Artificial Intelligence: Australia's Ethics Framework

University of Melbourne Response

June 2019

Executive Summary

The University of Melbourne welcomes the opportunity to respond to the Department of Industry, Innovation and Science's Discussion Paper, *Artificial Intelligence: Australia's Ethics Framework*. The continuous development of new technologies has considerable ethical implications, entailing both risks and opportunities. We welcome the interest in this area shown by the Department and by the Data61 division within the CSIRO.

The Discussion Paper recognises that emerging technologies come with risks but also generate opportunities. We support the attempt to account for both. While there is often an understandable focus on the dangers that AI and related technologies represent, it is important that we not lose sight of the potential benefits that these technologies offer. In many cases, there is a clear ethical imperative to make use of automated systems e.g. where automated vehicles are likely to significantly reduce road fatalities. A key challenge in developing an AI ethics framework is responding to the risks posed by new technologies without undermining innovation in this area, thereby depriving Australia of the benefits. The Discussion Paper has attempted to achieve this balance.

We do, however, argue that there are significant issues with the framework described in the Discussion Paper, largely relating to the Paper's narrow focus in some areas. This submission addresses some of these issues. The following comments do not represent a comprehensive statement on what an ethics framework should look like. Instead, this submission outlines specific issues that the Discussion Paper has either overlooked or mis-characterised. These include:

- The discussion of 'privacy' captures only a subset of the ethical and legal issues associated with AI and related technologies and fails to acknowledge the other relevant public inquiries underway in this area.
- The discussion of individual consent for the use of data requires further development. There are a number of important considerations relating to consent that are not addressed in the Discussion Paper.
- The discussion of the significance of AI and related technologies to Indigenous Australians and to persons with a disability needs to be further developed.

As well as raising these and other issues with the content of the Discussion Paper, comment on each of the "Core principles for AI" identified in the Discussion Paper – and some suggested additional principles – is included in the Appendix of this submission.

Prior to coming to a response to the Discussion Paper, we offer a broad overview of the University of Melbourne's engagement with the legal and ethical challenges associated with AI and related technologies, outlining some of the key points made in our contribution to public consultations in this area.

We would welcome the chance to further discuss these issues in more detail with the authors of the report and with the Department of Industry, Innovation and Science.

For more information, please contact Professor Mark Hargreaves, Pro Vice-Chancellor (Research Collaboration & Partnerships) on 03 8344 4447 or m.hargreaves@unimelb.edu.au.

The University of Melbourne and Digital Ethics

The University of Melbourne has a deep engagement with the challenges associated with the evolution of new technologies. We refer to the University's response to the Australian Human Rights Commission's (AHRC) *Human Rights and Technology* Consultation Paper in 2018, and in response to the AHRC and World Economic Forum's *Artificial Intelligence: Governance and Leadership White Paper* earlier this year. In each case, the University's response drew from a community of researchers from across a range of fields. Key points made in those submissions include the following.

Human Rights by default and design

In our submission to the AHRC's 'Human Rights and Technology' Consultation Paper, the University promoted the general principle of 'Human Rights by default and by design' as a way of thinking about the human rights-related implications of new technologies. The motivating insight behind this principle is that human rights-related considerations ought to inform the development of new technology from the beginning. This contrasts with an approach that seeks to ensure that already mature technologies are brought into line with a human rights standard.

This basic principle – defined broadly in terms of 'ethics' rather than the narrower category of 'human rights' – is largely consistent with the approach suggested in the Discussion Paper. A clearly articulated principle such as this may nonetheless be useful in highlighting the integrated response that is needed in developing an ethical framework for the use of AI and related technologies. An ethics by default and design approach would help to ensure that the relevant considerations influence development in an ongoing way, instead of retro-fitting already developed technologies (while acknowledging that retrofitting will in some cases be necessary for established technologies). An ethics by design principle also underscores the importance of bringing a range of viewpoints to bear upon the design of new technologies, including users who are especially vulnerable to technology that is poorly designed.

A multi-disciplinary approach

As noted, the University has adopted a multi-disciplinary approach to investigating the legal and ethical challenges associated with AI and related technologies. This broad approach is appropriate for the work Data61 is undertaking to develop an ethics framework for AI. The challenges raised by emerging technology are not merely or primarily technological challenges. The response should therefore not be limited to those working in technology or related fields. In addition to technology researchers, an ethics framework should draw from a community that includes the humanities and social sciences, legal scholars and disability researchers.

A Technology Commissioner

The University of Melbourne suggests consideration of a new Chair – a Technology Commissioner – being established within the Human Rights Commission to have oversight of this area. The legal framework for protecting human rights in the context of new technology is dispersed across a range of Acts and instruments, including the Privacy Act, various anti-discrimination laws, and Australian Consumer Law. Attempting to re-build this legal framework from the ground up is neither desirable nor practically feasible. Moreover, in view of the rapidly changing context that laws and regulations need to grapple with, a "set and forget" approach to this area is inappropriate. A new Chair would enable a cohesive approach across Government to engaging with these challenges, working with industry and the research sector.

Comment on the Discussion Paper

As noted above, the University of Melbourne argues that there are considerable gaps in the ethics framework presented in the Discussion Paper, largely owing to a relatively narrow approach to some of the issues that it raises. The following comments address some of these gaps. The intention in these comments is not to identify everything that should be included in an ethics framework for AI, but to identify specific areas where the ideas raised in the Discussion Paper need further development.

A focus on 'AI'

The Discussion Paper makes 'AI' its specific area of focus, by implication placing beyond scope non-AI forms of technology. We have argued elsewhere that drawing a line between AI and non-AI technology for the purposes of a legal, regulatory or ethical framework is problematic.[1] Firstly, the distinction is on a basic level ethically relevant. AI is just one form of digital technology. There is no reason to single it out against other forms which raise the same ethical challenges relating to privacy, fairness, the avoidance of harm, etc. Secondly, creating a set of AI-specific guidelines may encourage organisations to seek to categorise a given application as something other than AI as a way of avoiding the relevant requirements.

Given this, an ethical framework should be developed independently to algorithmic implementation details, remaining agnostic with respect to the type of technology in question. Instead of a framework that is specific to AI, using a broader category such as 'automated decision-making' will help to avoid the problems identified above.

Privacy

Privacy-related issues are of central importance to the development of an ethical framework for AI and related technologies. While the Discussion Paper addresses privacy, there are significant flaws in its treatment of this issue. We refer to Salinger Privacy's submission in response to the Discussion Paper, which identifies a number of these flaws.[2]

For example, the Discussion Paper identifies 'privacy protection' as one of eight core principles for inclusion in the ethics framework. The broader articulation of that principle appears to assume that privacy law is only (or primarily) concerned with "private data". Since privacy law covers a much broader set of issues than this, couching the discussion in terms of private data risks sidelining important privacy-related considerations.

Anonymisation and re-identification

The possibility of individuals being re-identified through anonymised data is a crucial issue in the context of a discussion of privacy (raised in section 3.3 of the Discussion Paper). This possibility underscores the key point that de-identifying or anonymising data does not obviate the need to obtain the data owner's consent before sharing. Unfortunately, de-identification of detailed unit-record level data does not work without substantially reducing the information content of the data. Recent episodes in Australia have shown that sincere efforts at de-identification were insecure.

The Discussion Paper notes that the Government has sought to address this issue through the Privacy Amendment (Re-identification Offence) Bill 2016 which would prohibit the re-identification of data.[3] Unfortunately, the proposed amendments to *The Privacy Act* could deliver the worst of both worlds.

---

[1] See University of Melbourne, 2018, *Response to Australian Human Rights Commission Issues Paper*, p.7. https://about.unimelb.edu.au/__data/assets/pdf_file/0017/60146/UoM_submission_Human_Rights_and_Technology_Issues_Paper.pdf

[2] Johnston, Anna, 2019, "The ethics of artificial intelligence: start with the law" (Salinger Privacy). https://www.salingerprivacy.com.au/2019/04/27/ai-ethics/

[3] pp.30-31.

The amendments would prevent an open examination of problems, thus making them less likely to be discovered by Australian researchers. Bad actors could nonetheless continue to exploit opportunities to re-identify individuals. This represents a major risk. Even data that are not made open may be distributed to entities (such as an insurers or employers) with strong incentives to identify individuals. Hence, anonymisation should not allow the data holder to share other people's data without their consent.

Digital platforms inquiry

We also note that the Australian Competition & Consumer Commission's (ACCC) 'Digital platforms inquiry' is currently active and due to deliver its final report by June 2019. A number of the points raised in the Inquiry's Preliminary Report are directly relevant to privacy issues associated with the collection and storage of data. For example, in the context of the "information asymmetry between digital platforms and consumers", the ACCC offers the preliminary finding that:

> […] consumers are generally not aware of the extent of data that is collected nor how it is collected, used and shared by digital platforms. This is influenced by the length, complexity and ambiguity of online terms of service and privacy policies. Digital platforms also tend to understate to consumers the extent of their data collection practices while overstating the level of consumer control over their personal user data.[4]

While information asymmetry is clearly relevant to the issues of privacy and consent, this issue is not addressed explicitly in the Discussion Paper. More generally, the ethical framework under construction will be more robust the better integrated it is with other work being done in this area.

Consent and "reasonable expectation"

The issue of consent is crucial to the conversation on the ethical use of AI and related technology. While this issue surfaces in the context of the discussion on privacy and data breaches, it is noteworthy that consent does not feature explicitly in the principles proposed in the Discussion Paper. A principle of transparency insists that people should be informed when an algorithm that impacts them is being used, but that consenting to that use is a further issue.

The discussion would also benefit from engaging with the conceptual question of what it means for an individual to consent to their data being used. The idea of a 'reasonable expectation' may be useful in capturing what is morally significant in any agreement for the use of an individual's data, emphasising that it should only be used for purposes and under conditions that those affected have reason to expect and accept as appropriate. This approach would help avoid the problems with a narrow approach that focuses on a discrete act of consent performed at a specific time:

- **Power asymmetries:** There are cases where an individual has little choice but to consent to the suggested use of their data, e.g. where an application is required for work. A discrete act of consent is clearly insufficient in these circumstances. Couching consent in terms of 'reasonable expectations' makes it clear that *what* an individual is consenting to must itself be reasonable.
- **Information asymmetries:** As noted above, an individual may "consent" to their data being stored and used without properly understanding the nature of this use. Defining consent in terms of reasonable expectations provides some protection in the face of information asymmetries.
- **Secondary uses of data:** Data that are collected for a given purpose may be put to additional or new uses at a later stage. An ethical framework should make clear the conditions under which such uses are appropriate. The notion of 'reasonable expectations' is useful in helping to define these conditions. The Discussion Paper comes close to this point in noting the requirement from

[4] Australian Competition & Consumer Commission, 2018, *Digital Platforms Inquiry: Preliminary report*, p.8.

The Privacy Act that consent be "current and specific".[5] However, this issue is worth addressing directly in view of the scope for secondary use of data.

- **Data use affecting others:** The storage and use of data often impact not only the user themselves but also others, e.g. friends, family etc. The use of reasonable expectation helps to safeguard the rights and interests of other affected parties.

Indigenous Australians and AI

Section 6.5 of the Discussion Paper addresses Indigenous communities and their relationship to AI, identifying three interrelated issues for consideration: the need to comply with Indigenous cultural protocols; the importance of AI development and use being guided by cross-cultural collaborative approaches; and the need for transparency that ensures that "Indigenous people and organisations are clear about how AI learning is generated and why this information is used to inform decisions that affect Indigenous estates and lives."[6]

This discussion would benefit from a sharper focus on the active role that Indigenous communities should play in the development of new technology. The Discussion Paper has a significant focus on the ways in which Indigenous persons and communities may be impacted by emerging technology, with much less attention given to the role of Indigenous persons as users and potential innovators. (The reference to "cross-cultural collaborative approaches" points to the importance of Indigenous community involvement in the collection and analysis of data). An ethics framework should emphasise the need to support Indigenous *participation* and *agency* in shaping the AI agenda, recognising the importance of including Indigenous communities in tech development and in the design of policies, and the matter of Indigenous data sovereignty.[7] AI and related technologies have a key role to play in advancing solutions to complex issues affecting Indigenous Australians. The opportunities will go unrealised if Indigenous people are not included in the development of these technologies.

We should also note that Indigenous communities are especially vulnerable to privacy-related risks that come with (for example) the collection and storage of data on individual persons. The risk of individuals being re-identified (see below) through anonymised data is heightened when dealing with minority groupings and with sparsely distributed populations. The heightened risks for Indigenous communities and other marginalised groups is worth addressing directly in an ethics framework.

Accessibility

The impact of new technologies on persons with a disability is an issue that deserves more attention than it receives in the Discussion Paper. In discussing the issue of fairness, the Paper rightly identifies persons with a disability as one of the cohorts of Australians who are potentially vulnerable to discrimination where algorithms are biased or rely on unrepresentative input data.[8]

While important, issues relating to discrimination are only part of the picture when it comes to the ethical implications of new technologies on persons with a disability. An additional set of issues concern the need for 'accessible technology' i.e. ensuring that new technologies are accessible to persons with a disability. In our submission to the AHRC consultation on Human Rights and Technology, the University of Melbourne noted that digital technology is proving to be a powerful

[5] p.28.

[6] Discussion Paper, p.56.

[7] See, for example, Kukutai and Taylor (eds.) 2016, *Indigenous Data Sovereignty: Toward An Agenda*, Canberra, ANU Press; Indigenous Data Sovereignty Symposium 2017, Indigenous Studies Unit University of Melbourne https://mspgh.unimelb.edu.au/research-groups/centre-for-health-equity/indigenous-studies/indigenous-data-sovereignty-symposium ; Indigenous Data Sovereignty Communique 2018, Maiam nayri Wingara Indigenous Data Sovereignty Collective and the Australian Indigenous Governance Institute)

[8] See p.39 and p.41.

enabler for persons with a disability.[9] However, if poorly designed, new technology can further contribute to the marginalisation of those with a disability.[10] In addition to discrimination-related issues, the need for inclusive design warrants emphasis in an ethics framework for AI.

Accountability and collaborative research

Section 7.1.6 discusses the relationship between business and academia, and appropriately identifies the benefits promised by deeper ties between industry and research. Industry-research collaboration should have a central place in Australia's innovation agenda and is central to the ethical development of AI and related technology.

However, the Discussion Paper does not address the accountability-related issues that are associated with collaborative research. If we are to incorporate "ethics by design", starting when projects are initiated, then there needs to be clarity at the beginning of funding relationships about which party is accountable for the ethical use of the research product. The basic point that technology is often put to new uses – i.e. different from that initially intended – is relevant here, as researchers involved in development may be unaware of those uses. The lines of responsibility need to be clearly drawn in such cases.

Robustness

An important ethical dimension for AI and machine learning omitted in the Discussion Paper concerns the robustness of automated systems. While minute changes to a small number of an image's pixels may not alter a human's perception of the image, such changes regularly fool state-of-the-art machine learning systems.[11] A number of related attacks on AI systems are well known in the study of adversarial machine learning[12] and have been demonstrated in the real world.[13] In many such cases, human decision-makers would not make these errors, and would be able to continue to make good decisions in situations that are unlike those they have seen before. In machine learning, related issues of poor data hygiene lead to overfitting systems to the data they have been trained on, and broader problems of replication crises widely publicised in the sciences.[14] This is further compounded by the fact that AI systems struggle to identify that they are presented with a problem that they are not trained for. Comprehensive frameworks for ethics in AI must recognise the importance of robustness and the need to support it through practices that require human oversight and intervention.

---

[9] See *Response to Australian Human Rights Commission Issues Paper*, pp.11-12.

[10] See, for example, Haxton, Nancy, 2017, "Blind groups push for CBA to find solution to 'inaccessible' touchscreen EFTPOS terminals", ABC. https://www.abc.net.au/news/2017-07-28/blind-groups-push-for-solution-to-inaccesible-eftpos-terminals/8751366

[11] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199*(2013).

[12] Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, and J. D. Tygar. *Adversarial Machine Learning*. Cambridge University Press, 2018.

[13] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. "Adversarial Examples in the Physical World." In *Artificial Intelligence Safety and Security*, pp.99-112. Chapman and Hall/CRC, 2018.

[14] Andrew Gelman, and Eric Loken. "The statistical crisis in science: data-dependent analysis--a" garden of forking paths"--explains why many statistically significant comparisons don't hold up." *American Scientist* 102, no. 6 (2014): 460-466.

Contributors to this submission

Dr Chris Culnane, Lecturer, School of Computing and Information Systems

Mr Mark Fallu, Digital & IT Advisor, Chancellery Research

Mr Assyl Haidar, Director, Digital and Data, Chancellery Research

Professor Mark Hargraves, Pro Vice-Chancellor (Research Collaboration & Partnerships)

Professor John Howe, Director, Melbourne School of Government

Associate Professor Reeva Lederman, Academic, School of Computing and Information Systems

Associate Professor Tim Miller, Academic, School of Computing and Information Systems

Professor Scott McQuire, Academic, School of Culture and Communication

Professor Jeannie Paterson, Academic, Melbourne Law School

Professor Megan Richardson, Academic, Melbourne Law School

Associate Professor Ben Rubinstein, Academic, School of Computing and Information Systems

Professor Liz Sonenberg, Pro Vice-Chancellor (Digital & Data)

Associate Professor Mark Taylor, Deputy Director, Centre for Health, Law and Emerging Technologies (HeLex @Melbourne)

Associate Professor Vanessa Teague, Academic, School of Computing and Information Systems

Professor Monica Whitty, Professor of Human Factors in Cyber Security, School of Culture and Communication

Appendix

Comment on "Core principles for AI"

	Principle	Comment
Principles identified in Discussion Paper	Generates net-benefits	Given the uncertainty typically associated with new technology, there are likely to be cases where an innovation is reasonable but where an organisation cannot be sure that it will generate benefits greater than the costs. The definition of 'net-benefit can arguably be arbitrary and contested in any event – a 'net benefit' from whose perspective, for example. The key point may be best captured in terms of an intent to cause harm and of foreseeable consequences. The 'Do no harm' principles articulates this point.
	Do no harm	Support.
	Regulatory and legal compliance	Support. We note that this leaves open the extent to which the legal and regulatory framework itself is adequate.
	Privacy protection	While the principle itself is supported, the ethical issues related to privacy are broader than those addressed in the Discussion Paper.
	Fairness	Support. We note, however, that 'fairness' is discussed exclusively in terms of non-discrimination. While this is important, it is only one aspect of fairness. There are separate questions as to whether the use of automated systems is fair in the circumstances e.g. in a criminal justice setting. Also, there is a need to determine what counts as 'unfair discrimination' and who determines this.
	Transparency and explainability	Support.
	Contestability	Support.
	Accountability	Support. We note that there is a need to address the accountability issues arising out of collaborative research.
Suggested additional principles	Consent/Reasonable expectations	The storage and use of data should be dependent upon the consent of individuals *and* should reflect reasonable expectations of users and other affected parties.
	Accessibility	New technology should be accessible to persons with disability and should not further contribute to their marginalisation.
	Enforceability	Where the framework identifies legal and regulatory requirements, there should be mechanisms for enforcement and penalties for non-compliance.
	Remedy/recourse mechanism	Those unfairly affected by AI should have a cost-effective, timely and non-litigious remedy available to them.