); Terry Carney and Gaby Ramia, *From Rights to Management: Contract, New Public Management and Employment Services* (Kluwer Law International, 2002). More recently, see Considine, Nguyen and Siobhan O'Sullivan (n 28) 1187, citing Mark Considine, 'The Reform That Never Ends: Quasi-Markets and Employment Services in Australia' in Els Sol and Mies Westerveld (eds), *Contractualism in Employment Services: A New Form of Welfare State Governance* (Kluwer Law International, 2005) 41, 49. The recommended new funding model was trialled in two regions from July 2019 before national rollout from July 2022: 'New Employment Services Model', *Department of Education, Skills and Employment* (Web Page) <https://www.dese.gov.au/new-employment-services-model>.

rules applying digital reporting as the default mode,[86] inadequate investigation of suitability of digital reporting,[87] and the issue of a tighter DSS 'list' of acceptable excuses — led to an unintended outcome of grave import for more vulnerable clients. It resulted in many Jobactive providers simply shifting responsibility for determining the reasonableness of compliance away from *caseworkers* onto front desk clerical staff unskilled in doing anything other than apply rigid rules. One consequence of the resultant poor-quality decision-making and adverse consequences for vulnerable sole parents was that Centrelink was obliged to abandon pursuit of nearly 50,000 warning strikes or potential suspensions.[88]

Such unintended policy consequences within complex systems and administrative settings are avoidable if comprehensive AI and technology plans are devised in accord with best practice agile co-design principles. As ParentsNext demonstrates, when there are several variables or moving parts in play, the use of AI and technology can too easily miscarry if design is inadequate, creating pathogenic vulnerabilities for users. In this instance, despite the complexities of the setting, the vulnerability was a simple one. Too many young and already at-risk sole parents experienced added stress from accumulation of dubious demerit points unable to be reviewed or corrected until they crystallised into a formal sanction. The AI system automated rendering of demerit points as the 'default' setting, excuse grounds became narrowly mechanical, and there was a lack of sufficiently skilled Jobactive staff to assess reasonable excuses for non-compliance.

Adequacy of review and accountability of AI decisions therefore continues to be critical.

B *Meeting the Accountability Challenges of AI*

As with any technological development, AI poses a number of accountability challenges for vulnerable social security clients.

1 *Opacity of Decisions*

One legitimate fear is that AI decisions affecting citizens (and the aggregate policies pursued) will be rendered too opaque to be the subject of meaningful merits review.

[86] Simone Casey, 'Social Security Rights and the Targeted Compliance Framework' (2019) 5(1) *Social Security Rights Review* <http://www.nssrn.org.au/social-security-rights-review/social-security-rights-and-the-targeted-compliance-framework/>.

[87] This is another of the pervasive risks identified in the Special Rapporteur's recent report on the risks of digital welfare: *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc A/74/493 (n 8) 15 [45].

[88] Approximately 48,500 such warning traffic light and other decisions over the period July 2018 to August 2019: Luke Henriques-Gomes, 'Jobseekers Had Payments Suspended for Breaching Rules in Faulty Job Search Plans', *The Guardian* (online, 25 October 2019) <https://www.theguardian.com/australia-news/2019/oct/25/jobseekers-had-payments-suspended-for-breaching-rules-in-faulty-job-search-plans>.

In social security however the opacity risk is most in evidence for vulnerable social security clients when seeking to understand their *initial* ('primary') decision, rather than a difficulty for the AAT when conducting merits reviews.[89] Tight means-testing and other policy settings inevitably generate considerable complexity,[90] but the robodebt example demonstrated the way automation tended to further reduce transparency and ease of comprehension of the way primary decisions presented to clients. Communications from Centrelink to current or former clients about the supposed debts merely *described* in passing that averaging from ATO data was being applied and may lead to error,[91] but failed to explain either how that average often led to vastly different outcomes to that under the required fortnight-by-fortnight calculation of rates,[92] or explain the legal basis for raising a debt in that manner.[93]

This lack of explanation is understandable because reasoned explanation is neither easy to program to auto-generate nor core to AI technology. Providing *adequate* reasons for decisions to the standard expected for legal review is therefore a real challenge for AI systems designers. This is because their brief tends to be merely ensuring 'explainability' of the workings of the AI system of rules, algorithms and so forth, rather than production of 'reasoned justification' in the individual case.[94] One suggested accountability answer to lack of transparency and adequacy of reasons is to allow AI to operate but insist on retaining both a 'human-in-the-loop' and a 'reason generating' element, as in Europe. Under the European Union's ('EU') widely acclaimed privacy regime, the *General Data Protection Regulation* ('*GDPR*'), special protections accrue if a decision is exclusively made

[89] For merits review it is a qualitative rather than quantitative change to the degree of difficulty posed, because material generated by a human hand already presents in a form which is difficult to decipher: Carney, 'Automation in Social Security: Implications for Merits Review?' (n 16). The complexity of rate calculations also means that any required recalculation is usually returned to be made by Centrelink in exercise of the AAT power to set aside with directions: *Social Security (Administration) Act 1999* (Cth) s 177(a)–(b), previously *Social Security (Administration) Act 1999* (Cth) ss 149(2)–(3), as repealed by *Tribunals Amalgamation Act 2015* (Cth) sch 3 item 43, and originally *Social Security Act* (n 17) s 1253(2).

[90] Neville Harris, *Law in a Complex State: Complexity in the Law and Structure of Welfare* (Hart Publishing, 2013); Carney, 'Conditional Welfare: New Wine, Old Wine or Just the Same Old Bottles?' (n 1).

[91] The Ombudsman's follow-up report in 2019 found less than adequate realisation of this goal: Stephen Easton, 'Commonwealth Ombudsman "Pleased" by Robodebt Changes, Legal Challenges Remain', *The Mandarin* (online, 4 April 2019) <https://www.themandarin.com.au/106766-commonwealth-ombudsman-pleased-by-robodebt-changes-legal-challenges-remain/>.

[92] See generally on the way digital engagement interfaces can fudge review pathways: *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc A/74/493 (n 8) 11 [29].

[93] Since the government ultimately abandoned all attempts to argue in the Federal Court that there was any credible legal foundation, this would have exposed that ultimately fatal defect much earlier.

[94] Veale and Brass (n 26) 131.

by AI without any human input.[95] First, these forms of decision-making must be backed by law. Second, under UK law, citizens must also be alerted to this AI quality when the decision is communicated, and they then have a month in which to ask either for its review or for a new decision that includes human input.[96]

There is general acceptance that this hybrid approach to human-machine interaction ought to be a bedrock design principle, and that unsupervised or unmediated AI decision-making is unacceptable. All this may prove to be mere window dressing, however, unless the public servant performing that human role is armed with sufficient material and adequate confidence in their own decision-making ability as to become a *genuinely* independent 'sceptic' of the AI output, rather than merely a rubber stamp regurgitator of what the algorithm has generated. Robodebt for example in theory did retain a human being in the chain between production of the 'debt' and its incorporation in a letter to the client, but that human element (admittedly less than *GDPR* compliant) proved to be an illusory protection, at least in that instance.

The challenge of accountability is further heightened when moving from automation to second wave AI such as machine learning (or systems verging on that density and complexity). This is because it then becomes difficult even to provide adequate explanations of how the overall *AI system itself* 'works',[97] let alone explain how any individual decision was made. This contributes to loss of trust in the system, in addition to any concerns about mistakes in the individual instance. Rebecca Williams captures it nicely when she writes:

> The key difference, then, occurs when the algorithm is no longer transparent. This might simply occur because the decision-maker does not release the whole decision tree to public scrutiny, or it may be because machine learning (ML) is involved in generating the decisions.[98]

This is where the risk of opacities of deliberate secrecy (including any compounding from contracting-out design and operation), technical illiteracy (compounded by inadequate explication) and the unknowability of say machine

[95] *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 22 ('*GDPR*'). For a discussion of the significant weaknesses of un-supplemented 'human-in-the-loop' protections (treated as a bedrock principle by the Commonwealth): see Jake Goldenfein, 'Algorithmic Transparency and Decision-Making Accountability: Thoughts for Buying Machine Learning Algorithms' in Cliff Bertram, Asher Gibson and Adriana Nugent (eds), *Closer to the Machine: Technical, Social, and Legal Aspects of AI* (Victorian Information Commissioner, 2019) 41, 47–50.

[96] Veale and Brass (n 26) 139–40.

[97] At least with robodebt people could grasp that an ATO average was being substituted when each and every fortnight's actual earnings was called for.

[98] Rebecca Williams (n 14) 4.

learning's 'complex learning technique[s]', all come to the fore.[99]

As explained, to date in social security such sophisticated machine learning has mainly been confined to *systemic* matters such as 'risk profiling' for client compliance and investigative purposes,[100] rather than to make the *substantive* decision affecting an individual client. The lack of transparency around use of AI for such systemic purposes already does raise important equity arguments (eg whether tax compliance gets a soft pass compared to welfare 'fraud').[101] But it is when AI combines with 'investment state' welfare programs that it currently is most problematic in Australia, as already shown for ParentsNext. This is because the human capacity-building promise of a true investment rationale[102] is prone to degradation to a regressive form of 'actuarial reductionism' solely aimed at minimising *outlays* on those at high risk of welfare dependence. Measures such as compliance sanctioning of clients for instance are favoured in place of adequately funded, individually tailored training and support programs.

This actuarial turn has been seen to have been unduly prominent within the ParentsNext program. As already explained, the accountability challenge is yet further compounded when services are contracted-out to the private sector. This is because contracting-out is a governance form that thrives on and requires AI and other quantification modes of administration as a central part of its governance DNA.

2 *The Challenge of Adequate Design*

A common theme across all of the Australian social security examples considered in this article so far has been the lack of investment in or selection of appropriate governmental *design processes* for AI. In particular there has been no evidence of understanding or application of the previously mentioned principles of 'agile co-design' seen to be central to sound AI program development in public administration.[103] The key features of grassroots consultation with the

[99] Zalnieriute, Bennett Moses and George Williams (n 6) 441–3, citing Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) *Big Data and Society* 1.

[100] Wilcock, 'Policing Welfare: Risk, Gender and Criminality' (n 37) 120–1.

[101] Scarlet Wilcock, 'Official Discourses of the Australian "Welfare Cheat"' (2014) 26(2) *Current Issues in Criminal Justice* 177, 183. For the recent history of shifts towards an emphasis on compliance and a decline in prosecutions: see Tim Prenzler, 'Reducing Welfare Fraud: An Australian Case Study' (2017) 30(2) *Security Journal* 569; Scarlet Wilcock, '(De-)Criminalizing Welfare? The Rise and Fall of Social Security Fraud Prosecutions in Australia' (2019) 59(6) *British Journal of Criminology* 1498.

[102] This was first advanced in a somewhat traduced form in the 2015 McClure Report on Welfare Reform Australia: McClure Report (n 72). For elaboration of 'true' social investment welfare: see Don Arthur, 'Investment Approach to Welfare' (Research Paper, Parliamentary Library, Parliament of Australia, May 2015) <http://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/BudgetReview201516/Welfare>; Christopher Deeming and Paul Smyth, 'Social Investment after Neoliberalism: Policy Paradigms and Political Platforms' (2015) 44(2) *Journal of Social Policy* 297; Paul Smyth and Christopher Deeming, 'The "Social Investment Perspective" in Social Policy: A Longue Durée Perspective' (2016) 50(6) *Social Policy and Administration* 673.

[103] See above n 47 and accompanying text.

members of the relevant sections of the public directly affected by the program; experimentation and modification of programs; and the multiple feedback loops and adaptations involved in agile co-design — all have been almost totally missing (and when present have been token and after the event). Instead, AI continued to be a top-down bureaucratically controlled and driven process.

This lack of adherence to agile design principles might not have had such adverse outcomes in social security had other industry protocols been known and applied. In May 2019 a discussion paper on ethics and AI commissioned by the Australian federal government from the CSIRO was released. It built on overseas initiatives in the EU, UK and USA, setting out eight core principles for AI design, namely that it: (i) generate net benefits; (ii) do no harm; (iii) comply with regulatory and legal obligations; (iv) protect privacy; (v) provide fairness; (vi) be transparent and explainable; (vii) be able to be challenged (contestable) and (viii) provide accountability.[104] Similar principles have been expressed by others,[105] but operationalising them is more challenging.

This lack of investment in proper governmental AI design processes in social security has come at an obvious cost to all concerned. Certainly the degree of difficulty was increased in three ways: by Services Australia and Centrelink's already lean administrative overheads, which meant few in-house policy design experts (itself a product of digitisation); by its significant outsourcing of client contact to private sector job matching providers (the centrepiece of 20 years of 'activation' of people of working age) in further distancing Centrelink from a detailed appreciation of the needs of clients when designing AI; and by other contracting out, such as outsourcing the collection of overpayment debts to private sector agencies (an efficiency measure that substituted hard-edged commercial logics for welfare appreciation of discretion and individual human circumstances). Most fundamental of all, however, is the resultant lack of appreciation of how especially susceptible and vulnerable are social security clients.

In short, poor or inappropriate AI design by government was experienced by a segment of society with among the highest incidence of people at risk of being harmed by any design deficiencies.[106] A most potent and unfortunate combination. The obvious question is what can be done to reduce that risk.

[104] Dawson et al (n 10) 6.

[105] Luciano Floridi et al, *AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations* (Report, November 2018) 15–21 <https://www.eismd.eu/wp-content/uploads/2019/03/AI4People%E2%80%99s-Ethical-Framework-for-a-Good-AI-Society.pdf>.

[106] Design failures of the kind exemplified by robodebt are not isolated ones, as shown in the analysis of UK and Australian examples in Jane Millar and Peter Whiteford, 'Timing It Right or Timing It Wrong: How Should Income-Tested Benefits Deal with Changes in Circumstances?' (2020) 28(1) *Journal of Poverty and Social Justice* 3.

3 *Legal Avenues as Correctives*

Adoption of the previously mentioned principle of agile co-design in theory should prove to be a major corrective because it is driven by end-user interests. But AI design currently is decided and delivered by executive government. Consequently, even designs *purportedly* being developed pursuant to best practice processes of agile design can be debased, such that the co-design becomes little more than 'consultation' about AI regimes already pre-determined by the upper echelons of the administration.[107]

Indeed it has been argued in this article that this is precisely the case under the currently applicable AI policy applying in social security (Part II(B)(1)). A partial counterbalance may be provided by external accountability bodies such as the Ombudsman, Productivity Commission,[108] or the Audit Office (through performance or efficiency audits).[109] However robodebt exemplified the failure of these and all other equivalent bodies[110] until it was finally brought down in its entirety by Federal Court settlements in November 2019 and 2020 (the class action). So, the protection cannot always be relied on.

Bringing *government processes* for AI design or aspects of it *within* the purview of legal accountability by courts or tribunals would be very difficult, however. Leaving aside debates about whether relevant design principles could be rendered sufficiently concrete as to be justiciable (especially challenging given the fluidity and lack of specificity of 'agile design') the experience with the NDIS, such as in determining what is a 'reasonable and necessary support' or in defining the line between disability specific and general health services, suggest that law is not well suited to the accountability task.[111] Some store instead has been placed by some commentators in the remarks of the majority decision of the Full Federal Court in *Pintarich v Deputy Commissioner of Taxation* ('*Pintarich*'),[112] that a valid decision

[107] See above n 47 and accompanying text.

[108] Its work in assessing implementation of the NDIS is a good example: see Productivity Commission, *National Disability Insurance Scheme (NDIS) Costs* (Position Paper, June 2017) <http://www.pc.gov.au/inquiries/current/ndis-costs/position/ndis-costs-position.pdf>.

[109] For two relevant examples: Australian National Audit Office, Department of Human Services (Cth), *Management of Selected Fraud Prevention and Compliance Budget Measures* (Report No 41, 28 February 2017) <https://www.anao.gov.au/sites/g/files/net4981/f/ANAO_Report_2016-2017_41a.pdf>; Australian National Audit Office, Department of Social Services (Cth), *The Implementation and Performance of the Cashless Debit Card Trial* (Auditor-General Report No 1, 17 July 2018) <https://www.anao.gov.au/sites/g/files/net4981/f/Auditor-General_Report_2018-2019_1.pdf>.

[110] Carney, 'Robo-Debt Illegality: The Seven Veils of Failed Guarantees of the Rule of Law?' (n 64). The future program of work advanced for consideration by the Australian Law Reform Commission lists automated decision-making and administrative law as its first recommended topic for a reference: Australian Law Reform Commission, *The Future of Law Reform: A Suggested Program of Work 2020–25* (Report, December 2019) 10, 24–30 <https://www.alrc.gov.au/publication/the-future-of-law-reform-2020-25/>.

[111] Carney et al, 'National Disability Insurance Scheme Plan Decision-Making: Or When Tailor-Made Case Planning Met Taylorism & the Algorithms' (n 24) (indeed the NDIS currently lacks a provision deeming an electronic decision to be a decision); Veale and Brass (n 26) 125.

[112] (2018) 262 FCR 41, 53–75 (Moshinsky and Derrington JJ, Kerr J dissenting at 42) ('*Pintarich*').

in law calls for the exercise of a human mind as a component of that decision-making process.[113] But this too proves to be something of a dry gully, as now discussed.

Pintarich is little comfort even within its limited sphere of operation in judicial review of decisions with a discretionary component,[114] on a number of grounds. First, it is too easily satisfied through injection of purely *pro forma* human input. As already mentioned, even robodebt was claimed by government Senators and others as being an improper label for the program because there were humans 'involved' in the decision-making process before an actual debt was formally raised (as distinct from it being put to the person for their refutation).[115] Moreover, since acceptance of a need for human input leads to a *purely* AI decision being found to be unreviewable, '[i]ronically, the majority's decision would encourage further automation of government decision-making processes, in order to reap the benefits of those determinations not being subject to review under the [*Administrative Decisions (Judicial Review) Act 1997* (Cth)]'.[116] Second, it is arguably wrong in law (in inappropriately extrapolating principles to apply to any and all decisions, instead of looking closely at the context of the *particular* decision in issue, misreading prior lines of authority),[117] and certainly is completely out of kilter with contemporary administration, as the powerful dissenting analysis of Kerr J rightly concluded.[118]

Finally, and most tellingly from a social policy perspective, to give colour of credibility to such a proposition about the power of human involvement is to reify a policy fallacy.[119] The fallacy being that putting a 'human-in-the-loop' (or somewhere in the decision chain) is any substantive protection at all.[120] Instead

[113] Ibid 67 [141], applying (and following) the Full Court's endorsement in *Semunigus v Minister for Immigration and Multicultural Affairs* (2000) 96 FCR 533, 536 [11] (Spender J), 540 [55] (Higgins J), 546 [101] (Madgwick J) of the remark of the primary judge (Finn J) that a decision necessarily entails, in addition to an overt act of decision, the bringing to bear a human mental element in its making: *Semunigus v Minister for Immigration and Multicultural Affairs* [1999] FCA 422, [19].

[114] Those problems all lie in areas of discretionary powers, where factors such as undue haste may attract invalidation for failure to 'exercise' the power, or it may lack a necessary element of completeness, or perhaps see any automated but erroneous decision *functus officio*, precluding human recall: Bateman (n 17) 10–17. One such required adjustment may be according less 'deference' in judicial review (ie a more inquisitorial 'referee' than the traditionally impassive 'enforcer of the rules', to use the UK's popular sporting metaphor for portraying deference): Rebecca Williams (n 14); Marion Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power' (2018) 376(2128) *Philosophical Transactions of the Royal Society A* 20170359:1–20.

[115] See, eg, Evidence to Senate Community Affairs References Committee, Parliament of Australia, Melbourne, 9 October 2019, 15 (Hollie Hughes, Senator).

[116] Ng and Maria O'Sullivan (n 17) 32–3.

[117] This critique is compellingly made by Ng and Maria O'Sullivan: ibid 28–9.

[118] *Pintarich* (n 112) 48–9 [40]–[52].

[119] For an advocate of human involvement: see Aziz Z Huq, 'A Right to a Human Decision' (2020) 106(3) *Virginia Law Review* 611.

[120] See above n 95 and accompanying text; Henrik Palmer Olsen et al, 'What's in the Box?: The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration' (iCourts Working Paper No 162, Faculty of Law, University of Copenhagen, June 2019) 4, citing Elin Wihlborg, Hannu Larsson and Karin Hedström, '"The Computer Says No!": A Case Study on Automated Decision-Making in Public Authorities' (Conference Paper, Hawaii International Conference on System Sciences, 5–8 January 2016).

it may give a veneer of human involvement to decisions that in substance are rubber-stamping the AI outputs. While some very marginal benefits may accrue based on EU experience with this aspect of the *GDPR*, commentators suggest that greater dividends therefore may result from pursuit of *other* aspects of the *GDPR* framework, such as conducting multi-layered impact assessments.[121]

Other as yet unexplored mechanisms of accountability of AI at law may be unearthed by a systematic inquiry of the type foreshadowed by the Australian Law Reform Commission in its wish list of references.[122] However for the immediate future it appears from the above analysis that attention must turn back to any viable forms of accountability outside the law.

4 *Extra-Legal Correctives for Fidelity of AI Design and Operation*

The most effective initial extra-legal measure may well be to revive the operation of the Administrative Review Council (‘ARC’). This is the body which in 2004 laid down a most far-sighted and comprehensive set of guidelines and principles for implementing AI (duly ignored in the design of robodebt),[123] but which the Callinan Report into AAT amalgamation found was unlawfully ‘abolished’.[124] Short of measures such as this with some proven track record of influence, a healthy dose of scepticism appears warranted regarding other extra-legal protections, however.

Certainly, ‘trust me, I’m an IT design expert’, may ultimately yet prove to be the least worst of the management tools for optimising AI design, as Desai and Kroll have claimed.[125] After all, the alternative of relying on traditional legal values of fair hearing and due process protections[126] do encounter major impediments to their operationalisation in an AI context. Most algorithms and AI system logics are simply too complex to be explicated, tested and validated (or not) in such

[121] For discussion of other potentially more promising protections in the *GDPR* (n 95), such as multi-layered impact assessments: see Margot E Kaminski and Gianclaudio Malgieri, ‘Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations’ (Research Paper No 19–28, University of Colorado, 18 September 2019).

[122] Australian Law Reform Commission, *The Future of Law Reform: A Suggested Program of Work 2020–25* (n 110).

[123] Administrative Review Council, *Automated Assistance in Administrative Decision Making: Report to the Attorney-General* (Report No 46, November 2004).

[124] The legislation establishing the ARC was not repealed and the process by which it ceased to meet was found to be improper: Ian Callinan, *Review: Section 4 of the Tribunals Amalgamation Act 2015 (CTH)* (Report, 19 December 2018) 19–20 [1.27].

[125] Deven R Desai and Joshua A Kroll, ‘Trust But Verify: A Guide to Algorithms and the Law’ (2017) 31(1) *Harvard Journal of Law and Technology* 1 (favouring adaption of industry quality assurance methods).

[126] Cary Coglianese and David Lehr, ‘Regulating by Robot: Administrative Decision Making in the Machine-Learning Era’ (2017) 105(5) *Georgetown Law Journal* 1147 (favouring standard administrative checks and balances).

settings.[127] However, trusting the IT experts surely is much harder to take seriously in Australia given the scale, intransigence and harm shown to be wrought by robodebt. Outsourcing of such fidelity-to-purpose work to the private sector would merely heighten those reservations. As Simon Chesterman observes, the central values at stake here are the 'legitimacy' of treating citizens as a means rather than an end, and the limits on outsourcing government responsibility.[128] Whether the Australian public service is more sensitive to the risks of outsourcing of AI design and loss of control of key data sets as experienced in the USA[129] remains moot, but the record so far is not very promising.[130]

Equally, the case for requiring AI to meet legal standards such as of 'explainability', while it has led to a whole new sub-field in computing design (termed 'XAI'), may also prove to be a mirage. This is despite some commentators retaining faith in judge-led common law oversight,[131] or the more cautious optimism expressed about rendering AI compliant with traditional administrative law principles about fair hearings, lack of bias, and avoidance of jurisdictional error.[132] This scepticism is because as Goldenfein observes:

> XAI has the potential to *entrench* problematic automated decision-making by *narrowing* the types of reasons that are given for decisions, *therefore narrowing the grounds for contesting them*. Being subjected to automated decisions without understanding how or why that decision was made may be problematic; but *receiving automated explanations that do not provide a premise on which to base an appeal or contest* — and simply justify the decision — *might be worse*.[133]

Consequently it may indeed be more profitable to instead turn attention to applying or developing some of the other mechanisms found in the *GDPR*, such as the previously mentioned idea of multi-layered impact assessments and/or multi-layered explanations.[134]

[127] Harry Surden, 'The Ethics of Artificial Intelligence in Law: Basic Questions' in Markus D Dubber, Frank Pasquale and Sunit Das (eds), *Oxford Handbook of Ethics of AI* (Oxford University Press, forthcoming) 16 (discussing the 'interpretability problem' and judicial deferral to system outcomes).

[128] Simon Chesterman, 'Artificial Intelligence and the Problem of Autonomy' (2020) 1(2) *Notre Dame Journal on Emerging Technologies* 210, 248–50.

[129] For discussion of some of the risks of private proprietary firms setting the AI agenda: see Richard M Re and Alicia Solow-Niederman, 'Developing Artificially Intelligent Justice' (2019) 22(2) *Stanford Technology Law Review* 242.

[130] As Alston rightly observes, Australia's cashless welfare card gives pause for thought given the role played by commercial firms Indue and VISA: *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc A/74/493 (n 8) 19 [68], 20–1 [72]–[74].

[131] Ashley Deeks, 'The Judicial Demand for Explainable Artificial Intelligence' (2019) 119(7) *Columbia Law Review* 1829.

[132] Ng and Maria O'Sullivan (n 17) 33–4.

[133] Goldenfein (n 95) 58 (emphasis added).

[134] See Kaminski and Malgieri (n 121) 3–4, 11–13. See also their suggestion of a proposed expansion to produce 'multi-layered explanations' by 'deliberately widen[ing] the lens from algorithms as a technology in isolation, to algorithms as systems embedded in human systems': at 21.

IV CONCLUSION

The heavily rule-based content of Australian social security made it a prime candidate for deployment of AI. For several reasons the record so far has been undistinguished.

The Australian Government overreached with expectations of rapid catch-up under the Digital Transformation Office of mid-2015, before seriously weakening the agency's capacity to deliver by way of public service status downgrades and other organisational reallocations (Part II(B)). Political pressure to preserve budget revenue and savings assumptions (including finance department rules setting savings targets) led to over-hasty roll-out of the very poorly designed robodebt AI data-matching algorithm for raising and recovering overpayment debts, no doubt bedazzled in part by the nearly $4 billion the government originally anticipated it would recover over its life (Part III(A)(1)). Piecemeal implementation of reforms without adequate regard to unintended outcomes (lack of funding leading to degraded testing of reasons for non-compliance), potentiation of risk from system interactions (between ParentsNext and TCF), and lack of attention to the needs and capacities of people on the wrong side of the digital divide (the technology poor), also led to unsatisfactory outcomes for those two major initiatives (Part III(A)(2)).

The overall balance sheet, then, is not an easy one to draw up in the case of AI, whether within (expert systems) or 'as' administration (machine learning). This is as unsurprising for this new technology, as it was for past technological innovations. As Tomlinson concludes regarding the 'ongoing incursion' of AI in administrative justice, '[i]t is essential that … [it] is not seen as some distinct field of interest and activity, but as *part of the core business* of those concerned with public law and administrative justice'.[135] Or in the words of Genevieve Bell:

> [M]y argument … is really to say, as we think about any technology, whether it is a robot, whether it is an algorithm, … there are those questions that you have to ask every time. … What is its purpose? What is its form? What is its level of agency? And what will be the consequences of that? Because … it's not as simple as just making technology. We are also always and already in the business of making culture. Because you can't set about to make a piece of technology without intersecting with 200–2000 years of stories about what it means to make

[135] Tomlinson (n 7) 89 (emphasis added). Surden lists the values as: 'equal treatment under the law; public, unbiased, and independent adjudication of legal disputes; justification and explanation for legal outcomes; outcomes based upon law, principle, and facts rather than social status or power; outcomes premised upon reasonable, and socially justifiable grounds; the ability to appeal decisions and seek independent review; procedural fairness and due process; fairness in design and application of the law; public promulgation of laws; transparency in legal substance and process; adequate access to justice for all; integrity and honesty in creation and application of law; and judicial, legislative, and administrative efficiency': Surden (n 127) 3, citing Garland Publishing, *The Philosophy of Law: An Encyclopedia*, vol 1743 (at 1999).

> life.[136]

That 'life' of AI for vulnerable social security clients has been shown in this article to still very much be a work-in-progress, the balance sheet for which is uncertain in the sense that the risk of dystopian outcomes remains unduly high.

When designed, monitored and refined in accordance with best practice standards, AI in administration undoubtedly can enhance client accessibility, deliver more accurate and responsive decisions, and contribute to greater efficiency. AI can also be a force for ensuring citizen entitlement, such as to identify and contact people otherwise missing out on an entitlement (addressing low take-up).[137] Poorly designed AI systems however put at risk values of procedural fairness, the rule of law, and government accountability; with much further work needed even on questions such as how to *define* fairness in the context of AI systems,[138] and how to recognise and overcome discrimination in AI administration.[139] Although ensuring that hybrid AI-human system and the option of asking for human generated reasons is a minimum design principle, it has been shown in this article that this is a necessary but by no means a sufficient remedy for vulnerable social security clients. Because in practice it may simply prove to be window dressing rather than a substantive protection.

As the UN Rapporteur writes, digital technologies and AI in welfare

> could also make an immense positive difference by improving the well-being of the less well-off members of society, but this will require deep changes in existing policies. The leading role in any such effort will have to be played by Governments through appropriate fiscal policies and incentives, regulatory initiatives *and a genuine commitment to designing the digital welfare state not as a Trojan Horse for neoliberal hostility towards welfare and regulation but as a way to ensure a decent standard of living for everyone in society*.[140]

Recognising the imperative to do so, and then taking constructive steps to turn the neoliberal Trojan horse into that *positive force* for the welfare of the most vulnerable, is however an opportunity yet to be recognised, much less realised in

[136] Genevieve Bell, 'Making Life: A Brief History of Human-Robot Interaction' (2018) 21(1) *Consumption Markets and Culture* 22, 34.

[137] *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc A/74/493 (n 8) 12 [32]. For the Australian case: Committee for Economic Development of Australia, *Disrupting Disadvantage: Setting the Scene* (Report, 2019) 41–4 <https://www.ceda.com.au/Research-and-policy/All-CEDA-research/Research-catalogue/Disrupting-disadvantage-setting-the-scene>.

[138] See, eg, Reuben Binns, 'Fairness in Machine Learning: Lessons from Political Philosophy' (2018) 81 *Proceedings of Machine Learning Research* 81:149–159 <http://proceedings.mlr.press/v81/binns18a/binns18a.pdf>.

[139] See, eg, Pauline Kim, 'Auditing Algorithms for Discrimination' (2017) 166(1) *University of Pennsylvania Law Review Online* 189.

[140] *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc A/74/493 (n 8) 12 [32] (emphasis added).

Australia. The government's December 2019 commissioned roadmap for AI[141] it appears has much catching up to do if those challenges and opportunities of AI in welfare are to be met.

[141] Stefan Hajkowicz et al, *Artificial Intelligence: Solving Problems, Growing the Economy and Improving Our Quality of Life* (Report, Data61 CSIRO, 2019) <https://data61.csiro.au/en/Our-Research/Our-Work/AI-Roadmap>.