

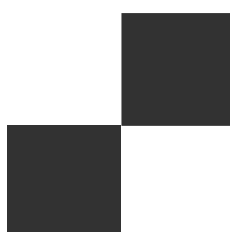


Human Rights and Technology Issues Paper: UTS Submission



Contents

02	–	Foreword
04	–	Project leads
05	–	Section leads
07	–	Contributors
08	–	Section 1: Executive summary
12	–	Section 2: Preamble
24	–	Section 3: Responses to AHRC’s questions
73	–	Section 4: Case studies
89	–	Section 5: Recommendations
99	–	Section 6: References
107	–	Section 7: Compiled recommendations



Foreword

Professor Attila Brungs

Vice-Chancellor and President

On behalf of UTS, it is my pleasure to submit a whole-of-university response to the Australian Human Rights Commission's Human Rights and Technology Issues Paper.

I commend the AHRC for this important body of work. This comes at a critical time as policy makers, ethicists, human rights activists, technologists and concerned citizens the world over, grapple with the impact of technology on every facet of our lives. Collectively, we need to do the hard work now to ensure we get the balance right; to yield the great benefits from technology and innovation while ensuring social cohesion and wellbeing is improved through appropriate approaches to privacy, inclusion and the ethical use of technology and data.

At UTS we truly believe that collaboration is key to success and making a greater impact in the world. UTS is playing a significant leadership role in the growing national and global debate about data ethics, privacy and technology.

We are particularly proud to be the only academic partner for this project and we believe it is a perfect fit for UTS because it's actually about technology, human impact and public benefit.

This work aligns with our core belief that universities exist for the public good. At UTS we are fundamentally interested in positive social impact and public benefit, and the use of technology to drive that.

We see the positive impacts that technology can have but we also see that without proper design, inclusivity and proper analysis of what can happen, then you cannot get positive social benefit.

As we experience the current era of technological change, it is important for us to identify gaps, create adaptive governance models and promote the development of technology in ways that protect human rights.

There is deep complexity in the issues raised in this project, and that's why UTS has ensured we take a transdisciplinary approach in response. I wish to thank and acknowledge the UTS academics and students who have contributed their valuable time participating in the conference and preparing submissions for the UTS response to the HRT Issues Paper.

I want to especially thank the submission's Leads – Dr Nicole Vincent, Professor David Lindsay and Monique Potts.

With contributions from every faculty at UTS – this collaboration has ensured we have approached the issues from a truly transdisciplinary perspective, giving rise to some fascinating insights and recommendations for the AHRC to pursue.





Foreword

The Hon. Verity Firth

Executive Director, Social Justice

A full appreciation of the impact of emerging technology on human rights must go beyond simply outlining risks and opportunities.

The solutions we seek need to take into account infrastructure which is human and social, as well as physical and technical. This requires a multi-disciplinary, collaborative approach. Universities are crucially placed to be convenors of discussions of this nature, as institutions which provide a neutral space in which to examine and analyse the impact of emerging technologies.

Contributing to this project has been a welcomed opportunity to engage UTS's strengths in grappling with an issue with far-reaching consequences for social justice in the 21st century. I am proud of the role which the UTS Centre for Social Justice and Inclusion has played in facilitating the transdisciplinary process, as the convening and driving force behind the university's social impact agenda.

Young people will experience the fullest implications of the decisions we make now to shape technological development, and it is important that their perspectives are taken into account. Included alongside expertise from academics in each of the faculties at UTS, are contributions from our students as well, submitted to the Australian Human Rights Commission as a complementary supplement to this submission.

Although technology is often described as a force beyond our control, that is far from the case. By being proactive in assessing the principles with which we will govern our technology, we can increase the probability that technology can be harnessed to bring about an equal, inclusive society that delivers benefit to more of its citizens. While the risks cannot be ignored, there are exciting pathways ahead.

The concepts, mechanisms and structures defined in the following pages outline how guiding principles can be applied to ensure that our machines continue to benefit humans. We look forward to continuing the process with the Australian Human Rights Commission and deepening this exploration.



Project Leads



Dr Nicole Vincent

In 2018 Dr Vincent joined the Faculty of Transdisciplinary Innovation at UTS as Senior Lecturer. She has an extensive career in academic and teaching roles in Australia and overseas including UNSW, University of Adelaide, Macquarie University, Swinburne University, Latrobe University, University of Auckland, Delft University of Technology and Georgia State in the US. Nicole's research is funded by more than \$1 million of external grants. She has published 40 peer reviewed articles, delivered 90 academic talks, and organised 19 conferences. Dr Vincent's research is eclectic and spans neuroethics, neurolaw, bioethics, philosophy of law, philosophy and ethics of emerging technologies, political philosophy, public policy, media, feminism, gender and happiness. Dr Vincent also heads up NeurolawAU, a consultancy that provides expert advice, research, education, as well as workshop/seminar/conference organisation, to allied legal and health professionals in private practice, government, and tertiary education sectors.



Professor David Lindsay

Professor David Lindsay joined UTS Law in 2018 after previously working at Monash University. David is an expert in law and technology, and is widely published in the areas of copyright, privacy, cyberlaw and communications law. He is the author of *International Domain Name Law* (Hart, 2007) and co-author of *Copyright's Public Domains* (CUP, 2018). At UTS he teaches Equity and Trusts, Copyright and Designs, and is the convenor of the Applied Project in Law, Innovation and Technology. David is General Editor of the *Australian Intellectual Property Journal* and a board member of the Australian Privacy Foundation.



Monique Potts

Monique is a thought leader in education, innovation and entrepreneurship with a mission to reimagine education for a knowledge economy. Currently Director of Strategic Projects at University of Technology Sydney (UTS) she works with teams across the university to create, develop and implement strategic projects to embed innovation and entrepreneurship into the fabric of the institution, culture and educational experience.

Section Leads

Dr Theresa Anderson

Dr Theresa Dirndorfer Anderson is Course Director of the Master of Data Science and Innovation in the UTS Connected Intelligence Centre. In her teaching and her research, Theresa engages with the ever-evolving relationship between people and emerging technologies, conceptually engaging with notions of risk, uncertainty and creativity. As a socio-technical researcher, she applies a transdisciplinary approach and value-sensitive participatory methods to explore human entanglements with emerging technologies and information practices. As an information ethicist, she is particularly interested in the interaction between creative and analytic thinking and doing and in examining ways information systems and institutional policies might better support both creative and analytic activities. Internationally she is leading discussion about these issues as chair of the Information Seeking in Context international research community and founder of the Human-Centered Data Science Network. Prior to UTS, she served as a diplomat, technical writer and environmental education officer.

Professor Simon Buckingham Shum

Simon Buckingham Shum is Professor of Learning Informatics, and Director of the Connected Intelligence Centre (CIC). CIC is an innovation centre for UTS, building the capacity of staff and students to gain insights from educational data science applications. Prior to joining UTS (Aug 2014) Simon was Professor of Learning Informatics and Associate Director (Technology) at the UK Open University's Knowledge Media Institute, a 70-strong lab researching the future internet for the knowledge society, and applications into the OU.

Dr Phillippa Carnemolla

Dr Phillippa Carnemolla is an industrial designer specialising in the design and evaluation of inclusive environments, products and information. Her research investigates the breadth of health, care and social impacts resulting from inclusive design approaches, including smart cities, ageing in place and disability housing models. In her role as Senior Research Fellow in Faculty of Design Architecture and Building at UTS, Dr Carnemolla is working on a diverse range of projects which investigate the impact of the inclusive and participatory design practice on service provision, caregiving and quality of life for older people and people living with disability.

Passiona Cottee

Passiona is a recent graduate of the UTS Master in Data Science as well as graduating previously from Law at UTS. She is now a co-lecturer in Data Science for Innovation at UTS and works as a data scientist for the NSW Treasury and the Commonwealth Bank. She is an experienced product manager, team player and data ethicist that specialises in the delivery of data-intensive products and services which traverse and simplify vast amounts of information to create meaning and compel action.

Dr Barbara Doran

Barbara Doran (PhD) specialises in identifying creative opportunities and putting them in action. She is an experienced speaker, mentor, educator, project innovator and artist.

Her skills are drawn from working across the arts, the tertiary sector, and with project experience in the private, public and community sectors which bring together practical, scholarly, imaginative, playful and collaborative outcomes. She works with all kinds of people and organisations including private, public and community organisations and one to one with accomplished and aspiring creative individuals. In addition to lecturing at UTS she is a consultant to NIDA and runs a creativity consultancy. She previously lectured at WSU and UNSW.

Dr Kirsty Kitto

Kirsty Kitto models the many ways in which humans interact with information, and how this can change as a result of the different contexts in which people find themselves. She is working towards providing unified mathematical and computational models of contextuality, which often results in apparently complex and unpredictable human behaviour. During 2010-2012 she was supported by a prestigious ARC Postdoctoral Fellowship for a Discovery Project investigating this problem. She has collaborated on projects with people from a wide range of fields, including Physics, Experimental Psychology, Cognitive Science, Computer Science, Social Psychology, Education and Computational Linguistics. She publishes papers in journals covering fields such as Complex Systems Science, Cognitive Science, Psychology and Computer Science.

Giedre Kligyte

Giedre is a Lecturer within the Faculty of Transdisciplinary Innovation. She brings her Design background and 12 years of expertise working in academic, educational and teaching in Higher Education contexts to the unique challenge of educating students for the future. Giedre brings an education perspective to transdisciplinary teams designing novel learning experiences within the transdisciplinary degrees. She believes that universities require creative, participatory and transdisciplinary approaches to curriculum design and teaching in order to develop graduates who are able to address the ill-defined, situated and inherently social problems in the complex world today. Giedre teaches into the Bachelor of Creative Intelligence and Innovation, supporting students in developing their critical thinking and research capabilities.

Dr Susanne Pratt

As a researcher, educator, artist and techno-scientific muser, Dr Susanne Pratt explores how creative practice can foster social and environmental responsibility, with an emphasis on improving environmental health and collective flourishing. She is currently based in the Faculty of Transdisciplinary Innovation, at UTS, where she co-founded the xFutures Lab. Susanne has over 12 years of experience working at globally recognised Universities in Hong Kong, New Zealand, the United Kingdom, and Australia. Her creative work has been internationally exhibited in various forms, including digital storytelling, convergent media installations, site-specific sound works, urban design proposals and participatory events.

Dr Hamish Robertson

Hamish is a health and medical geographer with 18 years of experience in population ageing and aged care. He has experience in a number of related fields such as multicultural health, disability and diversity work. His professional interests include some health informatics work, geographic information systems, data visualization and the like. He is also active in the museums sector and has written a number of pieces in this field. He is currently a Senior Research Fellow and Postdoctoral Research Fellow at UTS.

Kirsten Thorpe

Kirsten Thorpe (Worimi, Port Stephens NSW) is a professional archivist, who has led the development of protocols, policies, and services for Aboriginal and Torres Strait Islander peoples in libraries and archives in Australia. Kirsten's professional and research interests relate to Indigenous self-determination in libraries and archives. She has been involved in numerous projects that have involved the return of historic collections to Indigenous peoples and communities, and advocates for a transformation of practice to centre Indigenous priorities and voice in regard to the management of data, records, and collections. Kirsten joined the UTS Jumbunna Institute for Indigenous Education and Research as Cultural and Critical Archivist where she will continue research and engagement in relation to Indigenous protocols and decolonising practices in the library and archive field. Kirsten is an advocate for the 'right of reply' to records, as well as capacity building and support for the development of local Indigenous digital keeping places.

Matthew Walsh

Matthew Walsh is an Anaiwan man from northern NSW. He is the Executive Manager of Research at the Jumbunna Institute for Indigenous Education and Research at the University of Technology Sydney. Prior to taking on his current role, he was the manager, Indigenous Employment within the Office of the Pro Vice-Chancellor (Indigenous Leadership and Engagement) at the University of Technology Sydney where he was instrumental in positioning UTS as a leader in Aboriginal and Torres Strait Islander Academic and Professional staff employment. As an expert in institutional change and Indigenous policy engagement and implementation, Matthew has also led a number of projects in the Government, higher education, corporate and not-for-profit sectors.

Section contributors

Dr Wayne Brookes (Engineering and IT)
Professor Simon Darcy (Business)
Nik Dawson (Engineering and IT)
Dr Deborah Debono (Health)
Catherine Donnelley (Design, Architecture and Building)
Samantha Donnelly (Design, Architecture and Building)
Dr Jane Hunter (Arts and Social Sciences)
Professor Jennifer Loy (Design, Architecture and Building)
Dr Simon Knight (Transdisciplinary Innovation)
Dr Sacha Molitoricz (Centre for Media Transition)
Anjana Regmi (accessUTS)
Dr Philippa Ryan (Law)
Rosemary Sainty (Business)
Dr Olga Shimoni (Science)
Dr Linda Steele (Law)

Centre for Social Justice and Inclusion project leads

Verity Firth
Mitra Gusheh
Tida Tippapart
Jacqueline White

1 Executive Summary

This is the University of Technology Sydney (UTS) institutional response to the Australian Human Rights Commission (AHRC) Human Rights and Technology (HRT) Issues Paper. UTS would like to thank the AHRC for the opportunity to respond to the important questions posed in the HRT Issues Paper. This submission is based on comprehensive transdisciplinary dialogue bringing together over thirty researchers and experts from across the university. This summary conveys key insights and strategic priorities that emerged from this dialogue both within UTS and in conversation with the AHRC during the UTS/AHRC Roundtable on Human Rights and Technology.

1.1 New technologies and human rights

New technologies pose significant challenges to, and opportunities for, human rights. These are illustrated by the practical examples and case studies set out in this submission. The submission agrees with the HRT Issues Paper that a human rights approach is the best framework for analysing the threats and challenges of new and emerging technologies. We build on this position, proposing that a human rights approach, supplemented by insights gained from the transdisciplinary perspective, can lead to richer understandings of the problem space and potential pathways forward. The addition of the transdisciplinary perspective allows for a holistic and wider ranging approach that, for example, considers the interaction between technologies, thereby bringing into focus systemic influencers.

In making our recommendations, we have emphasised the importance of establishing and supporting processes for the ongoing assessment and evaluation of the social, political and ethical implications of new technologies, especially from a human rights perspective. In our view, the challenges and opportunities are best addressed by establishing a new regulatory body, the Technology Assessment Office (TAO), which will have functions involving the assessment of technologies, the coordination of regulatory responses, and the development of innovative forms of regulatory responses to complex, rapidly changing technologies.

1.2 Technologies as complex ecosystems

A key challenge identified in this submission is finding a suitable balance between identifying specific technologies which pose human rights risks, such as the twelve new technologies identified in the HRT Issues Paper, and understanding the broader impact of new technologies on human rights and values within an Australian context. Thus, in addition to our direct answers to the AHRC's specific questions, we have also proposed an approach which views technologies as complex ecosystems. This approach allows for the consideration of composite or hybrid technologies (i.e. combinations of both new and old technology, described at a fine level of granularity), together with human and contextual factors, to better assess the range of impacts of new technology and better understand the complexity of interrelationships between human and non-human actors. In Section 3 we offer answers to the ten questions posed by the AHRC in its HRT Issues Paper. However, some of the issues that new technologies raise for human rights cannot be easily discussed within the framing of those particular questions. For this reason, in Section 4 we present a number of case studies to draw out some of those other issues in specific contexts.

1.3 Effects of technologies and their significance

The effects of technology can be categorised by reference to three distinctions: intended vs unintended, temporally close vs distant, and hard vs soft. Intended, temporally close, hard impact style effects are easier to protect ourselves against as they are easier to imagine, to predict, and to evaluate. By comparison, a significant degree of nuance and sophistication is needed to even notice let alone predict or evaluate unintended, more temporally distant, soft impact style effects, and these effects often have incredibly weighty implications for human rights. For this reason, we dedicate a significant portion of our discussion to identifying the latter category of effects; explaining what is needed to identify them; and offering proactive mechanisms for responding. We argue a broad approach is required to conceptualise technology and its impacts (see Questions 1 and 5) and recognise that there is a considerable amount of legal and regulatory uncertainty that requires further research in order to strengthen our understanding of the impacts of technology on rights (see Question 3).

1.4 Regulation is not enough

Responsive regulation, co-designed with a broad range of stakeholders and based on principles-based approaches such as PANEL (Participation, Accountability, Non-discrimination and equality, Empowerment and Legality) can provide a level of protection for human rights. However, this approach is only responsive to specific and significant breaches by organisations such as technology companies and governments. The complexity and uncertainty of technological, economic, and social change make it increasingly difficult to predict the impact of technology, which limits the effectiveness of regulation as a tool for protecting human rights. Technology also moves much too fast for the law or for regulators to respond to problems in a timely fashion. For such reasons, we see great potential for the AHRC to support and advocate for multi-stakeholder partnerships to drive a more proactive, creative, participatory approach to technology and innovation (see Section 5).

1.5 The importance of value sensitive design

In part due to the challenges of predicting the impacts of technology on society and the limits of purely regulatory responses, we propose integrating ethics into the design process in order to protect human rights. A range of methodologies including Socially Responsible Innovation (SRI), Default Choice Architecture (DCA), and Value Sensitive Design (VSD) may be employed as part of technology design, implementation and maintenance. VSD, in particular, is a theoretically grounded approach to the design of technology – it includes such examples as rights by design, safety by design, and accessibility by design – that factor in human values in a principled and comprehensive manner throughout the design process. VSD allows for the involvement of ethicists, designers and developers, and diverse stakeholders including vulnerable and at-risk groups, in the development of new technology, in order to most effectively identify intended opportunities and also surface potential detrimental unintended consequences such as infringements on human rights. A focus on diversity in the workplace of technology design and incentives to drive this are also needed.

We argue (especially in our response to Question 4) that it is crucial that human rights and transdisciplinarity are taken into account at the technology design stage. By this, we mean that in designing, deploying, and monitoring technologies:

- a broad range of stakeholders should be involved through distributed shared agency (i.e., that agency should be distributed and shared across the involved stakeholders).

- the complex ecosystem of technology should be recognised, as discussed in detail in response to Question 1.
- as introduced in Section 2, various kinds of uncertainty – including those regarding the emergent structural impacts of technologies, and the effects they produce – should be recognised. As such, an ongoing iterative process of evaluation and decision-making should be considered.
- dialogue should be fostered between key stakeholders, in neutral spaces such as universities, to foster distributed shared agency.
- education is central to equipping citizens with the new literacy that is needed to live and work with emerging technologies.
- and, finally, that work is required to further develop ethical and philosophical frameworks for the broad range of stakeholders to work with and understand the impacts of technologies.

1.6 Distributed and shared agency

There is a need for public ethical assessment of emerging technologies by all those involved in and impacted by technological development and deployment. Platforms for dialogue and debate are needed to ensure all sectors of the community have an opportunity to contribute actively to this dialogue. Government and the education sector have a responsibility to equip citizens with the new literacy that is needed to live and work with emerging technologies.

In our response to Question 2, and the associated case studies in Section 4, we argue that to realise the potential of technology to protect and promote human rights, stakeholder engagement is required, with a broad understanding of the stakeholders impacted by technology, ensuring that engagement is accessible to stakeholders including via education. The positive potential of technology to promote human rights will be supported by an explicit targeting of positive outcomes (rather than simply regulating negative outcomes).

1.7 A transdisciplinary approach

We propose that as part of the public development of emerging technologies that protect and promote human rights, we must utilise transdisciplinary approaches (engaging multiple disciplinary approaches, as well as industry and users, to understand and develop solutions to complex problems) and creativity to imagine futures where emerging technology can promote human flourishing.

In the final section (Section 5) we argue that to identify and properly engage with the full range of issues that new technologies raise for human rights – that is, the issues and technologies we discuss in Sections 3 and 4 – a transdisciplinary approach is needed.

As we explain in the context of presenting our response to Question 6, regulatory approaches should supplement and build on this transdisciplinary design approach, drawing on experience from other jurisdictions and the work of non-government bodies in this space. Indeed, as we discuss in our response to Question 7, future regulatory development would benefit immensely from embracing a transdisciplinary approach. As we further elaborate in our responses to Questions 8-10, technology can have a profound impact on the human rights of people with disabilities, and the approach we recommend offers a wealth of opportunities in this context.

1.8 Technology assessment office

A major recommendation of this submission is for Australia to establish a new advisory body – a Technology Assessment Office (TAO) – to address a gaping absence in the Australian legal framework for new technologies, and to provide a platform for coordination, education, advocacy, and proactive public engagement. This recommendation is elaborated upon throughout the submission, and in Section 5 which contains our recommendations.

The process of framing a submission to respond to these complex and critical questions has enabled UTS, as an organisation, to develop a comprehensive appreciation of the threats and opportunities of emerging technologies and human rights. We are firmly committed to integrating human rights and ethics into the core of our education offerings and practices and see this as a critical foundation for promoting social justice and improving society. We look forward to working closely with AHRC and partners to address the challenges, and to realise the opportunities of these dynamic times.

2 Preamble

This document provides UTS's institutional response to the AHRC's HRT Issues Paper. In this preamble we cover six topics: (i) how this document is organised, (ii) key recurring themes in our responses to the AHRC's ten questions (iii) limitations to our responses, approach, and methodology, (iv) limitations to the AHRC's approach which warrant further attention and (v) the institutional process we undertook to develop this document.

2.1 How this document is organised

Section 2, following this organisational note, provides an overview of the approach taken by UTS in compiling this document, including the key themes that inform our submission.

Section 3, offers responses to the ten questions posed in the AHRC's HRT Issues Paper. These responses were led and written by a transdisciplinary team of Key Section leads with input from experts across faculties and from industry.

Section 4, presents case studies which bring together the themes and responses to the question, with reference to specific sectors and communities, to demonstrate the complexity of impact on lived experience of Australians in these areas. The first case study is an extended case study drawing on our expertise of AI and data analytics in an education environment. Three other case studies offer snapshots of impact of emerging technologies, in particular AI, on the disability sector, Indigenous communities and on people with intellectual disabilities.

Finally Section 5, which follows on from the case studies, outlines a set of **key recommendations** based on the research, analysis and findings of the 30 interdisciplinary experts who contributed to this response. Following on from this is an exploration of how a Technology Assessment Office (TAO) might fill a gap in oversight and education for emerging technology. The proposed TAO is contextualised in relation to best practice organisations emerging internationally with a remit for technology assessment and responsible innovation.

2.2 Key themes in our transdisciplinary framing

2.2.1 Hard impacts, soft impact, and guiding visions of human flourishing

Emerging technologies promise/threaten to have far-reaching disruptive effects on how we live, how we understand ourselves and others, and on the shape of society. Some of these technologies' effects can be predicted and evaluated with relative ease and certainty; these are sometimes called 'hard impacts'. For example, a hard impact of self-driving cars might be a reduction in the number and severity of motor vehicle accidents or reduced traffic congestion.¹ But other effects with more of a social and cultural impact sometimes called 'soft impacts' are harder to foresee, and even more difficult to evaluate. For instance if self-driving vehicles become the social norm we may no longer be able to have the independence to take out a non-autonomous vehicle due the increased risk and inability to get insurance. Despite their importance, soft impacts are extremely difficult to imagine, predict and evaluate.

¹ Tsjalling Swierstra, "Identifying the Normative Challenges Posed by Technology's 'Soft' Impacts," *Etikk i Praksis - Nordic Journal of Applied Ethics*, no. 1 (May 9, 2015): 5–20, <https://doi.org/10/48b>.

Kudina and Verbeek point out, while early in the development of a technology, its impacts can be hard to assess, later ‘when the implications for society and morality are clearer, it is more difficult to guide the development in a desirable direction’, because technologies become entrenched.²

Given the importance of being able to predict soft impacts, in this document we have attempted to identify and draw attention not only to the more easily predictable hard impacts of emerging and new technologies – ones which are most likely to have clear human rights implications on account that we are more likely to have noticed them and thus to have created legislation to protect them – but also to predict and evaluate potential soft impacts of new and emerging technologies. Where it is not clear precisely what human rights those impacts implicate, we instead reference the important human or social interests involved.

Lastly, what’s still left out by an approach that focuses solely on identifying potential threats to human rights from emerging and new technologies – even if the potential threats identified include soft impacts as well as hard impacts – is that while such an approach might help steer society away from uses of technology that would have negative impacts, it offers little guidance about how to design, evaluate, and regulate emerging and new technologies so that we may advance towards distinctly positive or desirable outcomes. For this reason, as we explain at length in Section 3, we also argue that what’s needed is direction on how emerging and new technologies should be developed, tested, evaluated, and regulated so that they may promote human and ecological flourishing.

2.2.2 Digital citizenship and agency

We often think about technological advancement simplistically by identifying commercial technology developers as the key actors driving technological progress. Opportunities for **distributed or shared agency** that arise in the contexts of technology use are frequently overlooked. Our response proposes several avenues for possible action to protect and promote human rights in fields as diverse as education, public sector, peak professional bodies, community organisations, in addition to the typical demands for accountability from companies developing technologies. We argue that there is a need to ‘complexify’ the ways we think about the field of emerging technologies. In particular, we propose that we need to create more **opportunities for citizen participation in evaluation of technologies and decision-making** in order to shape desirable futures for technologically enhanced societies that support and protect human rights to ensure human flourishing for all.

As an active global digital citizen there is a need to control your own digital identity and to have some transparency about what and where your personal data is being stored and how it is being used. This relates to the human rights to liberty and privacy. The recent introduction of the GDPR with a focus on data protection principles and data minimisation has had a significant impact for digital technology vendors, requiring many to rethink and revise their approach to personal data collection and privacy. However, there are few other legislative frameworks to protect people from long ranging data records that could have a very real impact upon the choices that are then made available to them throughout their life. This becomes increasingly significant with the growth in the use of decentralised and immutable ledgers such as blockchain technologies. This requires a rethink of the role of government or civil society in providing a stable and secure digital identity with greater transparency over data collection, value, exchange and destruction.

A core value of democratic, civil society must be acknowledged in making visible the

² Olya Kudina and Peter-Paul Verbeek, “Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy,” *Science, Technology, & Human Values*, August 21, 2018, 0162243918793711, <https://doi.org/10.1177/0162243918793711>.

frameworks which an organisation uses for data protection and data governance. Making this visible entails that citizens can and should expect to have a voice in the decisions being made by their government (and government agencies acting on their behalf) and the way that data practices shape government activities, especially decisions that impact on the everyday life of its citizens; ultimately, it involves showing the data ‘layer’ in government and organisational practice.

2.2.3 Education and awareness

Education is a key theme that will be explored throughout this paper. UTS has lived experience of working with AI and new learning technologies and seeks to establish models for best practice in this area. As a university, UTS also has a role to play in relation to the right to education of the public around the implications, opportunities and threats of emerging technologies. This education on technology, ethics and human rights is needed at all levels from pre-school to university education to education of technology industry in applying ethical and human rights principles within the design processes.

2.2.4 Reactive vs proactive responses

It is helpful to think of responses to identified concerns and problems related to new and emerging technologies in terms of two intersecting distinctions – namely, between reactive vs proactive responses, and between regulation vs design.

On the first distinction, because the concerns and problems that emerging and new technologies create are often only identified after the technology has been introduced into society, it may sometimes be necessary to wait until after the concerns and problems start occurring, and then to generate a response. However, on other occasions, it may be possible to predict at least some of the potential issues and problems that a technology may create, and in such cases, we may take proactive steps before the issues and concerns materialise in the hope of preventing them from doing so. We take it that, when important concerns and problems can be identified by engaging in predictive exercises, it is preferable to take the proactive rather than reactive approach, if only to avoid the concerns and problems manifesting in someone suffering avoidable harms.

2.2.5 Human rights by design and value centred design

When it is possible to predict identifiable concerns and problems, we propose that it is preferable to design the technology in ways that avoid harms. In this context, the most beneficial use of regulation is to mandate that technology designers and those who commercialise them have a legal duty to engage in adequately documented efforts to predict what potential concerns and problems the technology may give rise to, and then to employ one of three well-developed methods designed to create better technology (better in the sense of being less prone to giving rise to concerns and problems): Value Sensitive Design (VSD), Default Choice Architectures (DCA), and Socially Responsible Innovation (SRI).

Since a fuller discussion is provided in section 5 below, here we offer just a one-sentence summary for each of these methods. Firstly, Value Sensitive Design involves designers treating considerations like privacy, equality, responsibility, and accessibility – to name just a few values – as equally important to technical specifications like, for example, power consumption, features, materials, etc. Secondly, the core idea behind Default Choice Architecture when applied to technology design is to configure technologies in such a way that, by default, their normal use will generate positive outcomes. Thirdly, the point of Socially Responsible Innovation is to involve stakeholders at all stages of design, testing, and regulation (on an ongoing basis) of a technology, not only so that the values that society endorses can be ‘baked into’ the design of that technology (by using VSD and DCA), but also

as a way to help identify potential uses, potential problems, and potential opportunities which we may otherwise have overlooked.

When a proactive approach is taken, regulations mandate that designers shall use methods like VSD, DCA, and SRI to do their best to predict and to take measures to avoid their technologies causing adverse outcomes, or at least to minimise the risk of this happening. The approach of ‘human rights by design’ sits within a Value Sensitive Design framework alongside safety by design, accessibility by design etc.

2.2.6 Keeping the human in the loop – data humanism

A growing recognition of the value of ‘data humanism’ is countering the push to remove human hands from big data, drawing connections between humans and the (data) representations of them³. Concerns that data and AI-informed technologies truly serve humans (as individuals and collective groups) is fuelling government and social consideration of systems and functions that may need to be created to mitigate the damage caused by information and data asymmetries. Within the wider community, there is a growing consciousness about the vulnerability of data to misinterpretation, misuse and misappropriation. These concerns are intertwined with emerging concerns in public and professional contexts about increasing applications of and dependencies on AI and data-driven decision making in ever-increasing contexts.

There is also a growing cry to ‘turn data around’ and design data systems that take into account the well-being of the very people whom the data is obtained from, and represents, in the first place⁴. There are two synergistic calls in this space. First, for better algorithmic governance, that actively engages the public in long-term visions for data and AI practices which more explicitly protect human rights. Second, a growing interest in frameworks by which governments and organisations that gather, hold and use data, and apply algorithmically-informed decisioning can obtain a ‘social license’ to operate. Establishing co-design frameworks and participatory models and mechanisms for ongoing feedback with sufficiently diverse participants, especially including vulnerable groups, is one way to work towards this future.

2.2.7 Role of the public sector in demonstrating best practice

The delivery of what were once key public services of health, education, social services and emergency services, are increasingly shifting towards a private/public partnerships model. This creates a unique opportunity to prototype and demonstrate best practice in design and delivery of emerging technologies. A proactive approach to ensure the best public outcomes, which take into account ethical principles and public rights, is essential. This will support the development of public trust in public institutions and services and avoid a loss of trust as in the recent case of ‘Robodebt’ with Centrelink customers. This opportunity to work closely with the technology industry to proactively shape the process of designing, testing and implementing new technologies will enable new models and frameworks to be developed for use within both the public and private sector and will lead to a better understanding of the resources implications and requirements of applying ‘human rights by design’ methodologies.

³ Giorgia Lupi, “Data Humanism: The Revolutionary Future of Data Visualization,” Print Magazine (blog), January 30, 2017.

⁴ Jer Thorp, “Turning Data Around,” Memo (Random) (blog), November 18, 2016, <https://medium.com/memo-random/turning-data-around-7acea1f7479c>.

2.2.8 Challenges in assessing technology impacts

The task of prediction is complicated, in part, simply because we rarely know all the relevant factors and complex interactions that influence what effects will be produced. Furthermore, our scientific theories, models, and methods used in making predictions, are rarely if ever complete and fully accurate. Finally, many of the important effects that new technologies produce fall into the category of so-called ‘soft impacts’. For instance, effects on society, or on our views and attitudes, or emergent cultural forces that in turn shape the way people behave, which adds a further layer of complexity.

2.2.8.1 Inadequate theories, extraneous factors, and ‘hard impacts’

To see where the challenges that beset prediction stem from, consider three examples of new technologies – CRISPR Cas-9 gene editing, so-called ‘smart drugs’ or ‘nootropic’ medications, and autonomous vehicles – which may plausibly have, among their respective intended effects, such things as fewer genetic diseases or disorders, better learning outcomes in education and improved productivity and extended wakefulness at work, and a reduction in the number and severity of motor vehicle accidents.

The first group of predictive challenges is common to many if not all technology assessment (TA) attempts. Put simply, as long as everything goes according to plan – e.g. our science is sound, our tools are reliable, mistakes aren’t made, and no extraneous factors intervene – then the intended effects will be produced (for the moment, we leave aside the question of what other effects those effects might produce, which we return to below). However, if our scientific theories are flawed or incomplete, if we overlook important factors, if corners are cut in the manufacturing process, if products are not tested properly, if identified problems are ignored or covered up rather than reported and fixed, or if the products get used in novel ways or under novel conditions that we did not anticipate, then not only might our intended effects not manifest, but a range of unintended effects might materialise. For instance, we might inadvertently create new genetic diseases or disorders, smart drug users may develop insomnia or become addicted to them, and software or hardware malfunctions in the control systems of autonomous vehicles may cause accidents and traffic jams. Under such conditions, foreseeable but nevertheless unexpected effects, or ‘hard impacts’, may come about.

History is replete with examples that demonstrate this point. For instance, despite the fact that the pharmaceutical industry is heavily regulated, extremely well resourced vis à vis finances, and has an intricate and thorough system of methods – e.g. laboratory experiments, clinical trials, and studies of epidemiological data – unexpected adverse reactions and longer-term side-effects (like those that occurred in the diethylstilbestrol and thalidomide tragedies) still occur. And despite decades of experience in the design of computer software, as well as extensive public test beta cycles intended to spot and fix bugs, even very financially prosperous companies like Apple and Google still end up releasing computer hardware and software that crashes, results in data loss, and opens up glaring security holes that can be exploited by hackers.

Even in industries that have highly developed methods designed to discover and tie down all the relevant factors, immense financial resources, and tight regulation, predictions of such obvious and easy-to-imagine hard impacts are still often plagued by a lack of certainty, and even by the certainty that unexpected and unwelcome effects will come about. Given such mundane but undeniable facts, there is every reason to be conservative and maybe even pessimistic when it comes to estimating our ability to accurately predict the hard impact style effects of new technologies.

2.2.8.2 Soft impacts stretch the imagination and involve much complexity

However, even if we set aside the above concerns and suppose that everything goes according to plan – i.e. that the intended effects are produced because we correctly anticipated all the relevant factors and our theories were sufficiently robust etc. – there is still a good chance that the three example technologies mentioned above (CRISPR Cas-9, smart drugs, and autonomous vehicles) may produce a plethora of additional flow-on effects, many of which are way more difficult to even imagine let alone predict.

For instance, with a healthier population, safer roads, improved learning, and greater work productivity, the need for some medical, educational, and other professional services may reduce. This may lead to closure of some hospitals, schools, and businesses, and that in turn may lead nurses, doctors, teachers, and other professionals to lose their jobs. Consider three more similar examples. If fewer people fall ill due to the increasing eradication of genetic diseases and disorders, perhaps we may become less tolerant of workers who do fall ill and need to take time off work on sick-leave, because their parents did not have the sense to ensure that they edited their child's genes. Similarly, if safe, effective, and inexpensive smart drugs are developed and made available to everyone, rather than helping us get through a day's work more quickly so that we have more free time on our hands to relax, this development might, somewhat paradoxically, instead create an even more competitive and demanding work environment. Given the competitive nature of work in industrialised societies, if even some people decide to use smart drugs not to get their work done faster and then relax but to work even longer and become more competitive, then whoever they put at a positional disadvantage by doing this may feel the need to also start using smart drugs in order to not fall behind. This, in turn, will put even greater pressure on the rest of the population to follow suit, and the more people do so, the more the pressure on those who haven't yet done so to start using smart drugs just to avoid falling behind. Finally, if autonomous vehicles indeed turn out to be much safer than human-operated motor vehicles, then humans might eventually lose the right to take their manually operated car out on the roads, or to get on a motorbike and enjoy a ride through winding roads on the weekend.

In all of the above examples, the imagined scenarios depict clear examples of effects – i.e. job losses, less tolerance of sickness, an even more work-obsessed society, and the loss of a right to drive and ride manually operated motor vehicles. However, none of these effects are easy to imagine – at least not without the benefit of hindsight – in part because they are more indirect and temporally further removed, in part because these effects are of a qualitatively different kind to the effects which we originally intended to bring about, and also in good measure because of the staggering number of factors that the production of these effects are contingent upon. Whether any of these effects would occur is anybody's guess. However, our point is simply that effects such as these 'soft impacts' are incredibly difficult to even imagine, since they are probably the last thing that we will be thinking about when we set out to produce the intended effects.

In case the above examples seem too fancy or unrealistic, history is again replete with concrete examples of such soft impacts. For instance, who would have imagined that the introduction of the mobile phone would eventually result in job losses – e.g. of receptionists who were no longer needed by tradespeople who could now field their own calls from prospective clients while out on jobs. Who would have imagined that decades after the mobile phone was introduced, the clientele of taxi drivers would be diverted to ride-sharing services like Uber, or, equally, that restaurants would acquire new business by preparing meals for people at home that would be delivered by yet further services like Deliveroo, Foodora, and Menulog? And did it ever occur to anyone that not even a decade after the mobile phone and email were introduced, people would start to get annoyed if they did not get a reply to

their messages within a few hours or even minutes? Other examples of difficult to predict soft impacts might include the way that the introduction of electricity resulted in extended working hours in factories due to the availability of electric light, and how the introduction of motor vehicles impacted on the design of cities, that it would enable the urban sprawl, traffic congestion, and create countless business opportunities for service stations, car cleaning businesses, and the like, as well as enable fresh produce to be grown on farms in warmer climates far away from the cities where people live.

By comparison to the unintended hard impacts – genetic diseases and disorders, insomnia, drug addiction, and road accidents – many soft impact style effects involve changes to our norms and values, to the expectations we have of one another, and to how society is organised. Our point is simply that such effects, though clearly important, are extremely difficult to imagine let alone to predict, because we are unlikely to even be looking out for them.

2.2.8.3 Prediction is complex

Another critically important factor, which applies equally to hard and soft impacts, is that in few if any cases, will most of the effects simply be the results of some individual new technology. Rather, as we detail in our response to Question 1, most effects will be produced by different combinations of many older technologies (e.g. GPS, cellular telephony, motor vehicles, and the internet), new technologies (e.g. high resolution touch screens, high capacity electric batteries, low power consumption and high-powered microprocessors), as well as countless human decisions and interactions, which in turn operate within social contexts (e.g. market pressures and opportunities) that in turn exert their own further influence on how people subsequently behave.

The sheer complexity produced by the countless interactions, and the way that these interactions create emergent properties – e.g. the cited market pressures and changes to individual values and cultural norms – which produce their own effects, means that predicting even a small portion of the unintended (and even the intended) effects of new technologies presents seriously daunting challenges.

At a minimum, a holistic approach is needed to make accurate predictions – an approach that is capable of taking into account a broad range of factors and interactions, as well as emergent properties of systems and how those systems in turn create their own effects.

2.2.9 Evaluation

In this section we argue that evaluating the effects of new technologies is also very complicated.

2.2.9.1 We cannot evaluate what we cannot predict

Our first point is simply that, given the predictive challenges (see above), whatever effects we cannot predict we also will be unable to evaluate.

However, even with those effects which we can predict, as the next two sections will argue, there are still important constraints on our ability to evaluate them.

2.2.9.2 Experience and entrenchment of technologies

Sometimes we simply need to experience the effects of a technology before we can evaluate it. When social media platforms like Facebook first appeared, some people viewed them as frivolous, shallow, and superficial forms of interaction – i.e. something that was not valuable. After experimenting with social media, though, many people changed their minds.

However, given that technologies, once fully deployed, can become socially entrenched, this creates a double-bind. On the one hand, we cannot properly evaluate some technologies until after we have experienced their effects. Thus, to experience their effects we may need to allow them to be deployed. However, by deploying them, the technology may become socially entrenched and impossible to withdraw from use in society.

This double-bind, known as the ‘Collingridge Dilemma’, means that the process of evaluating a technology’s effects may itself entrench that technology in society, and thus undermine the aim of evaluating it and potentially taking measures to address problems prior to it being deployed.

2.2.9.3 Experience and transformative change

Relatedly, the need to experience the effects of new technologies to evaluate them also creates another challenge. Namely, that in the process of using that technology our values can change, and thus that we will use the changed values not the original ones to evaluate the technology. The above Facebook example demonstrates this, since people who originally viewed the forms of interaction available on Facebook as frivolous, shallow, and superficial, eventually developed an interest in interacting this way.

What makes this a problem specifically is that if experience is needed to evaluate effects of new technologies, but yet having the experience will transform our values, then by the time we get around to evaluating the technology, we may no longer have the same values. Consequently, since the aim of evaluating a new technology is to make an informed choice about whether it has adverse or beneficial effects, this aim will be undermined if the process of experiencing the new technology’s effects itself alters our values rather than just giving us new data to evaluate.

2.3 How this document was developed

The approach to developing a response to the AHRC’s HRT Issues Paper at UTS was somewhat experimental in the number of contributors from different disciplines involved in order to provide a truly transdisciplinary response. The response to the questions were shaped by small teams of academics from different disciplines with self-nominated team lead/s. These contributors then met at regular points to discuss key themes, research findings and recommendations. As a result a number of the responses to the questions below may have a slightly different disciplinary skew and style and tone of writing. Where possible the responses have been edited to ensure consistency and readability while avoiding losing the specific disciplinary insights.

A draft response to the Issues Paper was delivered to the AHRC in order to provide a foundation of discussion and dialogue for a successful Roundtable between UTS and AHRC which enabled UTS to provide an initial response to the AHRC’s burning questions as well as identify areas of additional research and focus for this final submission.

2.4 Limitations to UTS’s approach

The volume and speed of emerging technologies that are impacting Australians can sometimes feel overwhelming at both an individual, community, and societal level. Even the assumption that we can talk about a category of ‘emerging technology’ or ‘new technology’ is problematic as technologies are increasingly combined and used in tangential ways as noted in AHRC’s Issues Paper.

It is impossible and indeed perhaps unnecessary to attempt to address all of the emerging technology categories and instead this paper will focus on particular technologies where the writers have direct experience or research knowledge of human rights and ethical implications.

Through developing this paper it has become clear that no one size fits all approach will apply for legislation or regulation of emerging technologies. Each technology and applications and permutations of that technology must be considered within the specific context in which the technology is applied. For the reasons noted below in response to Question 1, we find it unhelpful to work with a fixed list of broad categories of technologies. Instead, to properly identify and characterise salient issues, we find it more helpful to focus on concrete examples. That is, examples of particular technologies (or combinations thereof) used in specific contexts for determinate purposes.

Although over 30 people from UTS contributed to producing this document, this is a small proportion of the entire 3,600 staff and 44,900 student population. Efforts were made to be as inclusive as possible and to include responses from all those who expressed an interest, ideally this document would be circulated once more to obtain wider feedback from the university, both in writing and at a public university forum. We note this to acknowledge that the views expressed may not be representative of the whole of UTS's community and to gesture at what steps could be taken to further improve it.

2.5 Limitations to AHRC's approach

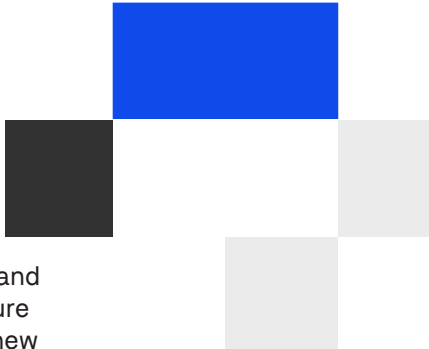

Despite our support for a human rights approach, an exclusive focus on human rights has some limitations. First, a focus on the effects of specific rights can tend to ignore more subtle effects, which we refer to as 'soft impacts'. Secondly, an exclusively 'rights-based' focus can lead to threats to human interests which may not, strictly speaking, be defined as human rights being overlooked. Thirdly, a human rights approach can tend to lead to too much of a focus on protecting rights against threats rather than on providing guidance on how to steer to a positive future. These three potential limitations on a human rights approach are discussed in detail below.

We offer three examples to explain why sometimes an exclusively human rights approach may either delay us from noticing, may lead us to overlook, or may result in us misunderstanding the specific character of the adverse effects that new technologies have.

However, before we do this, let us clearly state that this is not intended as a criticism of a human rights approach, but as a way of highlighting why this approach could benefit from being supplemented with a broader transdisciplinary approach. Clear advantages of a human rights approach include the greater certainty and protection than what is offered by ethical frameworks – for instance, like those being developed in regards to AI – in good measure due to being well grounded in established and widely endorsed international human rights law. A human rights approach also provides clear, well tested, thorough, and comprehensive methods for mitigating potentially detrimental human impacts of new technologies, by recourse to a broad range of substantive and procedural rights.

However, at the same time, as the next three examples demonstrate, it may sometimes be necessary to step outside of a human rights approach in order to identify emerging vulnerable populations, to notice social effects which have important adverse flow-on effects for individuals, and to accurately understand the adverse character of some effects.

New vulnerable populations. Although Article 16 of the Universal Declaration of Human Rights (UDHR) grants men and women the right to marriage and family, transgender people may unfortunately be deprived of protections granted by this right. Although modern medical procedures have made it possible for transgender people to undergo medically-assisted gender transitions through such techniques as puberty-blocking medications, cross-sex hormone therapy, and gender reassignment/confirmation surgery, these procedures unfortunately cause sterility. Given that medical treatment with puberty blocking medications



and later with cross-sex hormones can start before children have gone through puberty – and thus before they can produce sperm and eggs which could be saved for later use – to ensure that transgender children can raise a family, the medical profession may need to develop new technologies to mitigate this effect. Admittedly, not all transgender people undergo medical transitions, since some identify as non-binary – that is, as neither women nor men. However, given that the rights granted by Article 16 are extended to men and women, and non-binary transgender people are neither women nor men, non-binary transgender people’s right to marry may at present not be protected.

Our point in citing this example is that prior to the development of medical technologies that enable people of one gender to transition medically to the other gender, and prior to recognition of people of non-binary genders, these issues are unlikely to have even occurred to anyone simply because these groups of people either did not exist or were not recognised. However, given that new technologies can create new social categories of vulnerable people to whom presumably these human rights protections should still be extended, it may occasionally be helpful to step outside of a human rights framework to investigate whether the use of new technologies might create new vulnerable groups.

Structurally produced infringements of rights. Some ways in which new technologies can compromise important human interests may not be noticed if we restrict ourselves to using an exclusively human rights approach. For instance, consider how the gradually increasing use of new neurotechnologies to predict and prevent antisocial behaviour in the criminal justice system, surreptitiously re-prioritises the relative importance attached to safety by comparison to liberty.

Brain scanning technologies are increasingly being introduced into courts around the world – including in Australia – to more accurately predict people’s risk of offending, calls to use neuro-intervention techniques to target the brain-based causes of antisocial behaviour are also gaining traction. The past two decades have witnessed incredible breakthroughs in the field of neuroscience, and as this trend continues these technologies will increasingly become more powerful and effective at making accurate predictions and at providing effective interventions. However, when courts are presented with increasingly accurate predictions of important risks and ways of addressing those risks, this leaves them with little choice but to do something about them. Knowing about an imminent risk and doing nothing would be reckless – in this way knowledge can be coercive, since discovering some things can create imperatives to do something about them. However, what this overlooks is that whatever gains society might eventually make in terms of safety, will ultimately be purchased with sacrifices to liberty. To become a safer society, more things will need to be monitored and controlled, and so individuals will gradually have less say over those things.

Our aim here is not to take a stance on whether this is a price we should be willing to pay, or whether freedom should not be traded away to purchase more safety. Rather, our aim is to highlight how technological developments can surreptitiously re-prioritise our values – in this instance, in favour of safety, and against liberty. This occurs not because anybody made a conscious choice to trade away some liberty for more safety, but because of how predictions can fixate our attention on safety, while failing to draw our attention to the fact that preventive measures will impact on liberty.

Because in such cases no concrete instance of the use of a new technology will adversely impinge on anybody’s liberty, human rights concerns are not likely to be noticed. On the face of it, these will appear like well-founded and rational choices – to take measures to protect ourselves from identifiable risks. However, at a structural level, some degree of freedom will have been sacrificed to purchase that increased safety, and this is precisely the sort of effect that we need to notice so that it can be opened up for public debate. For instance, to discuss

how much safety we want to secure, given that additional safety may come at the price of diminished liberty.

To even notice that new technologies can re-prioritise such core values as safety and liberty – values that lie at the heart of what human rights are designed to protect – paradoxically, we may need to employ something other than just a human rights approach.

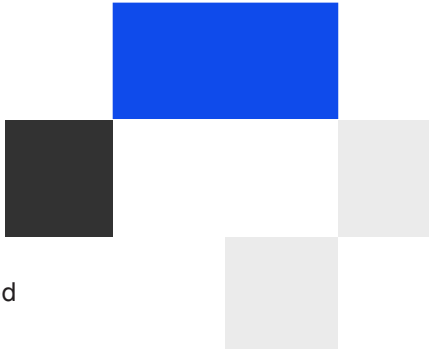

The evolving character of important human interests. Consider two distinct ways in which we can account for why the virtual world matters, and how these different accounts impact on how long it takes to give due recognition to the need to extend human rights protections into new domains.

On the first account, the virtual world matters only derivatively, because of how the things that happen in the virtual world impact on what happens in the physical world. The nexus where tangible human interests intersect with intangible technologies such as software and data is one example. Software and data play a prominent role in the operation of autonomous vehicles and in rendering professional advice – e.g. legal advice, or in medical diagnostics and treatment. The fact that autonomous vehicles can injure people, damage property, and cause pollution, and that lives may be saved, lost, or ruined when good (or bad) professional advice is rendered, clearly means that the virtual or intangible nature of software and data are no reason to disregard them. Technologies involved in robotics and the Internet of Things (IoT) offer other examples of how the code and data have concrete physical effects. The HRT Issues Paper also notes that factors which create or widen the digital divide by impeding some groups' access to the internet can have adverse effects by excluding people in the affected groups from participation in the digital economy and from equally deriving benefits from new technologies.

On this first account, the reason why the virtual world matters is derivative – its significance derives from the fact that what happens in the virtual world has spill-over effects in the physical world. However, it would be a mistake to only imbue the virtual world with such derivative significance. Not only does this overlook the distinct significance that virtual interactions and spaces have acquired, but it also delays how long it takes to extend human rights protections to new domains of importance to people.

On the second account, the virtual world has acquired significance to humans in its own right. Human interactions increasingly take place online. We do not do our shopping online, or pay our bills online, or interact with each other online – we just do our shopping, pay our bills, and read, comment on, and post updates on social media. Losing access to our social media accounts, or staying away from online forums – e.g. due to fear of online harassment and exposure to cyber-abuse – are significant even if they have no adverse spill-over effects in the physical world.

When harms such as cyber-bullying, cyber-hate, or cyber-abuse are only given recognition because of the flow-on impacts in the physical domain, this fails to give due recognition to the evolving character of what comprises important human interests – in this case, it fails to recognise the significance that the virtual world has acquired for humans. It is also telling that unlike the physical environment which is policed, regulated, and protected, in online environments, there are still unregulated spaces where people interact. Quickly created, disposable accounts are easy to set up and abandon, and abusers can benefit from being located in a different jurisdiction (e.g. in another country) where particular abuses are not recognised as criminal offences. There is also no easy way to report online abuse, it is not clear what evidence a victim would even need to present to the police to follow up online abuse, police are not trained to deal with online abuse, and despite serious personal consequences and locking out people from an increasingly important sphere of human



interaction, cyber-abuse such as online rape threats and revenge porn are still often played down as if they were more trivial, less important, or less worthy of the law's attention than physical abuses.

Our aim in citing these differences between how the physical and virtual interactions and spaces are regulated is not (just) to argue that greater protections are needed in the virtual domain – something that the AHRC already recognises – but that this is a symptom of a failure to recognise that, what constitutes an important human interest, is something that evolves over time. Human rights protections should be extended to the virtual domain not only because of the virtual domain's derivative significance, but also because of the novel significance that online interactions and spaces have acquired for humans. Put another way, even if what happens in the virtual world had no ramifications for what happens in the physical world, interactions and spaces in the virtual world should still be protected because new technologies have created these new important human interests. The concern is that by anchoring ourselves too firmly within a human rights approach, this may repeatedly delay recognising the evolving character of important human interests, which in turn will delay extending human rights protections to important new domains of human life as these domains come to be created through new technologies.

For reasons like the ones highlighted by the three above examples, we believe that the human rights approach could be helpfully supplemented by a transdisciplinary approach which is further expanded on in Question 4 below.

3 Responses to AHRC's questions

Each of the consultation questions posed by the AHRC's HRT Issues Paper raises issues of considerable complexity. The following captures some of the most salient elements of the multi-faceted responses of the transdisciplinary team contributing to this submission.

3.1 What types of technology raise particular human rights concerns? Which human rights are particularly implicated?

The HRT Issues Paper has a particular focus on AI, big data, and AI-informed decision making. We agree that features of many of the new and emerging technologies giving rise to human rights concerns involve the collection and analysis of large data sets, and decisions and other automated actions based on data analysis. We do, however, consider that it is important to take into account implications of a broader range of technologies, a focus on particular technologies in specific contexts used for concrete purposes, significant new uses of established technologies, and novel combinations of existing and new technologies. We agree with the HRT Issues Paper that the human rights particularly implicated by new technologies include the rights: to privacy; security, safety and the right to life; and the right to non-discrimination and equal treatment. It is, nevertheless, important to consider the broader effects of new technologies on human rights more generally, which have implications for rights as diverse as equality before the law and freedom of expression. We are especially concerned with the differential impact new technologies may have on vulnerable and at-risk populations.

In this response we identify a range of considerations that count in favour of revisiting the list of technologies which the AHRC listed in its HRT Issues Paper. In particular, we propose that it would be beneficial to consider not only new technologies, but also new uses of existing technologies in certain contexts, novel combinations of new and existing technologies, interactions between technologies, as well as interactions between technologies, regulations, laws, market mechanisms, and humans.

3.1.1 Assumptions and structure of our discussion

Our discussion proceeds from four considerations:

1. If new technologies pose threats to human rights or, if they create opportunities to protect them, then this is presumably because of their potentially detrimental or beneficial effects on humans or society.
2. We presume that the AHRC's call for submissions reflects the view that some effects of new technologies might be non-obvious and thus difficult to identify. For this reason, our discussion shall focus on 'the risk of unintended consequences'⁵ and 'mechanism[s] by which to identify, prevent and mitigate risk'.⁶

⁵ AHRC, "Human Rights and Technology Issues Paper (2018)" (Australian Human Rights Commission, July 24, 2018), 24, <https://www.humanrights.gov.au/our-work/rights-and-freedoms/publications/human-rights-and-technology-issues-paper-2018>.

⁶ AHRC, 17.

3. We shall thus also focus on more temporally distant and thus the more difficult to identify risks. This does not entail that current, imminent, and obvious technologically induced risks may be disregarded, but only that we wish to set our focus on explaining the challenges involved with identifying and taking effective measures against ‘the risk of unintended consequences’.⁷
4. We agree with the AHRC’s comment that ‘[n]ew technologies do not inevitably threaten human rights, but the problem of dual affordances, or multiple uses, is particularly acute with new technologies. Many such tools can be used to protect and violate human rights.’⁸ Since it is not feasible to investigate all technologies, though, our discussion will focus on technologies which raise concerns but not chiefly because they could intentionally be used nefariously, since effectively every technology could be intentionally used for nefarious ends.

Given the above assumptions, one approach to the AHRC’s initiative on human rights and technology could look like this. First, we need to *identify* the set of potentially relevant new technologies. Next, we must *predict* the potential effects of those technologies on humans and society. Then, we must *evaluate* the predicted effects to ascertain if they are detrimental, beneficial, or neutral *vis à vis* human rights. Finally, armed with evaluations, we should *devise measures* to protect the implicated human rights. Schematically, the four components or stages of this approach could be represented as follows:

identify new technologies → predict their effects → evaluate the effects → devise protective measures

The next four sections discuss challenges that each of these components must tackle, in light of the discussion regarding prediction and evaluation in the preamble.

3.1.2 Technology

The HRT Issues Paper states that ‘We need to set priorities for our response, and so it is critical to understand which forms of technology most urgently engage human rights. The World Economic Forum highlighted 12 types of technology that merit close attention.’⁹ We agree about the importance of priority setting, and believe that these twelve items present a helpful starting point, however we shall also argue that (a) some other important items should be added to this list, (b) novel combinations of new and existing technologies should also be considered, (c) as should new uses of existing technologies in different contexts for different purposes, and that we should also consider (d) interactions between technologies, humans, regulations, laws, and markets mechanisms, as well as (e) emergent effects produced by these interactions.

3.1.3 Omitted technologies

Some important technologies are absent from The World Economic Forum’s list. For instance, *nanotechnology* which has applications in medicine, cosmetics, and many other domains is an important example. Given the data collection technologies play a critically important role in AI-Informed decision making, devices such as *fitness trackers* (which are extremely common and collect incredible amounts of intimate biological data about us, including even when and how we sleep and wake), GPS and other *location tracking* devices and techniques (which again have become so commonplace that it is easy to overlook them, but yet the data they produce

⁷ AHRC, 24.

⁸ AHRC, 19.

⁹ AHRC, 18.

about our whereabouts, with whom we meet and thus our associations which we may prefer to keep private, is again very intimate and revealing), and other similar data collection devices should be added to the list. *Immersive games* including VR have become a mainstream form of entertainment, which may have a profound impact on culture, and social as well as personal values, so this may be another technology that should be considered.

3.1.4 Older technologies used in new ways

Other less ‘novel’ technologies are coming to be used in novel contexts and in novel ways.

For instance, there is mounting concern about the increasing use of *video conferencing* in parole board hearings in Australia, which adversely affects prisoners’ rights by creating significant communication barriers, and affects prisoners’ demeanour during hearings in ways that are likely to work against them. On the first point, often prisoners do not even have a monitor on which to view their interlocutors, or the camera is not adjusted to point at the prisoner’s face with the effect that the parole board cannot see their facial expressions and thus cannot judge their affect, and even the quality of the audio connection can be variable. On the second point, because of the jittery nature of the communication, prisoners not infrequently become nervous, act confused, and behave in ways that depart from how they would behave if they appeared in front of the parole board in person.

Another example, also from the criminal justice context, involves the use of medical interventions to maintain competence for punishment. Medicating criminal offenders to make them fit for execution is one example of a particularly concerning use of a medical technology that is practiced in the USA. However, treatments for medical conditions that people suffer as they age – e.g. dementia – have existed for quite some time, and these treatments may also be used to keep prisoners sufficiently mentally healthy so that they may be kept in prison longer merely to serve out their full sentence. The use of healing technology for this purpose, though, may be an inhumane and degrading treatment.

Video conferencing and dementia medications are clearly not new technologies. However, we offer them as examples that *new uses* of older technologies – especially when this happens in sensitive contexts – raises important concerns. In our view it would be wise to not overlook such new uses of older technologies merely because the technologies themselves are not new. Arguably, given the pervasiveness of older technologies, their potential to have negative effects on human rights when used in novel ways may pose an even greater risk than new technologies. We thus urge the AHRC to reflect on whether it may not also be important to investigate the risks posed to human rights by the use of older technologies in new ways.

3.1.5 Hybrid technologies

Technologies are also often combined in various ways to create what might be called new ‘hybrid technologies’. As the example of fitness trackers and location tracking devices (e.g. GPS) discussed above demonstrates, the introduction of new technologies can significantly alter the issues raised by older though widely deployed technologies. Consequently, it may help to look beyond the distinctness of the twelve items on The World Economic Forum’s list, and make room for appraising hybrid technologies.

Consider the distinction between new and old technologies, and the delineation of twelve different kinds of technology based on the core disciplinary field or industry from which they derive.

The problem with restricting the investigation to a fixed list of technology types is that hybrid technologies can combine technologies (some new, some old) which straddle different technology types. For instance, *smart pills* combine older medical technology (pharmaceuticals) with a data gathering/tracking technology that senses and records

information about when a patient took their medicine, and conveys this information to a smartphone or a computer. Another example is *synthetic biology*, which might either be viewed as a biomedical technology, or as a nanotechnology, depending on which element of the hybrid technology one focuses on. Finally, *autonomous vehicles* and *smart cities* are both clearly new technologies, but they too are hybrids, since they combine a wide variety of different technologies.

The city may be understood as a technical system, and as an expression of the cultural and civic aspirations of a people, indeed the essence of its polity, built on fundamental rights that underpin notions of public space, civic expression and contribution, and cultural development. Yet every aspect of our urban lives, from the quality of the water we drink, to the reliability of the public transport that moves us to the way we gather together, are being impacted by smart technologies created to optimise our urban environments by reducing waste, managing costs and risk, and ultimately aiming to improve the overall quality of life for us all. According to announcements from governments across South East Asia alone, close to 1 billion people in over 500 cities will be touched by the roll out of smart cities initiatives by 2025.¹⁰

But these same technologies create significant challenges to our lived experience of and very concept of urban life. Who, when and where we meet are now data points that tell a story of who we are and how we live. And as more and more data sets are linked and analysed using AI's and ML, our urban centres and patterns of living are being assessed and shaped in whole new ways with profound implications for our future. As we transition from a benign sense of the environment, to a sensing environment, one which now monitors us as well as our individual and collective patterns of activities and connections, the sense of control has never been more intimate, our sense of exposure never been more public.

Under a 24 hour surveillance, how are our rights to express our cultural diversity effected? How does algorithmic bias entrench inequality in our planning system? How do location based social media insulate us from meeting people with different views to our own? How is our access to the public domain protected?

3.1.6 Disciplinary lenses on technology impacts

As our second example, consider how grouping technologies by the core disciplinary field from which they derive can unhelpfully shape the sorts of effects that we search for when we set out to make predictions about the effects of those technologies. As we noted in the preamble, the task of making even short-range predictions is challenging.

When we shift to attempting to make longer-range predictions of consequences or effects of new technologies that manifest themselves in the social domain, the task of prediction will undoubtedly become even harder. For instance, effects like people becoming more competitive due to using smart drugs, or holding each other to more demanding expectations because of instant messaging and email created through mobile phones and email technologies, or in parents raising the stakes for one another's children by editing their own children's genes to give them a better start in life, or a reduction in traffic accidents and traffic congestion as people switch across to ride-sharing due to the introduction of autonomous vehicles. Predicting such effects will undoubtedly be more difficult than predicting the more easily imaginable short-range effects like adverse medical side effects of smart drugs, or potentially adverse effects of radiation released by mobile phones, or accidentally creating new genetic diseases or disorders, or software and/or hardware bugs and malfunctions that cause accidents.

¹⁰ <http://www.siemens.com/innovation/en/home/pictures-of-the-future/infrastructure-and-finance/smart-cities-facts-and-forecasts.html>

One of the limitations of delineating emerging technology based on the core disciplinary in industrial field from which they derive is that the impact in other disciplines of aspects of society may be overlooked. Take, for example, the case of smart drugs. Three values are prominent in the ongoing current academic and public debates about how smart drugs and other putative cognitive enhancement methods should be regulated – namely, safety, effectiveness, and equity of access. Firstly, in regards to effectiveness, concerns are raised about whether, for whom, and under what conditions smart drugs work – i.e. whether they indeed improve people’s ability to learn, their productivity, and extend their wakefulness and degree of alertness. Secondly, in regards to safety, concerns are raised about whether the use of smart drugs might have adverse medical side effects like increased blood pressure, cause seizures and overdoses, and whether users might become addicted to them. Thirdly, in regards to equity, almost unilaterally participants in this debate express concern that if smart drugs that are effective and safe became available, then they should be made as inexpensive as possible to make sure that they are available to everyone equally rather than, for instance, to being available to those who can afford them, to make sure that this does not increase social inequality further.

Now, from one perspective, all of these concerns are extremely well-grounded and pertinent. However, what is striking is that the notions of ‘safety’ and ‘effectiveness’ are construed in an extremely narrow way. Regarding safety, only the potential for adverse medical side effects is considered, but their potential to have adverse social side effects like the ones discussed in the preamble are not even viewed as a safety consideration. And regarding effectiveness, only the potential to improve competitively valuable traits is considered, but yet other important human qualities like honesty or compassion are simply left unmentioned. Likewise, although equality is usually a very important factor, providing effective and safe (in the respective narrow senses just described) smart drugs to everyone *equally* is precisely what, in this case, would paradoxically create the problem of an even more competitive and work-obsessed society. Our point here is just that an overly-narrow understanding of notions like safety and effectiveness, and of the significance of equality, is likely to result from a compartmentalised approach. To even see that safety in regards to medications might sometimes involve potentially adverse social side effects (or the respective understandings of ‘effectiveness’ and ‘equality’), we need to adopt a non-compartmentalised approach of the sort that is typical to transdisciplinarity.

We revisited these examples to demonstrate the need for greater nuance and complexity *within* the selection of technologies, predicting their effects, evaluating those effects, and taking appropriate measures. Technologies function as composites not as discrete entities. Bricks and mortar are not novel, but their novel arrangements can be used to intentionally shape human behaviour. Often the most important effects of a new technology are on the way it alters social relations by subtly changing incentive structures and subsequently how people behave in a competitive environment. The more we focus on predicting risks and developing strategies to mitigate them, the more we are likely to overlook what trade-offs in freedom we make to mitigate those risks and secure greater safety.

3.1.7 Broad categories

Similar issues arise with respect to broadness of categories like ‘*biotechnologies*’ and ‘*neurotechnologies*’.

When CRISPR Cas-9 gene editing technology is combined in a clinical setting with genetic screening technologies for the purpose of screening for- and editing the genes of embryos afflicted by genetic diseases, disorders, and susceptibilities (e.g. to depression, addiction, or just a weaker immune system), what we get is a distinct and specific hybrid technology that raises specific issues such as eugenics, mutual social coercion, and normalisation. In regards

to eugenics, if gene screening could be used to detect autism, intellectual disabilities, congenital deafness or blindness, or gender incongruence, should prospective parents be allowed (or encouraged, or discouraged, or expected) to use gene editing technologies to ensure that their children are not born with such conditions? In regards to mutual social coercion, if some parents use such technologies to ensure that their children do not have such conditions – or, perhaps, just to ensure that they do not have a disposition to develop depression, to become addicted, or to have a weak immune system – then might that result in other parents feeling pressured to use genetic screening and editing technologies to ensure that their children do not suffer positional disadvantage? And in regards to normalisation, if prospective parents are permitted (or perhaps encouraged or even required) to use these technologies to eradicate such medical conditions, then will that not result in the eventual disappearance of these genetic variations and a subsequent convergence or narrowing-in on a shared pool of genotypes and phenotypes? Concerns are clearly also raised about human dignity and disability rights, given that such hybrid technologies could be used to identify and edit out genetically dependent congenital disabilities. The potential of these two technologies to be used in ways that have serious human rights implications is clear, but while both might be categorised as biotechnologies, only the gene editing part of this hybrid technology is new.

Similarly, within the category of ‘neurotechnologies’ we find: *brain scanning* (i.e. diagnostic) techniques; *brain modification* (i.e. intervention) techniques; techniques that employ *neuro-interventions* as a neuro-diagnostic tool (i.e. another hybrid technology, for instance, providing a patient with a medication and watching their response to gather information about how their brain operates); as well as complex *data analysis* and *prediction techniques* (which could either be categorised as mathematical techniques, big data techniques, or prediction techniques). A range of these techniques – some of which are new, others not – are currently being considered as candidates for use within the criminal justice setting. For instance, there is a strong push in the discipline of neurolaw to medicalise the reform of criminal offenders *via* brain-based interventions referred to as ‘moral enhancement’ that are said to improve offenders’ moral judgment and self-control. However, this raises serious human rights concerns. For instance, the prospect of replacing punishment with a treatment that is administered at the discretion of medical technocrats rather than the criminal justice system, and that an offender’s personality may be altered – effectively treating them like a broken toy to be mended at our discretion – presents serious worries that are the focus of current scientific investigation and legal and ethical debate. The problem with the broadness of the category of ‘neurotechnology’ is that only some of the neurotechnologies that comprise the hybrid technologies that are proposed for use in the criminal justice setting are new, which again creates a problem for deciding whether this warrants including them in the AHRC’s investigation, if only human rights implications of *new* technologies should be considered.

To create a space in which precise issues can be identified, investigated, and framed in the right way and with the right degree of nuance and precision, we believe it is important to sometimes look at finer grained categories of technology.

3.1.8 Particular technologies, contexts, and purposes

A fine-grained and nuanced particularism about technology emerges from the foregoing discussion – i.e. the need to focus on particular technologies embedded within specific contexts and used for specific purposes. This particularism, rather than the reliance on a fixed list of broad types or categories of technologies, may in our view be one of the most important outcomes of the AHRC’s initiative on human rights and technology.

For the range of reasons noted above, we worry that it is unhelpful to work with a fixed list of broad categories of technologies or, equally, to focus on a narrow/singled-out technology in isolation from its context and its connections to other technologies and to human behaviour.

Instead, to properly identify and characterise salient issues, we believe it is necessary to focus on concrete examples. That is, examples of *particular technologies* (or combinations thereof) used in *specific contexts* for *determinate purposes*.

Consider the vastly different implications that a concrete technology like *artificial wombs* raises when it is used in different contexts. On the one hand, if a woman's life was threatened by continuing a pregnancy, then the potential to transplant a foetus to an artificial womb instead of having to abort it would happily resolve what is otherwise a morally difficult situation. In a way, in this context the new technology creates a new morally cleaner option by enabling us to avoid making a choice between two undesirable options – to end the life of a developing human, or to endanger the life or health of its mother. On the other hand, if a developing foetus could always be transplanted into an artificial womb, then how might that impact women's rights to have abortions, and might it require that we extend a right to life to foetuses? In the absence of such a technology, the viability of a foetus depends on it remaining in its mother's womb. However, the capacity to keep a foetus alive within an artificial womb might be taken by some to mean that mothers should not be permitted to abort foetuses when a viable option exists to keep it alive – namely, not to abort it and thus end its life, but to transplant it into an artificial womb. We do not intend to argue the case here either in favour of or against any position. Rather, we cite this example to demonstrate how a new technology can, in one context, resolve a difficult moral dilemma, but in another context it can create a new moral dilemma.

Also, importantly, contexts change over time and from place to place. We must therefore recognise the need to revisit and revise our stance on any given particular technology, since its impact on human rights will depend on the precise purpose for which it is being used, and on the precise context in which it is used.

3.1.9 Summary and Recommendations

In summary, although we believe that the twelve items cited by the World Economic Forum present a helpful starting point, as we explained (a) some other important items may need to be added to this list, (b) novel combinations of new and existing technologies should also be considered, (c) as should new uses of existing technologies in different contexts for different purposes, and we should also consider (d) interactions between technologies, humans, regulations, laws, and markets mechanisms, as well as (e) emergent effects produced by these interactions.

A more nuanced approach to technology is therefore needed – one which recognizes that effects of technologies are in fact effects of systems in which combinations of technologies both new and old, as well as human and contextual factors, generate the effects. No effects are ever just the product of any one specific isolated factor, since all effects are always produced by a combination of many different contributing factors. Failure to notice this will result not only in problematic analysis of the causes of human rights violations, but subsequently also in sub-optimal recommendations regarding how or where to intervene in order to protect human rights.

Recommendations:

- 1.1 **Consideration of the impacts of emerging technology on human rights should consider the specific impact on individuals and communities as well as broader impacts on society and values**
- 1.2 **A nuanced approach to technology should be adopted which recognises that effects of technologies are in fact effects on systems in which combinations of technologies both new and old, as well as human and contextual factors, generate the effects.**

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

New technologies, and new uses of technologies, have the potential to significantly expand and exacerbate social inequality. The large-scale unconstrained use of data analytics and automated decision making can increase social sorting and surveillance, and entrench biases. Moreover, unequal access to new technologies, which has already led to a persistent 'digital divide', can further disadvantage vulnerable and at-risk populations. To ensure equitable access, the needs of diverse social and cultural groups should be taken into account in technology design.

This question has been addressed in a series of case studies in Section 4 below which aim to address some key aspects of the lived experience and complexity of new technology. These case studies have been written by researchers in relevant disciplines to provide an insight into the benefit of combining both deep disciplinary analysis with a broader transdisciplinary vision and understanding.

The following case studies are provided

- AI and data analytics in education: ethics issues (extended case study)
- AI and data analytics in the disability sector: opportunities and ethical issues
- Health, AI and intellectual disability
- AI and Indigenous data: managing data ethically

The first of these draws on expertise at UTS in the application of Artificial Intelligence & Data Analytics (AIDA) to educational contexts. In it attention is drawn to the potential of AIDA to support learning, and some risks in this space. The second draws particular attention to the potential of AIDA for supporting independence and autonomy in the lives of people living with disability, including the ways that AIDA might break down entrenched physical, social, financial and political barriers. The third focuses on the general category of intellectual disabilities, and similarly differential outcomes that might be addressed – or exacerbated – by AIDA. The final case study draws attention to opportunities for Indigenous Australian peoples to engage with AIDA to support the realisation of self-determination, including opportunities relating to equitable outcomes in health and education, as well as broader cultural implications. Across these case studies key recommendations are highlighted. At the high level these recommendations can be summarised.

Recommendations:

- 2.1 A broad range of stakeholders should be involved in understanding the impacts of technology development and deployment across contexts
- 2.2 Accessibility of technologies and their uses must be a core consideration in their development and deployment, including physical, cultural, socio-economic educational and other barriers to access
- 2.3 There should be clear recognition of the broad range of stakeholders

impacted by technologies, and the ways in which they may be impacted both directly by the technology (hard impacts) and through more indirect means (soft impacts)

2.4 Positive outcomes for stakeholder groups should be an explicit aim in developing and deploying technologies

3.3 How should Australian law protect human rights in the development, use and application of new technologies?

Australian law should protect human rights in the development, use and application of new technologies by a combination of the human rights approach with the transdisciplinary approach recommended in this submission. Our response to this question should be read together with our responses to Questions 5 to 7, which deal specifically with the challenges of applying a human rights approach to artificial intelligence technologies. As this submission explains, we adopt a broad understanding to what we regard as new technologies, as well as a broad approach to human rights.

This submission considers that a human rights approach is the preferred framework for addressing the considerable challenges associated with new technologies. While a human rights approach is not perfect, it has considerable advantages over possible alternative approaches, including current initiatives aimed at developing ethical principles or guidelines for new technologies.

Human rights approaches to law and regulation are aimed at ‘turning human rights from purely legal instruments into effective policies, practices, and practical realities’.¹¹ Comprehensive human rights approaches focus on both the values underpinning human rights, as well as the substantive content of specific rights; and incorporate principles that relate to both substantive and procedural rights.¹² Such approaches have advantages over alternatives, such as ‘ethical’ approaches, in that: they take a holistic approach to the protection of human rights, ranging from prevention of infringements of rights to remedies; they are based on relatively well-accepted international laws, standards and norms; and may be enforceable.¹³ Further, as the Office of the United Nations High Commissioner for Human Rights (OHCHR) puts it:

A programme guided by a human rights-based approach takes a holistic view of its environment, considering the family, the community, civil society, local and national authorities. It considers the social, political and legal framework that determines the relationship between those institutions, and the resulting claims, duties and accountabilities. A human rights-based approach lifts sectoral ‘blinkers’ and facilitates an integrated response to multifaceted development problems.¹⁴

11 Australian Human Rights Commission, “Human Rights Based Approaches,” accessed October 22, 2018, <https://www.humanrights.gov.au/human-rights-based-approaches>.

12 The Human Rights, Big Data, and Technology Project (HRBDT), “The Human Rights, Big Data and Technology Project – Written Evidence (AIC0196), Submission to the House of Lords Select Committee on Artificial Intelligence,” 2017, <http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69717.html>.

13 The Human Rights, Big Data, and Technology Project (HRBDT).

14 Office of the United Nations High Commissioner for Human Rights (OHCHR), “Frequently Asked Questions on a Human Rights-Based Approach to Development Cooperation” (United Nations, 2006), 17, <https://www.ohchr.org/Documents/Publications/FAQen.pdf>.

A human rights approach to law and regulation, which is both holistic and proactive, is therefore well-matched to the transdisciplinary approach applied in this submission, which we suggest should supplement the human rights approach. As we explain in this submission, the transdisciplinary approach is needed to supplement the human rights approach as it allows for the identification of issues that may not be immediately obvious, such as the implications of interactions between different technologies and potential unintended effects of technologies, as well as the emergence of new vulnerable groups, and the reframing of rights discussed in our response to Question 1. There is, however, a pressing need for further research in elaborating on how the combination of human rights and transdisciplinary approaches might be brought to bear in developing concrete legal and regulatory regimes for complex new technologies. In addition, given the current emphasis on the development of ethical principles or guidelines for new technologies, as the UK Human Rights, Big Data and Technology Project has put it, there is a ‘need for further research into the nexus between human rights and ethics in the context of the digital age, focusing on potential areas of overlap that may lack clarity and/or produce tensions due to differing approaches’.¹⁵

In the following, we elaborate on the application of human rights and transdisciplinary approaches to new technologies in answering the specific sub-parts to this question.

3.3.1 In particular: What gaps, if any, are there in this area of Australian law?

The Australian legal system protects human rights by a variety of mechanisms, such as incorporation of rights in specific legislation, common law protections, and government authorities or agencies responsible with protecting or promoting human rights, such as the Australian Human Rights Commission (AHRC). Especially in the absence of an enforceable constitutional or statutory bill of rights, however, the Australian legal framework for protecting human rights is fragmented and undeveloped, leading to clear gaps in the law. For example, under the general law, Australia does not recognise a legal right to privacy, with protections being confined to a patchwork of specific legislation, as well as common law actions that are not specifically directed at the protection of privacy. Moreover, although the High Court has recognised an implied constitutional protection of freedom of expression, this is much more limited than a fully-fledged right to freedom of expression recognised in a bill of rights as, for example, it is confined to political speech. The limited protection of rights to privacy and freedom of expression under Australian law pose considerable challenges for the protection of human rights in the context of new technologies. The strengthening of the overall framework for the protection of human rights under Australian law would therefore go some way towards better promoting and protecting human rights in the development, use and application of new technologies.

The broad scope of new technologies, and the deficiencies of the Australian legal system in protecting human rights, are so extensive that it is impossible to be comprehensive or definitive about the gaps in legal protection in this area. Our response to this sub-question therefore focuses on identifying some specific areas where there are gaps in the law, which should be taken as illustrative of more general deficiencies.

- *Information privacy laws.* New technologies, such as the Internet of Things (IoT) and data analytics allow for the mass collection, processing and matching of data, including personal information. Australia’s information privacy laws, including the Privacy Act 1988 (Cth), were not drafted with these technologies in mind, and are

¹⁵ The Human Rights, Big Data, and Technology Project (HRBDT), “The Human Rights, Big Data and Technology Project – Written Evidence (AIC0196), Submission to the House of Lords Select Committee on Artificial Intelligence.”

ill-suited to the protection of privacy in the face of, for example, big data practices and algorithmic decision making. For instance, big data practices may rely on correlations across disparate data sets in order to draw inferences and, potentially, make predictions. As it is impossible, at the time that personal information is collected, to know how that data will be used, or what inferences or predictions may be drawn, it is difficult to apply the purpose specification principle, which requires the purpose for which personal data will be used to be specified and notified at the time of collection, to big data practices. Furthermore, one of the purported advantages of distributed ledger (or, colloquially, blockchain) technologies is the ‘immutability’ of data stored on the ledger. If personal information is stored, however, this effectively frustrates the core principle of data privacy law that a data subject has the right to correct or delete incorrect or incomplete personal information. Similarly, Australian privacy law does not clearly give a right to people to apply for the removal of search engine links to personal information that is incorrect or incomplete, which is available to citizens or residents of European Union states. All of this suggests that, given the state of development of significant new technologies, especially data-based technologies, there is a need for a fundamental review of Australian privacy laws to ensure they remain fit for purpose.

- *Liability issues.* A number of new technologies, such as autonomous vehicles, cryptocurrencies and 3D printing, involve complex interactions between hardware providers, software providers, service providers and users. It is therefore difficult to determine who might be held liable for harms, including breaches of human rights, which, in the absence of clear rules may threaten human rights, especially the right to an effective remedy.¹⁶ There is therefore a need for a review of cross-sectoral laws with a view to clarifying rules relating to liability for human rights abuses in relation to complex new technologies.
- *Technology auditing and assessment.* Although there are a wide range of standard setting processes, Australia does not have a comprehensive cross-sectoral system for the auditing or assessment of new technologies with potentially significant social effects. A major recommendation of this submission is the establishment of a new regulatory body, the Technology Assessment Office (TAO), and associated processes, to address this gap in the Australian legal framework for new technologies.

As we explain in this submission, the proposed new body, the TAO, should have responsibility for undertaking iterative reviews of legal and regulatory frameworks to ensure that they remain adequate and appropriate for the protection of human rights in the face of complex new technologies. Given the proposed broad remit of the TAO, assessments of regulations and technologies would be able to be undertaken across different technologies and different sectors.

3.3.2 In particular: What can we learn about the need for regulating new technologies, and the options for doing so, from international human rights law and the experiences of other countries?

Given the broad range of technologies, and the broad range of human rights, implicated by the AHRC inquiry, it is impossible to be definitive about how human rights approaches are being used in other countries to protect human rights in regulating new technologies. In our response to this question, therefore, we simply provide the following diverse examples which might be drawn upon in developing an Australian approach to regulating new technologies.

¹⁶ See, for example, European Commission, “European Commission Staff Working Document: Liability for Emerging Digital Technologies” (European Commission, 2018), <https://ec.europa.eu/digital-single-market/en/news/european-commission-staff-working-document-liability-emerging-digital-technologies>.

- In May 2018, the European Commission issued a communication on autonomous vehicles, with a view to developing an EU strategy for driverless vehicles.¹⁷ The communication stated that '[a]utomated vehicles will have to be safe, respect human dignity and personal freedom of choice'; and pointed to the establishment of an EU task force on ethical aspects of automated and connected driving.¹⁸
- In 2011, the United Nations Human Rights Council (UNHRC) adopted the United Nations Guiding Principles on Human Rights (UNGPs), which establish a global framework for addressing the risk of adverse impacts of business activity on human rights.¹⁹ The UNGPs imposes moral obligations on businesses in relation to socially responsible innovation. In the EU, this is being implemented by the European Commission's Strategy on Corporate Social Responsibility²⁰. Given the role of businesses in developing and applying new technologies, the UNGPs establish an important framework for developing principles relating to socially responsible innovation.
- In 2016, UNESCO published a comprehensive report on human rights and encryption, which focused on how encryption can support freedom of expression, anonymity, access to information, private communication, and privacy; and the importance of subjecting any limits on encryption to careful scrutiny²¹. The report indicated UNESCO's support for the 'ROAM principles', which refers to a '(human) Rights-based, Open and Accessible Internet that is governed by Multi-stakeholder participation²².
- In 2017, the Council of Europe adopted Recommendation 2115 on *The use of new genetic technologies in human beings* which, amongst other things, recommended the development of a 'common regulatory and legal framework which is able to balance the potential benefits and risks of [genetic] technologies aiming to treat serious diseases, while preventing abuse or adverse effects of genetic technology on human beings'.²³

Although the above developments and documents do not necessarily reflect the application of principles of international human rights law in national regulatory regimes, they do illustrate the scope of the legal and regulatory challenges posed by emerging new technologies, and the complexities entailed in developing human rights responses to regulating the

17 European Commission, "On the Road to Automated Mobility: An EU Strategy for Mobility of the Future," Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (European Commission, 2018), https://ec.europa.eu/transport/sites/transport/files/3rd-mobility-pack/com20180283_en.pdf.

18 European Commission, 16.

19 UN Office of the High Commissioner for Human Rights, "United Nations Guiding Principles on Human Rights" (United Nations, 2011), https://ec.europa.eu/transport/sites/transport/files/3rd-mobility-pack/com20180283_en.pdf.

20 See, for example, European Commission, "A Renewed EU Strategy 2011-14 for Corporate Social Responsibility," Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (European Commission, 2011), https://www.eurocommerce.eu/media/7237/position-csr-renewed_csr_strategy_2011-14-07.03.2012.pdf.

21 Wolfgang Schutz and Joris van Hoboken, "Human Rights and Encryption; UNESCO Series on Internet Freedom" (UNESCO, n.d.), <http://unesdoc.unesco.org/images/0024/002465/246527E.pdf>.

22 This is associated with the UNESCO MAPPING project (Mapping Alternatives for Privacy, Property, and Internet Governance)

23 Council of Europe, "Recommendation 2115: The Use of New Genetic Technologies in Human Beings," 2017.

technologies. In addition, the diverse developments indicate the importance, as emphasised in the transdisciplinary approach applied in this submission, of being sensitive to the particular issues associated with specific technologies. For example, the wider range of soft impacts discussed in our response to Question 1 with respect to genetic technologies. Finally, the selected examples illustrate the ongoing importance of Australia remaining abreast of cross-jurisdictional initiatives and developments in applying human rights approaches to diverse technologies, which could be a significant function of the proposed new body, the TAO. Further specific examples of what can be learned from the application of human rights approaches to new technologies in other countries are included in our response to Question 6, which deals with AI technologies.

3.3.3 In particular: What principles should guide regulation in this area?

Human rights approaches to regulation, which are aimed at translating abstract rights into concrete policies, practices and practical realities, apply a common set of principles known as the PANEL principles, which stand for Participation, Accountability, Non-discrimination and equality, Empowerment and Legality.²⁴

In this submission, we consider that the PANEL principles provide a useful high-level framework for the regulation of new technologies but may, where necessary, need to be supplemented by a transdisciplinary perspective. In accordance with our transdisciplinary approach, however, the PANEL principles need to be elaborated upon so that they are adapted to apply to specific technologies. In our response to Question 6, which addresses the regulation of AI technologies, we explain the need for the development of mid-level principles, based on international human rights, which adapt and apply the PANEL principles to particular technologies. In our response to that question, we also point out that the more detailed principles may be used for guidance of other forms of regulation, potentially including legislation. In other words, the development of mid-level principles, designed to apply to specific technologies, can be a first step in the development of more comprehensive and holistic regulatory responses to new technologies. While we acknowledge that elaborating regulatory principles that apply to the range of technologies identified in this submission is demanding and may be time-consuming, especially if an inclusive approach is adopted to developing the principles, we consider that it is a necessary stage in the development of adequate regulatory responses to significant new technologies.

In our response to Question 7, and throughout this submission, we explain how an adequate legal and regulatory response to new technologies requires a spectrum of regulatory responses, which can to an extent be guided by Braithwaite's regulatory pyramid, ranging from education through to legal sanctions, potentially including criminal sanctions. For example, in dealing with controversial and complex areas, such as the regulation of gene editing technologies or smart drugs, it seems important for proper regulatory design to take into account the full spectrum of regulatory options, including community information and education, rather than relying entirely on relatively blunt tools, such as prohibition or criminal law.

In our response to Question 7, we further identify the importance of encouraging potentially new forms of regulation, such as adaptive or anticipatory regulation, provided that such approaches remain grounded in protecting human rights, while not disproportionately inhibiting technological innovation. If properly implemented, the result of such an approach should be the development of responsible innovation, which remains firmly grounded in the protection and promotion of human rights.

24 Australian Human Rights Commission, "Human Rights Based Approaches."

As part of the holistic regulatory framework recommended in our response to Question 7, we emphasise the importance of adopting proactive forms of ex ante regulation, especially those based on ‘regulation by design’, including Value Sensitive Design and ‘human rights by design’. As technological developments in the areas identified in this submission are so fast-paced, ex post regulations and laws invariably struggle to keep up with technological change. Proactive forms of regulation by design, however, have the potential to prevent human rights harms before they occur.

In itself, however, Value Sensitive Design or human rights by design is not a panacea for human rights harms that may arise from new technologies. To succeed, these approaches must be properly supported, including by inclusive approaches to technology design which engage with a diverse range of perspectives, including those of vulnerable and at risk groups. Moreover, applying design-based approaches to complex technological challenges, such as those emerging in the area of neuroscience, requires adequate regulatory resources for research and assessment activities.

Recommendations:

- 3.1** Research must be conducted to elaborate how human rights and transdisciplinary approaches can be brought to bear in developing concrete legal and regulatory regimes that supplement the PANEL principles (Participation, Accountability, Non-discrimination and equality, Empowerment and Legality) for complex new technologies, particularly in understanding the ‘nexus’ between human rights and ethics.
- 3.2** A fundamental review of Australian privacy laws to ensure they remain fit for purpose
- 3.3** A need for review of cross-sectoral laws with a view to clarifying rules relating to liability for human rights abuses in relation to complex new technologies
- 3.4** The establishment of a new regulatory body, the Technology Assessment Office (TAO) and associated processes, to address the gap in the Australian legal framework for new technologies (see Section 5)

3.4 In addition to legislation, how should the Australian Government, the private sector and others protect and promote human rights in the development of new technology?

The challenges and opportunities arising from new technologies are such that legislative responses, while sometimes necessary, can never be sufficient. In our view, human rights must be promoted through a combination of ‘soft responses’, such as education, standards and self-regulatory codes, and ‘hard responses’, including appropriate legislation which establishes enforceable rights. Over and above this, following from our transdisciplinary approach, we emphasise the importance of understanding and regulating technologies in a holistic manner, taking into account interconnections between issues and technologies. We therefore consider that there may be considerable advantages in establishing a cross-sector body for promoting understanding, education and dialogue on the social, ethical, and legal implications of new technologies. A major component of any potential regulatory response should incorporate requirements for monitoring and auditing new technologies with potentially significant social implications. Importantly, it is vital to understand that technologies are not ‘neutral’, but have built-in values, which should be demystified and

subject to critical analysis. This suggests that rather than relying exclusively on ex post regulatory responses, there is a pressing need for human rights considerations to be taken into account at the design stage of technologies.

3.4.1 Distributed and Shared Agency

One of the key enablers for protecting and promoting human rights in times of rapid technological change is a well-informed public who see themselves as having a degree of shared agency in decision making around the use of technology in society, rather than being passive recipients of inevitable technological change.

A proactive approach is needed in order to develop this sense of distributed or shared agency with specific programs and targets for education and engagement across sectors and within the broader public sphere. The level of economic and political power leveraged by large technology companies, locally and globally, requires a civic counterweight empowering citizens to better understand the potential impacts of new technologies and provide avenues for influencing dialogue and change in this arena.

Many technology development environments lack the input of broader perspectives outside of technological expertise present in the development process (other than positioning individuals as ‘consumers’ or ‘users’ of technologies). We propose, again, that the field of emerging technologies needs to be ‘complexified’ through **engagement with a multiplicity of disciplines and perspectives** in order to better understand the impact of technologies on all those implicated in their development and use (e.g. decision-makers, developers and those affected by technologies).

3.4.2 Emerging technology as a complex ecosystem

Discourse about technology is often oversimplified. In this way, responsibility for technology’s effects is sometimes attributed wholly to technology and its developers, or wholly to its users. Thus, in our response we wish to move away from approaches that tend to place the responsibility over technological impacts upon a single player, either upon technology developers or the end users – the two opposing positions that can be respectively categorised as ‘determinism’ and ‘instrumentalism’. Technological determinists assume that users behave in accordance with technological dictates²⁵ whereas instrumentalists ‘downplay the power of technology, believing tools to be neutral artefacts, entirely subservient to the conscious wishes of their users’.²⁶ Seeking a position in between these extremes we wish to conceptualise the emerging technology field as a complex ecosystem of actors that act and are acted upon by each other through complex interrelationships. Thinking about emerging technologies ecologically allows us to identify which players in the ecosystem are best placed to protect and promote aspects of human rights, moving away from construing the sole of purpose of regulation as reigning in technological development.

One model for understanding the influence and interaction between technology and society is Lessig’s framework of four regulatory ‘modalities’: the law; social norms; the market; and the ‘code’ or technological architectures. These four regulatory modalities circumscribe technologically mediated behaviours.²⁷ We argue that technological developments are mostly considered in terms of market (which is often linked with an instrumentalist position) and

²⁵ Merritt Roe Smith and Leo Marx, *Does Technology Drive History?: The Dilemma of Technological Determinism* (MIT Press, 1994).

²⁶ Nicholas Carr, *The Shallows: What the Internet Is Doing to Our Brains* (WW Norton & Company, 2011).

²⁷ Lawrence Lessig, *Code: And Other Laws of Cyberspace* (ReadHowYouWant.com, 2009).

the technological architectures modalities, with attempts to curb undesirable developments through legislation (which often ties in with a determinist view). The broadly defined domain of social norms (which can include societal and professional norms, as well as commonly agreed roles and responsibilities) needs to be explored so that citizens (or consumers) as well as decision-makers can realise their agency in promoting and protecting human rights in technologically rich contexts of their lives. In other words, we suggest that protecting and promoting human rights in the development of new technology can benefit from broad consideration of not just technology and users, but a more holistic consideration of the role that can be played by social norms and the market, as a supplement to the role that can be played by the law and the coding/technology itself.

3.4.3 Engaging with uncertainty

There is a need for society to engage with uncertainty as a persistent and, as such, unexceptional feature of contemporary life in an increasingly complex world. Uncertainty is frequently framed as a condition to be avoided or in terms of averting or mitigating risks, due to our inability to accurately predict future outcomes of technological developments. Whereas policy-driven approaches can contribute to protecting and promoting human rights at the level where outcomes can be anticipated, we also need to begin to engage with approaches that allow us to collectively, creatively and productively explore aspects of uncertain future outcomes of emerging technologies. These may include trajectories to be avoided, but also previously unimagined opportunities for human flourishing.

Due to uncertainty and the complexity of interconnected issues (e.g. uncertainty in how systems utilising machine learning might evolve) there is a need for an ongoing iterative process of evaluation and decision-making. Public dialogue about emerging technologies and their impact on human rights needs to be an ongoing engagement, not a one off occurrence. Such dialogue also has to be future-oriented, and therefore, not only focus on critical analysis of the past or extrapolation of contemporary trends, but also to stimulate imagination about future possibilities.

Continuing with the idea of the importance of distributed or shared agency, it is clear that many issues arising through the use of technologies are not possible to address by any player in isolation. Returning to the Lessig's notion of social norms, society's views on what matters and what is of concern evolve and change all the time²⁸. Whilst technological or scientific disciplines are historically lacking in methodologies for understanding social contexts, socio-political theory, humanities and creative disciplines (as well as science fiction) have many tools and approaches that can help us explore the uncertain implications of technologies and ethical dilemmas in social contexts. Thus, we propose that as part of the public development of emerging technologies that protects and promotes human rights we must utilise transdisciplinary and creative approaches to imagining futures that can provide us with vivid, complex and empathetic understandings of future situations.

3.4.4 Fostering dialogue in a neutral environment

There is a need for public ethical assessment of emerging technologies by all those implicated in technology development and deployment as well as those impacted. A formalised process of dialogue (or a cross-sector body) could be established to foster distributed or shared agency in the emerging technology field, in order to move away from the dominance of purely technological innovation agendas. Given the diversity of agendas at play that include commercial interests as well as concerns about unintended consequences or potential harm to groups of people, there is a need for a neutral place for a dialogue to take place.

²⁸ Lessig.

Universities are well-positioned to act as a nexus that brings together governments, private sector organisations, communities and the academy, without a singular agenda (other than promoting and protecting human rights) to host such dialogue. Conversations held in a participatory format could then inform government legislative bodies about practices that might best support human flourishing and human rights for all. Such dialogue could also intersect with the broader public discussions through media and popular culture to build public awareness of impacts of technology. This would offer further opportunities for discussion, and contribute to the shaping of the social norms around the development and deployment of emerging technologies.

This dialogue and discussion around emerging technologies can be contextualised for specific sectors such as health or education as well as considering cross-sectoral and broader societal implications. Opportunities for intergenerational dialogue and bringing in groups who may lack opportunity to engage with public dialogue due to their age, socio-economic circumstances or location in remote/rural communities are important considerations.

3.4.5 The imperative of education

Society has a responsibility to equip its citizens with the new literacy that is required to live and work with emerging technologies. Education is one of the most important ways that society ensures continuity of human knowledge, cultures, values and social norms.

Education can be offered through formal qualifications or continuous professional development to current and new leaders that develop or deploy technologies that have the potential to affect human rights. If we see large technology companies as key drivers of technological development, we would argue that everyone involved in decision-making in such companies, e.g. CEOs and CTOs, be trained to consider the ethical dimension of their actions. The peak professional bodies in technology areas (e.g. Engineers Australia (EA), The Australian Computer Society (ACS), etc.) could play a role by embedding protection and promotion of human rights as a requirement they expect for professional practice in technology sectors. Ethics and human rights considerations could be introduced as a requirement for accreditation or be included in professional codes of ethics.

However, continuing the theme of complexity of the ways that technology is embedded into our daily lives, it is clear that individuals in government decision-making positions or those handling procurement in a variety of positions in both public and private sectors must also learn to make well-informed and ethical decisions. Therefore, there is a need to develop the capacity to understand the technology's impacts on human rights and make ethical decisions that affect others more broadly within the society. This means that a wide range of educational programs across a variety of professions and disciplines must include the development of the basic emerging technologies literacies (including ethics and human rights) as a necessary requirement.

We must teach the next generation how to ask the right questions about how AIDA functions, from school age, to citizens at large. Scientists, scholars, policymakers and business analysts will increasingly sense the world through the (always distorting) lenses of computational models consuming (partial) data feeds from (imperfect) sensors. They must be empowered to question automated recommendations, and reflect on the importance of seeing and acting with knowledge and integrity.

School education and teacher education, in particular, are positioned in a crucial role to influence how society reproduces or re-invents itself. Since the majority of the population goes through the formal schooling process, school educators have the ability to sensitise those entering adult life, professional world or university study to the importance of human rights in technological contexts. The changing nature of literacy and what it means to be a

literate citizen today must be explored in the school curriculum. This may include, for example, media literacy on privacy, but also ethical and technology literacies that would allow young people to form a stance on issues as part of a community and citizens.

As young people and adults learn to partner with intelligent agents (in their learning, and in life), and whether the AI is an assistant or encased as a robot or less anthropomorphically, it will be critical to learn a new set of 'interpersonal skills', analogous to those we cultivate for people, e.g. to judge an agent's areas of expertise, trustworthiness, social and emotional awareness, and other attributes. This is how we calibrate our interactions with others: what and how we choose to share, and how we interpret others' actions and advice.

3.4.6 Developing ethical and philosophical models

Finally, we also argue for the need for various players in society, including those involved in technological development, to build their capacity to examine new emerging developments using a variety of ethical and philosophical frameworks. In particular, we would like to advocate for the broad understanding and use of principles based approaches founded on notions of respect for human dignity, approaches which provide the foundation of human rights, rather than the prevailing consequentialist, or utilitarian, approaches, which seek to weigh up benefits and harms, and tend to be favoured by policy makers and legislators seeking neat prescriptive solutions and desirable outcomes. Further, we suggest that principles based approaches, with broad prescriptions to promote and protect dignity, fairness, privacy and more, are well placed to respond to the dynamic field of emerging technologies, as they can adapt and expand as needed.

3.4.7 Protecting and promoting human rights by technology design

The majority of implications on human rights come to our attention due to technology misuse. There is a tendency for commercial Research & Development (R&D) environments to privilege the technological (and commercial) aspects of innovation, whereby technologies are evaluated on their technical robustness, user acceptability and commercial viability before entering the market with little consideration for the broader effects these technologies might have in the contexts of their use. Many technologies are deployed directly from R&D environments into the world with no consideration for the ethical implications of their implementation in the society or their potential effects on human rights. Furthermore, due to the complexity of ways technologies are entwined with the fabric of society, many large technology companies end up carrying out large-scale deployment of technologies that have only been tested in labs with limited ability to anticipate their effects in complex live contexts. This is exemplified by the cases where cities or councils collaborate with large technology companies (e.g. Sidewalk Labs, a sister company of Google, being contracted to develop a 'smart' neighbourhood in Toronto) under a promise of innovative, data-driven 'smart' services without the ability to present a real case or proof of the benefits that particular technological interventions might bring beyond the reputational advantage to the community.

An alternative to punishing misuse where harm is discovered through monitoring or auditing the outcomes of technology is building in considerations for human rights by design. An emerging field of media and communication research, known as 'platform studies', offers a way to understand the 'affordances' of various technologies to affect human behaviours. An affordance is what a technology allows its users to do, which highlights the fact that technologies are not neutral, but have built-in values and judgements. Consideration for human rights, such as privacy, thus, could also built-in into technologies by design. This requires those building technologies as well as making decisions about their deployment to have the ability to apply ethical frameworks or methodologies to understand the social implications of their creations. This could be achieved by either requiring those practicing

in technology industries to have training in ethical judgement or by encouraging technology development teams to include ethicists, formally educated to engage with interdisciplinary teams and examine the impacts of technology deployment in society.

Recommendation:

4.1 Human rights considerations should be taken into account at the design stage of technologies. In order to do this, we outline a set of guiding transdisciplinary principles, for: Distributed and shared agencies; understanding emerging technology as a complex ecosystem; engaging with uncertainty; fostering dialogue in a neutral environment; the imperative of education; and developing ethical and philosophical models, elaborated on in Section 5.

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

This question, together with the following two questions, deal with issues involving a specific area of technology, AI-informed decision making (AIDM). As our responses to these three questions indicate, AIDM raises distinct human rights issues. However, in accordance with our transdisciplinary approach it is inadvisable to segregate the implications of one particular form of technology from other, potentially interacting, technologies.

3.5.1 AI-Informed Decision Making (AIDM)

AI-Informed Decision Making (AIDM) is not one technology, but involves the use of various technologies and data sources to perform a range of tasks with varying degrees of automation in diverse contexts.

In this submission we therefore adopt a broad understanding of AIDM, so as to ensure that threats to human rights and their sources are not overlooked, excluded, or misidentified. The main features of our broad approach are as follows:

- AIDM includes both *data processing* and *data collection* technologies;
- AIDM employs a very *broad range (or cluster) of technologies*;
- *the degree of automation* in AIDM ranges from *full autonomy*, to *human-supervised systems*, and systems that merely *provide advice* to humans;
- consequently, *humans are completely or partially responsible* as to that AIDM poses to human rights; and
- because *the range of applications for AIDM is incredibly broad*, so too is the range of contexts in which AIDM poses threats to human rights.

3.5.2 Human rights implications of AIDM

Although much of the discussion about the social and political implications of AIDM has been framed by initiatives for developing ethical standards for AI²⁹, there is growing interest in the application of human rights based approaches. Apart from the AHRC HRT Issues Paper, in March 2018, the Council of Europe released a study on Algorithms and Human Rights, which

29 AHRC, "Human Rights and Technology Issues Paper (2018)," 17.

was prepared by its committee of experts on internet intermediaries.³⁰ Even within some of the initiatives aimed at developing ethical frameworks, the importance of protecting human rights has been recognised with, for example, the Institute for Electrical and Electronic Engineers (IEEE) in its 2017 report on ethically aligned design for AI acknowledging the fundamental principle that AI should not infringe international human rights.³¹ Furthermore, some attention has been given to developing principles for applying human rights to AI, notably the May 2018 Toronto Declaration, drafted by Amnesty International and Access Now, which focuses on the implications of machine learning for the right to equality and non-discrimination.³² That said, much remains to be done in mapping the full effects of AIDM for human rights, and developing human rights based principles for regulating AI.

Like other multi-use technologies, AIDM has both negative and positive implications for human rights.

AIDM poses specific challenges to human rights due to:

- potential bias in data and algorithms;
- opacity, leading to difficulties in ensuring transparency and accountability;
- the ability of AI to process, analyse and match data that produces or reveals personal information; and
- the use of AI-powered bots to influence opinions and potentially elections.

These features of AIDM pose obvious challenges for particular human rights, especially the rights to equality and non-discrimination, privacy, political participation and freedom of expression, and a fair trial and fair hearing.^{33,34} Nevertheless, the effects of AIDM are so extensive that they potentially apply to a very broad range of human rights, including the rights to health, education, social security, participation in cultural life, equality before the law, and access to an effective remedy. The range of human rights affected by AIDM suggests that these technologies implicate the foundation principles underpinning international human rights law, namely autonomy and human dignity. Due to space constraints, however, this submission does not attempt to be comprehensive in its analysis of implications of AIDM for all human rights, but focuses on some of the main examples, as well as some less obvious examples.

While AIDM poses serious threats to human rights, it is also important to acknowledge its positive potential for promoting and protecting human rights, such as by the use of AI in

³⁰ Committee of experts on internet intermediaries, “Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications” (Strasbourg: Council of Europe, 2017), <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5>.

³¹ IEEE, “Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems (Version 2)” (IEEE Computer Society, 2017), https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_brochure_v2.pdf.

³² Amnesty International and Access Now, “The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems” (Access Now, May 16, 2018), <https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/>.

³³ Mark Latonero, “Governing Artificial Intelligence: Upholding Human Rights & Dignity” (New York, NY, USA: Data & Society, 2018), https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.

³⁴ AHRC, “Human Rights and Technology Issues Paper (2018),” 15–16.

improving accessibility for people with disabilities.³⁵ This submission is particularly concerned to emphasise the importance of promoting positive uses of technologies, especially in addressing difficult social problems, rather than focusing exclusively on the risks associated with new technologies. In the following section, however, in our response to the first part of Question 5, we do emphasise problems with the protection of human rights in the face of AIDM.

3.5.3 How well are human rights protected and promoted in AIDM?

The following details some of the main ways in which AIDM challenges the protection of human rights, as well as some potential effects that may be less obvious. In addition, it identifies some ways in which AIDM may be used to promote human rights.

3.5.3.1 Non-discrimination and equality

AIDM threatens to increase discrimination and inequality through biased data and/or biased algorithms. A significant amount of concern about the social implications of AIDM has focused on the bias that may be associated with machine learning AI, with its tendency to promote and reinforce discrimination and inequality. AIDM may reflect biases in the data sets that machine learning algorithms rely upon, as well as human biases of those responsible for developing algorithms. Data sets involving humans are not neutral, and will reflect historical inequalities, such as under- or over-representation of specific groups. If past human behaviour is taken as the standard from which AI is trained, whatever biases are present historically will become ingrained in the dataset from which machine learning AI learns. Moreover, the design of algorithms can also embody the programmer's implicit values, which are also a source of bias in AIDM. Bias is a particular concern where it is hidden because the AIDM is not transparent.

AIDMs can also offer new forms of protection against human bias and discrimination.

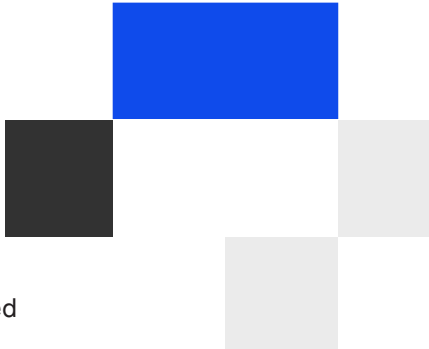

While AIDM is subject to bias, human decision making is also susceptible to bias, often unconscious. Decisions made by AI exist wholly in data form. As such, bias may be objectively discoverable in ways previously unimaginable, by examining patterns of decision making and analysing the extent to which it exhibits unwarranted bias. Moreover, if appropriate measures are taken to ensure transparency, bias can be tracked over time, and algorithms adjusted to re-train the AI and reduce or eradicate bias.

AIDM may perpetuate unequal access to technologies. A less obvious implication of AIDM for the right to equality is its potential to increase the 'digital divide' in cases where AIDM-facilitated service delivery requires people to have a digital presence, which may contain data about them. Services delivered through AIDM may therefore exacerbate inequality for those who do not have a sufficient digital presence, for example due to economic, cultural or geographic factors.

AIDM may, on the other hand, promote equality by enhancing access to technologies and services. For example, people who face obstacles in accessing technologies due to geographic or economic circumstances may have new opportunities for participation, such as via the use of digital teachers or AI-assisted education platforms. To be effective, however, AI-assisted measures for redressing disadvantage must be adequately resourced by, for example, training in the use of ICT and digital technologies.

AIDM may reinforce inequality by interacting with other technologies. AIDM is not necessarily a stand-alone technology, but may interact with other technologies in ways that affect the right to equality. For instance, distributed ledger technologies (colloquially known as 'blockchain') present ways of storing and processing data, which may be used in

35 Latonero, "Governing Artificial Intelligence: Upholding Human Rights & Dignity."



combination with AIDM. Blockchain technologies can have differential impacts which can reinforce social inequality. In particular, wealthy or privileged groups may have accumulated social capital, which means they do not need to disclose additional information in order to establish trust. The socially disadvantaged, however, may need to disclose information, such as credit information, to establish an equivalent level of trust, with associated unequal protection of privacy.

3.5.3.2 Non-transparent and unaccountable decision making

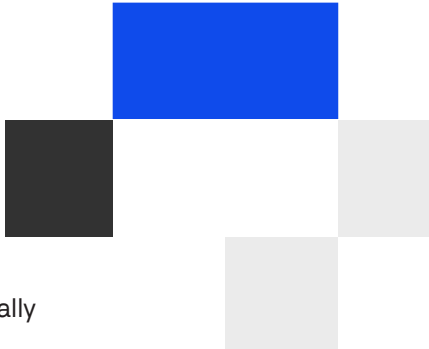

Significant human rights, such as the right to equality before the law and the right to an effective remedy, are premised on transparent and accountable decision making. AIDM, however, presents considerable challenges to these rights due to its potential opacity. Algorithms may be opaque either because they are protected as proprietary confidential data or due to difficulties in understanding the operation of an algorithm arising from the complexity, for example, of neural networks. If the operation of an algorithm is effectively undiscoverable, decisions based on the algorithm – such as decisions to refuse a welfare payment, not to approve a loan, or deny medical treatment – may be immune to challenge. Furthermore, those subject to AIDM may be denied the ability to correct inaccurate or biased information on which a decision is made. The protection of a number of significant human right therefore depends upon a right to know that a machine is involved in a significant decision, and a right to an explanation as to how a decision was made.

3.5.3.3 Privacy

AIDM threatens privacy by expediting and automating collection, matching and analysis of data. AIDM may be implemented at all stages of the information life cycle, from collection, through to analysing and matching, to disclosure. In the online context, an enormous amount of data is collected by technology companies, such as internet service providers, search engine operators or social media platforms, which can then be combined with other data, including metadata (such as an IP address or network ID), in order to profile users and target advertising. Search engine operators and social media platforms have been testbeds for the development of AI based on large datasets. While the aggregated data do not need to identify individual users to achieve the objectives of targeted advertising this, nevertheless, entails the collection of large amounts of often highly revelatory data that may be associated with an individual. Furthermore, much of the data is collected and processed without users necessarily being aware this is happening. Additionally, the data may be associated with individual identities where, for example, it is linked to an identifier, such as an email address or user account. Finally, AI-based data analysis may produce new information about an individual, such as psychological insights concerning, for example, interests, education, political views or sexual preferences.

AIDM may exacerbate privacy risks when combined with other technologies. With technological developments, such as the Internet of Things, the privacy risks associated with online data processing spill over into the offline world. For example, facial recognition algorithms applied to images collected by CCTV cameras, may be combined with other data as part of the compilation of significant revelatory information. Developments in machine learning have created considerable potential for personal information to be determined from videos, images and sounds by means of facial and voice recognition software. Moreover, the application of these technologies, whether online or offline, may enable the identification of individuals who are incidentally captured, for example, in photographs or videos.

AIDM may be associated with **blockchain technologies** by, for example, being applied to the analysis of data stored on a digital ledger, potentially to reveal personal information. While, ostensibly, blockchain technologies may promise greater privacy by, for instance, the encryption of data and the use of synonyms rather than names, disclosure of data concerning



a user's history may be necessary in order to ensure trust. Furthermore, blockchain technologies pose difficulties for the application of data privacy laws. For example, especially with open public ledgers, there is no clear entity that controls the data, and is therefore responsible for privacy infringements. Furthermore, the immutability of blockchain data effectively means that rights to correction or deletion of personal information cannot be applied.

3.5.3.4 Political participation and freedom of expression

AI is capable of analysing, mimicking and influencing human behaviour. By analysing data sets, such as data stored on social media platforms, AI is capable of providing psychological insights, such as individual preferences, weaknesses and desires. These insights may be used to generate automated communications by means of bots, that may spread disinformation which is designed to manipulate views or behaviours. The non-transparent use of bots to influence people for purposes ranging from marketing to voting in elections poses risks for rights such as the right to self-determination and the right to political participation.

Given the centrality of social media for political and social communications, the manipulation of content on social media by AI-powered bots, which may include false information, hate speech or other biased content, poses significant threats to the fundamental right to freedom of expression. In particular, these practices jeopardise two of the major objectives of freedom of expression, the promotion of informed democratic participation and the pursuit of truth.

3.5.3.5 Fair Trial And Fair Hearing

AIDM enables predictive policing programs, such as COMPAS, which are designed to predict the likelihood of a person committing an offence or reoffending. The use of machine learning AI as a predictive tool is, however, susceptible to bias which may, for example, discriminate against members of vulnerable groups. The use of AIDM as part of the criminal justice system, including by judges to estimate the probabilities of re-offending, threaten the rights to a fair trial and fair hearing. This is especially the case where the processes by which a prediction is made are not transparent.

On the other hand, AIDMs are not subject to the limitations or biases of human decision making, such as susceptibility to decision making fatigue (or 'ego depletion') or other unconscious biases. AI may be able to be used to detect patterns in decisions by humans, such as judges, and in appropriate cases supplement such decision making in order to reduce bias. Furthermore, AIDM has the potential to be used to alleviate endemic problems with access to justice by, for example, enabling more efficient decision making or more cases to be decided with fewer resources.

3.5.3.6 Freedom Of Association

Freedom of association can underpin other rights, such as freedom of political participation and freedom of expression. One technology used for generating data for AIDM is graph theory, which can be a way to visualise social networks. This methodology works by ascribing attributes to people within a network, and drawing inferences about people from data about other people in their network. AIDM may use this data in order to recommend or make decisions, for example, about access to welfare or credit. At the extreme, these systems can underpin society-wide measures for rating people, such as China's social credit system. Attempts to map and quantify social interactions can lead to large-scale social sorting, undermining freedom of association and, consequently, posing system-wide threats to, for example democratic participation.

3.5.4 In particular, what are some practical examples of how AIDM can protect or threaten human rights?

Given the breadth of the potential applications of AIDM, it is impossible to be anywhere near comprehensive in giving examples of how it can protect or threaten human rights. The following, however, provides some practical illustrations of the threats, as well as some potential benefits, of AIDM for the human rights identified above.

- A 2013 study by Latanya Sweeney found that searches for African-American names were 25% more likely to result in AI-generated advertising suggesting a criminal record than white identifying names, indicating systemic violence.³⁶
- In the United States, payday lenders such as ZestFinance are applying proprietary algorithms to large data sets, including public internet data and social network data, to determine creditworthiness. The data includes social graphs of a potential borrower's online social networks.³⁷
- The Allegheny Family Screening Tool (AFST), which is being used to predict child abuse and neglect, applies algorithms to data confined to families that use public services and, as a result, has been criticised for potential bias.³⁸ On the other hand, the tool was only implemented following an independent ethics review, and it has been estimated that it has led to a significant reduction in the percentage of low risk cases proposed for review.³⁹
- Independent analysis of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a commercial AI tool for estimating the likelihood of recidivism, found that black defendants were more likely than white defendants to be at a higher risk of recidivism.⁴⁰ A further study found that COMPAS fared no better than humans recruited using the Amazon Mechanical Turk.⁴¹
- In 2017, Stanford University researchers used images collected from online dating sites to train a deep neural network to 'predict' the sexual orientation of people, without obtaining their consent.⁴²

36 "Racism Is Poisoning Online Ad Delivery, Says Harvard Professor," MIT Technology Review (blog), 2013, <https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/>.

37 Mikella Hurley and Julius Adebayo, "Credit Scoring in the Era of Big Data," Yale Journal of Law and Technology 18, no. 1 (2017): 5, <http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1122&context=yjolt>.

38 Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018).

39 Dan Hurley, "Can an Algorithm Tell When Kids Are in Danger," New York Times 2 (2018), <https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html>.

40 Jeff Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm," ProPublica, May 23, 2016, <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

41 Issie Lapowsky, "Crime-Predicting Algorithms May Not Beat Untrained Humans," Wired, January 17, 2018, <https://www.wired.com/story/crime-predicting-algorithms-may-not-outperform-untrained-humans/>.

42 Yilun Wang and Michal Kosinski, "Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images," Journal of Personality and Social Psychology 114, no. 2 (2018): 246, <https://doi.org/10/gczpph>.

- Facebook has encountered significant difficulties in dealing with ‘hate speech’ directed against the Rohingya minority in Myanmar, which was exacerbated by its algorithm-based news feed.⁴³ On the other hand, the Europol Internet Referral Unit has flagged steps to improve its system for evaluating violent extremist content by introducing the Joint Referral Platform.⁴⁴

Recommendations:

As in the recommendations for Question 1, in the specific case of AI-informed decision making:

- 5.1** Consideration of the impacts of emerging technology on human rights should consider the specific impact on individuals and communities as well as broader impacts on society and values.
- 5.2** A nuanced approach to technology should be adopted which recognizes that effects of technologies are in fact effects of systems in which combinations of technologies both new and old, as well as human and contextual factors, generate the effects.

3.6 How should Australian law protect human rights in respect of AI-informed decision making?

The limitations of the Australian legal framework for protecting human rights, especially in the face of the challenges of rapidly developing technologies, are set out in our response to Question 3.

The responses to this question are directed at the distinctive human rights challenges posed by AIDM.

In our response to Question 5, we explained the broad approach we take to the scope of AIDM, and identified the main human rights implications of AIDM, including the threats posed by AIDM for human rights and the potential for AIDM to be used to promote human rights. While our response to this question focuses specifically on issues relating to the regulation of AIDM, it must be seen against our overall recommendations favouring a holistic approach to regulating rapidly developing technologies as well as new approaches to regulation.

Our responses to how best to regulate AIDM, including the appropriate mix of regulatory tools, are set out in our response to Question 7. For the purpose of this question, however, it is important to note that the protection of human rights in respect of AIDM requires a combination of regulatory responses, ranging from legislative responses to ‘soft law’. In addition, we propose that a new regulatory body be given responsibility for ongoing auditing processes to determine the extent to which cross-sector laws remain ‘fit for purpose’ for regulating AIDM.

⁴³ Steve Stecklow, “Why Facebook Is Losing the War on Hate Speech in Myanmar,” Reuters, 2018, <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/>.

⁴⁴ European Commission, “Communication from the Commission to the European Parliament, The European Council and The Council Delivering on the European Agenda on Security to Fight against Terrorism and Pave the Way towards an Effective and Genuine Security Union” (European Commission, 2016), https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/european-agenda-security/legislative-documents/docs/20160420/communication_eas_progress_since_april_2015_en.pdf.

3.6.1 What should be the overarching objectives of regulation in this area?

Regulation of AIDM should promote the public interest by maximising the benefits of AI while minimising the risks and harms.

As we explain in our response to Question 3, we consider that a human rights approach is the preferred framework for specifying the objectives of regulating AIDM. The relative advantages of a human rights approach are explained further in the response to Question 3. This means that, as a general proposition, the regulation of AIDM should be aimed primarily at maximising the protection of human rights while minimising the risks posed to human rights.

As we mention throughout this submission, the use of AIDM has some potential for promoting and protecting human rights. For example, it has been suggested that AI can be used to support the United Nations Sustainable Development Goals⁴⁵, and can potentially be used to analyse large data sets to detect human rights abuses. As explained in our response to Question 5, however, AIDM poses threats to a broad range of human rights. In addition, powerful technologies, such as AIDM, raise questions concerning the adequacy of the current human rights framework, requiring careful consideration of the extent to which it is adequate to protect the values underpinning human rights in the face of rapidly developing technologies, or whether further development of the framework is required. This is, in our view, a significant research question that merits further research.

While a human rights approach is, in our view, the best way to formulate the advantages and threats of AIDM, the application of abstract rights to concrete problems presents difficulties. In particular, translating the protection of human rights into concrete regulation in the context of rapidly developing technologies is complex, especially as poorly conceived or overly prescriptive regulation can inhibit innovation and, consequently, jeopardise social benefits.

The social benefits and disadvantages of regulating AI were canvassed in the recent report of the UK House of Lords Select Committee on Artificial Intelligence⁴⁶. Although the report did not expressly adopt a human rights approach, it explained and analysed the options for developing a regulatory framework for AI, identifying the following three broad approaches:

- that existing laws and regulations are adequate;
- that laws and regulations are inadequate and immediate action is needed; and
- that a cautious and staged approach should be adopted.⁴⁷

The arguments against regulation, or in favour of a cautious approach, are based on concerns, first, that premature or inappropriate regulation would deter innovation and, secondly, that laws and regulations may be simply unable to keep pace with rapidly evolving technologies.

Reflecting these concerns, the Committee concluded that, at this stage, blanket AI-specific regulation would be inappropriate.⁴⁸ In June 2018, in its response to the Committee's report, the UK government agreed that sector-specific regulation was premature, but also indicated that it was establishing a Ministerial Working Group on the Future Regulation to identify

⁴⁵ See, "AI for Good Global Summit 2017," June 7, 2017, <https://www.itu.int/en/ITU-T/AI/Pages/201706-default.aspx>.

⁴⁶ Select Committee on Artificial Intelligence, "AI in the UK: Ready, Willing and Able" (House of Lords, 2018), <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>.

⁴⁷ Select Committee on Artificial Intelligence, 112.

⁴⁸ Select Committee on Artificial Intelligence, 116.

‘areas where regulation needs to adapt to support emerging technologies such as AI’, and announced a £10 million Regulators’ Pioneer Fund to support new approaches to regulating emerging technologies such as AI.⁴⁹ Issues relating to potential new approaches to regulating AI are taken up further in our response to Question 7.

Applying a human rights approach to the regulation of AIDM entails recognising that human rights are not absolute but also that limits to human rights must be justified in accordance with human rights principles. The principle of proportionality has emerged as the preferred principle for balancing rights against other objectives, such as social utility, or against other rights.⁵⁰ If, as we suggest, a human rights approach is applied to developing a regulatory framework for AIDM this necessarily implies that any limits to the protection of rights must be justified in accordance with other legitimate objectives or human rights, and must not be disproportionate.

3.6.2 What principles should be applied to achieve these objectives?

The formulation of principles that reflect international human rights law and norms can assist in guiding the development of legal and regulatory regimes for AI-informed decision making.

As pointed out in the HRT Issues Paper, a human rights approach may be applied by means of the ‘PANEL principles’, namely participation, accountability, non-discrimination and equality, empowerment, and legality.⁵¹ As explained in our response to Question 3, we agree that the PANEL principles provide an appropriate high level foundation for guiding more detailed regulation. It is, nevertheless, important that the application of these high level principles to the particular context of AIDM be investigated in more detail so as to provide practical guidance on the regulation of AI technologies. As further explained in our response to Question 3, we support the development of mid-level principles that provide guidance as to how human rights may be implemented in laws and regulations that apply to specific technologies, which may be one of the most important outcomes of the AHRC process. Such mid-level principles, which Schauer refers to as ‘rules of weight’⁵² and Luizzi as ‘specific guides’⁵³ – may assist in bridging the gap between high-level principles, such as the PANEL principles, and practical context.

The Toronto Declaration, drafted by Amnesty International and Access Now, represents an attempt to spell out principles relating to the application of the right to equality and non-discrimination to machine learning.⁵⁴ For example, the Declaration states that:

49 Department for Business, Energy & Industrial Strategy, “AI in the UK: Ready, Willing and Able? - Government Response to the Select Committee Report” (Department for Business, Energy & Industrial Strategy, 2018), <https://www.gov.uk/government/publications/ai-in-the-uk-ready-willing-and-able-government-response-to-the-select-committee-report>.

50 See, for example, Alec Stone Sweet and Jud Mathews, “Proportionality Balancing and Global Constitutionalism,” *Colum. J. Transnat’l L.* 47 (2008): 72.

51 AHRC, “Human Rights and Technology Issues Paper (2018)”; See also, Scottish Human Rights Commission, “Human Rights Based Approach | Scottish Human Rights Commission” (Scottish Human Rights Commission, 2018), <http://www.scottishhumanrights.com/rights-in-practice/human-rights-based-approach/>.

52 Frederick Schauer, “Proportionality and the Question of Weight,” in *Proportionality and The Rule of Law: Rights, Justification, Reasoning*, Cambridge University Press, Cambridge, ed. C. Husfort, B.W. Miller, and G. Webber (Cambridge, UK: Cambridge University Press, 2014), 173–185.

53 Vincent Luizzi, “Balancing of Interests in Courts,” *Jurimetrics* 20, no. 4 (1980): 373–404, www.jstor.org/stable/29761723.

54 Amnesty International and Access Now, “The Toronto Declaration” For another example of a statement of principles for applying specific human rights see Article 19, Principles on protection of freedom of expression and privacy,

Inclusion, diversity and equity entails the active participation of, and meaningful consultation with, a diverse community, including end users, during the design and application of machine learning systems, to help ensure that systems are created and used in ways that respect rights – particularly the rights of marginalised groups who are vulnerable to discrimination.⁵⁵ Our response to Question 5 identifies the following features of AIDM as posing particular threats to human rights: bias in data and algorithms; opacity; the capacity to produce or reveal personal information; and the capacity to influence opinions and potentially elections. For each of these features, and for each of the rights implicated by AIDM, there is a need for the development of principles that assist in guiding how human rights, including the PANEL principles, can be applied to AIDM.

Work that is being undertaken in this area, which can assist in the development of human rights principles, include initiatives aimed at establishing ethical guidelines for AI. A number of ethical approaches to AI, for example, are based on the ‘FAT principles’, which refers to ‘fairness, accountability and transparency’.⁵⁶ For instance, Microsoft’s ethical framework for AI released in 2018, *The Future Computed*, incorporates the six principles of fairness, reliability and safety, privacy and security, inclusiveness, transparency and accountability.⁵⁷

Attention is particularly required for principles aimed at ensuring transparency and accountability of AIDM. Although it is generally acknowledged that there are problems with the opacity of AIDM, usually known as the ‘black box’ problem,⁵⁸ there are questions about the extent of the problem and how it might best be addressed.⁵⁹ In examining this issue it may be important to distinguish between the explainability of AIDM, which refers to how decisions are arrived at given certain input factors, and transparency, which entails disclosure of details of how an algorithm works.⁶⁰

In understanding issues relating to the explainability and/or transparency of AIDM, it is important to pay attention to both the nature of the AIDM and the reasons for an explanation or transparency. For example, there are, in general, two categories of explanation of machine learning algorithms: global or model-centric explanations and local or subject-centric explanations.⁶¹ Global, or model-centric explanations look at the AI model as a whole and describe the influential elements that shape the decision making of the model. Local or subject-centric explanations, on the other hand, focus on the individual decision and the elements responsible for that particular decision, such as the data sets used to train a

May 2016.

55 Amnesty International and Access Now, 21.

56 See, for example, the papers presented at the “Annual ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*),” accessed October 22, 2018, <https://fatconference.org/>.

57 “The Future Computed: Artificial Intelligence and Its Role in Society,” The Official Microsoft Blog (blog), January 18, 2018, <https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificial-intelligence-role-society/>.

58 Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press, 2015).

59 See, for example, Deven R. Desai and Joshua A. Kroll, “Trust but Verify: A Guide to Algorithms and the Law,” 2017.

60 Nik Dawson, “Bits & Atoms,” AI Policy White Paper (University of Technology Sydney, 2018).

61 Lilian Edwards and Michael Veale, “Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For,” *Duke L. & Tech. Rev.* 16 (2017): 18.

machine learning algorithm. In some circumstances what might be required may be a global explanation, while in other circumstances a local explanation may be warranted. In addition, considerable efforts are being expended on investigating technological means for assisting with the explainability or transparency of AIDM.⁶²

In elaborating on the PANEL principles, such as accountability and legality, so as to apply them to AIDM it is necessary to build a clear understanding of the state of the art concerning the transparency of AIDM, and of how legal requirements for explainability and transparency may be applied in practice. In this respect, some assistance may be derived from experience in implementing Article 22 of the GDPR, which applies to automated decision-making. For example, Annex 1 of the guidelines on Article 22 produced by the Article 29 Data Protection Working Party sets out good practice recommendations for the implementation of Article 22.⁶³

In any case, at a general level, the principles of transparency and fairness require that people affected by AIDM be informed when a decision that may significantly affect them is made with the assistance of, or by, AI technologies. In addition, where decisions are informed by or made by AI technologies, people affected by the decisions should have a right to an explanation as to how the decision was made.

While the rights to be informed and to an explanation are sound base-line principles, there will be occasions where other rights or interests may over-ride the rights to know that a decision is made by AI technologies or how a decision has been arrived at. In such circumstances, the rights to be informed and to an explanation must be subject to the proportionality principle. Therefore, applying the human rights approach to AIDM requires the application of rights, and limits to rights based on the proportionality principle, to specifying principles that may be applied in concrete contexts. For example, just as the Toronto Declaration distinguishes between the obligations of public sector and private sector actors in relation to the right to equality and non-discrimination, a similar distinction is likely to be needed in applying principles of transparency and accountability to AIDM.

The difficulties of ensuring transparency and accountability in AIDM suggest that any principles directed to these objectives should be supplemented by principles capable of embedding human rights in the design of AIDM, known as ‘human rights by design’. The advantages of ‘human rights by design’ approaches, especially as a means for proactively preventing human rights breaches, are addressed further in our response to Question 7.

3.6.3 Are there any gaps in how Australian law deals with this area? If so, what are they?

As explained immediately above in our response to part (b) of this question, the development of mid-level principles for applying human rights to AIDM may be used to develop legislation, or other forms of regulation, that applies to AIDM.

Accurately and comprehensively identifying gaps in how Australian law deals with AIDM is a challenging task, as AIDM may have effects across broad areas of the law, some of which are obvious and some of which are potentially not so obvious. Given the scope and complexity of these issues, we suggest that one of the functions of the proposed new regulatory body recommended by this submission should be to undertake an audit of laws that may need to be amended to appropriately protect human rights in relation to AIDM. In this section of our response we therefore focus on some obvious gaps in the law.

⁶² Joshua A. Kroll et al., “Accountable Algorithms,” U. Pa. L. Rev. 165 (2016): 633.

⁶³ Article 29 Data Protection Working Party, “Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679” (European Commission, 2018), http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

First, while Australian information privacy law, such as the Australian Privacy Principles (APPs) included in the *Privacy Act 1988* (Cth), may have some application to AIDM – such as notification requirements in relation to the collection of personal information – it does not incorporate rights to information or to an explanation in relation to automated decision making. While there are difficulties with the wording of Article 22 of the GDPR, experience in implementing the GDPR in the EU can provide some guidance as to the introduction of rights to information or to an explanation in jurisdictions such as Australia. Although, to date, no other jurisdictions include provisions comparable to Article 22 of the GDPR, there are some indications that governments are increasingly concerned with the lack of transparency with AIDM. For example, California has recently introduced a law requiring a level of transparency for bots used in commercial transactions, or to influence a vote in an election.⁶⁴

Secondly, there are considerable limitations in the extent to which current anti-discrimination laws are able to deal adequately with bias and discrimination arising from AIDM. One set of limitations arises from the potential opacity of AIDM, which may make it difficult to establish discrimination under relevant laws, and which presents another compelling illustration of the need for new transparency laws. One way to deal with this, in the absence of clear transparency requirements, might be to apply a version of an EU principle, according to which a presumption of discrimination arises where an automated decision making system is not transparent, meaning that the onus shifts to the service provider to establish there has been no discrimination.⁶⁵ Apart from this, anti-discrimination laws do not cover all potential grounds of discrimination. For example, AIDM that discriminates against poor or lower income people would not be caught by anti-discrimination laws.

Thirdly, fundamental principles of public law require that decisions by public authorities must be transparent and reviewable. While AIDM promises the potential for improvements in administrative decision making, key elements of administrative law, such as the Administrative Decisions (Judicial Review) Act 1977 (Cth) and freedom of information legislation require review to ensure they are fit for purpose to deal with the use of AIDM in the public sector.

Fourthly, there are, as we suggest above, a broad range of laws that are potentially affected by AIDM. For example, under Australian copyright law creative works generated by a computer and not by a human author is not entitled to copyright protection.⁶⁶ Given the increasing use of AI in producing creative works, and that such works may be protected in other jurisdictions, this is an area of Australian copyright law that requires review.

3.6.4 What can we learn from how other countries are seeking to protect human rights in this area?

Legal frameworks for the protection of human rights in relation to AIDM are in the early stages of development.

The most significant rights-based law dealing with AIDM is Article 22 of the GDPR. Article 22 does not prevent automated decision making or profiling, but gives individuals a qualified right not to be subject to purely automated decision making. In addition, it provides that the data controller should use ‘appropriate mathematical or statistical procedures for the profiling’

⁶⁴ artificiallawyer, “Declare Your Legal Bot! New California Law Demands Bot Transparency,” Artificial Lawyer (blog), October 3, 2018, <https://www.artificiallawyer.com/2018/10/03/declare-your-legal-bot-new-california-law-demands-bot-transparency/>.

⁶⁵ See, Case 109/88 Danfoss[1989] ECR 3199 (European Court of Justice October 17, 1989).

⁶⁶ Andres Guadamuz, “Should Robot Artists Be given Copyright Protection?,” The Conversation (blog), 2017, <http://theconversation.com/should-robot-artists-be-given-copyright-protection-79449>.

and take measures to prevent discrimination on the basis of race or ethnic origin, political opinions, religion or beliefs, trade union membership, genetic or health status or sexual orientation. Although Article 22 is not perfect, experience with the implementation of the GDPR in the EU can assist with the development of regulatory regimes in jurisdictions such as Australia.

In the United Kingdom, the Industrial Strategy, released in November 2017, identified artificial intelligence and the data economy as one of four Grand Challenges to which the UK needs to respond. Prior to this, in June 2017, the House of Lords appointed a Select Committee on Artificial Intelligence ‘to consider the economic, ethical and social implications of advances in artificial intelligence’. The Committee’s report, *AI in the UK: Ready, willing and able?* was released in April 2018, and included 74 conclusions and recommendations. The report did not expressly consider the human rights implications of AI technologies, but some conclusions drawn from its consideration of the ethical implications of AI are relevant to the protection of human rights in AI-informed decision making. In particular, the report concluded that:

... the developments of intelligible AI systems is a fundamental necessity if AI is to become an integral and trusted tool in our society. Whether this takes the form of technical transparency, explainability, or indeed both, will depend on the context and the stakes involved, but in most cases we believe explainability will be a more useful approach for the citizen and the consumer.⁶⁷

The UK government released its response to the Committee’s report in June 2018.⁶⁸ The government did not accept all of the Committee’s recommendations, but some of its responses are relevant to the human rights issues identified in this submission. For example, in response to potential algorithmic bias, the response indicated that:

Government recognises that one of the risks of automated decision-making is that the datasets which the algorithms learn from may reflect the structural inequalities of the society from which data are collected and that this can lead to the encoding of unintentional bias. We will work to ensure that those developing and deploying AI systems are aware of these risks, and the trade-offs and options for mitigation are understood. It is important that multiple perspectives and insights are represented during the development, deployment and operation of algorithms. To this end, we will work with the Alan Turing Institute, which has been working to address these issues.⁶⁹

The UK policy processes initiated by the UK government response to the House of Lords committee report include the establishment of an incipient regulatory framework, which is dealt with more fully in our response to Question 7. We recommend that the developments in the UK be monitored and taken into account in the formulation of an Australian response to the human rights challenges of AIDM.

Also in the UK, the Law Society of England and Wales has initiated a public policy commission to examine the impact of technology and data on human rights and justice.⁷⁰ The Technology Law and Policy Commission established by the Law Society is initially focusing on the use of algorithms in the justice system, and has commenced public consultations on identified

67 Select Committee on Artificial Intelligence, “AI in the UK: Ready, Willing and Able.”

68 Department for Business, Energy & Industrial Strategy, “AI in the UK.”

69 Department for Business, Energy & Industrial Strategy, 29.

70 The Law Society, “Using Algorithms to Deliver Justice – Bias or Boost? – The Law Society,” 2018, <https://www.lawsociety.org.uk/news/press-releases/using-algorithms-to-deliver-justice-bias-or-boost/>.

issues, including calling for written evidence on significant issues such as ‘how does bias in decision making cut across existing legislation?’ and ‘can we create an algorithm that is “ethical-by-design”?’⁷¹ We consider that a similar policy process directed specifically at the use of technology in the legal system would have merit in Australia. As part of its inquiry, however, the AHRC should monitor the findings of the process initiated by the Law Society of England and Wales.

In the European Union, the European Commission has published a communication on AI for Europe, which proposes that the EU takes the lead in developing a framework ‘which promotes innovation and respects the Union’s values and fundamental rights as well as ethical principles such as accountability and transparency’.⁷² The communication indicated that the Commission would bring relevant stakeholders together to, by the end of 2018, draft AI ethic guidelines, with due regard to the Charter of Fundamental Rights of the European Union. In developing draft guidelines, the communication indicated that it would take into account the work of the European Group on Ethics in Science and New Technologies,⁷³ as well as other efforts at developing ethical guidelines.⁷⁴ The communication further indicated that the Commission would support research in the development of explainable AI and implement the pilot project proposed by the European Parliament on Algorithmic Awareness Building, to gather an evidence base to support the design of policy responses to the challenges of AIDM, including biases and discrimination.

As part of the EU response, the EU Fundamental Rights Agency (FRA) is undertaking an assessment of the challenges faced by producers and users of new technology with respect to fundamental rights compliance. The FRA’s work is specifically aimed at bringing human rights ‘more strongly into the development of new technologies and provide data for implementing and developing policies’.⁷⁵ Apart from the European Union, the Council of Europe has a work program on AI and human rights, including work on developing a standard setting instrument based on its study of the human rights dimensions of automated data processing techniques.⁷⁶ Given the extent to which a human rights legal framework provides the foundation for policy development in both the relevant EU institutions and the Council of

71 The Law Society, “Technology and the Law Policy Commission - Algorithms in the Justice System,” September 11, 2018, <https://www.lawsociety.org.uk/policy-campaigns/articles/public-policy-technology-and-law-commission/>.

72 Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, “Artificial Intelligence for Europe” (European Commission, April 25, 2018), <https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe>.

73 European Group on, Ethics in Science and, and European Group on Ethics in Science and New Technologies, “The European Group on Ethics in Science and New Technologies, AI, Robotics and ‘Autonomous’ Systems” (European Commission, 2018), https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

74 The communication gave the following specific examples: Future of life institute, “Asilomar AI Principles,” Future of Life Institute, 2017, <https://futureoflife.org/ai-principles/>; Montreal Declaration for a Responsible Development of AI, “Declaration of Montréal for a responsible development of AI,” Declaration of Montréal for a responsible development of AI, 2017, <https://www.montrealdeclaration-responsibleai.com>; UNI Global Union, “10 Principles for Ethical AI,” UNI Global, 2017, <http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/>.

75 European Union Agency for Fundamental Rights, “Artificial Intelligence, Big Data and Fundamental Rights,” European Union Agency for Fundamental Rights, May 28, 2018, <http://fra.europa.eu/en/project/2018/artificial-intelligence-big-data-and-fundamental-rights>.

76 Council of Europe, “MSI-AUT Committee of Experts on Human Rights Dimensions of Automated Data Processing and Different Forms of Artificial Intelligence,” 2018, <https://www.coe.int/en/web/freedom-expression/msi-aut>.

Europe, it is essential for Australian policy processes to be informed by the processes initiated in Europe, especially processes aimed at embedding human rights principles in instruments such as the proposed draft AI ethical guidelines.

Beyond official government processes, it is important for policy responses to take into account developments in promoting human rights principles in AIDM led by non-government institutions, such as universities. For example, an audit toolkit for algorithmic fairness has been developed by the University of Chicago's Data Science for Social Good project.⁷⁷ In the broader context, work completed in the Netherlands by 4TU.Ethics group of universities on Socially Responsible Innovation and Value Sensitive Design may provide guidance on the incorporation of ethical principles in AI design, which is taken up further in our response to Question 7.⁷⁸

While the above brief survey of international and comparative developments in developing policy responses to AIDM does not purport to be comprehensive, it does indicate the extent to which governments are identifying and confronting similar issues in developing responses to meet the human rights challenges of AIDM. The AHRC process has the potential to make a significant contribution at a critical stage in the development of international legal and regulatory responses to AIDM, but in order to do so it should be informed by the work undertaken elsewhere on the incorporation of human rights considerations into principles for guiding the development of legal and regulatory frameworks.

Recommendations:

- 6.1** **The principles of transparency and fairness require that people affected by AIDM be informed when a decision that may significantly affect them is made with the assistance of, or by, AI technologies. In addition, where decisions are informed by or made by AI technologies, people affected by the decisions should have a right to an explanation as to how the decision was made.**
- 6.2** **Human Rights by Design principles can embed human rights into the design of AIDM to supplement the principles of transparency and fairness**
- 6.3** **We must learn from other jurisdictions and non-governmental organisations in developing human rights approaches for AIDM**

3.7 In addition to legislation, how should Australia protect human rights in AI-informed decision making?

The challenges of promoting and protecting human rights in AIDM are so significant that legislative responses, in isolation, are inadequate. As indicated by Braithwaite's work on responsive regulation⁷⁹, effective regulation requires engaging actors, with a 'pyramid' of regulatory strategies ranging from education through to 'command and control' regulation or criminal sanctions. Responsive regulation (including the regulatory pyramid) must not,

⁷⁷ Center for Data Science and Public Policy - University of Chicago, Bias and Fairness Audit Toolkit . Contribute to Dssg/Aequitas Development by Creating an Account on GitHub, Python (2018; repr., Center for Data Science and Public Policy - University of Chicago, 2018), <https://github.com/dssg/aequitas>.

⁷⁸ See, "4TU | Centre for Ethics and Technology," accessed October 22, 2018, <https://ethicsandtechnology.eu/>.

⁷⁹ Ian Ayres and John Braithwaite, *Responsive Regulation: Transcending the Deregulation Debate* (Oxford University Press, USA, 1995); John Braithwaite, "The Essence of Responsive Regulation," *UBCL Rev.* 44 (2011): 475.

however, be applied in a mechanical fashion since, as Braithwaite puts it:

Strategic use of the pyramid requires the regulator to resist categorizing problems into minor matters that should be dealt with at the base of the pyramid, more serious ones that should be in the middle, and the most egregious ones for the peak of the pyramid. Even with the most serious matters – flouting legal obligations for operating a nuclear plant that risks thousands of lives, for example – we stick with the presumption that it is better to start with dialogue at the base of the pyramid.⁸⁰

We agree with the assumptions of responsive regulation that a regulatory regime must be developed with the active engagement of the broadest range of actors and stakeholders, including government, the private sector, universities and NGOs.

The development of regulatory frameworks for AIDM poses difficulties over and above those associated with regulation in other areas, as it concerns regulating rapidly developing technologies. As explained in our response to Question 6, the UK House of Lords Select Committee on Artificial Intelligence, taking into account the extent to which regulation might deter innovation and concerns about the ability of law and regulation to keep pace with technological change, concluded that '[b]lanket AI-specific regulation, at this stage, would be inappropriate'.⁸¹ As we also pointed out, in its response to the Committee's report, the UK government agreed that sector-specific regulation was premature, but also indicated that it was establishing a Ministerial Working Group on the Future Regulation to identify 'areas where regulation needs to adapt to support emerging technologies such as AI', and announced a £10 million Regulators' Pioneer Fund to support new approaches to regulating emerging technologies such as AI.⁸² This suggests the need for new forms of regulation that are better designed for dealing with rapidly changing technologies than 'regulate-and-forget' approaches. In our response to this question we identify some emerging approaches to regulation that may assist with determining legal and regulatory responses to AIDM.

Considerable effort has been expended by regulatory scholars in distinguishing between different levels of granularity in stating forms of regulation – rules, standards and principles – and when each may be appropriate. Principles-based regulation means 'moving away from reliance on detailed, prescriptive rules and relying more on high-level, broadly stated rules or Principles to set the standards by which regulated firms must conduct business'⁸³ Principles-based forms of regulation have been proposed for dealing with rapidly changing technologies because they are more flexible and, consequently, potentially more durable in the face of technological change. The choices between forms of regulation are, however, more complex than the simple opposition between principles-based and rules-based regulation might suggest. As Julia Black has suggested:

In practice, characterising a regulatory regime as rules based or principles based does not take us very far, descriptively or normatively. It is hard to classify any one regulatory regime as being either entirely rule based or entirely principles based; the better question is what is, and should be, the relative roles of each. Neither principles nor rules usually function particularly successfully without the other. However debates on PBR are in fact rarely about the linguistic

80 Braithwaite, "The Essence of Responsive Regulation," 483.

81 Select Committee on Artificial Intelligence, "AI in the UK: Ready, Willing and Able."

82 Department for Business, Energy & Industrial Strategy, "AI in the UK."

83 Julia Black, Martyn Hopper, and Christa Band, "Making a Success of Principles-Based Regulation," *Law and Financial Markets Review* 1, no. 3 (2007): 191–206, <https://doi.org/10/gffdfq>.

structure of written norms. They are usually much more about the nature of regulatory practices, of regulatory relationships, and as to who should have the final say in interpreting the rule or principle. Moreover, it is the substantive nature of these relationships and practices which are far more relevant for understanding the operation of a regulatory regime than what the rulebooks look like.⁸⁴

An effective regulatory regime for AIDM should therefore include a combination of principles and rules. As explained in our response to Question 6, to date attention has focused on the development of principles for regulating AIDM. For example, the report House of Lords Select Committee on Artificial Intelligence included the following recommendation:

We recommend that a cross-sector ethical code of conduct, or ‘AI code’, suitable for implementation across public and private sector organisations which are developing or adopting AI, be drawn up and promoted by the Centre for Data Ethics and Innovation, with input from the AI Council and the Alan Turing Institute, with a degree of urgency. In some cases, sector-specific variations will need to be created, using similar language and branding. ... In time, the AI code could provide the basis for statutory regulation, if and when this is determined to be necessary.⁸⁵

While the UK government did not specifically endorse this recommendation, it has established a new Centre for Data Ethics and Regulation, with the response to the House of Lords Select Committee report indicating that:

The Centre will identify the measures needed to strengthen and improve the way data and AI is used. It will operate by drawing on evidence and insights from across regulators, academia, the public and business and translate these into actions that deliver direct, real world impact on the way that data and AI is used. Following the public consultation in the summer, the Centre, in dialogue with the Government will carefully prioritise and scope the specific projects in its work programme.⁸⁶

Furthermore, as we explained in our response to Question 6, the European Commission communication on AI for Europe indicated that draft ethical guidelines for AI will be developed before the end of 2018.

As a first stage in designing a human rights approach to the regulation of AIDM we therefore support the development of principles for promoting and protecting human rights in AIDM, which take into account principles and ethical guidelines formulated as part of policy processes in other jurisdictions, and which can be used as the basis for more detailed regulation, including statutory regulation. The responsibility for developing principles, by means of a public policy process, could be given to the proposed new regulatory body recommended by this submission.

84 Julia Black, “The Rise, Fall and Fate of Principles Based Regulation,” LSE Law, Society and Economy Working Papers, 2010, 17, <https://doi.org/10/fzn7k7>.

85 Select Committee on Artificial Intelligence, “AI in the UK: Ready, Willing and Able,” 74.

86 Department for Business, Energy & Industrial Strategy, “AI in the UK,” 38.

3.7.1 What role, if any, is there for: an organisation that takes a central role in promoting responsible innovation in AI-informed decision making?

A major recommendation of this submission supports the establishment of a new regulatory body, known as the Technology Assessment Office (TAO). We believe that such a body is needed to coordinate the regulation of new technologies, including conducting technology assessments and developing new forms of regulation. On the basis of our transdisciplinary approach, which emphasises the interconnection between technologies, we do not consider it advisable for the TAO to be confined to the regulation of AI, but should extend across the board to new transformative technologies. As we explain more fully below, the TAO should operate on a transdisciplinary model, drawing on expertise from government, industry, academia and the community sector, and should coordinate with existing regulatory authorities, such as the Office of the Australian Information Commission (OAIC) and the Australian Competition and Consumer Commission (ACCC).

Proposals for regulating algorithms are not new. For example, in 2016 Geoff Mulgan from the UK Nesta (National Endowment for Science, Technology and the Arts), proposed a new machine intelligence commission, which would not have formal regulatory powers of approval or certification, but would have powers of investigation and recommendation, extending to powers to develop algorithms and machine learning tools to interrogate AIDM.⁸⁷ As we explained above, the UK government is establishing a new Centre for Data Ethics and Innovation, whose activities are intended to be primarily investigative and advisory, but may extend to reviewing the existing regulatory framework to identify gaps in response to the uses of data and AI, and identifying steps to ensure that the law, regulation and guidance keep pace with technological developments.⁸⁸ The Centre for Data Ethics and Innovation is part of a new UK institutional structure designed to promote innovation in AI in the UK, which includes the AI Council and the Office for AI, with the three bodies having the following functions:

- *Centre for Data Ethics and Innovation* – supply government with independent, expert advice on the measures that are needed to enable and ensure safe and ethical innovation in data-driven and AI-based technologies.
- *The AI Council* –bring together leading figures from industry and academia to provide strategic leadership, and promote the growth of the sector.
- *The Office for AI* –the civil service secretary for the AI Council, which will drive implementation and lead coordination on AI within government.

The limitation of the Centre for Data Ethics and Innovation to investigatory and advisory functions clearly represents concerns both that regulation not unduly inhibit technological innovation and that the new Centre should not cut across the jurisdiction of existing sector-specific regulators, such as the Information Commissioner’s Office (IOC) and the Competition and Markets Authority (CMA).

In Australia, the News Corp Australia submission to the ACCC Digital Platforms Inquiry supported the establishment of an Algorithm Review Board to analyse and remedy distortions of competition by digital platforms.⁸⁹ This is an example of the need for understanding the

⁸⁷ Geoff Mulgan, “A Machine Intelligence Commission for the UK,” Nesta (blog), 2016, <https://www.nesta.org.uk/blog/a-machine-intelligence-commission-for-the-uk/>.

⁸⁸ Department for Digital, Culture, Media & Sport, “Department for Digital, Culture, Media & Sport, Centre for Data Ethics and Innovation Consultation,” GOV.UK, 2018, <https://www.gov.uk/government/consultations/consultation-on-the-centre-for-data-ethics-and-innovation/centre-for-data-ethics-and-innovation-consultation>.

⁸⁹ News Corp Australia, “Submission to the Australian Competition and Consumer Commission: Digital Platforms

interactions of technologies in specific contexts, in this case, interaction of AI and digital platforms in the context of competition in the media sector. In our submission, the proposed TAO should be responsible for issues relating to transparency and accountability for algorithms that significantly affect people across sectors, which is essential for understanding regulatory issues from a holistic perspective. Moreover, in not being confined to issues relating to AIDM, the proposed TAO would be able to examine issues that cut across technologies.

The key issue in regulating AIDM, from a human rights perspective, is ensuring that rights are appropriately protected while not disproportionately inhibiting innovation. In areas characterised by rapidly changing technologies, this may involve investigating the benefits of novel or creative approaches to regulation, some of which are being pioneered in the regulation of financial technology, such as adaptive regulation or anticipatory regulation. As Armstrong and Rae point out:

There are a number of elements that distinguish anticipatory approaches from other more reactive forms of regulation, namely they are proactive, forward-facing, flexible, iterative and more inclusive. A central goal of these emerging anticipatory methods has been to enable and support innovation around new technologies or business models in a ‘responsible’ way.⁹⁰

As yet, there is a lack of consistent terminology in this area, including for innovations such as ‘regulatory sandboxes’ or ‘regulatory testbeds’, and more theoretical work is required to properly understand the strengths and weaknesses of these approaches. Moreover, in relation to complex technologies such as AIDM, it is likely that a combination of forward-looking approaches will be needed.

Armstrong and Rae, for example, distinguish between ‘adaptive’ approaches to regulation, which are appropriate ‘when a regulator wants to help facilitate the development of new products or services but existing regulatory frameworks may have to be adapted to do so’,⁹¹ and ‘anticipatory’ approaches, which are designed to ‘better understand what the impacts of an emerging technology (which may not be developed enough for use) might be on the economy and society, and therefore what the potential regulatory needs will be’.⁹² On this approach, key distinctions between ‘adaptive’ and ‘anticipatory’ regulation are that anticipatory approaches are necessarily more iterative, with regulations under continual review, and must involve more inclusion and engagement with a broader range of stakeholders in regulatory activities. Furthermore, ‘anticipatory’ approaches can incorporate strategic visions with longer timeframes.

Potential challenges arising from the proposal to introduce a new body include the difficulties of coordinating with other sector-specific bodies, and the possibility that an additional body may simply increase regulatory and advisory complexity. In addition, considerable difficulties are likely to arise from attempting to embed human rights considerations into a regulatory framework while also promoting innovation. This might, for example, suggest the need for rigorous scrutiny of the potential impacts of new technologies before decisions are made to grant regulatory holidays or create regulatory sandboxes. Nevertheless, we consider that

Inquiry,” 2018, <https://www.accc.gov.au/system/files/News%20Corp%20Australia%20%28April%202018%29.pdf>.

⁹⁰ Henry Armstrong and Jen Rae, “A Working Model for Anticipatory Regulation: A Working Paper” (Nesta, 2017), 7, <https://www.nesta.org.uk/report/a-working-model-for-anticipatory-regulation-a-working-paper/>.

⁹¹ Armstrong and Rae, “A Working Model for Anticipatory Regulation”; See also, William Eggers D., Mike Turley, and Pankaj Kishnani, “The Future of Regulation” (Deloitte Insights, 2018), <https://www2.deloitte.com/insights/us/en/industry/public-sector/future-of-regulation/regulating-emerging-technology.html>.

⁹² Armstrong and Rae, “A Working Model for Anticipatory Regulation,” 8.

the benefits of a new form of regulation outweigh any of these possible disadvantages. First, the proposed TAO allows for problems to be identified across both regulatory sectors and technologies, enabling greater coordination and coherence in regulation. Secondly, the proposal new entity would support new approaches to regulation, drawing on the expertise of transdisciplinary collaborators. Thirdly, the proposed body would be based on broad engagement with government, business, academia and community groups, in recognition that the challenges of AIDM, and new technologies, must be met by drawing on a variety of perspectives and expertise.

3.7.2 What role, if any, is there for: self-regulatory or co-regulatory approaches?

As we explained earlier in our response to this question, the regulation of AIDM requires a variety of responses across the spectrum from education to potentially criminal sanctions. There is therefore scope for self-regulation, such as through standard setting or industry codes, as well as co-regulation, such as where industry-based regulation is backed by legislative standards.

A considerable amount of effort has been expended by scholars and policy-makers in building understanding of self-regulation, and of when different forms of regulation, ranging from self-regulation to command-and-control regulation, may be appropriate.⁹³ The insights obtained from traditional approaches must clearly be taken into account in developing regulatory responses to AIDM. That said, the challenges of new and emerging technologies, such as AIDM, are so significant that, as explained above, new approaches are required. To an extent, these new approaches must combine elements of traditional self-regulation and co-regulation, together with more experimental approaches. Over and above this, the new technologies do seem to raise issues that have been confronted in regulating technologies in particularly acute forms. For example, the question of how legislation, which is relatively difficult to change, can be combined with iterative approaches to regulation, where regulation is under continual review, has become more pressing. There is no all-purpose formula for managing the difficulties involved with balancing flexibility and certainty, or regulatory incentives and regulatory sanctions, in establishing regulatory regimes for new technologies, including AIDM. We do suggest, however, that while there is clearly a role for regulatory flexibility, establishing a set of mid-level principles for applying the human rights approach to AIDM can provide a reasonably stable framework (or benchmarks) for evaluating flexible approaches to the regulation of AIDM, and then refining those approaches.

3.7.3 What role, if any, is there for: a 'regulation by design' approach?

The difficulties in applying a human rights approach to regulation to technologies such as AIDM, even with the incorporation of new forms of regulation, suggests that ex post forms of regulation alone are inadequate, especially as they are only capable of dealing with human rights harms after the event. Moreover, it is important to understand that technologies are not 'neutral', but have built-in values, which should be demystified and subject to critical analysis.⁹⁴ This suggests that rather than relying exclusively on ex post regulatory responses, there is a need for human rights considerations to be taken into account at the design stage of technologies, with the aim of effectively embedding human rights protections within the design of AIDM. This approach is suitable for rapidly developing technologies as it can proactively incorporate human rights into future technologies, rather than relying entirely on

93 The literature is extensive. See, for example, Anthony Ogus, "Rethinking Self-Regulation," *Oxford J. Legal Stud.* 15 (1995): 97, <https://doi.org/10/cqhz4v>; Julia Black, "Constitutionalising Self-Regulation," *The Modern Law Review* 59, no. 1 (1996): 24–55, <https://doi.org/10/dcnkrb>.

94 Lessig, *Code*.

retrospective application of the law. As explained in this submission, a proactive approach to regulation fits within our overall transdisciplinary perspective.

In this submission, we support the adoption of ‘human rights by design’ within a framework of Value Sensitive Design, with principles developed for applying human rights to AIDM being used to guide a ‘regulation by design’ approach. Building on the evolution of ‘privacy by design’^{95,96} human rights by design requires human rights to be taken into account through the entire engineering process of the development of a technology. ‘Privacy by design’ is, itself, a development of a more general approach of ‘Value Sensitive Design’. ‘Value Sensitive Design’ is ‘a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process’.⁹⁷ It incorporates a structured, tripartite methodology involving conceptual, empirical and technical investigations. Conceptual investigations, for example, are aimed at clarifying the values to be embedded in technologies, including how competing values are to be balanced. Human rights by design can build upon Value Sensitive Design, by applying a human rights framework to clarify values, including the application of the proportionality principle where there are competing rights or values.

The application of a principle of human rights by design holds the promise of ensuring that those responsible for developing AI technologies take human rights into account in developing AI systems, rather than the protection of rights depending upon enforcement by individuals after the fact.^{98,99}

Applying Value Sensitive Design or human rights by design approaches are not panaceas for the human rights challenges posed by AIDM, and there are difficulties in embodying these approaches within regulatory regimes. Nevertheless, experience is being accumulated in the practical application of such approaches to concrete problems. For example, the principle of privacy by design is already embodied in law in the form of Article 25(1) of the GDPR, which provides as follows:

Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

Experience with the implementation of Article 25(1) should be taken into account in developing, and expanding, on privacy by design so that it more fully reflects the broader range of human rights that are implicated by AIDM that we have identified in our response

95 Ann Cavoukian, “Privacy by Design—the 7 Foundational Principles (2011),” 2011.

96 “Privacy and Data Protection by Design — ENISA,” Report/Study, accessed October 22, 2018, <https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design>.

97 Batya Friedman et al., “Value Sensitive Design and Information Systems,” in *Human-Computer Interaction in Management Information Systems*, ed. P Zhang and D Galletta (Springer, 2013), 55–95.

98 Edwards and Veale, “Slave to the Algorithm.”

99 Committee of experts on internet intermediaries, “Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications.”

to Question 5. The responsibility for developing ‘human rights by design’ so that it is appropriately adapted to AIDM could be given to the proposed new regulator, the TAO.

In implementing a human rights by design approach, it is important to ensure that values and groups that might easily be overlooked in the design of AIDM should not be neglected. As we explain in this submission, our transdisciplinary approach is intended to ensure that issues and people that may otherwise be ‘invisible’ are not ignored, but are appropriately factored into regulatory design. Furthermore, our approach fits well within the tradition of Value Sensitive Design as it is aimed at inclusively involving a broad range of participation and collaboration in technology design.

In summary, in implementing a human rights by design approach, we suggest that our transdisciplinary approach entails:

1. striving to build diverse teams and inclusive practices into design work;
2. appropriately articulating values and specifying how those values are translated into design, which should reflect a human rights approach;
3. realising that there will be unintentional and difficult to predict consequences from new technologies, such as AIDM, which are designed to be made more transparent by our transdisciplinary approach;
4. building proper mechanisms for involvement of educated communities, and for consultation and feedback; and
5. an iterative approach, as opposed to ‘regulate and forget’, which involves ongoing assessment and evaluation.

Given that human rights by design approaches are not perfect, and that there are likely to be unintended consequences even where such approaches are used, it is important for ‘regulation by design’ to be supplemented by other forms of regulation. In particular, human rights can be embedded into AIDM by requiring human rights impact assessments where AI technologies are ‘likely to result in a high risk’ to human rights.¹⁰⁰ Therefore, in addition to the principle of human rights by design, human rights in AIDM can be protected by requiring those responsible for developing AIDM to submit the technology to a structured human rights impact assessment undertaken by an independent third party, where the decision making poses a sufficient risk to human rights.

Recommendations:

- 7.1 A regulatory regime must be developed in respect of responsive regulation that has active engagement of the broadest range of actors and stakeholders**
- 7.2 We must learn from other jurisdictions and non-governmental organisations in developing human rights approaches for AIDM**

¹⁰⁰ Edwards and Veale, “Slave to the Algorithm”; Committee of experts on internet intermediaries, “Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications” By analogy, Article 35(1) of the GDPR provides that: ‘Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data’.

7.3 In implementing human rights by design a transdisciplinary approach should be adopted, entailing: (1) striving to build diverse teams and inclusive practices into the design; attending to hard and soft impacts consultation and education with stakeholders; and an iterative approach of ongoing assessment and evaluation, discussed in Section 5

7.4 Human rights in AIDM can be further protected by requiring those responsible for developing AIDM to submit the technology to a structured human rights impact assessment, undertaken by an independent third party, where the decision making poses a sufficient risk to human rights.

3.8 What opportunities and challenges currently exist for people with disability accessing technology?

The history of the relationship between people with disability and technology is a complicated one. Technological and medical developments have at times been imposed on people with disability both with, but also without, their permission. The Deaf community and people with Deaf children have been framed, for example, as (actual or potential) cochlear patients under a system that favours medical knowledge over the lived experience of people with disability. The continued screening of in *utero* fetuses for ‘defective’ genes and the results for groups such as people with Down Syndrome, are well-documented, if not well-advertised. Australia has one of the higher termination rates for this group. This complex history makes for a necessarily careful understanding of what and how emerging technologies might be poised to support, or be further imposed on people with disabilities.

Clearly, in both these examples there are arguments made on both sides – the issue becomes problematic, however, when for example parents of Deaf children are branded ‘irresponsible’ when they have made an informed decision to refuse a cochlear implant for their child.

In this context, a co-produced perspective is needed in order to develop a nuanced understanding of the human rights and ethical implications of emerging technologies (AI, Internet of Things, Machine Learning, robotics, haptics etc.) and the experience of living with disability in contemporary Australian society. The inclusion of people with a disability, and their advocates and supporters, in disability policy and service design and provision, including new technology applications, needs to be central if the judgements and mistakes of the past are not to be repeated under a potentially simplistic rubric of technological ‘innovation’.

Some of the challenges affecting access to technology for people with disabilities can also affect those who care for them, which can affect both the carer’s capacity to perform the caring role and also carers’ rights as individuals.

The four pillars of equitable service delivery are *access, utilisation, quality and safety, and outcomes*. A technological approach that is fully informed by human rights looks at each of these elements, and not just (as is often the case) access.

- Is access differentially distributed, and if so, who gets access, who has a more difficult time getting access, and who misses out altogether, and why?
- If access is assured, what do patterns of utilisation indicate? What aspect of the service, technology and its delivery (including implicit or explicit discrimination) stop some individuals and groups from using the service or see them dropping out? Are there historical or cultural elements, that are not immediately evident, that may be limiting both access and utilisation?

- What are the quality and safety issues for vulnerable groups? This assessment needs to include physical, social, psychological and cultural safety and to look at the quality of the service for all groups
- What are the outcomes for different groups? If they differ, how is this reconciled from a human rights perspective?

3.8.1 Opportunities

3.8.1.1 Access to services

Having a disability can require a considerable amount of time spent negotiating the complications and limitations of the health and social care systems. This is because these systems are most definitely not designed around the current, and growing, needs of people with disability. While the intention of the National Disability Insurance Scheme (NDIS) is to resolve at least some of these issues, the shift to remote technologies (online access, Chatbots, AI classification algorithms etc.) may represent a mix of opportunity and barrier for some people with disability. Many contemporary 'helplines', for example, are designed to cut costs rather than facilitate access for non-disabled customers. How such innovations affect people with a disability is unclear. In this context, design, oversight and regulation may be essential components of such innovations and their potential impacts on people with disability.

In the specific context of healthcare, the potential of remote access to expert knowledge and supports is substantial. This includes data and technology-driven strategies such as t-health, m-health and e-health, all developed to address a variety of access issues. However, health systems are often jurisdiction-based, with territorial areas (state, Local Health District (LHD), Private Healthcare Australia (PHA), etc.) of responsibility (and associated funding). In addition, resources, including expertise, are likely to be finite for many disability issues (where expert knowledge is required) meaning that shifts in technology may lack systemic backing and associated resources. This may mean that access is theoretically improved, while coordination of care and specific service access is not necessarily enhanced to the benefit of the person with a disability.

3.8.1.2 Access to supports

People with disability may well be able to access a wider variety of supports, assuming their sensory, cognitive and communicative skills are accommodated by emerging technologies. Groups such as those with an intellectual disability or sensory impairment are likely to need explicit consideration in the design, delivery and funding of such innovations. In addition, passive or active monitoring systems (such as for those living at home with a dementia or other chronic disease condition) need to be seen also as having a potential surveillance function. How such an individual is engaged, issues of consent (temporary or ongoing) need to be considered in the implementation of such technologies.

3.8.1.3 Access to information and knowledge

This seems the least problematic area of technology but participation in, for example, social media technologies can be a mixed experience. We know that various types of bullying and associated harms can occur for a variety of vulnerable groups in the community who use social media. The implications for people with some disability conditions need to be emphasised in this context too. Such access may provide very positive supports (e.g. shared experiences with people with similar conditions/situations) and this should not be minimised but there are associated risks. This too is an area where regulatory knowledge and experience is developing at a slower rate than the technology. While technology can play an enabling role

by providing increased access to information and knowledge it can also inadvertently have the opposite effect and become overwhelming for people to keep up with the volume and variety of information that an individual needs to navigate and keep up with.

3.8.1.4 Scalability

A key issue for the disability sector includes the scalability of connecting technologies. This could include scaling a pilot project upwards to a sector/industry level or modification and transfer to other areas of disability. Which of these technologies produces real, positive outcomes for different disability groups and categories? How does such translation occur, and how do we avoid the perennial problem of always a pilot never a fully funded program?

3.8.2 Barriers

3.8.2.1 Wilful ignorance

This concept describes when systems, groups and professions choose not to inquire on issues of the kind described in this document. That is, there is a deliberate strategy to not know about or understand a situation or a group of people (or individuals) in their context and the implications of access to, or denial of a service or product. This can include funding access to enable an individual and their supports to access or customise a service or product that might provide valuable support.

3.8.2.2 Loss of trust

As noted in the introduction, the care of people with disabilities has a fraught history. This has produced issues of systemic mistrust among a variety of groups and communities, including for example some Aboriginal and Torres Strait Islander people who may have very mixed feelings towards health and social care providers based on deep historical experiences or even more recent ones. The implementation of new technologies may occur in a context of unresolved trust issues and uptake or acceptability may be issues as a consequence of such histories.

3.8.2.3 Technology as a disability fix

Given that many disabilities have historically been framed as a 'lack' or 'failing' in medical and social policy paradigms, the risk exists that emerging technologies may be framed as simply potential 'fixes' for peoples' conditions. This is a potential barrier to the optimum utility of such technologies and their potential contribution to the lives of people with disabilities, their supports and the wider community.

3.8.2.4 Infrastructure

If the available infrastructure is poor or of variable quality then access will be compromised. This has historically been an issue for those in rural communities. It is clearly an issue for those in remote communities still and this includes Aboriginal and Torres Strait Islander communities whose cultural models of disability may not be accommodated or even acknowledged, and whose living situations may currently limit or negate the implementation of such technologies without additional infrastructure.¹⁰¹ This also raises the issue of alternative models of care and access, their consideration in the technological arena or lack thereof. Non-Indigenous people with a disability may also be isolated from the potential benefits of such innovations depending on their geographic and socio-economic circumstances. The

¹⁰¹ Avery, S Culture Is Inclusion: a narrative of Aboriginal and Torres Strait Islander people with disability, (First Peoples Disability Network Australia, 2018)

situation for people from CALD backgrounds in this mix is, and remains, very unclear, as does the impact on successive generations of immigrants (e.g. younger, older, English language competence etc.), refugees and associated groups.¹⁰²

3.8.2.5 Market models and funding

This brings us to the issue of market and marketised models of care. The NDIS is a marketised model of care underwritten by taxpayer funds. The person with a disability is framed as a fully 'informed consumer' of marketised services and products. Many of the emerging technologies indicated in this response operate on a market or semi-market model. This makes their motive largely one of for-profit service provision, often via a data transaction relationship (free software, monetised data). The implications of this for people with a disability, still one of the poorest groups in society, is also problematic.

In a related context, the funding of new technologies, upgrades to such technologies and issues such as training and support for new technologies, can all be potential barriers. Appropriateness may be another issue operating in this scenario, in which the electric wheelchair is all very well for urban environments with adequate street and suburban paving but far less optimal (unless modified) for rural and remote regions.

3.8.2.6 Knowledge production and translation

The great majority of disability research conducted in Australia has ignored women, rural and remote people, CALD groups, and Aboriginal and Torres Strait Islander communities.¹⁰³ This means that disability knowledge production and translation are limited at this stage and raises the issue as to how these emerging technologies will, potentially, change this inequality or, alternatively, reinforce existing biases in research focus.

Recommendations:

- 8.1 Increased focus and resources for research into disability service delivery for women, rural and regional, CALD and ATSI communities particularly with a focus on emerging technology**
- 8.2 Providing subsidised access to digital services and assistive technologies for platforms and tools regularly used by people with disabilities to protect privacy of personal data.**

¹⁰² Robertson, H. and Travaglia, J. Cultural Diversity Competency Framework (2015) https://www.researchgate.net/publication/282735649_Cultural_Diversity_Competency_Framework_2015 (viewed 17 October 2018)

¹⁰³ Centre for Disability Research and Policy, The University of Sydney, The Audit of Disability Research in Australia 2000–2013 <http://sydney.edu.au/health-sciences/cdrp/projects/auditresearch.shtml> (viewed 25 October 2018)

3.9 What should be the Australian Government's strategy in promoting accessible technology for people with disability? In particular:

- a. What, if any, changes to Australian law are needed to ensure new technology is accessible?
- b. What, if any, policy and other changes are needed in Australia to promote accessibility for new technology?

The relationship between human rights, disability, and technology ultimately needs to be understood in the context of the whole range of human rights, and cannot be reduced to the right to accessibility. As noted in response to the previous question the four pillars of equitable service delivery are *access, utilisation, quality and safety, and outcomes*. A technological approach that is fully informed by human rights looks at each of these elements, and not just (as is often the case) access.

It is vital to consider the full complexities of the human rights implications of technology for people with disabilities in the context of the fundamental principles underpinning the UNCRPD, including dignity, equality, inclusion and autonomy, and by reference to all of the rights protected by the CRPD and other international human rights instruments. While the right to accessibility in Article 9 of the CRPD is essential for addressing technical and material barriers to accessing public space, public infrastructure and ICT, for many people with disabilities the human rights issues they encounter are more profound and fundamental than mere technical and material barriers and fall outside of the scope of Article 9 and the possibilities of universal design.

For example, a homeless person with an acquired brain injury and mental illness might be excluded from public space not because of an absence of personal physical mobility technologies or audio announcements on public transport, but by reason of hostile architecture or CCTV surveillance technologies that result in over policing of public space and heightened enforcement of coercive mental health laws. Another example is an individual with a physical disability who is living in public housing, only receives Newstart allowance and is socially isolated might not be able to afford the ICT technology to be able to participate in online communities for people with disabilities. Or, a person with intellectual disability who is incarcerated in prison because of discriminatory forensic mental health laws (and potentially also the operation of biased AI risk assessment) might not be able to access new technologies because these are not available in prison. In these three examples, these individuals' relationship to new technologies cannot be reduced to an issue of technical and material accessibility (as per Article 9) but instead relates to complex structural circumstances and engages other rights in the CRPD including rights to equality and non-discrimination, liberty, equality before the law, personal integrity and adequate standard of living. Technology cannot be used as substitute for strategies aimed at addressing poverty, legacies and ongoing impacts of colonisation and eugenics, discriminatory systems of incarceration, and ongoing stigma.

a. What, if any, changes to Australian law are needed to ensure new technology is accessible?

It is increasingly important for Australian laws and policies to extend beyond addressing access to technologies. Technologies can also be used in relation to, rather than by, people experiencing disabilities. Technologies used for surveillance purposes by carers in homes, to limit movement, are offered as an example of a scenario that may fall into regulatory gaps and therefore needing attention. In response, again, we argue that complementary ‘soft’ and ‘hard’ approaches can work to respond to known challenges – issues of consent, security, ownership, and use of data, as examples – and incentivise innovations in this space.

A strategy which incorporates co-productive perspective we believe is equally relevant to the formulation of Australian Government legislation and policy in this space. Developing a nuanced and layered understanding would support the emergence of necessary alternate positions that challenge the current technical and economic boundaries that shape decision making.

The Disability Discrimination Act 1192 (Cth) (DDA) has played an important role in ensuring the rights of people with disability to access the built environment and participate freely are upheld. This law applies in all areas of public life, with a particular focus on access to public premises.

There have been examples of recent technology cases where a lack of inclusive design has led to litigation (signaling the important role of regulatory measures) and embarrassing and negative PR for private sector organisations. The Commonwealth Bank’s ‘Albert’ EFTPOS machine is a case in point.¹⁰⁴

These large cases are rare, and they expose a plaintiff to considerable financial risk. There is an opportunity to make changes to the disability discrimination act to make it easier to take Federal Court actions so that binding legal common law decisions can benefit the whole disability community. Even if a plaintiff wins, they may have costs awarded against them. This is a significant risk to consumers and a barrier to people with disability challenging discriminatory places, products and services.

b. What, if any, policy and other changes are needed in Australia to promote accessibility for new technology?

The cost to society, as well as the individuals involved, both in non-financial and financial terms, of non-inclusive technologies is that a sector of the population is effectively (i) locked out of participating in the economy, and (ii) becomes reliant on support from carers who would otherwise be able to also participate in other activities. The economic costs of non-inclusive technology, when looked at from this society-wide perspective, far outstrip the individual costs that designers would need to bear to design technologies that enable rather than disabling sectors of the population. Since this cost-benefit analysis applies to the level of society, Government incentives may need to form part of a solution to provide business with incentives to develop inclusive technologies that enable people to be self-sufficient and productive engaged members of society, rather than being locked out of engagement in socially productive and personally meaningful activity instead of becoming patients.

A proactive approach to assessing and quantifying the economic, social and human rights benefits in improving social and economic participation for people with disabilities through assistive technologies and universal design of emerging technologies is required. For example

¹⁰⁴ Naomi Selvaratnam and Sarah Farnsworth, “Blind Woman Takes Bank to Court over ‘inaccessible’ EFTPOS Machines,” ABC News, March 16, 2018, <https://www.abc.net.au/news/2018-03-16/blind-discrimination-lawsuit-cba-albert-eftpos-machines/9551458>.

the potential for autonomous vehicles to provide greater autonomy, mobility and quality of life for people with disabilities may factor in projections of health and financial implications for individuals and potentially reduced caring responsibilities for family.

Strategies and policies which promote universal design through more inclusive approaches to design, development and implementation of emerging technologies through meeting required standards for universal access also reduce the cost and effort of retrofitting technology which is found to be inaccessible. The process and cost of retrofitting government websites in an attempt to meet WCAG 2.0 Accessibility Guidelines is a good case in point here. If the government can provide a pool of resources and expertise ahead of time to support agencies and companies in considering and designing for the four pillars of equitable service delivery in emerging technology services (access, utilisation, quality and safety and outcomes) it will substantially reduce the cost and reputational damage of trying to retrofit services after they have been found to be inaccessible or discriminatory.

Recommendations:

- 9.1 Demonstrate best practice in relation to promoting diversity, universal design, value centred design and consideration of human rights in design, provision and procurement of government services.**
- 9.2 Incentives and assistance for curriculum development across all areas of technical and a higher education to introduce disability, access, inclusive design and universal design across disciplines**

3.10 How can the private sector be encouraged or incentivised to develop and use accessible and inclusive technology, for example, through the use of universal design?

In order to encourage and incentivise a significant shift towards a business culture of inclusion at all levels of our private sector organisation will require multiple 'soft' and 'hard' mechanisms and interventions.

Although there is an important role for legislated measures such as the DDA to regulate the accessibility of technologies in all their applications, these are a safety net to ensure consumer rights are upheld and aren't considered drivers for innovation. In order to initiate a cultural shift within private sector organisations towards inclusive design, sustainable drivers for change are more likely to be founded in positive reinforcement and education. Any interventions or drivers should be designed to educate the private sector on the multiple benefits of an inclusive, Universal Design approach, encourage more diverse workplaces and foster inclusive collaborations between business and community. The goal of this educational approach is to transform preconceived perceptions about the cost and value of accessible and inclusive technology, while also breaking culturally ingrained habits of practice across and within industry/organisations/private sector.

3.10.1 Conceptualising digital inequity

The first step to educating the private sector is supporting it to understand the barriers that prevent people from participating in, using and benefiting from technology. Digital inequality is a concept that can frame the barriers experienced by people in the community when technology is inaccessible.¹⁰⁵

¹⁰⁵ Mark Warschauer, *Technology and Social Inclusion: Rethinking the Digital Divide* (MIT press, 2004).

People can experience digital inequality and lack of access in multiple ways:

- Accessibility of technology/ mode of delivery/interface.
- Lack of bandwidth or lack of device.
- Autonomy of Access (can a person log on monitored/unmonitored, at will or specific times).
- Limitations in affordability and availability of technology (the impact of poverty on inclusion).
- Lack of awareness of specialised technology development that can support autonomy and independence in the community.
- Technical Skill, experience (opportunity), familiarity and knowledge.
- Purpose (why is the technology being used).

Solutions to these experiences of digital inequality are as diverse as the people who experience them - people with access to technology but without the skills, require different solutions to people with technical skill but no access to technology.

3.10.2 Introducing the concept of universal design

The second step is educating the private sector about inclusive design principles. These, such as Universal Design, can be applied as a means of identifying and minimising barriers experienced by people in accessing technology. The United Nations uses the term Universal Design, and defines it as ‘the design of products, environments, programmes and services to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design. *Universal design* shall not exclude assistive devices for particular groups of persons with disabilities where this is needed.’¹⁰⁶

Having an understanding of inclusive design through Universal Design principles is an important step towards understanding the ways design decisions can exclude particular groups, and establish links between design, access and human rights across all sectors of technology.

3.10.3 The business case for inclusive design

The third step is educating the private sector on the business benefits of a more inclusive, universally designed approach to innovation and technology.¹⁰⁷ Universal Design can contribute to business and social sustainability^{108,109}, growing new markets, and strengthening existing markets.¹¹⁰

106 Convention on the Rights of Persons with Disabilities,” A/RES/61/106, article 2 § article 2 (2008).

107 Sam Waller et al., “Making the Case for Inclusive Design,” *Applied Ergonomics* 46 (2015): 297–303, <https://doi.org/10/gffdf6>.

108 Mian M. Ajmal et al., “Conceptualizing and Incorporating Social Sustainability in the Business World,” *International Journal of Sustainable Development & World Ecology* 25, no. 4 (2018): 327–339, <https://doi.org/10/gffdf8>.

109 Tom Vavik and Martina Maria Keitsch, “Exploring Relationships between Universal Design and Social Sustainable Development: Some Methodological Aspects to the Debate on the Sciences of Sustainability,” *Sustainable Development* 18, no. 5 (2010): 295–305, <https://doi.org/10/c37dgs>.

110 Roger Coleman et al., “From Margins to Mainstream,” in *Inclusive Design* (Springer, 2003), 1–25.

Assistive technology is an important area of technological innovation that supports people with disability to access and utilise mainstream products and services, maximising autonomy and independence. Tech-businesses like Apple and banking organisation such as TD promote the benefits of assistive technology for ALL users, whether they identify as living with disability or not. Entrepreneurship in assistive technology development is being supported through online communities, and developer community hubs such as Microsoft’s Reactor.

Online communities and hubs that harvest and nurture ‘citizen knowledge’ have historically resulted in the development of open source assistive technology.¹¹¹

3.10.4 Workplace diversity

Finally, but possibly most importantly, we need to find ways of encouraging the private sector to improve workplace diversity.^{112,113,114} Inclusive decision making will be best made by diverse decision makers and supported by diverse teams.

Processes of Participatory Design and Co-Design should be implemented in all decision making and problem solving. This means setting up teams with lived experience who contribute all along the design and development chain. Unfortunately, often expert advisory panels are consulted at the ends of various design phases, and even more often are not remunerated for their contributions. This needs to shift so that expert panels are paid for their contribution and are included from the very first design meeting.

Recommendations:

- 10.1 Provide incentives and rewards to businesses who demonstrate best practice e.g. Human Rights Awards for inclusive business organisation, national star rating framework for sustainability and inclusion within organisations, tax incentives for Universal Design Departments such as those implemented in Japan.**
- 10.2 Incentivise and provide resources to support technology design approaches which focus on Universal Design and participatory design. Provide industry with panels of expert user groups of people with disabilities available to contribute to design, testing and review of emerging technology**
- 10.3 Educate the private sector about the importance of workplace diversity and implications for inclusive decision-making and outcomes. Encourage organisations to require every employee of an organisation to complete the Universal Design introductory course. Sponsor the development of further Universal Design Training for tech companies. Development of professional certification programs for architects, IT, engineers**

111 Erin Buehler et al., “Sharing Is Caring: Assistive Technology Designs on Thingiverse,” in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (ACM, 2015), 525–534.

112 Michalle E. Mor Barak, *Managing Diversity: Toward a Globally Inclusive Workplace* (Sage Publications, 2016).

113 Dariusz Turek, “What Do We Know about the Effects of Diversity Management? A Meta-Analysis,” *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie*, no. 4 (964) (2017): 5–25.

114 Subhash C. Kundu and Archana Mor, “Workforce Diversity and Organizational Performance: A Study of IT Industry in India,” *Employee Relations* 39, no. 2 (2017): 160–183, <https://doi.org/10/f9wbx4>.

4 Case studies

4.1 AI & data analytics in education: ethical issues - extended case study

UTS uses Artificial Intelligence & Data Analytics (AIDA) in its teaching, service delivery, and has an active Learning Analytics R&D program. Used well, AIDA opens up exciting possibilities to improve learning, but equally, it use raises ethical issues. Our expertise in this area enables us to respond to the AHRC consultation with specific reference to the education sector, for schools, colleges and universities.

AIDA is more than a set of computers, data and software, AIDA ‘infrastructure’ includes the entire human-computer nexus, including technological, psychological, and social characteristics. This complexity of these systems makes it difficult for AIDA to account for human attitudes, beliefs and goals, which can lead to violations of human rights. This situation is aggravated when humans are left out of the decision-making loop and therefore lose autonomy over the way in which AIDA is applied to the digital traces they leave. In education, this points to concerns in common with other sectors, but the complexity of learning heightens these tensions; using AIDA to improve learning requires that we model and assess distinctive aspects of the human mind. The difficulty of modeling the learner’s state of mind (e.g. what they are struggling with; whether they are disengaged; whether they are thinking critically) is a source of great innovation for AIDA based approaches, but it also raises the stakes should these models be poorly implemented.

So while on the one hand educational AIDA is an exciting evolution of humanity’s search for new tools to think with, clearly, there are some important principles at stake in this area. These include rights to: access educational data; object about its automated processing; port data to other systems and institutions; and to be forgotten over defined temporal periods. One of the key places where the dual affordances of educational AIDA arise is in the right to an explanation about data based decisions that can have a very real impact upon the lives of students, and the right to challenge/correct incorrect data points. Keeping an informed and empowered student in the decision making loop provides many ways in which potential misuses and abuses of educational data can be avoided.

4.1.1 AIDA in education

AIDA infrastructure is entering education in numerous ways, through commercial products, government schemes and ongoing research programs that use data to inform educational decision making. These infrastructures operate at multiple levels and have a wide range of stakeholders, purposes, and end goals. They are also the subject of increasing controversy, with the ‘datafication’ of education, and the impact this has on the system’s functioning, fast becoming the subject of close academic scrutiny.¹¹⁵

¹¹⁵ Ben Williamson, *Big Data in Education: The Digital Future of Learning, Policy and Practice* (Sage, 2017).

1. The datafication of education

There is significant interest in relatively conventional metrics that can be used to rank schools and universities at different geographical scales, from local neighbourhoods, to cities and eventually between countries. Even in their early forms metrics such as PISA,¹¹⁶ NAPLAN,¹¹⁷ and other national ‘high stakes testing’ schemes came with significant resource implications. Parents choose schools, governments to allocate funding on the basis of results and as such they are highly contested education spaces.¹¹⁸ However, such testing regimes are increasingly turning to methods reliant upon AI, with, PISA introducing AI-based problems to evaluate students’ collaboration and critical thinking, and NAPLAN rolling out an adaptive online testing methodology in 2018. Even before the move to these emerging AI based approaches, a number of issues arose with the application of data driven decision making into the context of education. Years later one would never have imagined that the impact of national and state testing systems could equate to high levels of student stress and elevated suicide rates.¹¹⁹ Accordingly, some claim that part of the promise of AIDA is the demise of traditional exams and marking workload, all of which could be made redundant by effective formative feedback and continuous assessment.¹²⁰ But in a regime increasingly reliant on standardised tests and the use of the resulting data in measures of institutional effectiveness and accountability this opportunity is lost.

2. Institutional analytics

With schools and universities generating and ingesting increasing amounts of data there is now a drive to use that data in deriving actionable insights to improve student outcomes and create better organisational efficiencies. Progress towards using business intelligence type analytics developed in the corporate world and applying it to derive educationally relevant insights is common. Such ‘institutional analytics’¹²¹ inform leaders about the concerns of administering an organisation such as resource usage, staff/student demographics, and student behaviours at the level of enrolments, grades, drop-out rates, satisfaction levels, and so forth. However, many institutions make the mistake of valuing what they can measure, instead of measuring what they value. With care and foresight it is possible to extract deep insights from educational datasets that drive curriculum reform and other improvements,¹²²

116 OECD, “Programme for International Student Assessment,” accessed October 22, 2018, <http://www.oecd.org/pisa/>.

117 ACARA, “National Assessment Program - Literacy and Numeracy (NAPLAN),” accessed October 22, 2018, <https://www.nap.edu.au/naplan>.

118 Sam Sellar, Greg Thompson, and David Rutkowski, *The Global Education Race: Taking the Measure of PISA and International Testing* (Brush Education, 2017).

119 Abbie Wightwick, “Is Exam Stress Driving Our Children to Mental Illness and Even Suicide?,” *WalesOnline*, April 27, 2018, <https://www.walesonline.co.uk/news/education/exam-stress-driving-children-mental-14582450>.

120 Rose Luckin, “Towards Artificial Intelligence-Based Assessment Systems,” *Nature Human Behaviour* 1 (2017): 0028, <https://doi.org/10/gc3gdj>.

121 George Siemens and Phil Long, “Penetrating the Fog: Analytics in Learning and Education.,” *EDUCAUSE Review* 46, no. 5 (2011): 30.

122 For example, the Open University in the UK has demonstrated marked relationships between student satisfaction and learning design, but only once they performed an intensive mapping of their teaching and learning strategies. For more details see: Bart Rienties and Lisette Toetenel, “The Impact of Learning Design on Student Behaviour, Satisfaction and Performance: A Cross-Institutional Comparison across 151 Modules,” *Computers in Human Behavior*

but it often much easier to follow a path that leads to few insights beyond random correlations and false positives, which are nonetheless sometimes accorded extraordinary importance. Importantly, techniques developed in the business world do not necessarily translate directly to the domain of education - the impact upon an individual of an incorrect classification is much smaller when it concerns what books they might like to buy in an online store. In an educational setting this incorrect classification might have ramifications throughout an individual's life. This means that the risks associated with using such models can be significantly greater in the educational context; it is essential that the AHRC recognise the way in which a tool that is largely innocuous in one context can have markedly different ramifications in another.

3. Learning analytics

Moving from the institution down to the level of the individual, a number of fields have made progress in understanding student progress, and giving better feedback.¹²³ These fields seek to improve the way a subject is taught, and to help students learn how to learn. Such goals require an understanding that crosses domains such as teaching and learning, data science, and psychology. Critically, such work is now well out of the academic research lab and in mainstream products, with huge investment from major technology platforms, publishers and myriad educational technology startups in applying various techniques at scale. Key questions might be: How are these technologies being used to help students learn? And what concerns might arise in their implementation?

Automated classification of students at risk. Intervening as early as possible to help struggling students seems an admirable goal, but the predictive power of AIDA models brings new risks, and current education policy has a mixed history in addressing equity and the needs of 'students at risk'. The shift is from monitoring the student's formal outputs, to monitoring many other features of the student's activity, since these leave data traces. Some of these traces may indeed help, but as more data is added to such models the risk of spurious correlations increases. Ethical issues that arise in this area include: (i) on what empirical and computational basis, over what timeframe, predictive classifications are made; (ii) once classifications are made, the nature of the intervention that might be run, indeed some have pointed out that institutions have an *obligation to act*¹²⁴ if their analytics suggest that students are likely to be wasting money and time in a chosen course of study; (iii) the right of students to view and potentially challenge their 'risk profiles'; (iv) the degree to which models are sensitive to different learning contexts, and how dependent they are on future learning contexts replicating the particular history on which the models were trained; (v) how long a risk profile is stored, and whether a student can demand that it be deleted; (vi) whether that profile should be exported to other institutions and learning contexts to inform future predictions.

Automated tutoring systems. Given a set of digital traces about the choices a learner makes in a system it is possible to construct profiles of what they are likely to know, like, benefit from

60 (2016): 333–341, <https://doi.org/10/gffdgb>.

123 This work has taken place over a number of years, and its transdisciplinary nature has meant that a number of different sub-communities have formed out of the various contributing fields. Terms such as Learning Analytics, Educational Data Science/Mining, Artificial Intelligence in Education, and Computer Supported Collaborative Learning have all produced substantial bodies of work. An internet search on these will reveal these research communities and the rich array of results that they have generated over more than 30 years.

124 Paul Prinsloo and Sharon Slade, "An Elephant in the Learning Analytics Room: The Obligation to Act," in Proceedings of the Seventh International Learning Analytics & Knowledge Conference (ACM, 2017), 46–55.

studying next, and to then adapt content, or to make personalised recommendations for what they should do next. Such systems also come under the name of intelligent tutoring systems, or adaptive learning, and they have been intensively studied in research labs for well over 30 years. They tend to work best when they are helping a student to master the content and skills in a narrow domain of knowledge, one that has been analysed and modeled in great detail. For example, the ASSISTments system¹²⁵ has been used with over 600 teachers from 42 states and 14 countries to help K-12 students master mathematics skills in their curriculum. However, if narrow conceptions of ‘efficient learning’ come to dominate priorities, numerous ethical issues arise, for example: (i) the replacement of teachers with non-human agents may require the transitioning of the workforce into new roles; (ii) forms of learning that cannot be tutored in this way might come to be devalued; (iii) what is lost from a holistic, social learning experience in authentic contexts (namely, anything that cannot be computationally modeled). Finally, it is important to note that the student models in such systems will only be as accurate as the trace data that they rely upon; students learn in a wide range of environments (both physical and online) and so a system that builds a model from the limited interactions that it witnesses is unlikely to develop a sound model. Ways in which to mitigate against this problem exist,¹²⁶ which empower learners to understand (and so perhaps modify) their learning processes for themselves instead of blindly following recommendations made by a system over which they have no control.

Automated grading and feedback. Some automated scoring systems can now provide more consistent, accurate grading than humans, for specific kinds of tightly defined summative task (including basic maths, and essay writing, where correct or good archetypical answers exist). Such systems can provide 24/7 formative feedback at a speed and scale that no human team can deliver, a marked improvement over the current state of affairs if used well.¹²⁷ Student facing dashboards can inform our learners about how they are progressing, providing helpful feedback, if designed well. Other automated systems enable the delivery of personalised feedback to classes with thousands of students based upon their participation in instructor defined learning activities, so providing a quality of communication between instructor and students in systems of mass education that have not before been possible.¹²⁸ However, some of the ethical issues that arise in applying these technologies are: (i) the trustworthiness of automated systems must be clearly communicated to different audiences (students; teachers; parents);¹²⁹ (ii) should students be given the option to opt out of automated assessments that professionals endorse? And; (iii) in what contexts can AIDA be used to make autonomous interventions? Choosing the right mode (automated feedback or human provided) depends upon the situation - not all feedback can be automated, and it takes a deep understanding of both educational systems and human psychology to choose the right mode, features driving the feedback, and ultimately, the intervention delivered.

125 Assistments (Worcester Polytechnic Institute, 2016), <https://www.assistments.org/>.

126 Such as the manual override provided by CogBooks, which allows the learner to either follow its recommendations, or to choose an alternative pathway: <https://www.cogbooks.com/>, and open learner models – see e.g. Susan Bull and Judy Kay, “Open Learner Models,” in *Advances in Intelligent Tutoring Systems* (Springer, 2010), 301–322.

127 Linda Corrin, “Supporting the Use of Student-Facing Learning Analytics in the Classroom,” *Learning Analytics in the Classroom: Translating Learning Analytics for Teachers*, 2018.

128 For example, OnTask (OnTask), accessed October 22, 2018, <https://www.ontasklearning.org/>.

129 The recent public debate around the equivalence claims made for paper-based and online NAPLAN testing is one instructive example of the challenge of transitioning a national system to an adaptive platform.

Lifelong linked data. Educational datasets are increasingly being linked up, enabling an understanding of how different factors affect one another in increasingly sophisticated models. As different educational providers move towards more advanced data policies and practices it becomes possible to link data such as: student demographics; attendance; participation in learning activities; extracurricular activities; and records of behavioural issues. This is an invaluable resource, helping institutions to create rich student models, and so derive actionable insights that can be used to support their students. However, this linking also opens up a wide range of possibilities that will quickly start to impact upon the rights of the students that these rich stores of data describe. There are existing cases that demonstrate that people are wary about linked data, and the wholesale sharing of student data, such as the collapse of inBloom.¹³⁰ In considering why this might be, some key questions are: (i) Who has access to and control of this data? (the vendors whose products create it? the institutions who buy those products? the students who it describes?); (ii) Are there data sets that should not be linked up?; and (iii) What rights does the data subject have to object to models created with this linked data?

The GDPR grants data subjects a wide range of rights (e.g. to access, erasure, rectification, portability, and to object),¹³¹ but implementing these in an educational context creates a number of thorny issues. For example, compliance with GDPR rapidly becomes very difficult to maintain with the trace data collected about participation in groupwork projects; what if one student demands erasure? Does this overrule the desire of another team member to port data about their group project to another environment?

A new kind of digital divide? History teaches us that technological advances do not in themselves lead automatically to a more egalitarian society. On the contrary, they often serve to exacerbate disparities. There is a growing understanding of how using biased data can train algorithms to replicate historical injustices in any sector, but specifically in education, it is not inconceivable that AIDA could open up a new kind of digital divide — an ‘educational AI divide’. Very careful judgements must be made around when there is sufficient evidence to fully automate a process, how it is monitored, and when to augment rather than automate human intelligence.¹³² Without policy oversight it is highly likely that only the wealthy will be provided with best practice teaching, and AIDA tools that have been well designed and validated, while the less advantaged among us could well serve as ‘the guinea pigs’ for testing those emerging technologies.

4.1.2 Policies and strategies helping educational institutions to protect human rights

Fortunately, many educational providers are leading the way in helping to protect human rights in the age of AIDA. Six lessons have been learned already, they are:

1. Develop clear policies around the use of student data

Universities and at least one national body have now developed clearly worded policies for students, which give clear indications about the ways in which student data, and the algorithms analysing it, will and will not be used.¹³³ In Australia, the recent report of Senate

¹³⁰ K.N.C., “Withered InBloom,” *The Economist*, April 30, 2014, <https://www.economist.com/schumpeter/2014/04/30/withered-inbloom>.

¹³¹ <https://gdpr-info.eu/chapter-3/>

¹³² Gavriel Salomon, David N. Perkins, and Tamar Globerson, “Partners in Cognition: Extending Human Intelligence with Intelligent Technologies,” *Educational Researcher* 20, no. 3 (1991): 2–9, <https://doi.org/10/c285rs>.

¹³³ University policies on the use of student data, designed for students to read: The Open University (UK), “Ethical Use of Student Data for Learning Analytics,” *Student Policies and Regulations - Open University*, 2018, <https://help>.

Select Committee on the Future of Work and Workers makes six important recommendations that reiterate the role of education policy to protect vulnerable groups in the Australian community, pointing to the need to address flexibility, post-school options, engagement, better integration between sectors (e.g. VET and higher education) and the reversal of funding cuts.¹³⁴

2. Prevent inappropriate exploitation of student data

Learning technology companies that deliver their products over the cloud are already harvesting and merging data from all institutions using their products to develop sophisticated large scale models of student behaviour. These emerging data markets could lead to substantial risks for students if identifiable data is traded among institutions and beyond - the permanent student record is something that would be of inordinate value to many organisations, but is this something that we should be encouraging? Educational institutions must negotiate acceptable terms with vendors, and establish best practice policies ensure that student data is retained in an identifiable form for only a specified period (e.g. while the student is enrolled or remains an active alum of the institution or for five years as per Human Research Ethics guidelines). Questions must be asked now about whether the student should have not just access to their learning data, but also control and potentially even ownership of this key 21st century resource.

3. Provide students with timely and considered support

As was discussed above, much work in the automated identification of 'at risk' students has helped leading institutions to target interventions to students who are of low socio-economic status, or first in family attendees of university. This work can both threaten the rights of those students to an education if poorly implemented (e.g. a student might drop out if told that they are likely to fail) or enhance their chances of success if used wisely (e.g. a student who would have left education could be encouraged to stay if a timely intervention is made). AIDA should only be considered as an assistant in these scenarios - humans must be left in the decision making loop.

4. Plan for the new burden of AIDA-based knowledge

As institutions learn more about student progress, this raises issues over the legal obligation to intervene. Recent work has considered whether universities are legally obliged to act in order to achieve 'effective allocation of resources to ensure appropriate and effective interventions to increase effective teaching and learning'.¹³⁵ However, staff workloads may limit an institution's capacity to act, apart from using semi/fully automated tools. Designed well, fully automated feedback can be extremely effective.¹³⁶ Poorly executed, it may be just as weak as the very limited feedback many students already receive from humans, it could even be worse, where it is factually misleading, or demotivating.

open.ac.uk/documents/policies/ethical-use-of-student-data; The University of Edinburgh, "Learning Analytics Principles and Purposes," 2017, <https://www.ed.ac.uk/files/atoms/files/learninganalyticsprinciples.pdf>; UK JISC: Joint Information Systems Committee code JISC, "Code of Practice for Learning Analytics," 2015, https://www.jisc.ac.uk/sites/default/files/jd0040_code_of_practice_for_learning_analytics_190515_v1.pdf.

¹³⁴ Select Committee on the Future of Work and Workers, "Hope Is Not a Strategy – Our Shared Responsibility for the Future of Work and Workers" (Commonwealth of Australia, 2018), https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Future_of_Work_and_Workers/FutureofWork/Report.

¹³⁵ Prinsloo and Slade, "An Elephant in the Learning Analytics Room," 46–55.

¹³⁶ Madeline Huberth et al., "Computer-Tailored Student Support in Introductory Physics," *PLoS One* 10, no. 9 (2015): e0137001, <https://doi.org/10/gffdfg>.

5. Prepare AIDA for the future of work and learning

A lifetime of learning (and now the GDPR¹³⁷) require ‘portable data’ as citizens transition between myriad learning/training platforms. This has huge implications for AIDA infrastructure, which must handle learners engaging with the university potentially many times over a lifetime, expecting that their ‘transcripts’ will be smoothly recognised and updated each time, and perhaps ported to and from the workplace to demonstrate competency and compliance with training regimes. Questions must be asked now about who will have access to this portable data, and what rights the data subject will have to access it. They might want to correct, amend, curate, and delete it – should this be allowed?

6. Involve stakeholders in the design process as early as possible

A long history of user-centered design shows that successful technology is distinguished by design processes that empower the diverse personnel who must eventually run and use the technology. Participatory/co-design methods bring the right stakeholders together, and give them a voice in conceiving the system, shaping prototypes from the earliest stages through to pilot testing. It is only through such methods that the unforeseen emergent properties of the whole human-computer system become apparent before full deployment.

In summary, we see the potential for AIDA to provide very significant benefits to students, helping them to become mature digital citizens who can engage in an increasingly datafied society in a critically aware and respectful manner. The European Union funded Learning Analytics Community Exchange (LACE) project conducted a *Visions of the Future project*¹³⁸ which maps out 8 possible futures for learning analytics, ranging from utopian ideals where all students ‘control their data’ through to dystopian ordeals where everything is tracked and the right to privacy is lost. What results from the potential of AIDA technology will depend heavily upon developing a population that is well informed enough, and empowered, to make sound decisions.

4.1.3 Critical roles for educational institutions

In this section, we consider the role of educational institutions more broadly in preparing society for AIDA and the dilemmas that it can create. In particular, we believe that *educational institutions have a critical role to play in raising AIDA literacy*.

Schools, colleges and universities must equip students (and indeed, staff) with the ‘functional literacy’ required to thrive as partners with AIDA: technology should be demystified, never taken for granted, and suitably transparent when questioned by different audiences.

AIDA is so pervasive that it is hard to identify subjects in which this should not be integrated. UTS, in addition to its formal degree programs, is already distributing AIDA learning resources for free, via Open Educational Resources and online courses.¹³⁹

137 GDPR Article 20: Right to data portability: <https://gdpr-info.eu/art-20-gdpr/>

138 Learning Analytics Community Exchange (LACE), “‘Visions of the Future’, Horizon Report,” Public Deliverable, 2015, http://laceproject.eu/wp-content/uploads/2016/02/LACE_D3_2.pdf.

139 For example, Kirsty Kitto and Simon Knight, “Journey through Data,” UTS Open, 2018, <https://open.uts.edu.au/datajourney.html>; Simon Knight and Kirsty Kitto, “What Does Facebook Know about You?,” UTS Open, 2018, <https://open.uts.edu.au/facebookknowyou.html>.

Three examples of what AIDA education might include are:

1. **Equipping the critical mind**

It is vital to develop critical perspectives on the emerging AIDA ‘infrastructure’, to puncture the tech vendor/media hype bubble. It might often feel as though AIDA is something that others design, or is too big to change, leaving us (schools; colleges; universities; students; educators; parents; leaders) as passive recipients. It is in fact up to schools and universities to shape AIDA infrastructure to serve their values, or others will do it for them. Educational institutions must lead the way by researching, training and teaching, shaping the policies that regulate their own behaviour, and that of societal entities, public and private, who influence education.

2. **Demystifying AIDA infrastructure**

AIDA is not magic. At every step in the process, from conception to deployment, people are making decisions about infrastructure at many different levels. Software and organisational design decisions always promote values, making the ethical dimension both inescapable, and something that we can all shape. We must teach all citizens to ask the right questions about how AIDA functions, from school students, to citizens at large. Scientists, scholars, policymakers and business analysts will increasingly sense the world through the (always distorting) lenses of computational models consuming (partial) data feeds from (imperfect) sensors. The public must be empowered to question automated recommendations, and taught how to see and act with knowledge and integrity.

3. **Cultivating ‘AI interpersonal skills’**

As young people and adults learn to partner with intelligent agents (whether the AI is a virtual assistant, encased as a robot or in some other form), it will be critical to learn a new set of ‘interpersonal skills’, analogous to those we cultivate for people, e.g. to judge an agent’s areas of expertise, trustworthiness, social and emotional awareness, and other attributes. This is how we calibrate our interactions with others: what and how we choose to share, and how we interpret others’ actions and advice.

In parallel, influential stakeholders who have the power to shape the AIDA infrastructure, such as educators, designers, lawyers and policy makers, need to be alert to technology ethics and its impacts. There is a huge opportunity to provide short courses for interested citizens, and executive/professional development, potentially with micro-credentials and fee based structures, but the wider role of universities in raising the AIDA literacy of all citizens should be considered and supported.

4.2 **AI & data analytics in the disability sector: opportunities and ethical issues - case study**

The advent of AIDA opens up previously unimagined ways of supporting independence and autonomy in the lives of people living with disability with innovations that are potentially cost effective and with implications on how disability support is provided. But the answers are not as simple as they may seem, and there are of course complex ethical considerations that need addressing.

People with disability continue to face physical, social, financial and political barriers. These barriers continue to prevent people with disabilities from enjoying their right to employment, education, health and autonomy. How might artificial intelligence help some of these barriers to be broken down?

4.2.1 Opportunities to improve accessibility across platforms

One of the advantages of the advent of AI in technology products is the opportunities to improve accessibility for greater audiences. One of the great opportunities for AI is that once developed, there is great potential for the benefits of improved accessibility to reach global populations across developed and developing nations, including places without service and support.

One of the most important ways that AI can support access and inclusion of people with disability is to increase the ways a person can communicate, and receive and input and information. Two examples of AI converting communication formats are speech to text and also image to text.

- Speech to text software has meant that deaf and hard of hearing audiences are included in the experience of watching film, TV and video. Google's machine learning technology has enabled YouTube to offer speech to text software automatically captioning speech in videos since 2009. More recently it rolled out algorithms that indicate applause, laughter and music in captions.¹⁴⁰
- Alt text is important meta-data that accompanies a photo or diagram that explains its contents in words; it ensures that people who are blind or have low vision are able to experience the content of a photo or image. Facebook recently launched a feature that automatically creates text descriptions of images – and can be shared with friends.

4.2.2 Imagining and realising new assistive technologies through co-design

As we continue to search and develop new assistive technologies, striving to break down barriers faced by people with disability, there is only one way forward – Inclusive Design. In order to create intelligent solutions that are insightful, ethical and that really do transform lives we need to be looking at research and development practices that are inclusive and diverse and include diverse decision makers, from the word go.¹⁴¹

4.2.3 The impact of AI on independence and disability support

AI has the potential to increase levels of autonomy and independence in the daily lives of people with disability. Where once a blind person might rely on a sighted companion to describe or help with navigation or identification of places and things, AI technology has the capacity to use in-built cameras in devices to 'see' and then identify and describe people, scenarios, and groceries. Not having to rely on a person to physically be present and help you to make decisions directly impacts levels of autonomy in the lives of people with disability.

4.2.3.1 Personalisation and diversity

AI has the potential to accommodate for uniqueness in the design of assistive technology. For example, for those who are unable to use text input devices such as keyboards and mouse, voice to text software has opened up new and accessible ways to input information and navigate device interfaces. Historically these voice to text software have taken time to set up

140 Tom Simonite, "Machine Learning Opens Up New Ways to Help People with Disabilities," MIT Technology Review, 2017, <https://www.technologyreview.com/s/603899/machine-learning-opens-up-new-ways-to-help-disabled-people/>.

141 Cecily Morrison et al., "Imagining Artificial Intelligence Applications with People with Visual Disabilities Using Tactile Ideation," in Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ACM, 2017), 81–90.

for individual voice, and starting afresh with new software can be a slow and frustrating time while the software 'learns' your unique voice. However continued development in AI and voice recognition has meant that software can more quickly learn and adapt to individual voices and unique patterns of speech further refining and improving the efficiencies of the way we input data.

These are technologies that again, are not only beneficial for people living with disability. These innovations will benefit a wide sector of the population and increase the data input and navigation options for all.

4.2.3.2 Artificial intelligence and intellectual disability

AI Technology can be harnessed such that information can be presented in different ways and understandable terms. There is an opportunity for AI to re-package complex information into easy read information for people with intellectual disability.^{142,143} This easily read information will also mean better access to information for people with low levels of literacy and English as a second language. Potential applications include legal documents, terms and conditions for software usage, manuals and instructions.

4.2.3.3 Data analytics

Many AIDA applications promise more freedom for people with disability as advances in machine learning and cognitive systems mean presenting people with new ways to see, hear and communicate. But most importantly, if people with disability are to benefit from AI they must be included in data capture. This highlights to importance of awareness of the ways people with disability might be excluded from data capture, for example in smart city applications, for reasons such as not being able to afford devices, not being able to access the city, not feeling welcome in certain areas.

4.2.3.4 Risk: coercion and data collection

Where a person with disability relies on a particular assistive technology to improve their independence, will they be coerced into providing data against their will? One risk is that people with disability will feel forced to share data they might not otherwise, but a reliance on a particular technology will mean they face coerced.

4.2.3.5 Risk: do not assume improving accessibility means improved access

It is important to recognise that access to new assistive technologies requires access to devices, and also the ability to maintain and upgrade software as required. This means that those who cannot afford devices or maintenance costs will not benefit from the benefits that new AI technology can bring.

4.2.3.6 Risk: surveillance technologies and people with disabilities

People with cognitive and psychosocial disabilities are overrepresented in criminal justice systems, including in prisons. Over the past decade there have been significant developments in the legal and service frameworks facilitating the punishment and control of people with cognitive and psychosocial disabilities in community settings (including disability group homes or through disability case management). In some circumstances people with such disabilities might avoid imprisonment but instead be placed under criminal law, forensic

142 David Grigoryan et al., "Creating Artificial Intelligence Solutions in E-Health Infrastructure to Support Disabled People," in International Conference on Computational Science and Its Applications (Springer, 2018), 41–50.

143 Ping Chen et al., "Automatic Text Simplification for People with Intellectual Disabilities," Artificial Intelligence 10 (2016): 9789813206823_0091.

mental health or guardianship orders that enable their control in the ‘community’ or at other times this control might fall outside of specific legal orders and instead be embedded in the service-service user relationship. Surveillance technologies (e.g. ‘smart homes’ and movement trackers) provide opportunities for control which do not involve the same levels of physical confinement and intervention as locked doors or physical and chemical restraint. However, this raises important questions about the risks and potential impact on the human rights of individuals including the following:

- What framework will there be for consent to surveillance?
- How will individual use of these surveillance technologies be regulated (if at all) through guardianship and mental health legislation?
- How will these surveillance technologies infringe on rights to equality and non-discrimination, privacy, community inclusion and liberty?
- Who will own the data generated by these surveillance technologies?
- What rights will residents who are the subject of these technologies have to access this data, control the uses of the data (including use by services in criminal justice processes or legal proceedings?)

4.2.4 Recommendations

1. People with disability must have access to, and control over, the data that is retained or not retained about them; they must not feel coerced into sharing data in exchange for access to assistive technology.
2. AI be harnessed to ensure that written and spoken information in all forms be translatable into easy read for people with intellectual disability.
3. Despite the many opportunities for inclusion that AI presents, it is important to acknowledge the pathways to exclusion that people with disability face, that often lie outside the technology itself.

4.3 Health, AI and intellectual disability - case study

4.3.1 Background

Intellectual disability is a general category of disabling conditions representing a very wide range of people and conditions, including complex, multifactorial scenarios. Conditions of intellectual disability can arise at birth and throughout the life course. One of the key representational factors is the sheer diversity of forms of ID and their varying impact on individuals, and their interactions with support, care and health systems. The focus in this case study is people with an intellectual disability (PWID) and their interactions with health care systems, a known area of inequity and highly differential outcomes for PWID. Healthcare has the same time a poor history of servicing PWID appropriately and is an environment in which the adoption of new technologies is both popular and variable. Oversight of such applications to PWID and their individual situations will be crucial.

The role of AI as a rapidly developing domain of both theoretical and practical application in many industries has particular resonance in healthcare environments. Vulnerable groups and communities have had and, in many cases, continue to have variable outcomes compared to the general population within the healthcare. For example, premature deaths in the comparator group were largely due to lifestyle factors, whereas those for people with learning

disabilities were mostly due to delays or problems with investigating, diagnosing, and treating illnesses and with receiving appropriate care.¹⁴⁴

In this context, we present this case study as an intersectional one examining the potential positive and negative contributions of AI to the intellectual disability environment and the opportunities and risks for PWID. Traditional top-down models of medical and wider healthcare technology design and implementation would be inappropriate in providing beneficial outcomes for PWID.

4.3.2 Opportunities

The current literature indicates four main areas in which ID is poorly addressed by existing healthcare policies, practices and procedures.¹⁴⁵ In the context of opportunities for AI to improve the healthcare environments and experiences of PWID, we explicitly add diversity and context to this mix.

4.3.2.1 Analytics and AI

Quality AI applications have the capacity to analyse and adapt both analytics and service interventions (e.g. chatbots) to the circumstances of clients and service providers to support improved engagement with and feedback from PWID. This is especially important in situations where cognitive and communication impairments may be present and can extend the options for improving successful outcomes across complex care environments. Properly designed AI supports do not exhibit the stereotypical social attitudes and prejudices that healthcare providers may possess in their engagements with PWID, what Dourish and Bell (2011) describe as the worldview inherent in any technology.¹⁴⁶

4.3.2.2 Organisational development

Healthcare environments are multiple, complex and often poorly connected. This can leave clients and their supports negotiating a complex system on their own merits with limited quality assurance and follow-up. AI applications have the potential to inform and support multi-system complexity and quality assurance potential for PWID.

4.3.2.3 Workforce

Expert knowledge is a key feature of successful engagement with and service provision for PWID. Such expertise often translates poorly outside of specific service environments. AI systems can support and/or compensate for the unequal distribution of expert knowledge helping to improve consistency and quality of outcomes for PWID.

4.3.2.4 Resource allocation

AI has a potentially significant role to play in areas such as: communication technologies for PWID; decision-making tools; resource-allocation processes; and the provision of safer environments and processes (place-making). These areas can be improved by better data analysis and complex, adaptive systems approaches enhanced by AI techniques.

144 Sheila Hollins and Irene Tuffrey-Wijne, "Meeting the Needs of Patients with Learning Disabilities," *British Medical Journal*, 2013, 1, <https://doi.org/10.1136/bmj.f3421>.

145 Joanne Travaglia, Deborah Debono, and Georgia Debono, "Capacity Building and Intellectual Disability Health Services: An Evidence Check Rapid Review Brokered by the Sax Institute (Www.Saxinstitute.Org.Au) for the NSW Ministry of Health" (Sax Institute, 2017), <https://www.health.nsw.gov.au/disability/Documents/evidence-check-cbidh.pdf>.

146 Paul Dourish and Genevieve Bell, *Divining a Digital Future: Mess and Mythology in Ubiquitous Computing* (MIT Press, 2011).

4.3.2.5 Diversity

As noted earlier in this case study, PWID are an extremely complex group of individuals – easily as complex as the general population. In addition, ID frequently intersects with other health conditions, making complexities of treatment and care central to successful health outcomes. We see AI as offering major capacity for identifying and intervening in the specific, individualised requirements of PWID so that health outcomes can be optimised.

4.3.2.6 Complexity

Human beings are not especially good at identifying or managing complexity, particularly in dynamic settings like acute healthcare. PWID commonly experience the consequences of this limitation in healthcare environments. AI offers a range of opportunities to improve complexity management, risk/harm mitigation and quality of outcomes improvement for PWID, if appropriately applied.

4.3.3 Risks

Generally speaking, the potential for AI to improve healthcare provision for complex PWID clients is the positive side of the risk equation coin in healthcare. Risks from AI applications remain constant for vulnerable groups in general but the complexities experienced by PWID make for additional risks in healthcare environments.

4.3.3.1 Programmed inequalities

Technological systems, including hardware and software environments, are designed to produce particular outcomes. Poorly designed AI interventions, or ones influenced by political or ideological agendas, have the capacity to reinforce existing inequalities, as is well-documented in the broader AI literature. PWID are a group where ethical and human rights perspectives, in addition to co-productive strategies, can inform AI-based healthcare interventions.

4.3.3.2 Category effects

With human beings, knowledge about particular groups is often assumed based on existing/prevaling social category productions and the values, positive or negative, attached to such groups. PWID have a history of poorer outcomes in healthcare environments, at least some of which is predicated on social category attributions and resulting effects. Positive AI implementations can support improvements in this space, and negative or poorly designed ones will likely exacerbate existing inequities.

4.3.3.3 Social categories

PWID represent both a complex, multidimensional social category and group of people, with equally complex and dynamic healthcare needs. Positive AI can improve the analysis and intervention in the needs of ‘categories’ of PWID, while negative AI implementations may even constrain improvements in this area by providing a ‘scientific’ (uncritical) basis for poor quality interventions, resource rationing, finite communication options and so on.

4.3.4 Recommendations

AI represents a mixed environment for healthcare interventions for PWID. The historical and present scenario has been very poor for this complex and diverse group. They are proven to be at greater than average risk in contemporary healthcare environments. Potential AI applications offer a variety of potential contributions to improving the scenario that currently exists but engagement with PWID and their key supports, including family and direct care providers, will be key to successful outcomes. The recommendations here aim at ensuring

AI-based interventions in healthcare have the potential to improve care and quality outcomes for PWID.

1. That explicit reference to the experiences of PWID in healthcare is included in any AI applications in healthcare environments, including direct (e.g. communication app, pain assessment tools etc.) and indirect (e.g. risk assessment tools, electronic medical record analyses) interventions.
2. Wherever possible, co-design, pre-test and post-test practices with PWID inform AI project implementations. The issue here is that such inclusion is not simply tokenistic but completes the design-implementation loop in every case.
3. That healthcare systems are proactive in using AI methods to overcome existing inequities and service provider limitations (e.g. prejudice and discrimination) for PWID and related vulnerable groups. Virtue statements are common in healthcare but quality of outcomes represent an objective measure which AI strategies can work with for improvement.
4. That representative strategies are developed to ensure PWID, their carers and key supports are included in healthcare system AI implementation plans. The research base for PWID remains insufficient to assume healthcare innovation will adequately represent the needs of PWID without formalised, accountable representation.

4.4 AI & Indigenous data: managing data ethically - case study

There are many opportunities for Indigenous Australian peoples to engage with AI and new technologies to support the realisation of self-determination. As well as the broader opportunities relating to health and education, AI could enable dynamic possibilities for Indigenous peoples to transmit stories, language and culture through animation and augmentation. The management and use of Indigenous data needs to be considered within an 'Indigenous Data Sovereignty' framework in order manage data with appropriate ethics and recognition of Indigenous Intellectual and Cultural Property rights. Indigenous decision making and participation in the management of data is critical to ensure the protection of human rights.

Indigenous Australian peoples have had a complex and troubled relationship with research and data collected by government, churches and other private organisations. Historically, information and archives – which can now be conceptualised broadly as data – have been used as tools and apparatus to support the intervention and control of Indigenous lives. Indigenous rights in data are fundamental human rights, this includes being able to have a say in how data is created and captured and managed through time, as well as ways in which people access, use or re-use historic data sources. Indigenous Australian people continue to face barriers in terms of access to information.

4.4.1 Opportunities for Indigenous self-determination and data

The concept of Indigenous Rights in Records¹⁴⁷ and Indigenous Data Sovereignty¹⁴⁸ frame aspirations for Indigenous Australian people to be in control of data, both current and historic, that relates to knowledge or information about themselves and their communities. Given the complex interactions and role of data as an instrument of control in the past, it is important that any Indigenous data creation, management or use is managed and supported by Indigenous decision making.

The following concepts offer opportunities for Indigenous people to be involved in decision making in regards to the use of Indigenous data in AI.

4.4.2 Indigenous data stewardship

The concept of ‘data stewardship’ encompasses the creation, care, management, preservation and use of digital data (including born digital, digitised materials, and research data). Indigenous Data Stewardship concerns the management and organisation of Indigenous data including managing and organising materials, describing and preserving materials, and providing access to materials. It also considers ongoing care and management of materials over time through archiving, ensuring that rights are managed over the long-term, or when data is no longer active.

Indigenous Data Stewardship enables opportunities for metadata to be captured on Indigenous Cultural and Intellectual Property, to include areas such as attribution and terms and conditions of use and re-use of materials. Examples of this could be in relation to use of Aboriginal languages in AI settings, or use of Aboriginal stories in digital animations or augmented reality. Indigenous Data Stewardship ensures that data is not removed from people and communities so that source data is accessible by its creators, so that any attempts for material to be used or reused is managed through ongoing informed consent.

4.4.3 Indigenous data repositories

Use of Indigenous Data in AI projects should be relational, with materials being held and managed in Indigenous data repositories. The opportunities for new technologies to leverage Indigenous Australian digital cultural heritage relies on the development of repositories that can both manage rights, and appropriate access to materials. For example, it may be the case that some materials are identified as being appropriate for public use and others may be closed for community access only. The development of Indigenous data repositories would enable reuse of materials through differing permission structures, whilst at the same time securing and caring for digital data over the long-term. A data repository could assist the protection of cultural objects in digital format over the long-term. There is currently a lack of infrastructure to support Indigenous data repositories and archiving.

147 See summary related to ‘Statement of principles: Australian Indigenous knowledge and the archives’ in, Shannon Faulkhead et al., “Australian Indigenous Knowledge and the Archives: Embracing Multiple Ways of Knowing and Keeping.,” *Archives and Manuscripts* 38, no. 1 (May 2010): 27, <http://search.informit.com.au/documentSummary;dn=201007444;res=IELAPA>.

148 The Australian Indigenous Data Sovereignty Collective Maiam nayri Wingara provide useful definitions and statements regarding priorities in this area: The Australian Indigenous Data Sovereignty Collective Maiam nayri Wingara, “Key Principles for Indigenous Data Sovereignty,” Maiam Nayri Wingara, 2018, <https://www.maiamnayriwingara.org/key-principles/>.

4.4.4 Historic Indigenous data for AI use

Many national and state collecting institutions hold materials that are being digitised and made accessible online. Other published sources are being transcribed and made accessible as datasets for reuse by new technologies. The success of any AI project relating to Indigenous cultural heritage materials or data, relies on the quality of the data that is being produced.

There are opportunities to redress historic inaccuracies or silences in data that have been created relating to Indigenous Australian peoples. Mechanisms such as a 'right of reply' or ability for Indigenous people to respond to data that has been misinformed, or not managed through appropriate ethics processes. Any AI developments should acknowledge this historic legacy and the ongoing impact of incomplete data in relation to Indigenous engagement. This is of particular relevance in cases where machine learning technologies are curating content that is inaccurate.

4.4.5 Digitisation and imaging of Indigenous cultural heritage materials

Not everyone wants to have their cultural heritage materials material open for re-use, this is of particular concern in relation to the care and protection of objects that are being digitised and imaged into digital format. Many national cultural heritage institutions are digitising and making materials accessible online, and there are potentials for this imaging to be enhanced to enable reproduction of materials in both digital and physical forms, for example through animations or 3D printing. Any successful repurposing of material needs Indigenous participation and decision-making. This relates to questions of even deciding what is appropriate for digitisation and public display.

4.4.6 Risks

- Rebirth of historic injustice through AI systems through use of biased data
- Indigenous Data has complex and troubled historic roots which impacts the potential for ongoing racial bias in AI systems. Use of AI in these contexts may reaffirm pre-existing bias.
- Indigenous people are not often in control of all of their data that relates to them. Without appropriate return of data, and infrastructure for long engagement there is risk that Indigenous people continue to be marginalised in these spaces.

4.4.7 Recommendations

1. Indigenous Australian people must have access to, and control over the data that relates to themselves and their communities by establishing a focus on Indigenous Data Stewardship.
2. Contemplation of historical narrative used during original collection of data which informs AI must occur to avoid historical bias reoccurrence.
3. Infrastructure be developed to enable Indigenous people to build self-determination around the stewardship, archiving and reuse of data in AI projects.
4. Mechanisms be developed to enable redress of historically inaccurate data sources relating to Indigenous Australian people, including management of materials that are inaccurate or which have been managed without appropriate ethical frameworks.

5 Recommendations

In Section 3 we offered answers to the ten questions posed by the AHRC in its HRT Issues Paper. However, some of the issues that new technologies raise for human rights cannot be easily raised or discussed within the framing of those particular questions. For this reason, in Section 4 we presented a number of case studies to draw out some of those other issues in specific contexts.

In this final section we shall argue that to identify and properly engage with the full range of issues that new technologies potentially raise for human rights – that is, the issues and technologies we discussed in Section 3, as well as the ones discussed in Section 4 – a transdisciplinary approach is needed. We will also recommend that for an ambitious and important project like the AHRC’s initiative on human rights and technology to succeed, the Australian government needs to set up a Technology Assessment Office (TAO).

As we have argued, a broad approach is required to conceptualise technology and its impacts (see Questions 1 and 5). There is a considerable amount of legal and regulatory uncertainty, therefore further research is required to understand the impacts of technology on rights (see Question 3). In our response to Question 2, and the associated case studies in Section 4, we have argued that in order to realise the potential of technology to promote human rights, stakeholder engagement is required, with a broad understanding of the stakeholders impacted by any technology, ensuring that engagement is accessible to stakeholders including via education. The positive potential of technology to promote human rights will be supported by an explicit targeting of positive outcomes (rather than simply regulating negative outcomes).

We argued (especially in our response to Question 4), that human rights and transdisciplinarity are needed at the technology design stage. By this, we mean that in designing and deploying technologies:

- a broad range of stakeholders should be involved through distributed shared agency (i.e., that agency should be distributed and shared across the involved stakeholders).
- that the complex ecosystem of technology should be recognised, as discussed in detail in response to Question 1.
- that, as discussed in Section 2, various kinds of uncertainty – including regarding the impacts of technologies – should be recognised, and as such an ongoing iterative process of evaluation and decision-making should be adopted.
- that dialogue should be fostered between key stakeholders, in neutral spaces such as universities, in order to develop distributed shared agency.
- that education is central in equipping citizens with the new literacy that is required to live and work with emerging technologies.
- and finally that work is required to develop ethical and philosophical frameworks for the broad range of stakeholders to work with and understand the impacts of technologies.

As we outline in our response to Question 6, regulatory approaches should supplement and build on this transdisciplinary design approach, drawing on experience from other jurisdictions and the work of non-government bodies in this space. Indeed, as we discuss in our response to Question 7, future regulatory development should adopt the very transdisciplinary approach described above. As we draw out in our responses to Q8-10, the impact of technology on human rights has significant potential for people with disabilities, which the approach we describe here may help tackle.

In the following section (Section 5.1), we therefore provide a more detailed explanation of the benefits of a transdisciplinary approach to developing approaches to understand and manage the impact of technology.

And in Section 5.2 we propose that a good way to advance the AHRC's aims is for Australia to set up a Technology Assessment Office (as introduced in our response to Question 3), and we sketch some of the functions that a TAO would perform and some of its features.

5.1 Taking measures to protect human rights: regulation and value sensitive practices

In this section we particularly draw attention to the recommendations highlighted in our response to Questions 3, 4, and 7, regarding approaches to regulation and practice that would protect and promote human rights in the development, use, and application of technologies including in the context of AI informed decision making.

Once we have predicted and evaluated the effects of a new technology – our foregoing concerns about technology, prediction, and evaluation notwithstanding – we have three options vis à vis taking measures: (i) to directly regulate the technology, (ii) to use an approach like Value-Sensitive Design (VSD) to ‘bake’ important values into that technology at the design stage so that it protects important human interests, and (iii) to mandate through regulation that designers must employ VSD. We may also use a combination of these strategies. Below we briefly discuss these three options.

5.1.1 Regulation

Regulation can be mischaracterised in two important ways. One, that it involves a regulator – i.e. someone other than us – exerting control over us. Two, that it involves either prohibition or permission. When regulation is characterised like this, it generates polarized reactions. On the one hand, sometimes regulation seems necessary to protect something important however on the other hand coercion can result in unnecessary red tape and inhibit innovation. Such polarised reactions do not help gather public support for regulation, and so our aim in this section is to dispel this unhelpful characterisation by mentioning some of the fine-grained texture that regulation can have – texture which regulators should reveal to the public to avoid resistance to regulation – and that it can also be gradual and revisable. In the conclusion of this section, though, we will suggest that a more fine-grained approach to regulation can help address some of the predictive and evaluative challenges discussed above. In what follows we will again employ the smart drugs and CRISPR Cas-9 gene editing examples to animate our discussion.

In light of the concerns about these two new technologies, our first point is simply that the real question is not whether gene editing, or smart drugs, or any other new technology should be allowed or banned, but rather how it should be regulated, since prohibition and permission are just two modes of regulation on an axis that includes at least the following:

prohibit – discourage – permit – encourage – require

Second, as the AHRC notes, in addition to primary, subordinate, or delegated regulation, especially in regards to the encouragement and discouragement modes, tax (dis)incentives can also be very effective ways of regulation in addition to legislation. Other approaches include professional/industry monitoring, oversight, and accreditation organisations, often coupled with explicit codes of ethics.

Third, there is also scope for variety in regards to precisely who the subjects of regulation might be. That is, who should be permitted, or encouraged, required, discouraged, or prohibited? Are we talking about scientists, physicians, the general public, or someone else entirely? Might we not encourage (through research grants) – at least for a while – scientists to conduct research into gene editing technologies, while prohibiting everyone else from doing likewise? Different kinds of regulation can apply to what goes on in research laboratories, in doctors' offices, and in people's homes.

Fourth, who should do the permitting, encouraging, requiring, discouraging or prohibiting? Government technocrats? Scientific organisations? Corporations that market gene editing technologies? Citizens via referenda? Nobody likes to be told what they can and cannot do by someone else. However, even the unhelpful view of regulation as something that bureaucrats, technocrats, or anybody else imposes on citizens can helpfully be dispelled by ensuring that regulators engage the public in deliberations. Or, if what's needed is an expert panel, that the need for the expert panel is something that the public understands and supports, in the same way that we turn to experts for advice about such things as what medical treatments to use for ailments. This helps make it clear that regulators' decisions are not something imposed from the outside but merely something that the regulator does on behalf of the public.

Fifth, and lastly, there is nothing to prevent regulations from being revised. For instance, a new technology could in theory initially be released to a small portion of the population in order to gauge its effects. Subsequently, if its effects do not seem troubling, regulations could be relaxed, and again the effects monitored.

5.1.2 Value-sensitive design (VSD)

The values embedded in specific emerging technologies may be partly intentional (e.g. increased productivity, accuracy, efficiency) and partly unintentional resulting from the culture of the company or organisation developing the technology and the values and worldviews of the individuals designing, building and implementing the technology.

As referred to earlier in this paper, there is growing complexity in the nature of relationships between intelligent agents and human agents (e.g. autonomous vehicles, carer robots, AIDM). Questions of values and ethics are closely intertwined with the development of these new forms of relationships and require new theoretical approaches to understand them that move beyond more traditional concepts of rights-based or merit-based approaches. While in the past it may have been possible to determine responsibility with regards to technology use by resort to notions of legal liability or informed consent, these approaches become less useful when intelligent agents and autonomous technology are involved. A complex combination of causal factors and moral responsibility will increasingly focus attention on the need to carefully consider values in the design, implementation and maintenance phases of technology development.

Values are often seen to be embedded in technology during the design phase and recent cases of perceived bias in AIDM such as the ProPublica analysis of the COMPAS highlight the potential impact of bias in the design, testing and iteration of AI tools.

A range of design approaches have been advocated to improve inclusivity and representation of the end users of technology in order to improve the outcomes for them including inclusive design, universal design, participatory design and co-design. While these approaches can

significantly improve the outcome and experience for end users of the technology it is still difficult to engage with the complex ethical and moral questions which emerging technologies such as intelligent agents raise.

As intelligent agents become increasingly autonomous in their interaction with human agents – for example, care bots, sex bots, smart homes, smart cities, and autonomous vehicles – the incentive to design these agents proactively to create the most benefit and least harm becomes increasingly important. One approach that is being advocated for this is Value Sensitive Design which takes into account important human values such as privacy, accountability, equality, and sustainability during the early stages of design. This approach, developed initially by Batya Friedman, treats human values and legal requirements as being on a par with other technical specifications at the design stage of technology. This makes it possible for values and legal aims to be ‘baked in’ to technologies, rather than needing to regulate how a technology that doesn’t contain those values may be used.

Value-sensitive design relies on three interdependent methodologies being applied throughout the design, testing and implementation of a new technology which are conceptual, empirical and technological investigations. The conceptual evaluation relates to philosophical and ethical considerations of the potential impact of a technology both positive and negative, intended and unintended. The second is an empirical investigation which explores measurable, quantifiable effects of proposed technology which may be similar to a user-centered design approach engaging with the range of potential users and stakeholders of a product early in the design phase. The third is a technological investigation involving the actual materials and nature of the technology and brings a values focus to this within the design phase. These three approaches are then combined under an umbrella of ‘universally held’ values to enable engineers and ethicists to ‘front-load’ values into the design of new technologies. It also provides a platform to receive feedback from stakeholders and potential technology consumers early in the design process.

Other fields such as ‘data humanism’ and ‘data advocacy’ are gaining traction and also have important contributions to considerations of how values are embedded in technology and the expectations regarding the stewardship of data in a civil society.

5.1.3 Combined regulation and VSD

It is also possible to combine regulation with VSD. For instance, to provide incentives in the form of tax breaks to technology producers to send their employees for training in how to deploy VSD in their work, or to legislate that universities include VSD and ethics courses in their curricula to ensure that the future developers of technology understand both the normative dimensions of technology – that technologies are infused with values – and that these values can be intentionally designed into technologies.

5.1.4 Conclusion

An important upshot of our discussion in Section 2 of this submission, and throughout our responses to the questions is that unless effective ways are found to tackle the challenges involved in prediction and evaluation of new technologies’ effects, then we will have poor quality data on the basis of which to make informed decisions about what measures are needed vis à vis regulation and design to protect human rights from or by new technologies.

This point applies equally regardless of whether a regulatory or a design-based approach is taken, since in both cases what we will need is to know whether there is even a reason to take any measures in the first place. Put another way, if we don’t know what effects new technologies might produce, and/or if we cannot figure out if those effects would be detrimental or beneficial to human rights, then any measures taken to protect human rights would not be adequately evidence-based.

However, a distinct advantage of viewing regulation in the fine-grained way that we discussed in Section 5.1 above, is that some of the predictive and evaluative challenges discussed earlier could be eased. For instance, rather than attempting to predict all of the temporally-distant and often difficult to imagine soft impacts, if a technology were deployed slowly and gradually to an isolated part of the population, this could make it possible to observe some of the technology's effects on that population's values and social arrangements without having to imagine or predict them. It could also enable the rest of society to engage in a dialogue with that part of the population to debate the merits of the changes without themselves having had their values altered by experiencing that technology's effects. As a form of social experimentation, however, this approach would itself require oversight to ensure that ethical and human rights concerns were not breached.

5.2 Australia needs a Technology Assessment Office (TAO)

As we explained in Section 2, technology assessment is a very complex and involved task which requires the resources of a government organisation. What Section 5.1 added was that this task needs to adopt a transdisciplinary approach in order to recognise the many sources of nuance and complexity. The selection of risky technologies, prediction of their effects, evaluating those effects, and devising measures to protect human rights, are tasks that need to be approached holistically. This is why for an ambitious and important project like the AHRC's initiative on human rights and technology to succeed, the Australian government needs to set up a Technology Assessment Office (TAO).

5.2.1 A demanding task with high-stakes

Although technology is not the only factor in the equation, it is nevertheless a very prominent contributing element behind the important issues outlined in the AHRC's HRT Issues Paper. The challenges involved in technology assessment (TA) such as predicting and evaluating the effects of technologies, and then deciding what measures should be taken vis à vis human rights, are indeed very steep. However, just because the challenges are steep, that does not mean that we should do nothing about them, since failing to exercise due diligence in this regard can cost exponentially more. It might sound dramatic to claim that society has been reckless with its attitude towards how technologies have been designed, deployed, and monitored in the past. However, progressing in the manner that we have to date would certainly be nothing short of careless. Considering our better appreciation of the risks involved, and our knowledge that there is something we can do about those risks, failing to take adequate measures now would be an extremely irresponsible way of moving forward. Even if we set aside the potential benefits of taking a proactive or promotive stance in regards to reasons for taking an active role in driving technological innovation, from a purely reactive or protective perspective we have an obligation to do better in regards to how we go about designing, deploying, and monitoring technologies than what we have done to date.

Given the challenges involved in predicting the unexpected effects of new technologies, in properly evaluating those effects, and in monitoring and regulating how technologies are designed, deployed, and used, this is simply not a task for any single technology-related entity. Rather, this is a task for an independent government-sponsored or government-affiliated body that works in conjunction with industry partners, academic institutions, domestic and international regulatory bodies, and most importantly the stakeholders involved — most notably, the public, taking special account of vulnerable or at-risk groups and individuals.

A comprehensive, cross-sector approach is beneficial in many ways, three of which are identified here. Firstly, an independent body is well positioned to investigate a full range of potential interconnected consequences that may extend beyond the effects of a single product or domain (e.g. medical, communications, transportation, etc.). Secondly, and related

to the first point, self-regulation requires producers of technology to hold concern about the social-fabric, the global and local political, economic, technological, legal, cultural, and environmental impacts that extend beyond the boundary of the product or service that they offer. Thirdly, and finally, a decisive benefit of an independent body would be its enriched capacity, specialised resources, and focused skills for the task at hand. This is especially true of a body that is charged with the task of asking challenging questions, given the authority to make observations and recommendations that an entity subject to political or commercial fallout may not always find easy to make.

Private sector entities, compelled by an economically driven and competitive landscape, have insufficient incentive to be proactive in this space. Smaller operators have limited in-house knowledge and resources to respond effectively. By comparison, an independent regulatory body would be well positioned to step beyond assessment, and to use its knowledge and resources to also educate and involve the community in active co-creation of a vision of the future.

5.2.2 A valuable and rewarding task

Returning now to the proactive or promotive aspect of what our transdisciplinary approach includes, in addition to TA being a challenging task, it is also an immensely valuable and rewarding task. Not only because important human interests need to be protected from potential technological threats (that expresses reactive or protective concerns), but also – stated more positively – because a well thought out approach to managing technological innovation can promote important human interests and help shape the future of society in a direction that supports a country’s specific needs, strengths, values, and aspirations. When done properly, TA helps to coordinate different arms of government and industry, to set goals, to monitor progress, and to adjust with the changing times and circumstances as may be required.

For the public to be well prepared to contribute to designing positive visions of the future, to creatively imagining and exploring possible new ways the world could be, it is imperative for the TAO to provide the right educational and involvement opportunities. Creativity, innovation and technology entrepreneurship must be underpinned by an education of citizens that enables them to envision a range of possible futures.

When done right, TA can be an incredibly valuable and rewarding enterprise – not just for those working at the TAO, but for the whole society – as well as an incredibly important duty. But in order to derive these benefits and to avoid the dangers, our country needs a TAO. At present, however, Australia lacks a TAO.

Since the 1960s governments around the world have embarked upon setting up technology assessment offices, bureaus, and institutes. David Banta, Emeritus Professor at University of Maastricht in The Netherlands who has written extensively on the topic of TA, writes:

The term ‘technology assessment’ came into use in the 1960s ... in the United States, focusing on ... the implications of supersonic transport, pollution of the environment, and ethics of genetic screening. The term is said to have first been used in the Subcommittee on Science, Research, and Development of the House Science and Astronautics Committee of the U.S. Congress [which] in a series of hearings and reports, examined issues surrounding technology and proposed technology assessment as an approach to problems surrounding technology, its development and use. [I]t was defined as a form of policy research that examines short- and long-term consequences (for example, societal, economic, ethical, legal) of the application of technology. The goal of technology assessment was said to be to provide policy

makers with information on policy alternatives ... The main accomplishment of the years of work of the Subcommittee was the establishment of the U.S. Congressional Office of Technology Assessment.¹⁴⁹

Under the Reagan administration, scathing criticisms of the Office of Technology Assessment begun to emerge – it was deemed unnecessary on account of duplicating government efforts – and eventually under President Clinton’s administration in 1995 it was disbanded. (ibid p. 8) This decision is consistent not only with an ultra-minimalist approach to regulation favoured in the U.S.A, it is also consistent with a much more general disinclination to entrust government offices with making policy founded on substantive evaluations. This analysis is also supported by the fact that the Food and Drug Administration was not disbanded – since, after all, regardless of taste or political inclinations, in matters that concern basic bodily function and a narrow rendition of ‘health’, all humans share the same interests.

Since the 1970’s in Europe and the UK a total of 17 TA institutions have been created, and more recently have come together under the auspices of the ‘European Parliamentary Technology Assessment’ office.^{150, 151} While the challenging nature of technology assessment is also among the reasons why, more recently, these offices, bureaus and institutes have engaged in international collaborations focused on identifying priority areas, setting directions for the future, and related agendas regarding future technological innovation, design, and regulation. For Australia to have a say on this international TA front about the direction in which technology develops, and to promote the interests of Australians in a distinctly positive way, as well as to protect our interests, our key recommendation is that to advance the AHRC’s initiative on human rights and technology the Australian government needs to set up a TAO.

5.2.3 Functions of a TAO

Ideally, an Australian TAO would perform a range of functions that support both the protective aspirations set out in the AHRC’s HRT Issues Paper, as well as helping to shape the future of technological innovation and design to support Australia’s distinct needs and strengths, as well as Australians’ values and aspirations. Below we have listed in point form and grouped under heading and sub-headings the main functions of a TAO:

- Targeted initial and periodic (re-)assessment of:
 - specific technologies
 - specific technology companies
 - specific technology applications/uses
 - certification and periodic re-certification

149 David Banta, “What Is Technology Assessment?,” *International Journal of Technology Assessment in Health Care* 25 Suppl 1 (July 2009): 7, <https://doi.org/10/bmpn3t>.

150 European Parliamentary Technology Assessment (network) (EPTA), “Parliamentary Technology Assessment in Europe. An Overview of 17 Institutions and How They Work.,” EPTA Booklet – Policy Briefs & Reports – EPTA Network (European Parliamentary Technology Assessment (network) (EPTA), 2013), <http://www.eptanetwork.org/database/policy-briefs-reports/218242-epta-2013-parliamentary-technology-assessment-in-europe-an-overview-of-17-institutions-and-how-they-work-epta-booklet>.

151 European Parliamentary Technology Assessment (EPTA) Network, “EPTA Network – Full Members,” List of members, See, accessed October 24, 2018, <http://www.eptanetwork.org/members/full-members>

- Holistic periodic research (either in-house or commissioned) into:
 - current and evolving technological environment
 - horizon scanning, trend scanning, and forecasting
 - identification of vulnerable groups especially:
 - Indigenous people
 - disabled people
 - people without a digital footprint
 - pressing social needs and upcoming challenges
 - society’s values, positive goals, and vision for the future
 - national strengths in science and technology
 - TA methods and methodologies
 - regulation and design methods
 - special topics as selected by its board and government
- Social outreach and education:
 - of the public
 - of technology companies
 - of technology designers
 - at primary and high schools
 - in higher education institutions
- Domestic collaboration and coordination with/regarding:
 - science and research bodies (e.g. Australian Research Council)
 - other organisations and government bodies such as:
 - Australian Human Rights Commission
 - Australian Competition and Consumer Commission
 - State based consumer rights bodies
 - medical and health technology regulator
 - autonomous vehicles regulator
 - communications regulator
 - energy regulator
 - public transport regulator
 - education policy
 - commerce and industry

- International collaboration:
 - review projects conducted by international TA organisations
 - ensure Australia has a voice on international TA panels
 - to avoid duplication of efforts
 - to cover topics which impact different on Australia
 - Incentivising socially responsible innovation and value-sensitive design:
 - to supplement certification which can quickly become out of date
 - to include the public at the design stage of technologies
 - to stimulate technology manufacturers towards good practice
 - rendering advice to government and regulatory bodies based on the above
- Rendering objective, authoritative, and simple analysis on selected topics;
- Publish bi-annual report on past activities, upcoming focus, and recommendations.

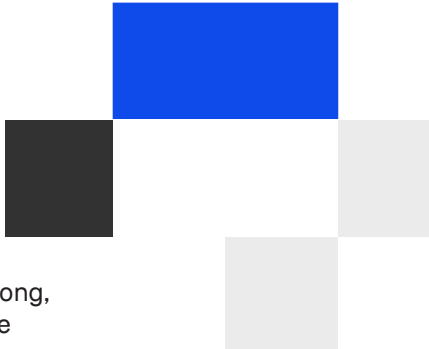

5.2.4 Notable features of our recommendation

5.2.4.1 Future-focused

Given that all regulation and design must take a stance on such substantive matters as which human interests are valuable and which states of affairs are worth promoting, instead of trying to propose a politically neutral approach – one that does not take a stance of different conceptions of the good – we instead opted for an approach that will actively consult with the public to ascertain its values. Given that the TAO should, in our view, be an independent office (not an arm of the incumbent government), and that it should render advice, and leave it up to government which advice to accept and how to follow through, these features would mitigate the risk of the TAO's showing partisan inclinations.

In its role as an advice rendering institution, we have suggested that a core focus of the TAO should be on how to create a desirable future – one way of putting this idea is 'a future in which people flourish with technology' – in addition to rendering advice that would aim to avoid problems such as human rights violations. There would also be a focus on periodically revisiting the challenges involved in predicting and evaluating unintended, overlooked, and difficult to recognize consequences. Our aim is to, as much as this is possible, in the first instance set technological innovation off in the right direction (as indicated by the country's needs, strengths, opportunities, values, goals, and visions for the future), and support this by checking for instances of things going wrong (in particular human rights infringements) or signs that they might go wrong (based on TA and forecasting strategies).

One part of the reason why this is our focus is because technology moves too fast for the law or for regulators to respond to current problems, or problems that are just around the corner, in a timely fashion. It would be counter-productive to devote most resources to bandaiding problems, when a better approach is to prevent the problems from occurring in the first place. The other reason why we propose this approach is because a sole focus on trying to avoid problems – e.g. infringements on human rights – will not yet guarantee that where technology leads Australians is a place worth going. For such reasons, we believe that it is instead better to encourage technological development in a direction that reflects the



overall goals of a society, while at the same time staying on the lookout for things going wrong, most importantly for human rights violations, and in particular taking account of vulnerable populations.

5.2.4.2 Participant-focused

Another core idea behind our proposal is to empower the public by giving it an opportunity to exercise their agency in the process of technology design. This is critically important to ensure that the Australian public can take charge of technology and use it to shape the environment which we co-habit, rather than allowing technology – or, worse, technology companies which have a tendency to exploit for profit – to shape how we live, by simply not noticing how technology shapes our lives and our choices. The aim is thus to not just use regulation, but also VSD approaches, and in particular to use regulation to incentivise and in some instances to mandate the use (and learning, for designers) of VSD.

In this context education is important in part to ensure that the public understands the issues, but also equally importantly to create a collective creative environment in which the public can participate in the creation of new visions for Australia's future. This requires the participatory co-design and co-evaluation of technology whereby an educated population – and not just technology assessors – is able to set goals, set pertinent questions, evaluate answers, and influence the shape of technology.

5.2.5 Conclusion

The problems that the AHRC sets out to tackle are complex. To effectively predict the potential impact of new technologies, a holistic system-wide approach capable of taking a broad range of considerations and interactions into account is required. Likewise, to properly evaluate the potential effects of new technologies, it is crucial to recognise the challenges we discussed, and to draw on methods and insights from multiple disciplines.

We have also emphasised that alongside ensuring that technologies do not have undesirable effects – most importantly, that human rights are protected and promoted, especially in regards to vulnerable populations – the guidelines for the design and regulation of technologies should be guided by how to create a future for Australia in which our people flourish with technology.

The complete absence of an organisation responsible for monitoring, assessing, and auditing new technologies, or promoting education and debate about the social, ethical or legal implications of such technologies, is something that needs to be addressed urgently. Addressing this will require a combination of technology design, technology regulation, legal regulation of technology design, as well as laws mandating transparency.

6 References

- “4TU | Centre for Ethics and Technology.” Accessed October 22, 2018. <https://ethicsandtechnology.eu/>.
- 109/88 Danfoss[1989] ECR 3199 (European Court of Justice October 17, 1989).
- ACARA. “National Assessment Program - Literacy and Numeracy (NAPLAN).” Accessed October 22, 2018. <https://www.nap.edu.au/naplan>.
- AHRC. “Human Rights and Technology Issues Paper (2018).” Australian Human Rights Commission, July 24, 2018. <https://www.humanrights.gov.au/our-work/rights-and-freedoms/publications/human-rights-and-technology-issues-paper-2018>.
- “AI for Good Global Summit 2017,” June 7, 2017. <https://www.itu.int/en/ITU-T/AI/Pages/201706-default.aspx>.
- Ajmal, Mian M., Mehmood Khan, Matloub Hussain, and Petri Helo. “Conceptualizing and Incorporating Social Sustainability in the Business World.” *International Journal of Sustainable Development & World Ecology* 25, no. 4 (2018): 327–339. <https://doi.org/10/gffdf8>.
- Amnesty International, and Access Now. “The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems.” Access Now, May 16, 2018. <https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/>.
- “Annual ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*).” Accessed October 22, 2018. <https://fatconference.org/>.
- Armstrong, Henry, and Jen Rae. “A Working Model for Anticipatory Regulation: A Working Paper.” Nesta, 2017. <https://www.nesta.org.uk/report/a-working-model-for-anticipatory-regulation-a-working-paper/>.
- Article 29 Data Protection Working Party. “Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679.” European Commission, 2018. http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.
- artificiallawyer. “Declare Your Legal Bot! New California Law Demands Bot Transparency.” *Artificial Lawyer* (blog), October 3, 2018. <https://www.artificiallawyer.com/2018/10/03/declare-your-legal-bot-new-california-law-demands-bot-transparency/>.
- Assistments. Worcester Polytechnic Institute, 2016. <https://www.assistments.org/>.
- Australian Human Rights Commission. “Human Rights Based Approaches.” Accessed October 22, 2018. <https://www.humanrights.gov.au/human-rights-based-approaches>.
- Ayres, Ian, and John Braithwaite. *Responsive Regulation: Transcending the Deregulation Debate*. Oxford University Press, USA, 1995.
- Banta, David. “What Is Technology Assessment?” *International Journal of Technology Assessment in Health Care* 25 Suppl 1 (July 2009): 7–9. <https://doi.org/10/bmpn3t>.
- Barak, Michalle E. Mor. *Managing Diversity: Toward a Globally Inclusive Workplace*. Sage Publications, 2016.

- Black, Julia. "Constitutionalising Self-Regulation." *The Modern Law Review* 59, no. 1 (1996): 24–55. <https://doi.org/10/dcnkrb>.
- . "The Rise, Fall and Fate of Principles Based Regulation." *LSE Law, Society and Economy Working Papers*, 2010. <https://doi.org/10/fzn7k7>.
- Black, Julia, Martyn Hopper, and Christa Band. "Making a Success of Principles-Based Regulation." *Law and Financial Markets Review* 1, no. 3 (2007): 191–206. <https://doi.org/10/gffdfq>.
- Braithwaite, John. "The Essence of Responsive Regulation." *UBCL Rev.* 44 (2011): 475.
- Buehler, Erin, Stacy Branham, Abdullah Ali, Jeremy J. Chang, Megan Kelly Hofmann, Amy Hurst, and Shaun K. Kane. "Sharing Is Caring: Assistive Technology Designs on Thingiverse." In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 525–534. ACM, 2015.
- Bull, Susan, and Judy Kay. "Open Learner Models." In *Advances in Intelligent Tutoring Systems*, 301–322. Springer, 2010.
- Carr, Nicholas. *The Shallows: What the Internet Is Doing to Our Brains*. WW Norton & Company, 2011.
- Cavoukian, Ann. "Privacy by Design—the 7 Foundational Principles (2011)," 2011.
- Center for Data Science and Public Policy - University of Chicago. *Bias and Fairness Audit Toolkit*. Contribute to Dssg/Aequitas Development by Creating an Account on GitHub. Python. 2018. Reprint, Center for Data Science and Public Policy - University of Chicago, 2018. <https://github.com/dssg/aequitas>.
- Chen, Ping, John Rochford, David N. Kennedy, Soussan Djamasbi, Peter Fay, and Will Scott. "Automatic Text Simplification for People with Intellectual Disabilities." *Artificial Intelligence* 10 (2016): 9789813206823_0091.
- Coleman, Roger, Cherie Lebbon, John Clarkson, and Simeon Keates. "From Margins to Mainstream." In *Inclusive Design*, 1–25. Springer, 2003.
- Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe. "Artificial Intelligence for Europe." European Commission, April 25, 2018. <https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe>.
- Committee of experts on internet intermediaries. "Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications." Strasbourg: Council of Europe, 2017. <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5>.
- Convention on the Rights of Persons with Disabilities, A/RES/61/106, article 2 § article 2 (2008).
- Corrin, Linda. "Supporting the Use of Student-Facing Learning Analytics in the Classroom." *Learning Analytics in the Classroom: Translating Learning Analytics for Teachers*, 2018.
- Council of Europe. "MSI-AUT Committee of Experts on Human Rights Dimensions of Automated Data Processing and Different Forms of Artificial Intelligence," 2018. <https://www.coe.int/en/web/freedom-expression/msi-aut>.
- . "Recommendation 2115: The Use of New Genetic Technologies in Human Beings," 2017.
- Dawson, Nik. "Bits & Atoms." *AI Policy White Paper*. University of Technology Sydney, 2018.

Department for Business, Energy & Industrial Strategy. “AI in the UK: Ready, Willing and Able? – Government Response to the Select Committee Report.” Department for Business, Energy & Industrial Strategy, 2018. <https://www.gov.uk/government/publications/ai-in-the-uk-ready-willing-and-able-government-response-to-the-select-committee-report>.

Department for Digital, Culture, Media & Sport. “Department for Digital, Culture, Media & Sport, Centre for Data Ethics and Innovation Consultation.” GOV.UK, 2018. <https://www.gov.uk/government/consultations/consultation-on-the-centre-for-data-ethics-and-innovation/centre-for-data-ethics-and-innovation-consultation>.

Desai, Deven R., and Joshua A. Kroll. “Trust but Verify: A Guide to Algorithms and the Law,” 2017.

Dourish, Paul, and Genevieve Bell. *Divining a Digital Future: Mess and Mythology in Ubiquitous Computing*. Mit Press, 2011.

Edwards, Lilian, and Michael Veale. “Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For.” *Duke L. & Tech. Rev.* 16 (2017): 18.

Eggers, William, D., Mike Turley, and Pankaj Kishnani. “The Future of Regulation.” Deloitte Insights, 2018. <https://www2.deloitte.com/insights/us/en/industry/public-sector/future-of-regulation/regulating-emerging-technology.html>.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin’s Press, 2018.

European Commission. “A Renewed EU Strategy 2011-14 for Corporate Social Responsibility.” Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. European Commission, 2011. https://www.eurocommerce.eu/media/7237/position-csr-renewed_csr_strategy_2011-14-07.03.2012.pdf.

———. “Communication from the Commission to the European Parliament, The European Council and The Council Delivering on the European Agenda on Security to Fight against Terrorism and Pave the Way towards an Effective and Genuine Security Union.” European Commission, 2016. https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/european-agenda-security/legislative-documents/docs/20160420/communication_eas_progress_since_april_2015_en.pdf.

———. “European Commission Staff Working Document: Liability for Emerging Digital Technologies.” European Commission, 2018. <https://ec.europa.eu/digital-single-market/en/news/european-commission-staff-working-document-liability-emerging-digital-technologies>.

———. “On the Road to Automated Mobility: An EU Strategy for Mobility of the Future.” Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. European Commission, 2018. https://ec.europa.eu/transport/sites/transport/files/3rd-mobility-pack/com20180283_en.pdf.

European Group on, Ethics in Science and, and European Group on Ethics in Science and New Technologies. “The European Group on Ethics in Science and New Technologies, AI, Robotics and ‘Autonomous’ Systems.” European Commission, 2018. https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

European Parliamentary Technology Assessment (EPTA) Network. “EPTA Network - Full Members.” List of members. Accessed October 24, 2018. <http://www.eptanetwork.org/members/full-members>.

European Parliamentary Technology Assessment (network) (EPTA). "Parliamentary Technology Assessment in Europe. An Overview of 17 Institutions and How They Work." EPTA Booklet - Policy Briefs & Reports - EPTA Network. European Parliamentary Technology Assessment (network) (EPTA), 2013. <http://www.eptanetwork.org/images/documents/news/EPTABooklet2013.pdf>.

European Union Agency for Fundamental Rights. "Artificial Intelligence, Big Data and Fundamental Rights." European Union Agency for Fundamental Rights, May 28, 2018. <http://fra.europa.eu/en/project/2018/artificial-intelligence-big-data-and-fundamental-rights>.

Faulkhead, Shannon, Livia Iacovino, Sue McKemmish, and Kirsten Thorpe. "Australian Indigenous Knowledge and the Archives: Embracing Multiple Ways of Knowing and Keeping." *Archives and Manuscripts* 38, no. 1 (May 2010): 27. <http://search.informit.com.au/documentSummary;dn=201007444;res=IELAPA>.

Friedman, Batya, Peter H. Kahn, Alan Borning, and Alina Huldtgren. "Value Sensitive Design and Information Systems." In *Human-Computer Interaction in Management Information Systems*, edited by P Zhang and D Galletta, 55–95. Springer, 2013.

Future of life institute. "Asilomar AI Principles." Future of Life Institute, 2017. <https://futureoflife.org/ai-principles/>.

Grigoryan, David, Avetik Muradov, Serob Balyan, Suren Abrahamyan, Armine Katvalyan, Vladimir Korkhov, Oleg Iakushkin, Natalia Kulabukhova, and Nadezhda Shchegoleva. "Creating Artificial Intelligence Solutions in E-Health Infrastructure to Support Disabled People." In *International Conference on Computational Science and Its Applications*, 41–50. Springer, 2018.

Guadamuz, Andres. "Should Robot Artists Be given Copyright Protection?" *The Conversation* (blog), 2017. <http://theconversation.com/should-robot-artists-be-given-copyright-protection-79449>.

Hollins, Sheila, and Irene Tuffrey-Wijne. "Meeting the Needs of Patients with Learning Disabilities." *British Medical Journal*, 2013. <https://doi.org/10.1136/bmj.f3421>.

Huberth, Madeline, Patricia Chen, Jared Tritz, and Timothy A. McKay. "Computer-Tailored Student Support in Introductory Physics." *PloS One* 10, no. 9 (2015): e0137001. <https://doi.org/10/gffdgf>.

Hurley, Dan. "Can an Algorithm Tell When Kids Are in Danger." *New York Times* 2 (2018). <https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html>.

Hurley, Mikella, and Julius Adebayo. "Credit Scoring in the Era of Big Data." *Yale Journal of Law and Technology* 18, no. 1 (2017): 5. <http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1122&context=yjolt>.

IEEE. "Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems (Version 2)." IEEE Computer Society, 2017. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_brochure_v2.pdf.

JISC. "Code of Practice for Learning Analytics," 2015. https://www.jisc.ac.uk/sites/default/files/jd0040_code_of_practice_for_learning_analytics_190515_v1.pdf.

Kitto, Kirsty, and Simon Knight. "Journey through Data." UTS Open, 2018. <https://open.uts.edu.au/datajourney.html>.

K.N.C. "Withered InBloom." *The Economist*, April 30, 2014. <https://www.economist.com/schumpeter/2014/04/30/withered-inbloom>.

- Knight, Simon, and Kirsty Kitto. "What Does Facebook Know about You?" UTS Open, 2018. <https://open.uts.edu.au/facebookknowyou.html>.
- Kroll, Joshua A., Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. "Accountable Algorithms." *U. Pa. L. Rev.* 165 (2016): 633.
- Kudina, Olya, and Peter-Paul Verbeek. "Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy." *Science, Technology, & Human Values*, August 21, 2018, 0162243918793711. <https://doi.org/10.1177/0162243918793711>.
- Kundu, Subhash C., and Archana Mor. "Workforce Diversity and Organizational Performance: A Study of IT Industry in India." *Employee Relations* 39, no. 2 (2017): 160–183. <https://doi.org/10/f9wbx4>.
- Lapowsky, Iessie. "Crime-Predicting Algorithms May Not Beat Untrained Humans." *Wired*, January 17, 2018. <https://www.wired.com/story/crime-predicting-algorithms-may-not-outperform-untrained-humans/>.
- Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, May 23, 2016. <https://www.propublica.org/article/how-we-analyzed-the-compass-recidivism-algorithm>.
- Latonero, Mark. "Governing Artificial Intelligence: Upholding Human Rights & Dignity." New York, NY, USA: Data & Society, 2018. https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.
- Learning Analytics Community Exchange (LACE). "'Visions of the Future', Horizon Report." Public Deliverable, 2015. http://laceproject.eu/wp-content/uploads/2016/02/LACE_D3_2.pdf.
- Lessig, Lawrence. *Code: And Other Laws of Cyberspace*. ReadHowYouWant. com, 2009.
- Luckin, Rose. "Towards Artificial Intelligence-Based Assessment Systems." *Nature Human Behaviour* 1 (2017): 0028. <https://doi.org/10/gc3gdj>.
- Luizzi, Vincent. "Balancing of Interests in Courts." *Jurimetrics* 20, no. 4 (1980): 373–404. www.jstor.org/stable/29761723.
- Lupi, Giorgia. "Data Humanism: The Revolutionary Future of Data Visualization." *Print Magazine (blog)*, January 30, 2017. <http://www.printmag.com/information-design/data-humanism-future-of-data-visualization/>.
- Montreal Declaration for a Responsible Development of AI. "Declaration of Montréal for a responsible development of AI." Declaration of Montréal for a responsible development of AI, 2017. <https://www.montrealdeclaration-responsibleai.com>.
- Morrison, Cecily, Edward Cutrell, Anupama Dhareshwar, Kevin Doherty, Anja Thieme, and Alex Taylor. "Imagining Artificial Intelligence Applications with People with Visual Disabilities Using Tactile Ideation." In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 81–90. ACM, 2017.
- Mulgan, Geoff. "A Machine Intelligence Commission for the UK." *Nesta (blog)*, 2016. <https://www.nesta.org.uk/blog/a-machine-intelligence-commission-for-the-uk/>.
- News Corp Australia. "Submission to the Australian Competition and Consumer Commission: Digital Platforms Inquiry," 2018. <https://www.accc.gov.au/system/files/News%20Corp%20Australia%20%28April%202018%29.pdf>.

OECD. “Programme for International Student Assessment.” Accessed October 22, 2018. <http://www.oecd.org/pisa/>.

Office of the United Nations High Commissioner for Human Rights (OHCHR). “Frequently Asked Questions on a Human Rights–Based Approach to Development Cooperation.” United Nations, 2006. <https://www.ohchr.org/Documents/Publications/FAQen.pdf>.

Ogus, Anthony. “Rethinking Self-Regulation.” *Oxford J. Legal Stud.* 15 (1995): 97. <https://doi.org/10/cqhz4v>.

OnTask. OnTask. Accessed October 22, 2018. <https://www.ontasklearning.org/>.

Pasquale, Frank. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.

Prinsloo, Paul, and Sharon Slade. “An Elephant in the Learning Analytics Room: The Obligation to Act.” In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 46–55. ACM, 2017.

“Privacy and Data Protection by Design – ENISA.” Report/Study. Accessed October 22, 2018. <https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design>.

“Racism Is Poisoning Online Ad Delivery, Says Harvard Professor.” *MIT Technology Review* (blog), 2013. <https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/>.

Rienties, Bart, and Lisette Toeteneel. “The Impact of Learning Design on Student Behaviour, Satisfaction and Performance: A Cross-Institutional Comparison across 151 Modules.” *Computers in Human Behavior* 60 (2016): 333–341. <https://doi.org/10/gffdgb>.

Salomon, Gavriel, David N. Perkins, and Tamar Globerson. “Partners in Cognition: Extending Human Intelligence with Intelligent Technologies.” *Educational Researcher* 20, no. 3 (1991): 2–9. <https://doi.org/10/c285rs>.

Schauer, Frederick. “Proportionality and the Question of Weight.” In *Proportionality and The Rule of Law: Rights, Justification, Reasoning*, Cambridge University Press, Cambridge, edited by C. Huscfort, B.W. Miller, and G. Webber, 173–185. Cambridge, UK: Cambridge University Press, 2014.

Schutz, Wolfgang, and Joris van Hoboken. “Human Rights and Encryption; UNESCO Series on Internet Freedom.” UNESCO, n.d. <http://unesdoc.unesco.org/images/0024/002465/246527E.pdf>.

Scottish Human Rights Commission. “Human Rights Based Approach | Scottish Human Rights Commission.” Scottish Human Rights Commission, 2018. <http://www.scottishhumanrights.com/rights-in-practice/human-rights-based-approach/>.

Select Committee on Artificial Intelligence. “AI in the UK: Ready, Willing and Able.” House of Lords, 2018. <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>.

Select Committee on the Future of Work and Workers. “Hope Is Not a Strategy – Our Shared Responsibility for the Future of Work and Workers.” Commonwealth of Australia, 2018. https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Future_of_Work_and_Workers/FutureofWork/Report.

Sellar, Sam, Greg Thompson, and David Rutkowski. *The Global Education Race: Taking the Measure of PISA and International Testing*. Brush Education, 2017.

Selvaratnam, Naomi, and Sarah Farnsworth. "Blind Woman Takes Bank to Court over 'inaccessible' EFTPOS Machines." ABC News, March 16, 2018. <https://www.abc.net.au/news/2018-03-16/blind-discrimination-lawsuit-cba-albert-efpos-machines/9551458>.

Siemens, George, and Phil Long. "Penetrating the Fog: Analytics in Learning and Education." *EDUCAUSE Review* 46, no. 5 (2011): 30.

Simonite, Tom. "Machine Learning Opens Up New Ways to Help People with Disabilities." MIT Technology Review, 2017. <https://www.technologyreview.com/s/603899/machine-learning-opens-up-new-ways-to-help-disabled-people/>.

Smith, Merritt Roe, and Leo Marx. *Does Technology Drive History?: The Dilemma of Technological Determinism*. Mit Press, 1994.

Stecklow, Steve. "Why Facebook Is Losing the War on Hate Speech in Myanmar." Reuters, 2018. <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/>.

Sweet, Alec Stone, and Jud Mathews. "Proportionality Balancing and Global Constitutionalism." *Colum. J. Transnat'l L.* 47 (2008): 72.

Swierstra, Tsjalling. "Identifying the Normative Challenges Posed by Technology's 'Soft' Impacts." *Etikk i Praksis - Nordic Journal of Applied Ethics*, no. 1 (May 9, 2015): 5–20. <https://doi.org/10/48b>.

The Australian Indigenous Data Sovereignty Collective Maïam nayri Wingara. "Key Principles for Indigenous Data Sovereignty." Maïam Nayri Wingara, 2018. <https://www.maïamnayriwingara.org/key-principles/>.

"The Future Computed: Artificial Intelligence and Its Role in Society." The Official Microsoft Blog (blog), January 18, 2018. <https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificial-intelligence-role-society/>.

The Human Rights, Big Data, and Technology Project (HRBDT). "The Human Rights, Big Data and Technology Project – Written Evidence (AIC0196), Submission to the House of Lords Select Committee on Artificial Intelligence," 2017. <http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69717.html>.

The Law Society. "Technology and the Law Policy Commission - Algorithms in the Justice System," September 11, 2018. <https://www.lawsociety.org.uk/policy-campaigns/articles/public-policy-technology-and-law-commission/>.

———. "Using Algorithms to Deliver Justice – Bias or Boost? - The Law Society," 2018. <https://www.lawsociety.org.uk/news/press-releases/using-algorithms-to-deliver-justice-bias-or-boost/>.

The Open University (UK). "Ethical Use of Student Data for Learning Analytics." Student Policies and Regulations - Open University, 2018. <https://help.open.ac.uk/documents/policies/ethical-use-of-student-data>.

The University of Edinburgh. "Learning Analytics Principles and Purposes," 2017. <https://www.ed.ac.uk/files/atoms/files/learninganalyticsprinciples.pdf>.

Thorp, Jer. "Turning Data Around." Memo (Random) (blog), November 18, 2016. <https://medium.com/memo-random/turning-data-around-7acea1f7479c>.

Travaglia, Joanne, Deborah Debono, and Georgia Debono. "Capacity Building and Intellectual Disability Health Services: An Evidence Check Rapid Review Brokered by the Sax Institute (Www.Saxinstitute.Org.Au) for the NSW Ministry of Health." Sax Institute, 2017. <https://www.health.nsw.gov.au/disability/Documents/evidence-check-cbidh.pdf>.

Turek, Dariusz. "What Do We Know about the Effects of Diversity Management? A Meta-Analysis." *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie*, no. 4 (964) (2017): 5–25.

UN Office of the High Commissioner for Human Rights. "United Nations Guiding Principles on Human Rights." United Nations, 2011. https://ec.europa.eu/transport/sites/transport/files/3rd-mobility-pack/com20180283_en.pdf.

UNI Global Union. "10 Principles for Ethical AI." UNI Global, 2017. <http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/>.

Vavik, Tom, and Martina Maria Keitsch. "Exploring Relationships between Universal Design and Social Sustainable Development: Some Methodological Aspects to the Debate on the Sciences of Sustainability." *Sustainable Development* 18, no. 5 (2010): 295–305. <https://doi.org/10.1002/sd.374>.

Waller, Sam, Mike Bradley, Ian Hosking, and P. John Clarkson. "Making the Case for Inclusive Design." *Applied Ergonomics* 46 (2015): 297–303. <https://doi.org/10.1016/j.apergo.2014.12.006>.

Wang, Yilun, and Michal Kosinski. "Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images." *Journal of Personality and Social Psychology* 114, no. 2 (2018): 246. <https://doi.org/10.1037/xap0000111>.

Warschauer, Mark. *Technology and Social Inclusion: Rethinking the Digital Divide*. MIT press, 2004.

Wightwick, Abbie. "Is Exam Stress Driving Our Children to Mental Illness and Even Suicide?" *WalesOnline*, April 27, 2018. <https://www.walesonline.co.uk/news/education/exam-stress-driving-children-mental-14582450>.

Williamson, Ben. *Big Data in Education: The Digital Future of Learning, Policy and Practice*. Sage, 2017.

7 Compiled recommendations

1. What types of technology raise particular human rights concerns? Which human rights are particularly implicated?


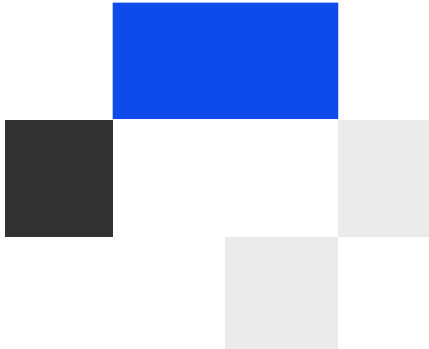
- 1.1 Consideration of the impacts of emerging technology on human rights should consider the specific impact on individuals and communities as well as broader impacts on society and values.
- 1.2 A nuanced approach to technology should be adopted which recognises that effects of technologies are in fact effects on systems in which combinations of technologies both new and old, as well as human and contextual factors, generate the effects.

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

- 2.1 A broad range of stakeholders should be involved in understanding the impacts of technology development and deployment across contexts.
- 2.2 Accessibility of technologies and their uses must be a core consideration in their development and deployment, including physical, cultural, socio-economic educational and other barriers to access.
- 2.3 There should be clear recognition of the broad range of stakeholders impacted by technologies, and the ways in which they may be impacted both directly by the technology (hard impacts) and through more indirect means (soft impacts).
- 2.4 Positive outcomes for stakeholder groups should be an explicit aim in developing and deploying technologies.

3. How should Australian law protect human rights in the development, use and application of new technologies?

- 3.1 Research must be conducted to elaborate how human rights and transdisciplinary approaches can be brought to bear in developing concrete legal and regulatory regimes that supplement the PANEL principles (Participation, Accountability, Non-discrimination and equality, Empowerment and Legality) for complex new technologies, particularly in understanding the 'nexus' between human rights and ethics.
- 3.2 A fundamental review of Australian privacy laws to ensure they remain fit for purpose.

- 
- 
- 3.3 A need for review of cross-sectoral laws with a view to clarifying rules relating to liability for human rights abuses in relation to complex new technologies.
- 3.4 The establishment of a new regulatory body, the Technology Assessment Office (TAO) and associated processes, to address the gap in the Australian legal framework for new technologies (see Section 5).
4. **In addition to legislation, how should the Australian Government, the private sector and others protect and promote human rights in the development of new technology?**
- 4.1 Human rights considerations should be taken into account at the design stage of technologies. In order to do this, we outline a set of guiding transdisciplinary principles for: distributed and shared agencies; understanding emerging technology as a complex ecosystem; engaging with uncertainty fostering dialogue in a neutral environment the imperative of education; and developing ethical and philosophical models elaborated on in Section 5.
5. **How well are human rights protected and promoted in AI-informed decision making? In particular, what are some practical examples of how AI-informed decision making can protect or threaten human rights?**
- As in the recommendations for Question 1, in the specific case of AI-informed decision making:
- 5.1 Consideration of the impacts of emerging technology on human rights should consider the specific impact on individuals and communities as well as broader impacts on society and values.
- 5.2 A nuanced approach to technology should be adopted which recognizes that effects of technologies are in fact effects of systems in which combinations of technologies both new and old, as well as human and contextual factors, generate the effects.
6. **How should Australian law protect human rights in respect of AI-informed decision making?**
- 6.1 The principles of transparency and fairness require that people affected by AIDM be informed when a decision that may significantly affect them is made with the assistance of, or by, AI technologies. In addition, where decisions are informed by or made by AI technologies, people affected by the decisions should have a right to an explanation as to how the decision was made.
- 6.2 Human Rights by Design principles can embed human rights into the design of AIDM to supplement the principles of transparency and fairness.
- 6.3 We must learn from other jurisdictions and non-governmental organisations in developing human rights approaches for AIDM.

6.4 The TAO should be responsible for developing a set of principles for applying human rights to AIDM, which may be used to develop legislation, or other forms of regulation, that applies to AIDM.

7. In addition to legislation, how should Australia protect human rights in AI-informed decision making?

7.1 A regulatory regime must be developed in respect of responsive regulation that has active engagement of the broadest range of actors and stakeholders.

7.2 We must learn from other jurisdictions and non-governmental organisations in developing human rights approaches for AIDM .

7.3 In implementing human rights by design a transdisciplinary approach should be adopted, entailing: (1) striving to build diverse teams and inclusive practices into the design; attending to hard and soft impacts; consultation and education with stakeholders; and an iterative approach of ongoing assessment and evaluation, discussed in Section 5.

7.4 Human rights in AIDM can be further protected by requiring those responsible for developing AIDM to submit the technology to a structured human rights impact assessment, undertaken by an independent third party, where the decision making poses a sufficient risk to human rights.

8. What opportunities and challenges currently exist for people with disability accessing technology?

8.1 Increased focus and resources for research into disability service delivery for women, rural and regional, CALD and ATSI communities particularly with a focus on emerging technology.

8.2 Providing subsidised access to digital services and assistive technologies for platforms and tools regularly used by people with disabilities to protect privacy of personal data.



9. What should be the Australian Government's strategy in promoting accessible technology for people with disability? In particular:

a. What, if any, changes to Australian law are needed to ensure new technology is accessible?

b. What, if any, policy and other changes are needed in Australia to promote accessibility for new technology?

9.1 Demonstrate best practice in relation to promoting diversity, universal design, value centred design and consideration of human rights in design, provision and procurement of government services.

9.2 Incentives and assistance for curriculum development across all areas of technical and a higher education to introduce disability, access, inclusive design and universal design across disciplines.



10 . How can the private sector be encouraged or incentivised to develop and use accessible and inclusive technology, for example, through the use of universal design?

- 10.1 Provide incentives and rewards to business who demonstrate best practice, e.g. Human Rights Awards for inclusive business organisation, national star rating framework for sustainability and inclusion within organisations, tax incentives for Universal Design Departments such as those implemented in Japan.
- 10.2 Incentivise and provide resources to support technology design approaches which focus on universal design and participatory design. Provide industry with panels of expert user groups of people with disabilities available to contribute to design, testing and review of emerging technology.
- 10.3 Educate the private sector about the importance of workplace diversity and implications for inclusive decision-making and outcomes. Encourage organisations to require every employee of an organisation must complete the Universal Design introductory course. Sponsor the development of further Universal Design Training for tech companies. Development of professional certification programs for architects, IT, engineers.

