Consultation response to the Australian Human Rights Commission Discussion Paper on Human Rights and Technology, March 2020

The Australian Academy of Science (AAS) welcomes the opportunity to respond to the consultation by the Australian Human Rights Commission on the published discussion paper "Human Rights and Technology". This submission has been prepared with advice and expertise from the AAS Fellowship; AAS National Committee for Data in Science; AAS National Committee for Information Communication Science; the Australian National University's (ANU) Humanising Machine Intelligence institute; and Professor Michael Barber AO FAA FTSE, Co-Chair of the Academy's ARC LASP report on Big Data.

Introduction and key messages

The Discussion Paper put forward by the Human Rights Commission (HRC) outlines the HRC's preliminary views and proposals in three key areas:

- Regulation, leadership and good governance amid the rise of new technologies
- The use of artificial intelligence (AI) in decision making
- Accessibility of new technologies for people with disability.

The discussion paper is a thorough, sensitive and value-aware proposal that takes the issues posed by new technology seriously, makes nuanced distinctions that respect the complex nature of the phenomenon under discussion and does not shy away from suggestions and demands to protect the interests of the citizens of Australia. We particularly encourage the current direction proposed in the discussion paper that looks towards making effective use of existing regulation. We have identified several avenues for further development.

Responses to the consultation proposals and questions by the Australian Human Rights Commission

We have provided direct responses to Proposals 2, 5 – 8, 11, 13, 16 and 27 and Question A. In addition to these, some general comments are offered below:

Public trust

The report reiterates the need to build public trust in technologies. Building public trust should be about building **trustworthy systems** and not solely about developing the public's trust in technologies. Appropriate actions noted in the report that could assist in this include adequate testing processes (page 119), the establishment of a new expert body (page 133) and developing clear standards and auditing processes.

Specifying duties and duty-holders

The paper focuses on "enforceable human rights". The Academy recommends that discussion should focus on the corresponding obligations of these rights as these duties are enforceable i.e. the corresponding duties, which lead to human rights, are enforceable and not 'human rights' in and of itself. To determine what these obligations are, the HRC must identify who can and does have these responsibilities i.e. clearly articulate who the "government, companies and others" are.

Privacy

The Academy encourages the HRC to include a more detailed discussion on the issues regarding privacy, as there are many different types of privacy harm. For example, a significant harm of AI technology in the medium-term is the violation of decisional autonomy (i.e. an individual's free will to determine their own behaviour autonomously and according to their own interest) through the use of AI in the business of advertising. This was demonstrated during the 2016 Presidential campaign where Cambridge Analytica used Facebook data to target users with advertisements and campaign materials[1]. Big data analytics and AI were used to manipulate the population's decisional autonomy. The "principal function of [AI] at present is to capture personal information, create detailed behavioural profiles and sell us goods and agendas[2]", losing our decision autonomy, which "is intimately linked to free-will[3]", is worthy of further attention in the report.

Five Safes Principles

The Academy would like to draw the HRC's attention to the Five Safes Framework[4,5] designed to reduce the risk of sensitive data being used incorrectly, developed by the United Kingdom's Data Service. Data is the central pillar required for the development and operation of many modern technologies (e.g. AI tools) and the Five Safes Framework provides guidance on how to design, implement and assess data systems to mitigate data mismanagement and consequent harm.

Many Australian Government agencies generate and hold onto to vast amounts of data with significant potential to inform policy development and contribute to Australia's economic growth. Of note, the Office of the National Data Commissioner (ONDC) have developed their Data Sharing Principles[6], based on the UK's Five Safes Framework, which are designed to enable safe and appropriate data sharing:

1. Projects: Data is shared for an appropriate purpose that delivers a public benefit.
2. People: The user has the appropriate authority to access the data.
3. Settings: The environment in which the data is shared minimises the risk of unauthorised use or disclosure.
4. Data: Appropriate and proportionate protections are applied to the data.
5. Output: The output from the data sharing arrangement is appropriately safeguarded before any further sharing or release.

Proposal responses

Proposal 2: The Australian Government should commission an appropriate independent body to inquire into ethical frameworks for new and emerging technologies to:

(a) Assess the efficacy of existing ethical frameworks in protecting and promoting human rights
(b) Identify opportunities to improve the operation of ethical frameworks, such as through consolidation or harmonisation of similar frameworks, and by giving special legal status to ethical frameworks that meet certain criteria.
(c) Resource education and training for government, industry and civil society.

Though it is tempting to establish a new regulatory body, given the serious risks posed by new and emerging uses of AI, we recommend that the HRC considers proposals to integrate regulatory mechanisms for new and emerging technologies into already established bodies. The issues raised by AI are best understood as a complex grouping of problems requiring interdisciplinary and cross-disciplinary expertise, and while they can be grouped together, there are reasons to reconsider setting up a new regulatory body or commission specifically for this grouping.

First, is there evidence from other jurisdictions and sectors noted in the report that these specific kinds of interventions have, in other areas, achieved the goals set out?

Second, as AI affects every area of society, we ought to integrate a human rights approach to AI within existing regulatory bodies, encouraging them to change and adapt rather than introduce a new body to oversee the application of AI within each domain.

If the HRC is committed to establishing a new body, we recommend one whose remit is to take a holistic review of the end-to-end production and use of emerging technology and whose recommendations were used to inform government use as well as the emerging duties of care of other departments and government bodies.

Proposal 5: The Australian Government should introduce legislation to require that an individual is informed where AI is materially used in a decision that has a legal, or similarly significant, effect on the individual's rights.

The Academy agrees to this recommendation in principle. However, we would advise the HRC to consider what would be the cost of making this information available and accessible, and in doing so, conducting a cost-benefit analysis of implementing AI technologies for various uses – as per Proposal 6a.

Proposal 6: Where the Australian Government proposes to deploy an AI-informed decision making system, it should:

(a) Undertake a cost-benefit analysis of the use of AI, with specific reference to the protection of human rights and ensuring accountability
(b) Engage in public consultation, focusing on those most likely to be affected

(c) *Only proceed with deploying this system, if it is expressly provided for by law and there are adequate human rights protections in place.*

Under (a) – we would like to suggest that accountability for AI-informed decision making will be improved by more rigorously applying existing laws, as well as some targeted reform. For AI-informed decision making to be accountable, it must be:

- Lawful, complying with existing laws and having legal authority where necessary
- Transparent, encompassing the notion that affected individuals are notified of AI being a material factor in decision engaging their human rights, as well as transparency regarding government use of AI technologies
- Explainable. Requiring a meaningful explanation for an AI-informed decision (see response to Proposal 7)
- Used responsible with clear parameters for liability
- Subject to appropriate human oversight and intervention.

This is not dissimilar to the FAIR data principles[7], which also include provisions for protecting peoples' rights (see "Accessible" section).

Under (b) – In addition to consulting with individuals most likely to be affected, we recommend that the HRC is also informed of the biases present in algorithms, which emanate from unrepresentative or incomplete training data. The HRC should also note that in some cases, the individuals most likely to be affected by AI-informed decision-making systems could also be discriminated against due to these biases.

Proposal 7: The Australian Government should introduce legislation regarding the explainability of AI-informed decision making. This legislation should make clear that, if an individual would have been entitled to an explanation of the decision were it not made using AI, the individual should be able to demand:

(a) *A non-technical explanation of the AI-informed decision, which would be comprehensible by a layperson, and*
(b) *A technical explanation of the AI-informed decision that can be assessed and validated by a person with relevant technical expertise.*

In each case, the explanation should contain the reasons for the decision such that it would enable an individual, or a person with relevant technical expertise, to understand the basis of the decision and any grounds on which it should be challenged.

The discussion of explainability presented in the paper is careful and acknowledges the need for a differential approach i.e. that not all occasions require an explanation, and that both technical and non-technical explanations have value. We have identified some issues for further consideration:

- An AI-informed decision that is unexplainable may be deemed unlawful. With the known complexities of deep learning models and the difficulties in their explainability, will we be able to communicate this?

- Is it the system or the decision that must be explainable?
- Where there is a legitimate case for an individual to demand an explanation, should they be entitled to both a technical and non-technical explanation?
- Is an explanation (especially a technical explanation) always needed if we know that the outcomes of a process are meeting the metrics of success? For example, if we know that a process is giving fair outcomes, do we need an explanation of how it works?
- Auditability and technological due process may be more important than explanations.
- The paper's flow chart makes clear that no explanation will be given for 'positive' outcomes. However:
    - This requires us to determine for whom the outcome is positive. For example, suppose in a case of theft an algorithm is instrumental in recommending sentencing times. The outcome might be positive for the defendant but not for the victim or vice versa. Is this a positive outcome or a negative one? If negative outcomes can include indirect, downstream or third parties, this should be made explicit.
    - If 'negative outcomes' are reserved for a narrower class of cases, then explanations may be important for positive outcomes too. For example, suppose that a specific group is the recurring recipient of a benefit. While this may have no serious significant impact on any particular person directly, an explanation may still be due as to why benefits keep flowing into the hands of this specific group.
- While much focus has been on the explainability of AI systems, often as individuals what we are really after is a justification. Ideally, we would be offered a reason that both explains and justifies. For example, being told "you did not get parole because you are black and this algorithm is biased" is a full explanation but no justification. Being told that "this AI says you will default, it has been extremely accurate in the past, and everyone, including you, values accuracy" might be a good justification but gives no explanation.

Proposal 8: Where AI-informed decision-making system does not produce reasonable explanations for its decision, that system should not be deployed in any context where decisions could infringe the human rights of individuals.

The HRC should consider that by completely restricting the deployment of AI-informed decisions to only those that are explainable could disadvantage and hinder the development and adoption of emerging technologies in Australia. "In today's world, the real task for [AI] regulators is to create a rules structure that both protects the public and promotes industry innovation – not to trade off one against the other"[8].

Generally, the Academy does favour technology specific prohibitions that could unreasonably impede scientific inquiry, unless it can be demonstrated that there are strong national security or societal wellbeing implications.

Proposal 11: The Australian Government should introduce a legal moratorium on the use of facial recognition technology in decision making that has a legal, or similarly significant, effect for individuals, until an appropriate legal framework has been put in place. This legal framework should include robust protections for human rights and should be developed in consultation with expert bodies including the Australian Human Rights Commission and the Office of the Australian Information Commissioner.

We applaud the argument and stance put forward by the HRC on the use of facial recognition. However, facial recognition is only one form of surveillance and the arguments against its use apply equally to other instances. We recommend that in this instance, and going forward, that it would be more productive to focus on practices (e.g. surveillance) rather than techniques and technologies (e.g. facial recognition).

Proposal 13: The Australian Government should establish a taskforce to develop the concept of 'human rights by design' in the context of AI-informed decision making and examine how best to implement this is Australia. A voluntary, or legally enforceable, certification scheme should be considered. The taskforce should facilitate the coordination of public and private initiatives in this area and consult widely, including with those whose human rights are likely to be significantly affected by AI-informed decision making.

The report advocates for both human rights by design and human rights by impact assessments. For these to be effective, they must be formally linked into a single, continuous production lifecycle. That is, there must be clear criteria at all stages for an AI product that adheres to human rights standards. Defining and measuring these standards with a uniform tool ensures that the goal of design converges with product evaluation. The framework of affordances is perhaps a useful one to adopt (see Davis, 2020[9], and Davis and Chouinard, 2016[10]).

The mechanisms of affordance refer to the actions and social dynamics a technology requests, demands, encourages, discourages, refuses, and allows. These variables may differ for the same technology from person to person and context to context. In short, what a technology *requests* of one person, it can *demand* of another. What it *allows* an individual, it may *refuse* another. When designing human rights centred AI, producers will begin by determining what they want the technology to request, demand, encourage, discourage, refuse, and allow for various populations encountering the technology under a range of circumstances. Assessment will entail evaluations of the extent to which the produced outcomes meet design expectations. That is, does this AI request, demand, encourage, discourage, refuse, and allow as the designers intended?

For instance, one may begin the design process by asking:

- How can we build a technology that demands lawfulness under all circumstances?
- Under what conditions would this technology refuse transparency?
- How do we materialise processes that request oversight and intervention?

Assessment would then include questions such as:

- Does this technology demand lawfulness under all circumstances?
- Does this technology refuse transparency under any circumstances? If so, what are those circumstances?
- Does this technology encourage oversight and intervention?
- Are there circumstances in which oversight and intervention are discouraged or refused?

In short, a mechanisms and conditions framework will help designers determine how the technology should ideally operate (be designed) and evaluate the extent to which it operates in practice as expected (assessment).

Even if 'affordances' is not the specific framework, **design and assessment should be formally linked**.

In addition to formally linking 'human rights by design' and 'human rights by impact assessment', the Academy stands ready to assist the HRC to establish the taskforce, which will go on to develop this framework, and can draw on the vast expertise and talents of its Fellowship and other expert Australian scientists. The Academy also encourages the HRC to consult further with other Learned Academies (which may be consulted through the Australian Council of Learned Academies[11] (ACOLA)) when developing the taskforce and framework.

Proposal 16: The proposed 'National Strategy on New and Emerging Technologies' (see Proposal 1) should incorporate education on AI and human rights. This should include education and training tailored to the particular skills and knowledge needs of different parts of the community, such as the general public and those requiring more specialised knowledge, including decision makers relying on AI data points and professionals designing and developing AI-informed decision making systems.

We highlight the school resources provided by the Digital Technologies Hub[12], a Government initiative supported by Education Services Australia and the Australian Government Department of Education. Specifically, the Hub provides some educational resources that touch on topics regarding AI and human rights e.g. modules exploring the positive and negative social impacts of AI[13], anti-bullying[14] and the limitations and biases of data[15,16,16].

Proposal 27: Professional accreditation bodies for engineering, science and technology should consider introducing mandatory training on 'human rights by design' as part of continuing professional development.

In addition to accreditation bodies providing training and continuing professional development, we advise that funding bodies and ethics committees responsible for reviewing research grant applications regarding human rights and technology must also be appropriately equipped. These bodies must be confident in understanding the intricacies, complexities and implications of the research they are assessing.

Question responses

Question A: The Commission's proposed definition of 'AI informed decision making' has the following two elements: there must be a decision that has a legal or similarly significant, effect for an individual; and AI must have materially assisted in the process of making the decision.
Is the Commission's definition of 'AI-informed decision making' appropriate for the purposes of regulation to protect human rights and other key goals?

We have several concerns and suggestions regarding the appropriateness of the HRC's proposed definition.

Human rights are necessarily individualistic. For human rights to be violated someone must be affected significantly. This explains the focus of the definition above. However, there are many issues that affect us as a collective, especially some of the new harms arising out of big data (such as manipulation, see response under section 'Privacy' and below under 'Low-level marketing decisions'). The definition might be good to capture the kinds of issues that lie within the remit of the HRC. However, it is not good as a definition of ethically relevant AI decision making.

In its definition of 'AI-informed decision making', the Academy has concerns regarding the first element, "legal, or similarly significant, effect for an individual":

- It should be specified what is meant by the term "legal", is the Australian Government or are states and territories also considered.
- What counts as "similarly significant" effects aside from legal ones?
- How is "significant" to be assessed?
- There are many cases that fall under the broad scope of concerns for individuals noted in the paper. However, several key kinds would not meet the definition of AI-informed decision making. For example:
    - Pricing insurance or risk pooling: This practice can leave some people un-insurable. The decisions are not about individuals and yet would have significant downstream effects for individuals.
    - Low-level marketing decisions: These decisions are not significant for any particular individual and yet can have serious and concerning effects in the aggregate, especially on individual autonomy, though perhaps not to the extent that it constitutes a significant violation of any particular individual's rights. In the Commissioner's foreword, he states that "Time and again people told us, 'I'm starting to realise that my personal information can be used against me'". However, there is no explicit discussion of autonomy-based harms in the Discussion Paper.

These concerns highlight that the human rights approach is insufficient to deal with all the issues that technology raises. This should be acknowledged explicitly. Human rights are the bare minimum or baseline level and are not intended to cover all the areas of concern in law and ethics.

"AI must have materially assisted in the process of making the decision"

There are two different cases that fall under this scope:

1) Where AI allows someone to make a decision that they could have otherwise made, but allows them to make it faster, with more confidence, etc.
2) Where AI allows someone to make a decision that they could *not* otherwise have made, perhaps because it would have been infeasible to do otherwise.

Both (1) and (2) should be explicitly identified and included within the scope of the definition.

Where the two parts of the definition come apart

The reality of distributed decision-making provides a challenge to the definition and shows some of its limitations. For example, suppose that company X uses AI to produce a credit ranking of individuals. The actions of company X alone provides no effects for individuals and so does not meet the first part of the definition, "*effect for an individual*". Suppose that company Y then uses that credit ranking to award loans to individuals. Company Y does not use AI and therefore does not meet the second part of the definition, *"AI must have materially assisted in the process of making the decision"*. However, taken as a whole, this seems exactly the kind of case that the discussion paper is concerned about. Yet the definition as it stands is not sufficient to capture these cases.

The HRC's discussion paper on Human Rights and Technology carefully presents the diversity of human rights issues that developing and emerging technologies could threaten. The Academy has highlighted some issues in the discussion paper regarding the clearer articulation of definitions, addressing the indirect consequences of some proposals (e.g. privacy, data management and the negative impact upon scientific discovery and R&D if technology specific prohibitions are in place) and drawn attention to some considerations regarding the methods of developing a framework and establishment of a taskforce. We encourage the HRC to consider including these suggestions to further strengthen the proposals presented in the discussion paper.

We are grateful to the Fellows and Associate Members who contributed to this response. For further information about this response, please contact Mr Chris Anderson, Director Science Policy at the Australian Academy of Science (Chris.Anderson@science.org.au).

References

1. Hern, A. Cambridge Analytica: how did it turn clicks into votes? *The Guardian* (2018). Available at: https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie. (Accessed: 20th February 2020)
2. Manheim, K. & Kaplan, L. Artificial Intelligence : Risks to Privacy and Democracy. *21 Yale J. Law Technol. 106* 106–188 (2019).
3. Williamson, R. C. The AI of Ethics. *MITPress* (2019).
4. UK Data Service. Regulating access to data. Available at: https://www.ukdataservice.ac.uk/manage-data/legal-ethical/access-control/five-safes.
5. Desai, T., Ritchie, F. & Welpton, R. *Five Safes: Designing Data Access for Research*. *University of the West of England* (2016).
6. Cabinet, C. of A. D. of the P. M. and. *Best Practice Guide to Applying Data Sharing Principles*. (2019).
7. Australian National Data Service. The FAIR data principles. Available at: https://www.ands.org.au/working-with-data/fairdata. (Accessed: 3rd February 2020)
8. MacCarthy, M. AI needs more regulation, not less. *Brookings* (2020).
9. Davis, J. L. *How Artifacts Afford: The Power and Politics of Everyday Things*. (MIT Press, 2020).
10. Davis, J. L. & Chouinard, J. B. Theorizing Affordances: From Request to Refuse. *Bull. Sci. Technol. Soc.* 36, 241–248 (2016).
11. Australian Council of Learned Academies. Available at: https://www.acola.org
12. Digital Technologies Hub. Available at: https://www.digitaltechnologieshub.edu.au/.
13. Digital Technologies Hub. AI & Me.
14. Digital Technologies Hub. Anti-Bullying AI: Integrating Digital Technologies. Available at: https://www.digitaltechnologieshub.edu.au/teachers/lesson-ideas/integrating-digital-technologies/anti-bullying-ai.
15. Digital Technologies Hub. Data Bias in AI. Available at: https://www.digitaltechnologieshub.edu.au/teachers/lesson-ideas/integrating-digital-technologies/data-bias-in-ai.
16. Digital Technologies Hub. AI Image Recognition: Exploring limitations and bias. Available at: https://www.digitaltechnologieshub.edu.au/teachers/lesson-ideas/integrating-digital-technologies/ai-image-recognition-exploring-limitations-and-bias.