Human Rights and Technology Project Discussion Paper submission

Thanks for the opportunity to provide further comment on these issues. The Discussion Paper represents a major and considered advance on the White Paper and provides some concrete and commendable recommendations. I agree that using the human rights lens as a starting point for the analysis of the evolution of technology such as AI is most appropriate, and I support the broad based approach suggested and its inclusion of support for capacity building and a multi-faceted regulatory approach (including law, co-regulation and self-regulation).

I think that this is an appropriate way to start, alongside a few targeted areas for justifiable intervention: such as facial recognition technology; and in those areas where government can take a lead and make an impact as a system developer /purchaser and user as well as standard setter. In my view the current calls for more automation in government and the production of "machine readable laws" / "rules as code" highlight the need to focus attention on the application of AI within government itself as a starting point.[1]

Below I provide some specific comment on a few issues. Where I don't comment on a proposal in the Discussion Paper I am broadly supportive of it.

Proposal 1: The Australian Government should develop a *National Strategy on New and Emerging Technologies*. This National Strategy should:

(a) set the national aim of promoting responsible innovation and protecting human rights
(b) prioritise and resource national leadership on AI
(c) promote effective regulation—this includes law, co-regulation and self-regulation
(d) resource education and training for government, industry and civil society.

I agree that the Government should have such a national strategy. There are a few bodies that are already charged with responsibility in these areas, including Innovation and Science Australia, the National Science and Technology Council, Australia's Chief Scientist, the Department of Industry, Science, Energy and Resources and CSIRO. I note that there is already some reference to new and emerging technology issues in the context at least of the Government's "Australia's Tech Future" strategy – primarily set in the context of the digital economy.[2]

I note that the label for this Strategy is very broad: "New and Emerging Technologies", but that the discussion paper (including element (b) of this proposal) seems to focus almost exclusively on AI, which is but one form of new and emerging technology. Is the proposal for

[1] CSIRO Submission 19/691 – to the Senate Select Committee on Financial Technology and Regulatory Technology - https://www.aph.gov.au/DocumentStore.ashx?id=2cb4fd49-f41e-4b62-9947-f6ec9a41ca0a&subId=675332
[2] https://www.industry.gov.au/data-and-publications/australias-tech-future

a general remit or specific focus? If specific to AI, then I suggest it should be relabelled. If general, then perhaps AI could be a "first cab off the rank" area for focus.

More fundamentally, why focus on AI? As the Discussion paper notes The Centre for Policy Futures pointed to the blurring of boundaries across scientific fields, and between digital and non-digital technologies (e.g. synthetic biology may have boundary issues with AI). As discussed later, there are tensions here in terms of the level of focus: there is a danger that regulatory efforts absent context will be ill fitting, and there is also a danger of a splintering of regulatory endeavour by the object of regulation (e.g. technology) that would then create a bewildering array of different regulators covering sub elements of human rights in particular technology fields.

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.

These aims may be achieved by encouraging co-operation between existing bodies (professional bodies, learned societies) and through funding additional research into ethical issues, alongside investment into technology development itself. Indeed, we see this approach already occurring in other countries in relation to AI.[3] Australia does not yet seem to have made this step in a significant way or to fully recognise the supportive relevance of such guiding frameworks for responsible and effective technology development.

The breadth intended by "new and emerging technologies" (as discussed above) may also have a significant impact on the best approach here, as there are many nuances and contextual issues that apply in different fields that may not be easy to assemble into an overarching approach (except at a very high and generic level that may not be that useful as clear guidance). For example, simply the field of genetic profiling and counselling has been the subject of many voluminous reports, dating back decades.[4]

Still, there may be some benefit in providing an overarching guide or checklist for different bodies to apply within the domain of their knowledge and expertise. How to best approach this area might be part of what the ALRC could consider if it was briefed to advise on these matters as suggested. I suggest that ongoing funding for a diversity of activity and research may be a better way to proceed than simply a one-off review, as these matters will continue to evolve.

[3] see eg summary of ethics and policy developments at https://www.loc.gov/law/help/artificial-intelligence/asia-pacific.php

[4] see eg the 2003 ALRC report "Essentially Yours: The Protection of Human Genetic Information in Australia" (ALRC Report 96)

Proposal 3: The Australian Government should engage the Australian Law Reform Commission to conduct an inquiry into the accountability of AI-informed decision making. The proposed inquiry should consider reform or other change needed to:

(a) protect the principle of legality and the rule of law
(b) promote human rights such as equality or non-discrimination.

I am supportive of this recommendation, as with the later recommendation around undertaking a review of existing use of AI in government (which might be an informative precursor to reference to the ALRC). Query whether this should be an ALRC only project or a joint project. It would also be useful to map out how the different taskforce and review activities suggested in the Discussion paper might inter-relate and the different bodies that may contribute to them.

Having a sound understanding of the differing contexts and affordances of the new technology is clearly critical. As Chris Reed comments in relation to the sub-issue of transparency "In the absence of a wide range of real-life examples of AI in widespread use, it is difficult to identify any fundamental principles which could be used to determine whether a transparency obligation should be imposed, and if so, what kind of transparency".[5] He also notes that, dependent on how they are framed, requirements for explainability may effectively amount to a prohibition on certain neural network approaches – so these matters will need to be considered carefully.

There is much international material to be considered as the discussion paper outlines. In case you are not already familiar with it then I highlight the work of the Berkman Klein Center Working Group on AI Interpretability for some useful discussion of one element of accountability.[6] This draws a distinction between "explanation systems" and " AI systems":

> "regulation around explanation from AI systems should consider the explanation system as distinct from the AI system. From a regulatory perspective, this opens the door to regulation that requires that an AI system be explainable some proportion of the time or in certain kinds of contexts—rather than all the time. Loosening the explanation requirement in this way may allow for the AI system to use a much more complex logic for a few cases that really need it."

The paper also notes the potential downsides of imposing non contextual one size fits all models of accountability and transparency, as this may result in "companies employing suboptimal but easily-explained models… [r]equiring every AI system to explain every decision could result in less efficient systems, forced design choices, and a bias towards explainable but suboptimal outcomes…[a]s we have little data to determine the actual costs of requiring AI systems to generate explanations, the role of explanation in ensuring

[5] Chris Reed, "How should we regulate artificial intelligence?" Published:06 August 2018 https://doi.org/10.1098/rsta.2017.0360

[6] Finale Doshi-Velez,* Mason Kortz,* Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Shieber, James Waldo, David Weinberger, Adrian Weller, Alexandra Wood "Accountability of AI Under the Law: The Role of Explanation" https://arxiv.org/ftp/arxiv/papers/1711/1711.01134.pdf

accountability must also be re-evaluated from time to time, to adapt with the ever-changing technology landscape."

That paper and Chris Reed's brief article have many other useful things to add around ex post and ex ante analyses, empirical evidence, theoretical guarantees and legal implications and they are best left to speak for themselves rather than be further re-hashed here. Approaching these issues in a truly effective manner requires a deeper look at systems design not only for AI enabled decisions but for human decisions – a significant undertaking.

As these issues are approached I think a useful perspective may be to look at the levels of trust in decisions, and the different factors that influence this (such as trust in underpinning data or understanding, trust in system design, trust in system operation, trust in review/appeal processes). We are never going to achieve perfect systems, whether human, AI enabled or hybrid. There will always be trade-offs to consider, and there may be conflicts between different human rights for different subjects that AI enabled decision making might influence.

Proposal 4: The Australian Government should introduce a statutory cause of action for serious invasion of privacy.

I am supportive of this recommendation and note that it has also been put forward by the ALRC and the ACCC. While there may be some special issues raised by AI in this context, I think a relatively broadly defined action (cf s18 of the ACL) would be best, allowing for evolution of its application through caselaw in better relation to shifting circumstances. Again, there are complexities here around the extent to which the right might be defined more by reference to informational elements of privacy (which is largely the frame of reference of existing laws), but I think a broader approach might be superior. This is really a matter for considered action in response to the extensive prior work conducted by other 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.

Is this proposal intended to have effect throughout Australia, at Federal, State and local level and in relation to all dealings with public and private entities of different sorts (wherever located – including internationally domiciled)? If so then there will obviously be the usual constitutional issues and limitations to consider. Matters such as new and emerging technologies certainly underscore the dated nature of current Commonwealth constitutional powers.

I realise that this requirement is directed at transparency, and I don't quibble with that, but I wonder whether it may have unintended effects in that some recipients might find this disempowering rather than informative. In other words, some may receive it as a case of

"computer says so" / Lex machina /deus ex machina to which the response may be resigned acceptance of a decision from an apparently impartial and omniscient technology. Even combining the notification with other theoretically enabling information around differential appeal paths (for example) for AI enabled decisions may not avoid this danger. In this regard I conjecture that it is reasonable to draw a parallel with some of the very interesting findings outlined in the recent ASIC report "Disclosure: why it shouldn't be the default".[7]

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.

Elements (a) and (b) seem very sensible. I note that compliance with such reasonable requirements has been an issue in the past (for example in relation to impact statements for new legislation and regulation), though this is clearly not a reason for *not* introducing such a requirement. However, I am less sure of the breadth and impact of (c) as this may be unnecessarily limiting in some situations - if it is intended that such systems require legislative backing for example.

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 lay person, 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.

I think this is a reasonable proposal on its face, though I query whether:

[7] https://download.asic.gov.au/media/5303322/rep632-published-14-october-2019.pdf To paraphrase some of its key observations around disclosure being necessary, but not sufficient: " Disclosure does not solve ... complexity", "Disclosure must compete for ... attention", "One size does not fit all – the effects of disclosure are different from person to person and situation to situation", " In the real world, disclosure can backfire in unexpected ways

- the explanation should be embedded with the decision in an accessible fashion rather than having to be demanded. The format could be potentially "laddered" – akin to the "scaled advice" guidelines ASIC has issued.[8] This could highlight key and potentially contestable inputs on which the decision has been based, actively probe for understanding and alignment.

- AI might create different affordances than existing human based decision-making systems. For example. the system might be able (dependent on context) to come to a tentative only decision or highlight areas that need further enquiry – which might be facilitated (by legitimate data matching in the case of government) or through engagement and query with the subject of the decision prior to making a decision. It could consider the impact on the individual across different dimensions and services: being person centric, rather than simply to the point of the decision. Naturally this same capacity would entail further need for safeguards around legitimate data sharing, confidentiality and human rights, but it may afford the capacity for an improved outcome using a collaborative service model.

- The technical explanation will be of much assistance in most cases. This is not to argue against the proposal but simply to query whether it will result in much benefit for the majority, and as discussed above may in some circumstances have unintended negative impacts. Individuals should be informed but putting the onus on them of trying to engage with the detail of the justifications – which may require engaging an expert to interpret (at whose expense? and how accessibly?) – does not seem sufficient safeguard. This feels to me like the sort of solution that will only be of much assistance to a small cohort. I think there will need to be additional dynamic systems of proactive audit (not simply post hoc ANAO style audit) designed to pick up problematic cases for further review.

I feel it is not appropriate to put the prime emphasis on the individual to self-help and try to navigate through technical complexity. Rather I think it is preferable to design systems with a capacity for doubt. This may not occur in the short term, but I believe it would be useful if systems were designed to identify grey areas, and trigger further enquiry pre decision if issues might be finely balanced. We might then build systems that are designed to draw out complexity and competing arguments from multiple perspectives, not just produce a "decision". This may include building multiple systems separately that take different stances or input. Ultimately it could include levels of meta governance to address the outputs of such "federated" AI enabled systems that help people resolve conflicts between those differing recommendations or perspectives (in the case of conflict).

I note that there is a broader issue at play here around how humans make decisions. I note that there is an opportunity to reflect on and improve overall accountability and system design in important decision-making processes, not just regulate against potential AI abuses,

[8] Australian Securities & Investments Commission, 'Providing digital financial product advice to retail clients' (August 2016) <https://download.asic.gov.au/media/3994496/rg255-published-30-august-2016.pdf> at RG 255.91-99

or simply mirror existing requirements. That is of course a much bigger task that extends well beyond the remit of this discussion paper and would require extensive effort.

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

Setting to one side the point that the generation of the explanation might in theory be created by a system separate from the decision making system (perhaps including in cases where there is a component of neural network based machine learning), putting a blanket ban on deployment of such systems because of the contingent possibility that there might be a human rights impact seems over-reach. There may be some contexts where there are over-riding factors or conflicting human rights interests that might point to permitting use.

However, especially from a governmental systems perspective I think the proposal is a reasonable starting point. There would need to be a very clear impact case analysis, rationale and post adoption review for deploying otherwise.

Question B: Where a person is responsible for an AI-informed decision and the person does not provide a reasonable explanation for that decision, should Australian law impose a rebuttable presumption that the decision was not lawfully made?

I sympathise with the intent behind this query, but I wonder whether it might have unintended side effects where there is no substantive issue with the decision (and perhaps there has been a simple failure to provide a readily producible explanation)?

Proposal 9: Centres of expertise, including the newly established Australian Research Council Centre of Excellence for Automated Decision-Making and Society, should prioritise research on how to design AI-informed decision-making systems to provide a reasonable explanation to individuals.

I am supportive of this proposal but don't believe that efforts in this regard should be relegated only to such centres of expertise, as there are many different areas and disciplines that will be relevant to assisting in this effort, whether around human machine interaction issues or around human centric design or regulatory compliance or human rights issues.

Question C: Does Australian law need to be reformed to make it easier to assess the lawfulness of an AI-informed decision-making system, by providing better access to technical information used in AI-informed decision-making systems such as algorithms?

There may be inherently more capacity to address these issues in relation to government funded/deployed systems - through procurement policies and guidelines requiring

transparency (even if with reason on occasion subject to commercial in confidence restrictions on further disclosure). Applying such a reform across the board to commercial systems – especially those developed and hosted internationally and used in the private sector – would obviously strike some significant hurdles (including those around jurisdiction and IP).

Technical information in this regard may also be quite extensive and unwieldy – especially in relation to neural network systems where nodal weightings and training data (along with issues around confidentiality and privacy, potentially cross border) would need to be considered (as opposed to more prosaic implementation of a decision tree in a clear algorithm for example, though even that might get quite complex in some situations).

I'm not sure that this is the big issue though – I think we are likely to see more proportionate progress through improved design up front than through post hoc review approaches (which I posit are likely to be seldom used, even where arguably they should be). The inherent jurisdiction of the court system would be available in appropriate cases to access relevant information, and already has protocols for handling commercial in confidence material.

Question D: How should Australian law require or encourage the intervention by human decision makers in the process of AI-informed decision making?

This is a complex and problematic case and presupposes that Australian law should do this. One obvious way of achieving the "how" element of this question is through making people legally responsible for AI informed decision making, which is the subject of a separate recommendation.

I suppose at a practical level in implementation in situations where there are people in the loop that to avoid the supposed human deference to AI systems that has been referenced by some in this field you could design systems to inject false "poor" decisions into the loop of those being reviewed to detect whether humans are effectively screening decision outputs – much as is done with false positive images on security screening for baggage for example.

But query whether we really want to design systems like that? What it does to the people involved? There is already a range of commentary on the potential harm caused to the human in the loop, who might be turned into a "moral crumple zone", and be reluctant to intervene.[9]

[9] see e.g. DGI(2019)05 A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework Prepared by the Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT) available at https://rm.coe.int/responsability-and-ai-en/168097d9c5

I think transparency and review in implementation, along with "failure reports" discussing shortfalls to educate people would be sensible measures that wouldn't necessarily require any legal intervention.

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.

This is a sensible recommendation given the extensive work highlighting current problems with this technology and its implementation. I note also the recent publicity around the potential use of the Clearview AI system by law enforcement in Australia.[10] Indeed given the discussion about the apparent use of that system by individuals inside law enforcement without their organisations necessarily having been aware[11], I think it is important that there also be procedures or administrative directions to staff that they should not use third party systems on their own initiative.

I think that such a moratorium is particularly important in view of abuses elsewhere such as the widely reported scope creep on use of meta data by public authorities for applications not linked to serious crime.[12]

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 in 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.

Perhaps a taskforce approach could also address the development and scope of any related sandbox concepts.

Proposal 14: The Australian Government should develop a human rights impact assessment tool for AI-informed decision making, and associated guidance for its use, in consultation with regulatory, industry and civil society bodies. Any 'toolkit for ethical AI' endorsed by the Australian Government, and any legislative framework or guidance, should expressly include a human rights impact assessment.

[10] https://www.abc.net.au/news/2020-01-23/australian-founder-of-clearview-facial-recognition-interview/11887112
[11] https://www.buzzfeed.com/hannahryan/clearview-ai-australia-police
[12] https://www.abc.net.au/news/2018-10-19/authority-creep-has-more-agencies-accessing-your-metadata/10398348

I think it is important that Government should undertake this even only for its own use – and this might be a matter on which it would be good to get joint COAG agreement. Clearly ensuring compatibility with similar developments internationally would be preferable. I would be cautious about embedding this in a legislative framework at this stage, while noting that government standards can still have a considerable influence even if not mandated, given the significance of government as a purchaser/developer and implementer of such complex systems, and the extent to which these may then influence service providers and other purchasers and indirectly set development standards which may become a reference point for the appropriate standard of care.

Question E: In relation to the proposed human rights impact assessment tool in Proposal 14:

(a) When and how should it be deployed?
(b) Should completion of a human rights impact assessment be mandatory, or incentivised in other ways?
(c) What should the consequences be if the assessment indicates a high risk of human rights impact?
(d) How should a human rights impact assessment be applied to AI-informed decision-making systems developed overseas?

See generally comments above. I suggest that its initial application should be within the public sector. It is difficult to be definitive in relation to consequences absent context. I think sub element (d) could be problematic, but may again at least be relevant to government purchasing – or require that such systems have been through equivalent processes internationally.

Proposal 15: The Australian Government should consider establishing a regulatory sandbox to test AI-informed decision-making systems for compliance with human rights.

I think the notion of using regulatory sandboxes, dependent on their design, is very interesting, and compatible with sound design thinking approaches. But I think we need to be realistic about what we are trying to achieve in this regard. If we have reasonable expectations and appropriate design, I am supportive.

In terms of those expectations, it will usually be impossible to conduct robust testing that will satisfy all use cases and remove all risks – so we need to remove those as expectations. However, we should encourage people to undertake such testing and we should expect them to make reasonable efforts to seriously engage with, hunt out, explore and try to cater for human rights issues.

Of course there may be many instances where there are conflicting human rights – again an area where there may be a need to cope with uncertainty, grey factors and discretion and

associated substantive and procedural review mechanisms (which may or may not be AI enabled) to address these issues (without hoping to perfect them).

Presumably there are some aspects of existing sandbox practice (whether around the ASIC fintech examples or otherwise) that can be drawn on to assist.

Question F: What should be the key features of a regulatory sandbox to test AI-informed decision-making systems for compliance with human rights? In particular:

(a) what should be the scope of operation of the regulatory sandbox, including criteria for eligibility to participate and the types of system that would be covered?
(b) what areas of regulation should it cover eg, human rights or other areas as well?
(c) what controls or criteria should be in place prior to a product being admitted to the regulatory sandbox?
(d) what protections or incentives should support participation?
(e) what body or bodies should run the regulatory sandbox?
(f) how could the regulatory sandbox draw on the expertise of relevant regulatory and oversight bodies, civil society and industry?
(g) how should it balance competing imperatives eg, transparency and protection of trade secrets?
(h) how should the regulatory sandbox be evaluated?

This is a big series of questions, on which I offer just a few thoughts – that are certainly not comprehensive. Features that might be considered include:

- initially at least making this opt-in, and not trying to cover all fields – perhaps starting out in some more active areas with a few stakeholders, in order to approach the sandbox design itself from an iterative design thinking perspective and then adjust as needed

- I would be cautious initially of providing any specific protections or incentives other than the intrinsic incentive offered of having multiple stakeholders (see further below) engage with the product/service in question in order to improve its human rights related features, don't try to make it formal clearance oriented (avoiding giving any guarantee)

- Use it to promote interaction between diverse stakeholders, especially to engage human/consumer rights/disability advocate bodies in the product/service development phase pre-release (i.e. in prototyping phases). This could in some senses look a little like "tiger team" assessment - to use a security testing analogy - but without necessarily going open source on the code

- Having been engaged in technology commercialisation for several decades, I think sometimes too much is made of the competing imperatives (transparency vs protection of trade secrets). It is often possible to enable interaction and

improvement without necessarily going "under the hood". Having said that, my personal feeling is that we are unlikely to see good design through obscurity – in the same sense that information security analysts are usually suspect of "security through obscurity" approaches.

- I think the best evaluation of the success of such a system is whether people use it and see benefits in terms of systems improvements flowing from it. I don't think improving human rights elements should be seen simply as a zero sum trade off game of compliance costs – I believe that in many situations improving human rights features is likely to result in a system that has better adoption, useability, reach, reception and results.

Proposal 17: The Australian Government should conduct a comprehensive review, overseen by a new or existing body, in order to:

(a) identify the use of AI in decision making by the Australian Government
(b) undertake a cost-benefit analysis of the use of AI, with specific reference to the protection of human rights and ensuring accountability
(c) outline the process by which the Australian Government decides to adopt a decision-making system that uses AI, including any human rights impact assessments
(d) identify whether and how those impacted by a decision are informed of the use of AI in that decision-making process, including by engaging in public consultation that focuses on those most likely to be affected
(e) examine any monitoring and evaluation frameworks for the use of AI in decision-making.

Starting with a comprehensive review is a sensible proposal that will provide a better understanding of context to enable appropriate measures to be implemented. It is the sort of work that has already been done elsewhere – including in New Zealand.[13]

Proposal 18: The Australian Government rules on procurement should require that, where government procures an AI-informed decision-making system, this system should include adequate human rights protections.

I support this proposal and note its linkage to my response to other elements including Question E.

[13] "GOVERNMENT USE of ARTIFICIAL INTELLIGENCE in NEW ZEALAND" (2019) - Final Report on Phase 1 of the New Zealand Law Foundation's Artificial Intelligence and Law in New Zealand Project https://www.cs.otago.ac.nz/research/ai/AI-Law/NZLF%20report.pdf

Proposal 19: The Australian Government should establish an AI Safety Commissioner as an independent statutory office to take a national leadership role in the development and use of AI in Australia. The proposed AI Safety Commissioner should focus on preventing individual and community harm, and protecting and promoting human rights. The proposed AI Safety Commissioner should:

- **(a) build the capacity of existing regulators and others regarding the development and use of AI**
- **(b) monitor the use of AI, and be a source of policy expertise in this area**
- **(c) be independent in its structure, operations and legislative mandate**
- **(d) be adequately resourced, wholly or primarily by the Australian Government**
- **(e) draw on diverse expertise and perspectives**
- **(f) determine issues of immediate concern that should form priorities and shape its own work.**

The notion of an AI safety commissioner is a much more delimited and tractable proposal than that mooted in the white paper of a "responsible innovation organisation". However, I remain hesitant about the splintering effect of this type of approach. It may produce the 'safety commissioner' equivalent of the "law of the horse". That is, an overall system of regulation, which is hard for citizens to understand and navigate, and creates boundary issues and gaps when it is looked at in helicopter view. This is the inevitable outcome of having different regulators for different sub issues (e safety, cybersecurity. AI, not to mention dedicated areas referenced such as OGTR etc).

Such problems are a general problem for our broader regulatory environment. To my mind this is another reason to strengthen the existing underpinnings of core institutions protecting human rights and consumer interests (eg including but not limited to AHRC, ACCC, ASIC) rather focussing on shifting and proliferating regulatory objects such as technologies.

Perhaps though the functions or even role of an AI Safety Commissioner could be subsumed within one of those existing bodies. Resourcing such functions in this way may be a more flexible approach to changing demand than creation of a new office. It would also seem consistent with other recommendations that discuss resourcing in those other areas.

Finally, I remain concerned as already expressed about the capacity of a single statutory office to effectively discharge such responsibilities across what is an ever-broadening field. I think a plurality of strategies, from resourcing some type of co-ordinating function through research and broader engagement, are more likely to be productive than putting all our eggs in the one (big) office basket.

Proposal 26: Providers of tertiary and vocational education should include the principles of 'human rights by design' in relevant degree and other courses in science, technology and engineering. With appropriate support, the Australian Council of Learned Academies

should undertake consultation on how to achieve this aim most effectively and appropriately within the tertiary and vocational sector.

As elaborated on further below in response to question H, I am concerned that this proposal is a little narrow in its scope and furthers an incorrect view that AI and related developments are the responsibility of, and influenced only by, STEM graduates.

Firstly, at a broader level the concept of "STEM" is too narrow. Australia continues to use it while in many other jurisdictions (not least the US), broader concepts such as STEAM are used. Indeed, the STEAM approach is exemplified by the success of products and services that incorporate many inputs and are the result of the co-operation of many disciplines (as indeed almost all are) – such as the oft referenced Apple suite of products and services.

Diversity of engagement with design is a sound principle – not only to bring many different lenses to bear on identifying the true problems and creating effective solutions to them, but also because it is more representative of the diverse user base that will interact with these systems.

There will be the inevitable issues of curriculum balance when raising the introduction of new matters (such as design thinking) into what might already be a rather crowded – if occasionally conservatively constructed – curriculum. However, there will usually be some treatment of these matters inside some topics and courses which can be expanded on. Good practises can be highlighted and shared, along with the provision of additional suggestions from learned bodies such as ACOLA.

I note also the potential cross linkage of human rights by design issues, as part of overall design thinking approaches, to other emerging debates such as the "right to repair" debate – which has consumer rights and intellectual property dimensions that are being highlighted by developments here and overseas.[14]

Question H: What other tertiary or vocational courses, if any, should include instruction on 'human rights by design'?

I think it is important to take a broad view of this matter and not box it down into a "STEM" responsibility category. This is because, as discussed above, product and service innovation, including increasingly AI products and services, spread across the economy and are not solely the preserve of a "STEM" cohort.[15]

While we may not yet see too many topics and courses with explicit "human rights by design" labels, "design thinking" approaches are becoming more common in many institutions. Many of these will already effectively include some elements of human rights

[14] locally for instance in some of the work being done by Prof Leanne Wiseman & Dr Kanchana Kariyawasam, who on 4 February 2020 hosted a Law Futures Centre Forum "Can we fix it? A Right to Repair for Australia?", and reportedly the ACCC is watching closely – see e.g. https://www.abc.net.au/news/2019-03-03/does-australia-need-a-right-to-repair/10864852

[15] cf the reference in the Discussion paper to moving beyond STEM to "to truly interdisciplinary learning models" (fn 700 referencing Professor Lyria Bennett Moses)

considerations in elements of what is taught or what is drawn out in experiential workshops, as design thinking is intrinsically a human centric discipline. No doubt however, the extent to which human rights issues (as opposed to human problems) are explicitly made part of the frames of reference used in the design process can be strengthened. As discussed above this can be assisted by sharing information about approaches and developments.

To provide an example of a different non-STEM discipline area where design thinking approaches are being introduced through topics and courses, consider the law itself. The law is fundamentally concerned with constructing and protecting human rights issues, and so it is a very appropriate area to consider in this context. This is not only because lawyers are part of interdisciplinary teams that are engaged in developing products and services across all areas (even if in some examples only from a regulatory review angle), but also in particular because lawyers are now actively engaged in driving the development of innovative – including AI based – systems. Such systems will often have – by their very nature – the capacity to have significant human rights impacts. Some of this may be positive, in terms of improving access to justice, and many new services tout this benefit, which should not be downplayed. However, there is clearly also the potential for such systems to produce negative impacts, and this maybe across many areas where citizens interact with public and private services. Think of examples such as robo-debt and automated migration review issues, then extend this across core service areas including, but not limited to, the justice domain. So training new law student cohorts with an awareness of the importance of sound design approaches – including human rights issues as part of the explicit frames they use – makes sense in order that they might help design better systems which attempt to cater in a more complete way to human needs and rights.

More broadly in the legal profession there is a wide recognition of the impact of new technologies and the importance of taking a human centred (and human rights aware) approach to service and professional redesign (whether or not AI enabled).[16]

A number of Universities are now offering design related electives across their law courses, some along the lines of the type of approach pioneered at Stanford d-school by Margaret Hagan.[17] Flinders has now introduced multiple design thinking topics as part of the core of its new law program (topics such as INNO1100 Legal Innovation and Creative Thinking; INNO2100 Innovation for Social Justice Impact). Some are introducing coding classes (including Melbourne, UTS, Flinders). More "traditional" electives – for example in technology law related areas offered in many law schools– now also use a human rights lens as part of the analysis of their discussion of the impact of new and emerging technologies such as (but not limited to) AI. These approaches are also useful.

[16] Report on the Commission of Inquiry into the Future of Law and Innovation in the Profession https://www.lawsociety.com.au/sites/default/files/2018-03/1272952.pdf; Disruption, Innovation and change: The Future of the Legal Profession http://www.vgso.vic.gov.au/sites/default/files/publications/Disruption%20Innovation%20and%20Change.pdf

[17] See e.g. https://law.unimelb.edu.au/alumni/mls-news/issue-22-november-2019/how-better-design-can-improve-the-law

While the above discussion highlights the law as one area not included in the traditional "STEM" bucket, similar arguments can be made in many other areas, and there are already a number of institutions attempting to provide more broadly based opportunities for students to be exposed to design thinking approaches and then work in interdisciplinary teams on projects of their choosing. These include work in UTS, Flinders, RMIT, UNSW and many other Universities, and range from optional extra-curricular events (such as "hackathons") through to many for credit options, and now some institutions introducing components of such instruction and work as core. UTS is one interesting example that has made a concerted effort to design and embed innovation training across many different fields.[18] Flinders has also made innovation topics available across many courses, including making them compulsory in some.

In closing I wish the AHRC all the best in continuing its important work in these and other fields

Robert Chalmers

[redacted]

3 March 2020

[redacted]

[18] see e.g. https://www.uts.edu.au/partners-and-community/initiatives/entrepreneurship/courses-subjects-degrees/studying-innovation