Centre for Policy Futures
31 May 2019

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Submission to the Australian Department of Industry, Innovation and Science's 'Artificial Intelligence: Australia's Ethics Framework' Discussion Paper

Introduction

The Ethics Framework *Discussion Paper* is a very welcome contribution to the active consideration of implications of new and emerging digital technologies in general, and Artificial Intelligence (AI) in particular.

A real strength of the *Discussion Paper* is its location of ethics and AI within a much longer trajectory of automation and digitisation. This has enabled the paper to acknowledge that important aspects of the regulation of ethical AI are already in place, such as data protection, privacy and accountability around automated decision making by government agencies. The paper also provides a very helpful overview of the rapidly multiplying (grey) literature relating to the topic, including what other jurisdictions are doing.

Given that AI is not a technological rupture, but part of an evolution of digital technologies, past experiences and policy settings help to provide some engagement with the proposed ethical framework principles, their adequacy and limitations, and their ability to achieve real world outcomes. Indeed, there are a wide range of algorithms in use by both the government and private sectors that would fail to meet many of the principles articulated in the proposal ethical framework (e.g. Centrelink's 'robodebt'[1]; Facebook and Cambridge Analytica). This in itself is not necessarily a criticism of the principles themselves. On the one hand, these (revisited) principles can help to strengthen current practices, to improve current practice. On the other hand, historical experience can also provide a check on the feasibility and/or suitability of how a lofty principle may be enacted in practice.

It is from this context that the feedback to some of the consultation questions is provided.

Responses to Consultation Questions

1. Are the principles put forward in the discussion paper the right ones? Is anything missing?

On the whole, the principles proposed provide a strong coverage of the ethical considerations. The provide an excellent basis for further refining. Specific comments on select principles are as follows:

> *2. **Do no harm**. Civilian AI systems must not be designed to harm or deceive people and should be implemented in ways that minimise any negative outcomes Government as a model developer and user of ethical AI.*
>
> This principle sounds laudable, but it is not clear how this will work in practice. For example, an automated decision to cease payment of disability support pension (paid by Centrelink) based on failure to respond to an electronic form would do harm. Would it be a breach of this principle if it is only enacting the law? Advertising in general and political advertising in particular regular have misleading information. No doubt the former is covered under consumer laws, but not the latter. I have great sympathy for improving the standards of commercial and political communication. Is it intended that the principle is seeking a higher standard than current conduct among humans?
>
> *5. **Fairness**. The development or use of the AI system must not result in unfair discrimination against individuals, communities or groups. This requires particular attention to ensure the "training data" is free from bias or characteristics which may cause the algorithm to behave unfairly*
>
> This is an important principle, however, it is far from unclear what counts as 'unfair discrimination'. A key outcome of algorithms in general and AI in particular is to more accurately respond to the individuality of people; products and services can be made more niche. Prices can be increasingly differentiated to different people, such as advertising higher airfares for people using an Apple IOS

[1] Henman 2017; Senate 2017; Carney 2018.

platform compared to a MS Windows system based on knowledge that the former are typically high paid. Similarly, people calculated to be ‘at high risk’ can be subject to more surveillance than others. This in turn fundamentally recasts the notion of a ‘level playing field’ to highly differentiated ones. One challenge, as alluded to, is that these differentiations can reproduce discrimination/bias, but also can be seen as being justified based on being ‘objective’.[2] It is also important to recognise that the nature of fairness can be ‘in the eye of the beholder’, and that there are different expectations of equal treatment from government and public services, than in the private sector.

6. ***Transparency & Explainability****. People must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions.*

Considering the reality of algorithmic decision making already in place today, the notion that people be informed of ‘when an algorithm is being used that impact them’ it completely infeasible. Referring to the case of the disability support pensioner mentioned earlier, what does it matter if the decision was made by a person or an algorithm? It is the basis of the decision that is more important, and the ability to appeal it that is more significant. Moreover, how might Facebook or Google enact such a principle advising people that search results for news feeds have been derived by an algorithm? More work is needed to think about what ‘transparency’ is in practice.[3]

The issue of explainability is a very important one and one that is getting a lot of academic interest. This principle should involve more than providing people subject to decisions with “information the algorithm uses to make decisions”, but also a basis for the decision. Returning to robodebt to illustrate, people receiving a debt notice should reasonably expect to have an explanation of how a debt was calculated, how information was used, not just what was the information used[4]. This is a basic principle of administrative justice and procedural fairness. Indeed, the EU has acknowledged a right to explanation in its General Data Protection Regulation[5]. (There are, of course, cases when we may expect that the notion of transparency and explainability as articulated above to legitimately not operate; for example in the USA has long used an algorithm in air passenger screening to allocate passengers traffic light risk categories and treat passenger groups differently.) Accompanying this, there is now research being conducted about trying to discern the key information used by machine learned algorithms in making particular decisions (given that the relationship between input data and output data is not designed by humans). The Department is strongly encouraged to develop a deeper understanding of algorithmic ‘explainability’.[6]

7. ***Contestability****. When an algorithm impacts a person there must be an efficient process to allow that person to challenge the use or output of the algorithm.*

This is an important principle; however, the terminology in public administration is more typically ‘appeal, review, correction and redress’. The need for correction and redress is important and also relates to principle 8. Redress not only relates to compensation, but ensuring that the harm that has been done due to an error can be undone as much as possible. For example, if an erroneous AI decision has led to a chain of automated data flows and actions (as in the case of Ibrahim Diallo’s experience of being fired by a machine[7], or if a wrongful allegation of child abuse is recorded on a system flowing through to automated decisions to remove a child) than the ethical principle is not just contestability, but also the error be corrected and the damage undone.

[2] C.f. Henman 2005
[3] Pasquale 2015; Watcher et al 2017
[4] Senate 2017; Henman 2017
[5] Edwards & Veale 2017; Watcher et al 2017
[6] Doran et al 2017
[7] https://idiallo.com/blog/when-a-machine-fired-me

*8. **Accountability**. People and organisations responsible for the creation and implementation of AI algorithms should be identifiable and accountable for the impacts of that algorithm, even if the impacts are unintended.*

It is suggested that this principle be renamed 'Responsibility and Accountability', as responsibility is different from, but related to, accountability and is used widely in the literature. Also, it would help to have '**use** of AI algorithms' be clearly included as implementation and use are not equivalent. It needs to be clear, for example, that if a state child protection agency makes use of a commercial risk assessment tool[8], or a state government uses an off the shelf payroll system[9], that it is not just the developer of the tool that is accountable for its operation, but also the agency that decides to deploy it. This is turn means that government agencies and organisations must undertake due diligence in deciding to adopt that software.

2. Do the principles put forward in the discussion paper sufficiently reflect the values of the Australian public?

3. As an organisation, if you designed or implemented an AI system based on these principles, would this meet the needs of your customers and/or suppliers? What other principles might be required to meet the needs of your customers and/or suppliers?

4. Would the proposed tools enable you or your organisation to implement the core principles for ethical AI?

5. What other tools or support mechanisms would you need to be able to implement principles for ethical AI?

6. Are there already best-practice models that you know of in related fields that can serve as a template to follow in the practical application of ethical AI?

7. Are there additional ethical issues related to AI that have not been raised in the discussion paper? What are they and why are they important?

Data retention and destruction. Accompanying the fourth principle of *privacy protection* are related principles about data retention and destruction. In terms of the latter, greater consideration needs to be given to the 'right to forget'[10] and when personal data is not appropriate to be used for AI input and decision making, in a parallel way criminal convictions are spent after so many years after the sentence has been served.

Thinking more broadly, the proposed ethical principles do not capture a wider, more sociological perspective about the longer term or more structural element of AI decision making and society. A key aspect of AI is its ability to increasingly treat individuals *qua* individuals based on personal characteristics and behaviours, or what one Chinese employee has described as "a scalpel"[11]. This can have many benefits for personalised services and information, but can fundamentally fragment the sense of society being a shared existence with shared experiences. We can see this in the rise of echo chambers[12], which can be attributed to an increasingly politically partisan public sphere. It can also be observed in the differential treatment for denial of services, such as in China's social credit system[13], and growth of inequality. Having an awareness of the systemic social effects is also an important consideration of the ethical dimensions of AI.

[8] Gillingham 2017; 2019; Gillingham and Humphries 2009

[9] http://www.healthpayrollinquiry.qld.gov.au/

[10] Rosen 2011; Floridi 2015

[11] https://livetrendynews.com/ai-is-a-scalpel-chinas-automated-censors-cutting-deep-ahead-of-tiananmen-anniversary/

[12] Flaxman et al 2018

[13] Engleman et al 2019

About UQ’s Centre for Policy Futures and Author

The University of Queensland, Centre for Policy Futures

Created in 2017, The University of Queensland’s Centre for Policy Futures (CPF) aims to enhance the University's position as a key source of ideas and insights on the policy priorities that matter to Australia and the Pacific region. It does this through robust, rigorous and timely research and sustained policy engagement. The Centre's researchers, affiliated senior associates and visiting fellows pursue a vibrant research program focused on independent and peer-reviewed research, as well as commissioned reports, discussion papers, and policy briefs. Working closely with governments, international organisations, and key stakeholders, the Centre specialises in three policy areas:

- Science, Technology and Society
- Sustainable Development Goals and Capacity- Building
- Trade, Foreign & Security Policy

In addition to its research program, the Centre provides policy engagement and studies, as well as executive education involving academics across UQ and beyond. This approach enables the Centre to be flexible and responsive to policy matters as they arise.

The Centre is leads a multi-million dollar **CSIRO-UQ research collaboration 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. At UQ, this collaboration involves a Principal Research Fellow, a Postdoctoral Research Fellow for Digital Human Rights, a Postdoctoral Research Fellow on the governance and regulation of synthetic biology, and eight PhD students involved in various projects relating to responsible innovation of new and emerging technologies being developed by CSIRO.

Associate Professor Paul Henman

Paul Henman is Associate Professor of Digital Sociology and Social Policy, School of Social Science, and Principal Research Fellow, Centre Policy Futures at the University of Queensland. In the latter role is leads the Science, Technology and Society research program, and the CSIRO-UQ Responsible Innovation partnership. As outlined below, he is ideally placed to provide expert advice into this White Paper process.

Paul has over 20 years of active research interest in digital technologies and public governance. His 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 Paul’s research has focused on governments’ use of digital technologies, his work also provides insights for the private and NGO sectors.

In particular, Paul’s research has investigated the ways in which new digital technologies have shaped the types of policy and services that can be and are enacted. His work predates current concerns about algorithms in profiling and targeting by over a decade. In the early 2000s, he identified the policy, social and ethical dynamics associated with digital technologies’ disruption of public policy and administration principles, often leading to increased inequalities (e.g. Henman 1997; 1999; 2002; 2004; 2006; 2010; Henman & Adler 2003)

Significantly, Paul’s 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). This has provided him with insights not typically open to people without such interdisciplinary training. To date, he has managed almost $7 million in research funding, including from the Australian Research Council, IBM, CSIRO, and the former National Office for the Information Economy. He has published 4 books and over 70 academic papers. He is currently lead an international comparative study of government web portals in 10 countries.

Importantly, Paul has also worked in government as a policy analyst (1996-99) thereby providing him with important insights into the way in which governments operate. Consequently, he has 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*, the Parliamentary Joint Committee on Intelligence and Security *Identity-matching Services Bill 2018* Inquiry, and the Australian Human Rights Commission's consultations on *New Technology and Human Rights* and *AI Governance and Leadership White Paper.*

References

Carney, T. (2018). Robo-debt illegality: The seven veils of failed guarantees of the rule of law?. *Alternative Law Journal*, 1037969X18815913.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv*:1710.00794.

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, *16*, 18.

Engelmann, S., Chen, M., Fischer, F., Kao, C. Y., & Grossklags, J. (2019, January). Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines Good and Bad Behavior. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 69-78). ACM.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1), 298-320.

Floridi, L. (2015). The right to be forgotten”: a philosophical view. *Jahrbuch für Recht und Ethik-Annual Review of Law and Ethics*, *23*(1), 30-45.

Gillingham, P. (2017). Predictive risk modelling to prevent child maltreatment: insights and implications from Aotearoa/New Zealand. *Journal of public child welfare*, *11*(2), 150-165.

Gillingham, P. (2019). Decision Support Systems, Social Justice and Algorithmic Accountability in Social Work: A New Challenge. *Practice*, 1-14.

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. (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* [now *Journal of Information Technology & 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, *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.

Pasquale, F. (2015). *The black box society*. Harvard University Press.

Rosen, J. (2011). The right to be forgotten. Stan. L. Rev. Online, 64, 88.

Senate. Community Affairs References Committee. (2017). *Design, scope, cost-benefit analysis, contracts awarded and implementation associated with the Better Management of the Social Welfare System initiative*, Canberra: The Senate.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, *2*(6).

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76-99.

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

Contact details

Associate Professor Paul Henman
T +61 7 **3443 3142**
M +61 **402 734 218**
E p.henman@uq.edu.au
W https://policy-futures.centre.uq.edu.au/

CRICOS Provider Number 00025B