Submission to the Australian Human Rights Commission, *Human Rights and Technology Issues Paper* consultation

Context of the submission and about the author

I am a University of Queensland (UQ) academic with over 20 years of active research interest in digital technologies and public governance. This research covers the use of digital technologies ***by*** government for the operation of government (including policy making, service delivery, governance of agencies), as well as the use of digital technologies ***for*** governing and governance. Whilst my research has focused on governments' use of digital technologies, my work also provides insights for the private and NGO sectors.

In particular, my research has investigated the ways in which new digital technologies have shaped the types of policy and services that can be and are enacted. For example, over a decade before current concerns about algorithms in profiling and targeting, I identified the policy, social and ethical dynamics associated with digital technologies' disruption of public policy and administration principles, often leading to increased inequalities. (See Appendix for a list of my relevant publications.)

Significantly, my research rests on interdisciplinary training in computer science (holding an award winning first class honours degree, 1989), and in sociology of technology and social policy (PhD, 1996), which provides me with insights not typically open to people without such interdisciplinary training. To date, I have received almost $3 million in research funding, including from the Australian Research Council, IBM, and the former National Office for the Information Economy. I published 4 books and over 70 papers. I currently lead an international comparative study of government web portals in 10 countries.

I am Principal Research Fellow in the UQ Centre for Policy Futures, where I lead a multiyear, $4 million collaboration with CSIRO on responsible innovation. This work covers questions of regulation relating to a wide range of emerging technologies, including AI and digital technologies, synthetic biology and DNA manipulation, hydrogen and nuclear energy cycles, and health monitoring and detection technologies.

Importantly, I have also worked in government as a policy analyst (1996-99) thereby providing me with important insights into the way in which governments operate. Consequently, I have regularly contributed to government and independent inquiries regarding regulation of new technologies, including the Australian Law Reform's 2003 inquiry into genetic testing, the 2009 Government 2.0 Taskforce, and the Parliamentary Joint Committee on Intelligence and Security *Identity-matching Services Bill 2018* Inquiry.

1. ***What types of technology raise particular human rights concerns?***

Digital technologies continue to raise a range of human rights concerns, particularly around the ways in which personal data is captured, held, shared and used for making decisions. In addition there are a range of current and emerging technologies relating to gene technologies that raise similar questions in relation to people's DNA. Given my expertise in digital technologies, this submission focuses on these.

The Commission has raised a number of questions regarding the development of Artificial Intelligence (AI) and its particular impact on human rights. I would strongly encourage the Commission to avoid using the term 'Artificial Intelligence' as it rather vague in what it actually means. Rather, it is more a broad concept (not a specific technology) and often used as a phrase in marketing. What scientists regard as AI has changed over the last 50 years as our digital technologies have become more sophisticated and as such has challenged us to think more carefully about what constitutes (human and machine) intelligence. Alan Turing famously articulated what we now call the 'Turing Test' to determine AI based on whether or not a series of textual responses from an algorithm could not be distinguished from a human. Yet this Test is often being challenged and rethought. For example, Joseph Weizenbaum, in his 1976 book *Computer Power and Human* Reason, demonstrated how a simple textual algorithm could mimic a human counsellor/psychiatrist and seem quite humanlike. Indeed, some people preferred the program over a real person. More recently, when algorithms have beaten masters of chess (1997) and Go (2016) we again have asked, what exactly is human intelligence that computers might reproduce.

In summary, I would encourage the Commission to be more explicit in its use of AI terminology as the above demonstrates that it is very unclear what 'AI' means and thus very hard to engage in questions about its impact on human rights and how for it to be governed. The discussion paper's mention of learning algorithms (or machine learning) is a more specific version of AI and a productive approach results from focusing on that.

In the academic and advocacy literature machine learning/learning algorithms have generated considerable interest regarding their ethical and regulatory implications, and there is an element of 'magical' thinking by some commentators about such algorithms that reflects a lack of understanding of how they operate. Compared to more conventional algorithms where human programmers specifically design the operation, learning algorithms operate in a more macro way by adjusting variables to best match input with desired output. In short, the human designers do not need to work out the processes by which such matching occurs, but provides an 'evolutionary' way in which the algorithm will identify its own way. When compared with conventional programs, what this means is that with learning algorithms it is much harder to identify how an algorithm produced the result it did; in other words, what the underlying 'logics' are. I would argue that the difference that such algorithms have for human rights compared with more conventional programs is actually relatively small, for two reasons:

- First, it is already very hard to understand the reasons/rationale for outcomes that many complex, yet conventional, algorithms produce. This is a consequence of the complex and rapid interactions algorithms have with their environmental inputs, as well as a result of the extremely complex and multi-authored nature of most

computer algorithms (again a point that Weizenbaum made in his chapter 'incomprehensible programs' in the mid 1970s!).

- Secondly, while providing a precise explanation for why a learning algorithm produced a certain result is not possible, it is technically possible to design a process with learning algorithms whereby key factors that have contributed to the result on a probabilistic basis can be indicated. This form of reverse engineering might be a very useful regulatory requirement for the deployment of human rights critical learning algorithms.

As a result of these points, I argue that a focus on the human rights impact of AI (a la machine learning) that emphasises new innovations is not particularly helpful. Rather, it would be better to focus on the human rights impact of computer programs in general (learning and more conventional) based on what has already occurred, but have been poorly engaged with and/or addressed. In short, the dynamics that learning algorithms induce within society are rather similar to the more conventional algorithms that have proceeded them.

To illustrate, consider the AI informed decision making based on learning algorithms (discussed in more detail below) and how they compare to more conventional algorithmic based decision making. I am currently unaware of the use in Australia of machine learning for making government administrative decisions. However, such algorithms are being used or in the process of being developed and deployed in areas such as:

- Identifying the risk of child abuse or neglect following a child protection notification (in the USA and New Zealand);
- Calculating an appropriate criminal sentence to inform judges (in the USA)
- Making recommendations about risk of recidivism for parole decisions (in the USA)
- Risk assessment for air passenger screening (in the USA)

These machine learning algorithms do not operate any differently within an organisational context to ones that have been built based on more traditional statistical analyses (such as multiple regression techniques). Such conventional algorithmic approaches have been deployed in Australia for child protection services (Gillingham 2006; Gillingham & Humphries 2009), employment services (Caswell et al 2010; MacDonald et al 2003), and compliance testing and review in taxation and social security (Henman 2004; 2010). Furthermore the human rights challenges that the use of such risk based and categorisation tools for administrative decision making are not substantially different to those of computerisation of policy and administrative practice.

Which human rights are particularly implicated?

There is long standing acknowledgement of the need to protect people's privacy through their personal data, and the need for strong protections on that data including restrictions on sharing. Key principles include '**data collected for one purpose is not to be used for another purpose without the individual's permission**', right to '**correct wrong information**', and '**control over personal data**'.

Despite these principles and matching policies that purport to uphold these principles, the Australian formal legal settings and in their practice have considerably fallen short of public

expectations and understandings. Indeed, Australian governments have breached these rights, and underfunded the institutions and processes to seek remedy.[1] The Human Rights Commission would do well to propose stronger data protection and privacy laws, such as those enacted in the EU's General Data Protection Regulation (GDPR). In addition, a much stronger independent agency that reports to Parliament (not the Executive) would be a better design.

In addition to weak legal protections, there are also additional practical challenges in upholding these digital human rights (in addition to poorly resourced governance agencies). In particular, it has been regularly shown that people typically have little understanding of what information is held by them, and thus have the capacity to identify and correct errors. Exacerbating this there is also a widespread cultural assumption that digital data is 'objective' and 'true'.

The growth in digital data networks, storage facilities and proliferation of digital devices for collecting data only exacerbates these human rights challenges, thereby heightening the need to get these right.

In addition to the aforementioned digital rights principles the **right to forget** could be better advanced and articulated in Australian law and regulatory practice, both in the public and private sector domains. The vastly increased capacity to capture and store information has created a dynamic where much more personal information can be stored and held close to indefinitely. Indeed, even when someone seeks to withdraw their remarks or correct the record (by removing a social media post), the original record often is still present. This creates an inability for people to 'move on' from the past, which is a principle encoded in criminal law by extinguishment of past offences. The digital world makes the bar for an 'offence' (in the court of public opinion) much lower. Given that the 'court' of public opinion does not follow basic human rights principles, it is appropriate to explore greater avenues for people to have their past personal information more secure or removed from records.

[1] For example, in the context of the 2017-18 Centrelink Robodebt debate, the Department of Human Services provided the media with personal information about a former client to ostensibly correct the record, when in fact what was provided went well beyond that (http://www.abc.net.au/news/2018-05-31/privacy-precedent-what-can-the-government-reveal-about-us/9816700). In 1999, the then Howard government breached privacy laws during caretaker period by using personal Centrelink data to send Age Pensioners information about a proposed (not actual) government policy.

2. ***Noting that particular groups within the Australian community can experience new technology differently, what are the key issues regarding new technologies for these groups of people (such as children and young people; older people; women and girls; LGBTI people; people of culturally and linguistically diverse backgrounds; Aboriginal and Torres Strait Islander peoples)?***

A key observation of my research is that as digital data has increased, so too has the ability to profile and delineate between individuals and sub-groups based on personal characteristics. Importantly, such delineation does not equate to 'personalised' or 'individualised' treatment, but treatment based on quite refined group memberships (see e.g. Henman 2005). When personal characteristics based on wider forms of inequality or discrimination (e.g. ethnicity, gender, sexuality, religion) are included in designing algorithms (be they based on learning algorithms or statistical analyses) the capacity to reinforce pre-existing social fissures is heightened. There is a corresponding long-standing debate in the USA about racial profiling.

Treating people differently based on the personal characteristics can be justifiable, but a range of human rights, social and ethical principles come into play:

- Differences in treatment need to be justified (not just on science, but on what it is that is being achieved).
- Thus, selecting people (which may include consideration of disadvantaged group membership) for differentiated treatment should result in treatment to alleviate (rather than control or coerce) those groups. A parallel principle rests in our Human Rights Law (and Constitution?) whereby discrimination based on indigenous identity can only be undertaken to ameliorate disadvantage, not exacerbate this.
- Treating people differently can politically undermine collectively provided services, precipitating a dynamic where services to the poor will become poor quality services.

Significantly, algorithmic profiling undermines the visibility of group membership as a basis for administrative decision making. This is because the algorithm appears 'objective', based on the 'truth' of mathematics. Secondly, profiling takes into account a wide range of personal attributed/characteristics and derives an outcome resulting from complex calculations using these attributes. Thus, the way in which social disadvantages are used and reproduced by algorithmic informed decision making is very unclear and invisible. In short, it is very hard to say the algorithm is 'racist' or 'sexist', because it can combine all of these different facets of disadvantage into a 'black box' calculation.

5. How well are human rights protected and promoted in AI-informed decision making? In particular, what are some practical examples of how AI-informed decision making can protect or threaten human rights?

The above text provides some illustrations of how AI-informed (or algorithmic) decision making can challenge human rights.

In the operation of algorithmically informed decision making, it is necessary to strongly emphasise the distinction between *de jure* and *de facto* protection of human rights, between human rights in theory versus those in practice. This is because the experience of human rights by citizens/service users in the world often differs from formal policy and law. For example, from my research I have encountered several occasions when the operation of an algorithm has been inconsistent with the law. Sometimes this is ignored, in other times the law is adjusted to be consistent with the algorithm, and in other times the algorithm is amended to reflect the law. The (increasingly) complex nature of algorithms make the gap between formal and experienced human rights increase. The increased gap resulting from automation is often invisible because administration by algorithm is typically thought of as equivalent to administration by human, albeit the decision maker is different. Often this is not the case, as the case of Robodebt illustrates (Henman 2017; see also Senate inquiry).

Some key points where algorithm and AI informed decision making challenge human rights (particularly in a government administrative context) are:

- Lack of responsibility and accountability. When organisations make decisions using or based on algorithms, there is typically a **lack of responsibility and accountability** by those organisations for problematic decisions and their negative social and economic impact on individuals. This is initially because of the 'black box' nature of algorithms (protected in the private sphere by commercial intellectual property, and protected in the public sector by a lack of providing public review of their algorithms). In the private sector there is also a lack of accountability based on the complex legal terms of service contact that individuals (and organisations) agree when registering for such services. As is evident from the past, it has taken very public debates to have social media companies to enhance their responsibility and accountability processes (such as increased moderating of content, increased capacity to remove content and appeal organisational decisions about removal; increased capacity for right to forget). However, consideration of legal settings that enhance these opportunities by service users would strengthen a culture of responsibility and accountability practices by digital companies and manufacturers of digital products.

- Related to responsibility and accountability is the **right to appeal and review** of decisions. Within the private sector there is no such capacity, and it is largely up to the digital provider's willingness to do this. Within government, the right to appeal and review sits within wider frameworks, but can also be overridden or greatly undermined in practice. This is well illustrated by Centrelink's Robodebt system, whereby people advised as having a debt were excluded from appealing through standard Centrelink communication processes, and were only able to lodge an

appeal through the computer system. This example demonstrates that just as governments should continue to offer multiple communication channels, so too should the rights (including appeal and review) be available through multiple communication channels.

- A key practical element to realise the right to appeal, is a **right to explanation** for a decision. This is an area that is currently poorly provided by computer informed decision making in government, and will only become worse with learning algorithms. If one does not understand how a decision to raise a debt (in the case of Robodebt) or a decision to not be allow to fly (in the case of passenger screening tools) was made, then how can a person make a case that the decision was erroneous or problematic. Without a right to explanation the right to due process is vacuous. Development of administrative law to enshrine such a right to explanation should be proposed by the Commission.

- The capacity for **redress** is also poorly protected in Australian private and government administration. Unfortunately, the civil courts seem to be the main opportunity for seeking redress for detrimental outcomes based on poor algorithmic decision making.

It is possible for AI-informed decision making to enhance human rights by being designed to remove forms of discrimination. However, how these play out in practice importantly relates to the way in which algorithms are used within organisations. To what extent, for example, do they override human decisions or do they inform human decisions. An important determinate of whether human rights are enhanced or not relates to whether the final decision (whether made by person or algorithm) is more consistent with the human rights of the person than the decision had it made by the other actor. For example, if a judge overrides an algorithm to suggest a longer sentence, than this arguably reduces the rights of the defendant. Similarly, if AI soldier overrides a human decision to not shoot when the human would have shot, then this arguably is more consistent with human rights. What matters is if the decision (whether human or algorithm made) errs on the side of the human rights of the subject or not (see also Henman 2005).

6. ***How should Australian law protect human rights in respect of AI-informed decision making? In particular: a) What should be the overarching objectives of regulation in this area? b) What principles should be applied to achieve these objectives? c) Are there any gaps in how Australian law deals with this area? If so, what are they? d) What can we learn from how other countries are seeking to protect human rights in this area?***

7. ***In addition to legislation, how should Australia protect human rights in AI-informed decision making? What role, if any, is there for: a) An organisation that takes a central role in promoting responsible innovation in AI-informed decision making? b) Self-regulatory or co-regulatory approaches? c) A 'regulation by design' approach?***

Given I am not a lawyer I will deal with these two questions together. There may be case for having different legislation for government and other use of AI-informed decisions due to the intrinsic difficulties with commercial intellectual property.

- A key principle for government AI (and algorithm) informed decision making is to **require the algorithm to be made public** for scrutiny and testing to ensure the algorithm does what it supposed to do. In the case of learning algorithms, it is also essential to provide supporting information for how the algorithm was trained. What data was used to train the algorithm as input and output? This is essential to identify if the algorithm uses indicators of social disadvantage (e.g. gender, ethnicity, indigeneity) to determine outcomes, the presence of which may, due realities of past discrimination, only serve to reinforce discrimination and disadvantage (see for comparison my paper on computer modelling and democratic processes, Henman 2002).

- Another key principle is that there remains a process of **inbuilt human override** to enable intervention when AI decision making might be regarded as problematic. This could be inbuilt into the design of the IT system, or in the organisational process surrounding the use of the system (such as informing – not determining -human decision making).

- **Government should act as an example of best practice** in digital human rights, and consider leading industry awards (such as it did with e-government awards) to publicise digital human rights innovation. In a parallel manner, government has led workplace gender equity.

- Government could also fund research that develops technical frameworks for **digital human rights by design**, paralleling 'privacy by design' and 'digital ethical framework' design and production processes.

a) An organisation that takes a central role in promoting responsible innovation in AI-informed decision making?

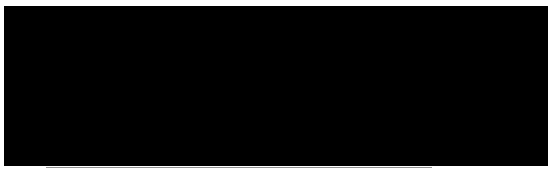
There are already a range of public sector agencies that provide oversight of government decision making that ostensibly ensure human rights in administrative law and other matters, for example, Ombudsmen, Privacy Commissioner, Human Rights Commission, AAAT and Right to Information/Information Commissioners. Rather than creating a new body, it would make sense to variously embed and strengthen digital human rights into these processes.

b) Self-regulatory or co-regulatory approaches?

As the recent Royal Commissions on Banking and Financial Services, and on Aged Care demonstrate, self-regulatory approaches have demonstrated limited effectiveness in addressing governance failures. A stronger approach would be preferable. This would also

provide greater steering for new innovations, rather than try to intervene after the negative effects have been felt.

I trust my contributions have been helpful to your deliberations, and would be very happy to discuss any of these matters further.



Paul Henman BScHons (computer science), PhD, GCEd
Principal Research Fellow, Centre for Policy Futures
Associate Professor of Digital Sociology and Social Policy
University of Queensland
[redacted]

2 October 2018

References

Caswell, D., Marston, G., & Larsen, J. E. (2010). Unemployed citizen or 'at risk'client? Classification systems and employment services in Denmark and Australia. *Critical Social Policy*, *30*(3), 384-404.

Gillingham, P. (2006). Risk assessment in child protection: Problem rather than solution?. *Australian Social Work*, *59*(1), 86-98.

Gillingham, P., & Humphreys, C. (2009). Child protection practitioners and decision-making tools: Observations and reflections from the front line. *British Journal of Social Work*, *40*(8), 2598-2616.

Henman, P. (2004). Targeted! Population segmentation, electronic surveillance and governing the unemployed in Australia. *International Sociology*, *19*(2), 173-191.

Henman, P. (2010) *Governing Electronically: E-government and the reconfiguration of policy, public administration and power,* Basingstoke: Palgrave.

McDonald, C., Marston, G., & Buckley, A. (2003). Risk technology in Australia: The role of the job seeker classification instrument in employment services. *Critical Social Policy*, *23*(4), 498-525.

Appendix: Relevant publications by Paul Henman

Henman, P. (1997). Computer technology–a political player in social policy processes. *Journal of Social Policy*, *26*(3), 323-340.

Henman, P. (1999). The bane and benefits of computers in Australia's Department of Social Security. *International journal of sociology and social policy*, *19*(1/2), 101-129.

Henman, P. (2002). Computer modeling and the politics of greenhouse gas policy in Australia. *Social science computer review*, *20*(2), 161-173.

Henman, P. (2004). Targeted! Population segmentation, electronic surveillance and governing the unemployed in Australia. *International Sociology*, *19*(2), 173-191.

Henman, P. (2005). E-government, targeting and data profiling: policy and ethical issues of differential treatment. *Journal of E-government/Journal of Information Technology and Politics*, *2*(1), 79-98.

Henman, P. (2006). Segmentation and conditionality: technological reconfigurations in social policy. In C MacDonald and G Marston (eds) *Analysing social policy: A governmental approach*, Basingstoke: Palgrave.

Henman, P. (2010) *Governing Electronically: E-government and the reconfiguration of policy, public administration and power,* Basingstoke: Palgrave.

Henman, P. (2017) The computer says 'DEBT': Towards a critical sociology of algorithms and algorithmic governance, Paper presented at *Data for policy conference*, London. https://zenodo.org/record/884117#.WcTlEsh97IU

Henman, P., & Adler, M. (2003). Information technology and the governance of social security. *Critical Social Policy*, *23*(2), 139-164.