Submission in response to Australian Human Rights Commission Discussion Paper, *Human Rights and Technology*, April 2020

Professor Kimberlee Weatherall, [redacted]
Dr Tiberio Caetano, [redacted]

The authors thank the Commission for this opportunity to comment on the proposals and questions raised by the Australian Human Rights Commission in its Discussion Paper, *Human Rights and Technology,* published in December 2019. The comments in this response focus on specific recommendations around artificial intelligence. The authors thank other members of Gradient Institute for comments and conversations that inform this submission. We would be happy to answer any questions or elaborate further on the points made here if helpful.

Ethical frameworks

> ***Proposal 2:*** *The Australian Government should commission an appropriate independent body to inquire into ethical frameworks for new and emerging technologies to:*
>
> *(a) assess the efficacy of existing ethical frameworks in protecting and promoting human rights*
>
> *(b) identify opportunities to improve the operation of ethical frameworks, such as through consolidation or harmonisation of similar frameworks, and by giving special legal status to ethical frameworks that meet certain criteria.*

In our view this proposal misses the mark: first by proposing work that has already been done, and second by misunderstanding the role and purpose of ethics and ethical reasoning, as compared to the role of human rights and law.

We are not convinced that a single inquiry into ethical frameworks for new/emerging technologies across multiple technologies, professions, industries, individuals, and businesses will be fruitful. The goals and (detailed) content of ethical standards for medicine are different from those for law, or for engineering, or banking, or education. While there are some general principles, how those apply, or are operationalised, can be expected to vary across industries. This is a strength, not a weakness. Rather than seeing the applicability of multiple ethical perspectives and frameworks as a weakness in seeking to deputise a single body to examine ethical frameworks generally, perhaps the Commission could instead seek to encourage or facilitate more detailed, practical ethical and practical discussions within industries: with a view to increasing concrete, industry-specific guidance or codes of conduct available to participants

in those industries, and helping ensure that such guidance includes specifically guidance on means for protecting and promoting human rights.

Australia does in some laws (such as the *Privacy Act 1988* (Cth)) elevate industry-specific negotiated Codes of Practice, giving them legal status. But we do not agree with any suggestion that Australia should seek to give legal force to *ethical* frameworks**.** This risks conflating the different roles of ethics and law. Law and ethics are not points on a continuum of tools for influencing behaviour: they operate in different spheres; and perform different roles. The study of ethics concerns reflection on the justification of human decision-making; on how to define and pursue desirable goals. Acting ethically concerns making the *right* (justifiable) decision (whether we are talking about an individual, organisational or political choice, or a choice in legal or technical design). Ethical frameworks are designed to provide, not so much a rule, as guides regarding important considerations to take into account when acting, and methods for ethical reasoning. They do not provide 'answers' regarding how decisions or tradeoffs should be made in particular cases.

Certain elements of existing AI ethical frameworks *overlap* with international human rights and with Australian legal obligations: ethical principles demanding "fairness" overlap with anti-discrimination law; respect for privacy is an ethical obligation as well as a legal one under the *Privacy Act 1988* (Cth) and other laws. Acknowledging those overlaps, assessing the effectiveness of established legal rights in the context of new technologies, and giving effect to (even increasing) the legal obligations is not the same as attempting to give legal effect to ethical frameworks. Human rights should set a baseline of enforceable protection for every person that governments, businesses and individuals are expected to observe.

Giving legal force to an ethical framework would also involve regulators, decision-makers or courts in a controversial task. When enforcing the law, courts and regulators make rulings, not on whether something is definitively 'right' or 'wrong' but on whether it is *legal.* People remain free to disagree whether a legal ruling comports with morality. A ruling on whether something is *ethical* (comports with ethical guidelines) however, gets much closer to a ruling on whether something is moral, or right - which is not the role of the courts, or other state organs in a pluralistic democracy.

Finally we note that in terms of harmonisation, there has already been significant progress in 'consolidating' or at least harmonising principles in the field of artificial intelligence. Significant work has already been done analysing and comparing the *content* of AI ethical frameworks, identifying common themes and differences and similarities in text and scope within those themes (Fjeld et al 2020; Jobin et al 2019). There is no pressing need to repeat or expand on this work: the commonalities and differences among ethical frameworks are reasonably well understood. The *efficacy* of the frameworks in promoting human rights is an empirical question which, though explored in the literature to some extent (see eg Metcalf et al 2019), warrants investigation. But truly understanding the efficacy of these frameworks will require rigorous,

well-designed research which is unlikely to occur in the context of a single inquiry, however independent.

Definition of AI

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

We acknowledge the intention of the Commission to subject a *subset* of uses of AI to additional scrutiny and obligations. We acknowledge too that the goal is to include only those uses involving *decision-making with a significant impact on individual human beings,* whether the decision-making is fully automated, or augmented by using artificial intelligence (that is, per the Commission's definition, where the decision is *materially assisted* by the use of AI). There are however some potential problems with the Commission's framing.

First, as many argued at earlier stages of this inquiry, 'AI' remains an unclear term with no settled technical or legal meaning. It is possible that 'AI' (if understood to require 'machines that learn') could exclude automated decision-making based on entirely pre-programmed and straightforward application of rules to data, as occurred in, for example, #Robodebt. The definition offered by the OECD Expert Group, quoted by the Commission, seems to require some level of inference or learning on the part of the system.[1] If the intention of the Commission is to include automated, but non-learning systems that simply apply pre-programmed rules to data inputs, like #Robodebt, we suggest making that explicit. This highlights the danger, raised at earlier stages of this inquiry, of focusing on the regulation of particular *technologies* (AI) rather than conduct or phenomena (like automated or data-driven decision-making).

[1] The OECD Group of Experts defined an AI system to include a "machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. It uses machine and/or human-based inputs to perceive real and/or virtual environments; abstract such perceptions into models (in an automated manner, eg with ML or manually); and use model inference to formulate options for information or action. AI systems are designed to operate with varying levels of autonomy": OECD (2019).

Second, we note that there is room for disagreement - particularly across different disciplines - regarding the scope of what constitutes a 'decision'. Lawyers may not have the same view of what constitutes a 'decision' as computer or data scientists or engineers. Consider, for example, the hypothetical of a bank which establishes a system incorporating machine learning methods into determining whether/how much credit to offer applicants. From a technical perspective, any such system is likely to involve numerous AI-informed 'decisions', both at the individual level (whether *this* person is offered credit; how much credit this person is offered) and at the systems level (how much money does the bank have to lend today, what is the 'baseline' interest rate subject to adjustment for the individual etc etc). It seems likely that the Commission's focus is on the *final* decision whether or not to offer credit to the particular individual, although it is also possible that the term is deliberately broad so as to allow challenge at the individual or systemic level. This is not explained however in the Paper.

We also wonder whether there may be some activities *not* covered by focusing on decision-making: such as the collection of detailed profiles on people and the drawing of inferences about their interests, behaviour, or ability to perform some task. Compare the approach of the European Union, which (in the GDPR) distinguishes between *profiling* ("any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements"[2]) and *solely automated decision-making.* A focus on decision-making should capture *most* situations where there is an impact on individual rights or interests. But while in general it is appropriate to focus accountability on the actor who performs the last action in the causal chain of events reaching the individual, there may be cases where it is more efficient to seek a remedy against a source or platform that is enabling multiple other firms to harm an individual. For example, there may be cases where one entity constructs profiles and draws inferences about individuals, where the results of those inferences are used by multiple third parties (EU Art 29 Working Party 2018, 8). In such cases, confining remedies to the final decision-making (and the entity making decisions) would require the individual to wait until impacted by a decision, and deprive them of an effective remedy against the common source of harm - the profile.

There is also the question of AI-informed decisions that have a small effect on large numbers of people, or uneven effects, rather than a significant effect on certain individuals. It is not just the *size* of the (individual) impact that is important, but its *distribution*. It is possible to imagine decision-making that is consequential for human rights overall, but might not fall within the scope of AI-informed decision-making as set out by the Commission because the impact on

[2] GDPR, art 4(4).

individuals is small. Whether this matters depends on the purpose of the definition: decision-making with negative/uneven effects across populations is likely best dealt with, not via individual challenges, but by more systemic tools, such as algorithmic/data protection impact assessment, auditing etc. The Commission's definition is reminiscent of the GDPR's art 22, which creates an individual right of challenge: compare for example, GDPR articles 25, and 35, which target system-level requirements (data protection by design and by default; data protection impact assessments), 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'.

Accountability

> ***Proposal 3:*** *The Australian Government should engage the Australian Law Reform Commission to conduct an inquiry into the accountability of AI-informed decision making. The proposed inquiry should consider reform or other change needed to:*
>
> *(a) protect the principle of legality and the rule of law*
>
> *(b) promote human rights such as equality or non-discrimination.*

We agree that questions of accountability, and liability for harm caused when (as is inevitable) the system gets things wrong are critical. Proposal 3 is, however, quite broad and framed anthropomorphically. AI *systems*, and 'AI decision-making' are not accountable: *people* are. There are also the questions of accountable *in what way* (politically? Legally?) and *to whom* (individuals? Company boards? The voting public?). The proposal looks quite differently if reframed as an inquiry into the accountability of [firms, government, individuals, developers] to [citizens/people harmed by] AI-informed decision-making.

It also becomes evident, once restated in this way, that an inquiry so broad-ranging would be challenging – and risk being shallow. Echoing comments made earlier: we recommend considering and the mechanisms/laws in concrete contexts. Quite different considerations pertain to ensuring the accountability of the builders and users of AI systems to society as a whole (which goes to general, systems-level obligations such as transparency (discussed below)), as compared to ensuring accountability for harm, through the legal system or other appropriate mechanisms, like effective systems for complaint/redress, or insurance.

The ALRC has proposed looking at automation of government decision-making and administrative law (ie the rules around how government makes decisions; challenging those decisions) (ALRC 2019). In other words, the ALRC contemplates a narrower question of when a person affected by a decision made involving government use of AI-informed decision-making can challenge that decision and seek a different decision or reconsideration. The ALRC could also be well-equipped to consider other questions of legal accountability, including the much-neglected potential role of aspects of private and commercial law, including consumer protection law, banking law, equity, contract, and employment/workplace law, to control the application of AI/automated decision-making.[3]

We also note that the topics contemplated would seem to require, not just a legal, but a legal-technical analysis: that grapples with the need to bring legal obligations together with deep questions regarding what the technology can do. We therefore wonder whether there is there another option, for example, some kind of combined legal/technical taskforce, that could be set up in the short term to consider these issues at the boundary of law and technology? Ideas of this kind were discussed in the 2019 forum organised in conjunction with the World Economic Forum.

Transparency

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

We understand the desire to promote transparency, and to ensure people are aware of the use of AI, and we note that this obligation is limited to those cases where the use of AI is *material* and the relevant decision has a *legal or significant effect* on the individual's rights.

We would argue however that there is a tradeoff to be made: providing more information to individual consumers and citizens imposes costs on those same consumers/citizens (in terms of attention), which has to be traded off against the benefits of providing the information. An obligation of this kind would likely lead to a flurry of letters from providers like banks, insurance companies etc all informing Australians that those companies use AI-informed decision-making to determine offers and to personalise commercial offerings.

[3] It may not be appropriate however to try to address all of these issues in *one* inquiry however, as any one inquiry by the ALRC can only have a certain number of Commissioners and advisory board members, and broad expertise would be required to address all of these issues

Too much information, *without more,* is more likely to generate confusion, and increase distrust. The research clearly shows that too much information can be overwhelming even for professionals (eg, Tonekaboni et al 2019; Kaur et al 2020). As the AHRC acknowledges at [6.5], providing information has limited benefit unless there is something that the consumer/citizen can *do* with it. Australians could quite rightly wonder why they were being told about AI use, and what they were meant to do with this information. It might be argued that providing information ensures better operation of the market (ie, if you don't want your data processed in this way, you can seek out another provider). Letters that merely disclose the fact of AI processing are however *not* likely to provide sufficient information for an individual to judge which (if any) of the possible banks or insurance agencies is using AI in a way that will best serve their interests. The ACCC concluded in its *Digital Platforms* inquiry doubted that privacy policies inform consumers or enable them to make decisions/commercial choices based on privacy.

In our view the key is to identify a *purpose* for transparency, and tailor disclosure to that purpose. So, for example:

- If a particular AI-informed decision has been made regarding the legal rights of an individual, and there is a *right to challenge* that decision, for example by correcting or adding to the information on which the decision is made, active disclosure *to the individual* of the fact of processing could be accompanied by information about the information used, perhaps 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject' (per GDPR art 13) and mechanisms for challenge or appeal;
- Where AI-informed decision-making is part of medical diagnosis and recommendation for a course of treatment, information on the systems used should be tailored to meet the expectations of medical practitioners (which are not necessarily the same as for other professions: Tonekaboni et al 2019) and the requirements of informed consent.
- If the user of AI is a *government* or public sector entity, disclosure might take a different form, for example, in the form of an online, always-accessible register of significant proposed, and in-use AI systems together with any impact assessments and audit reports;
- Industries where the consumer data right (CDR) is in operation could integrate transparency obligations into management of the CDR, for example:
    - Data holders could be required to provide information, or links to information about AI-informed decision-making as part of the individual consumer's data dashboard; and
    - Entities requesting CDR data could be required to provide information about AI-informed decision-making to consumers and data holders (for inclusion in the dashboard).

Government use of AI

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

(a) undertake a cost-benefit analysis of the use of AI, with specific reference to the protection of human rights and ensuring accountability

(b) engage in public consultation, focusing on those most likely to be affected

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

Proposal 17

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

(a) identify the use of AI in decision making by the Australian Government;

(b) undertake a cost-benefit analysis of the use of AI, with specific reference to the protection of human rights and ensuring accountability;

(c) outline the process by which the Australian Government decides to adopt a decision-making system that uses AI, including any human rights impact assessments;

(d) identify whether and how those impacted by a decision are informed of the use of AI in that decision-making process, including by engaging in public consultation that focuses on those most likely to be affected

(e) examine any monitoring and evaluation frameworks for the use of AI in decision-making.

We agree with the goals of reviewing the current landscape of government use of AI (Proposal 17) and ensure appropriate legislative and human rights oversight and rules for the use of AI in decision-making (Proposal 6). In addition to the specifics in the proposals above, the Discussion Paper suggests that where the government deploys an AI-informed decision-making system, this should be expressly provided for in law, to ensure there is legislative oversight, *and* that

where Parliament decides to permit this use of AI, that it (ie the Parliament) would then set legal rules "regarding how AI-informed decision making is deployed in each specific context".

Proposal 6 would be a (positive) contrast to the current Commonwealth approach to legislating for computerised decision-making, in which the legislation has generally simply empowered the relevant Minister to '*arrange for the use, under the Minister's control, of computer programs for any purposes for which the Minister may, or must, under this Act or the regulations, (a) make a decision; or (b) exercise any power or comply with any obligation; or (c) do anything else related to making a decision or exercising a power or complying with an obligation'* (this quote is from s 48 of the *Australian Citizenship Act 2007* (Cth), but similar provisions are common elsewhere). Provisions of this kind do not in themselves require any particular standards in the development or deployment of AI, although we note that the general principles of administrative law still apply (Bateman, 2019; Bateman, 2020).

We think it would be possible, and beneficial, to specify standards (for example for transparency, explainability (as discussed below) and fairness) and due process protections appropriate for the use of AI-informed decision-making, akin to, although not identical to, those in the *Administrative Decisions (Judicial Review) Act 1977* (Cth).

We also agree that it is important for the government to actively assess the appropriateness of AI for use in particular contexts. We think however it would be important to avoid a naive cost-benefit analysis focusing only on the overall balance of positives and negatives. A proper assessment should examine, not only the overall balance, but the *distribution* of any harms/costs across society and the severity and risk of any projected harms for different groups/individuals/entities. For example, what is the worst thing that could happen to the most negatively affected person? Unless this is well understood, a system ought not be deployed. Or, what is the worst impact in different quartiles? When is a person "affected", and by how much?

We also note that while the emphasis in the proposal is on the decision-making process that leads to deployment, equally important is what happens next. What guardrails are built into the system to alert officials in charge of deployment regarding the outcomes of the system and its impact on vulnerable groups? What governance/auditing mechanisms are in place to ensure ongoing monitoring and responsibility? Technology of this type is not and should never be 'set and forget'. This is hinted at in the requirement in Proposal 6 that there be '*adequate human rights protections in place*', but it should be explicitly stated that human rights protections includes not only recognition of rights, and remedies for their breach, but transparent monitoring and auditing of what the system does, and how costs and benefits are distributed, on an ongoing basis to ensure that those standards are being met.

Explainability

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

(a) a non-technical explanation of the AI-informed decision, which would be comprehensible by a lay person, and

(b) a technical explanation of the AI-informed decision that can be assessed and validated by a person with relevant technical expertise.

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

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

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

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

We agree overall with the goal of ensuring that the operations of AI systems — particularly those used in high stakes decision-making affecting human beings' rights and interests — are able to be understood and where necessary, contested. We think however that the current formulation of the proposals could be improved.

There are some questions of scope, outlined immediately below. Further, while the goal of explainability is important, the complexity of achieving it should not be underestimated. None

of what follows is to argue that explanation isn't important, or that some right to an explanation shouldn't be legislated: rather, it is to point out that the framing of the obligation could be more nuanced, and that in order to make an obligation to explain work in practice we will need to undertake further research, translation of research, and education. Specifically, we are going to need:

- A more rigorous identification of *when* there is an entitlement to explanation: ie the intended scope of this obligation;
- In those cases: clear identification of the goals of, and audiences for any explanation;
- Discussion of and guidance on *how* we meet those goals for these audiences across the wide range of contexts where this is important;
- Development of, and guidance for the practitioners on, the range of different kinds of explanation that can be offered by existing and developing technical methods.

This can't all be included as part of a legislative regime (that would be too inflexible). But a legal obligation to provide an explanation *will* need to be nuanced. In finalising its proposals, the Commission will need to think about the level at which an explanation is required (system or individual); whether it is always the individual who can demand an explanation; and whether there should, for example, be exceptions to the obligation where other means of contesting decisions and evaluating systems are provided for.

As for the research/education side: there is a great deal of work being done on these questions, and of course room for more research, which the Gradient Institute and others hope to develop further. But there will also be considerable work to be done creating the necessary guidance for practitioners in the public and private sector: suggesting that any AI Safety Commissioner (or other body created as a result of this work) will need some capacity to commission further work to support its mission.

Scope and strength of the proposals

While interpretability is an important goal for ML systems, achieving it is not costless: involves tradeoffs. So any demand for explainability should be proportional to the importance of the system and its impacts, and the goal being served. Stated in its present, categorical form therefore, the obligation may be overly strong. A prohibition on systems that cannot be evaluated for their accuracy, fairness etc, where there is a significant risk of a material infringement of human rights, may be more appropriate.

In addition, as Selbst and Barocas (2018) point out, in general 'explanations' are not an end in themselves so much as a means to an end: either to (1) protect the human rights of individuals, in particular their right to contest decisions that impact on their rights, and (2) to guarantee

respect for human rights by ensuring systems can be evaluated for their overall impact. As they also point out, there are technical and legal methods to achieve these goals, other than providing capacity to explain particular individual decisions. Evaluation may be achieved by testing overall outcomes/impacts of a system. Individual rights could be secured by ensuring individual decisions can be challenged or reconsidered (see, eg, GDPR art 22). We therefore think that Proposal 8 is stated in overly broad terms: an outright ban on any system that does not produce individual explanations for individual decisions, and that *could* infringe human rights is excessive and will prevent use of systems with significant advantages.

The complexity of explanation at the legal-technical interface should not be underestimated

We think a general obligation to provide "an explanation" (or "reasonable explanations") is overly simplistic. It implies that there is "an" explanation, appropriate for multiple audiences and goals. As elaborated further below, explanation is a complex concept, highly dependent on context and goals (Miller 2019), and bringing the legal goals together with technical capabilities is not straightforward.

The Commission uses the term *explainability* to refer to the idea that an individual affected by a decision has the right to request a meaningful explanation for an AI-informed decision: "a person affected by an AI-informed decision should be able to understand the basis of that decision—that is, decisions should be explainable". Proposal 7 also focuses on explaining a decision *to the individual affected*. But as the Discussion Paper acknowledges, there is more to 'explanation' - or interpretability in computer science terms than providing reasons for individual outcomes. Other mechanisms (of explainability/transparency/interpretability) can focus on the *system level* operation of AI, its design and logic, and collective (not only individual) outcomes.

Explanation/interpretability is also important for reasons beyond legal entitlements. Although the Commission is rightly focused on legal questions and human rights, any legal obligation to provide an explanation needs to be drafted with consideration for the full context in which it will operate, because any explanation offered will have impact beyond the legal sphere. Computer science/AI research has long recognised that interpretability of AI systems is critical to *uptake* of those systems in commercial and other contexts: people won't use a system they don't trust, and won't trust a system that isn't comprehensible in ways that are important to them or the context (Lipton 2016). The political acceptance of AI across society is also dependent on trust, and hence on interpretability.

The goal is not *absolute* trust, but an *optimal* level of trust. The key is to *calibrate* trust so it is in accordance to the true degree of trustworthiness of a system. But to achieve even that, we need to be able to interrogate the systems we use.

AI/machine learning research has developed a wide range of techniques for providing interpretability/transparency/understanding. For example, as outlined at length in Lipton (2016), methods for interpretability in machine learning include:

1. *Designing for simplicity:* for example by considering limited variables; using linear models; or optimising for less complexity;
2. *Approximating complex models in simpler form:* allowing the model to be complex, but approximating it later with something more interpretable;
3. *Extracting the most important factors in a particular decision:* attempt to establish the importance of any feature to a particular decision by iteratively varying the value of that feature while holding the value of other features constant;
4. *Explanation by example:* explain the decisions of a model by reporting (in addition to predictions) which other examples the model considers to be most similar;
5. *Allowing interaction with the models to see how changes in inputs affect outputs*;
6. *Visualization*: which can provide insights in model behaviour (Yosinski et al 2015; Ming et al 2017).

Not all of these are possible in all circumstances. They also provide different kinds of explanation: some are more diagnostic, providing insights that can help improve a model. Some methods focus on overall logic or design of a system, which from a legal perspective, could help audit or evaluate the model and its human rights or other impacts as a whole.[4] Others give insight into the reasons for particular individualised outcomes and are more targeted at something a lawyer would recognise as 'reasons' for a decision.

The choice between methods involves tradeoffs: there is no one, perfect "explanation" to be generated from any given model or system. Methods for designing simple models focus on the explanation for a single outcome, so may give radically different explanations to different individuals/for different individual outcomes. They may also have accuracy tradeoffs: conventional wisdom in data science is that simpler models will tend to be less accurate (Caruana et al 2015). This matters: there are cases where we may be prepared to trade lower accuracy in favour of explanation. For example, if the question were access to/allocation of credit, and we were worried about overall fairness in access to credit across society, then we

[4] Methods that focus on overall *performance* and, for example, the distribution of outcomes, may not fall within 'explanation'. But we wonder where such methods do fit in the Commission's overall schema, which seems to designate **transparency** as meaning transparency *about the fact of the use of an automated or AI system* rather than any properties or performance of that system.

might prioritise interpretability, to make sure the system is operating fairly and not discriminating on the basis of protected attributes. If, however, the system were designed for medical diagnosis, we might place a higher priority on accuracy (in the context of wanting to know about how the system was trained, and on what data, and overall, whether there are groups/populations/situations in which the predictions will be less accurate).

On the other hand, generating *post hoc* explanations for more complex models, for example by seeking to identify the features important to a particular decision, may be misleading (by explaining one outcome without giving a sense of the overall model); can risk overloading the audience with too much information and may fail to provide actionable insights once the number of relevant factors grows beyond a handful (Selbst and Barocas 2018); can cause audiences to place undue trust or overconfidence in the system (Kaur et al 2019); and can be deliberately designed to produce 'acceptable' explanations that hide the underlying unfairness of the system (Aïvodji et al 2019).

So to make best use of the existing technical methods, mindful of their shortcomings and the tradeoffs involved, we will need a much better understanding of individual contexts demanding an explanation. This requires an analysis of *when* an explanation of AI-informed decision-making is required, and in each of these cases:

1. *Who* is, or who are, the intended audience(s) of the explanation(s);
2. What is or are the *purpose(s)* of the entitlement to/requirement for an explanation;
3. What explanation(s) will serve the intended purpose(s) and the needs of the intended audience(s).

Only if we understand all of these critical aspects of context, can we then work out either how technology can help to meet these legal requirements, *or* how the legal requirements could be adjusted to make better use of the technology available to meet the intended purposes. We therefore endorse the Commission's call for research, and emphasise the importance of genuinely interdisciplinary work in this space: involving technical, legal, philosophical and other social science expertise. Some of the issues needing more research are outlined below.

When is a person entitled to an explanation?

The Commission's proposal limits legal consequences of non-explainability to cases where the law *requires* that an explanation be given. It doesn't create a right of explanation for every important decision. An initial important question, then, is: are there circumstances currently where there is no 'entitlement' to an explanation, where there *ought* to be? Further research/analysis is required.

There is also an open question how far the proposal extends, because the law does not always use this language: you can't search the statute books to find all the cases where a person is entitled to an 'explanation'. Administrative law requirements to 'give reasons' may be the closest legal analogue to a 'right to explanation'. Selbst and Barocas (2018) have identified other contexts in American law, such as in credit scoring. But there are other legal contexts which are less clear. Is a person giving consent to a medical procedure 'entitled to an explanation' by reason of the fact that informed consent is required or the procedure will be an assault? Is a person 'entitled to an explanation' in the context of wrongful termination of employment, if they are judged by AI to fail standards? Is a person who suspects discrimination on the basis of a protected attribute 'entitled to an explanation' if refused credit/a job interview? Or is the Human Rights Commission 'entitled to an explanation' when investigating a complaint?

Also worth exploring is the extent to which there is already a *de facto* right to an explanation in relation to many cases where the grounds for a decision can be challenged before the courts, via ordinary processes of discovery. If, for example, a patient suffers harm as a result of medical treatment recommended by an AI system, and the patient claims their consent wasn't properly 'informed', that patient can sue for negligence: could the court in any event require an explanation of the system in order to determine whether reliance on it was reasonable in all the circumstances (Cuéllar 2019)? It is possible that there are a very large number of cases where there is at least some *ex post* entitlement to an explanation, via the courts.

Who is the audience for the explanation?

The intended audience of an explanation matters a great deal; with different levels of explanation appropriate to different audiences. The intended audience may govern whether it is appropriate to provide explanation at (1) the system level (eg overall distribution of credit as between men and women; overall approval for interview in CV screening program as between white and non-white applicants), or (2) individual decisions.

The current proposals of the Commission seem to focus on explaining individual decisions, to the affected individual. The form of the current proposal also seems to suggest that the **individual** is entitled to both the "lay" and the "technical" explanation. What are the circumstances where the full technical explanation is not appropriate to reveal (eg, because the system is designed to detect fraud, and greater transparency would enable avoidance/gaming of the system)?

We also think a focus on the individual and individual decisions is too narrow.[5] First, there are going to be circumstances where an entity or institution that is not an individual should be entitled to require an explanation: a professional advisor or medical professional; an industry ombudsman; or a court/adjudicator/Fair Work Commission or Australian Human Rights Commission could all be important in giving effect to people's rights in relation to AI systems. There might be cases where it would be *more* appropriate for an independent body to request explanation or the means to evaluate the impact of a system.

Second, even an individual will sometimes need information beyond an explanation of a particular outcome. In the context of anti-discrimination law, explaining individual decisions may not assist in identifying indirect discrimination: information on overall model performance and distribution of outcomes would be important too. And as Tonekaboni et al (2019) have shown in research relating to the use of AI in medical contexts, for some audiences, information at both the individual decision and systems level are important, and not only to the affected individual. Tonekaboni et al's study involved interviewing and surveying medical clinicians about their information needs relating to AI models. They found that clinicians:

- Wanted to understand the clinically relevant model features: ie the subset of features deriving a particular model outcome- in order to compare model predictions to their clinical judgment;
- Expected to see patient-specific as well as population-level variable importance;
- Needed information about the context in which the model operates and situations where the model could fall short, so as to decide parameters for the use of the model and its predictions.

In other words, clinicians needed *both* individual and systems-level information, tailored towards their purpose of justifying their clinical decision-making (for instance, to patients and colleagues) in the context of the model's prediction.[6]

In our view, too, more research is required around how the various kinds of audiences understand and act on different kinds of explanations. Kaur et al's study (2020) suggests that even expert intended audiences for existing interpretability tools - data scientists - may misunderstand and misuse those tools. In that study, data analysts over-trusted the tools:

[5] We note that the GDPR may raise the possibility of different kinds of explanation, both 'meaningful information about the logic involved' under art 13 (which suggests a more systems-level explanation) and the possibility of contesting a decision (art 22) which would suggest more fine-grained information about a particular outcome.

[6] Tonekaboni's study related to AI in medical use in ICU contexts - meaning there might be no immediate 'entitlement' to an explanation as such. But the general point holds, that in some cases, the same individual might need an explanation at more than one level. And if harm results from a medical decision based on the predictions of an algorithm, there might well be a demand for explanation in any subsequent litigation or legal claim.

relying even on manipulated 'explanations' rationalising clearly problematic predictions generated by a model. Work would also be needed thinking about what happens when two explanations appear to differ (eg lay and technical; individual and system; or different explanations given to different individuals) : what is the impact on the credibility of the system?

What is the purpose of the explanation?

Possible reasons for providing an explanation include:

- To identify errors (diagnostic)
- To enable an affected individual to take action to change the outcome (actionable insights);
- To enable an affected individual to decide whether to challenge the outcome (contestability);
- To enable evaluation of a system: eg to evaluate whether it operates 'fairly' (acknowledging the many possible meanings for that term) across individuals or social groups.

The purposes are not mutually exclusive: more than one can be served in a given scenario.

But importantly, the appropriate form/level for an explanation varies according to the purpose or goal of the entitlement to an explanation. Miller (2019) goes into some detail about the kind of explanation that is appropriate for different contexts. The purpose of an explanation may also provide important context that will govern how explanation should be provided. Tonekaboni et al (2019) for example found that in the context of AI systems in use in intensive care units, insights that are not clinically actionable are not useful; and explanations provided in terms of 'similar past cases' were similarly not useful (they refer to this as 'domain appropriate representation').

A rebuttable presumption?

As mentioned at the start, we think that the proposals of the Commission need nuance. A rebuttable presumption, rather than outright prohibition, may be one way to provide some nuance. Although we would query whether, if you create a rebuttable presumption that a person responsible for a decision will not have made a lawful decision if a reasonable explanation is not provided, are you implying that it is sufficient for lawfulness if a reasonable explanation can be provided *ex post?* How would an obligation to provide *ex post* explanations interact with, for example, a duty to take reasonable care at the time a decision is made, or public law requirements to take relevant matters into account/not take into account irrelevant information? It would be important to take care not to create unintended consequences eg to change the law from an obligation to engage in lawful/proper decision-making processes *ex ante* as opposed to being able to explain/justify them *ex post*. It is possible that the system

would need to be more nuanced, imposing some level of obligation to understand a system at the time a decision is made, with a later obligation to provide more detailed evidence if the decision is later challenged judicially or eg within AHRC/Fair Work Commission/Medical Tribunal proceedings.

Access to technical information

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

The question of access to "technical information" is related to the discussion of explanation. In thinking about access to information to assess lawfulness, as with the question of explanation, there needs to be nuance around (1) **what:** what 'technical information' is required; (2) **why:** for what purpose; (3) **to whom:** who should have access to what in order to achieve that purpose; and (4) **what next:** what is done once access is obtained (whether broadly published or subject to confidentiality orders).

It is unclear what the AHRC considers "technical information". Information about the operation of a system can be supplied at different levels including without the production of source code. Some information would be "technical", but some non-technical information (eg, information about internal governance processes) may be just as relevant to assessing responsibility for a system and/or what might have gone wrong. To have confidence in a system, it may be sufficient to have some combination of information (depending on the purposes, and who receives access) about:

1. The history and provenance of the system
2. Training data used;
3. Data and data provenance (sources; currency; levels of pre-processing; methods for labelling etc)
4. Model(s) used;
5. Information on what factors are important in decision-making by the model;
6. Governance mechanisms around the system including methods for testing prior to deployment and review/audit post deployment;
7. Results of any testing/review/audit;
8. Outputs/outcomes of the system in operation following deployment.

Confidence in a system has a lot to do with which types of outputs the system produces for the range of inputs it is expected to receive. The challenge of course is that directly estimating

outcomes is not always easy, so we use input like training data and internals like algorithms to reason more effectively about what can or can't go wrong in terms of outcomes. At the very least the monitoring and testing of a variety of ethically relevant metrics on the outcomes of a system would be relevant.

One possible proposal or recommendation the AHRC could consider would be that some body or taskforce be asked to produce guidance, to provide decision-makers, auditors/ombudsmen, and members of the public with a more nuanced understanding of the kinds of information that could be demanded or published *without* revealing genuine trade secrets. This could assist both government bodies and members of the public to assess the lawfulness of a system.

It is also worth noting that there are already ways in law to require access to technical information used in AI-informed decision-making, although these vary according to who seeks the information, and whether it is sought from the public or private sector. For example:

- If use/impact of AI is challenged in court, courts have the power to require the production of confidential information, including trade secrets, and the power, where necessary, to protect the confidentiality of that information: see *Australian Competition and Consumer Commission v MSY Technology Pty Ltd* [2011] FCA 204; *Yara Australia Pty Ltd v Burrup Holdings Ltd (No 2)* [2010] FCA 1304; *Australian Competition and Consumer Commission v Cement Australia Pty Ltd (No 2)* [2010] FCA 1082. Courts have a range of inherent powers to manage evidence, including the power to restrict access to confidential information to the legal advisers/experts of an opposing party in proceedings to ensure further protection.
- The Australian Human Rights Commission has the power to require the production of information or documents, and the power to examine witnesses, in the discharge of its functions: *Australian Human Rights Commission Act 1986* (Cth) ss21-22. It has the power to make orders to preserve privacy or confidential information: s 14. Other regulators such as ASIC or the ACCC have similar powers. Any body set up as a result of other recommendations of the AHRC - such as an AI Safety Commissioner - would need similar powers.
- An entity (government or private sector) procuring an AI system for its own use could require information to be provided as an obligation under the contract: and if educated as suggested above could potentially be quite sophisticated about what it required;
- Some information about the operation of a technical system used by government can be sought through freedom of information processes. Such processes have been used, for example, to obtain information about the operation of #Robodebt.[7]

[7] See eg https://www.righttoknow.org.au/request/robodebt.

These tools however are incomplete. There may be *other* actors - other than regulators, governments, or litigants - who should have access to some level of information about systems that are used and how they impact. Legal mechanisms for *public* access to information about public or private use of AI are limited (Handler 2018). Whistleblower protection for those who disclose confidential information to the public (for example, about potentially dangerous uses of data analytics) is also limited. And while governments could require transparency from their commercial partners about the operation of data-driven systems, it is by no means clear that such transparency would flow through to transparency to the public; commercial sensitivity is a common ground for refusing access to information under freedom of information laws.

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

We agree that simply inserting a human decision-maker into this process will not guarantee better decisions.

At a very high level what matters in any decision making scenario are two questions (1) How do we measure the quality of a decision in this scenario? (2) How do we operationalise within this scenario a decision that is sufficiently 'good' as measured by the quality measure? The reality is that in different scenarios the answers to these two questions will differ in various respects, including in how humans should be involved. Humans are an important part of a decision making system; the extent and way in which they should be involved should depend on how likely it is said involvement will lead to a better decision.

In order to design effective decision governance mechanisms, it is important to understand the relative strengths and weaknesses of humans and machines, as they are today. For instance, humans are good at common sense reasoning, and machines are good at things like consistency, precision, finding correlations in data, etc. This suggests that humans should be trained to ask machines questions that they can be competent at answering and vice-versa. A good governance system is likely to be one in which there is some form of 'dialogue' between people and systems so as to transfer information efficiently both ways about things each knows best, up to a point in which the humans in charge of governing the system have high confidence that the system 'knows enough' and can be used (with continuous monitoring).

References

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, Alain Tapp, 'Fairwashing: the risk of rationalisation', Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019

Article 29 Data Protection Working Party (2018). Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 17/EN, WP251rev.01, as revised and adopted 6 February 2018.

Australian Law Reform Commission (2019), *The Future of Law Reform: A Suggested Program of Work 2020-25*

Will Bateman (2019a). Algorithmic Decision-Making and Legality: Public Law Dimensions (October 1, 2019). Australian Law Journal (Forthcoming). Available at SSRN: https://ssrn.com/abstract=3496386

Will Bateman (2019b). Automating Discretionary Decision-Making in the Public Sector: Legal Dimensions (November 1, 2019). Available at SSRN: https://ssrn.com/abstract=3493433 or http://dx.doi.org/10.2139/ssrn.3493433

Cuéllar, Mariano-Florentino (2019), A Common Law for the Age of Artificial Intelligence: Incremental Adjudication, Institutions, and Relational Non-Arbitrariness (November 20, 2019). *Columbia Law Review*, Vol. 119, No. 7, 2019. Available at SSRN: https://ssrn.com/abstract=3522733

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1721-1730. ACM

Jessica Fjeld et al (2020). 'Principled Artificial Intelligence: Mapping Consensus in ethical and rights-based approaches to principles for AI', Berkman Klein Center Research Publication No 2020-1.

Michael Handler, Reconsidering the need for defences to permit disclosures of confidential copyright material on public interest grounds (2018) 12 *Journal of Equity* 195

Anna Jobin et al (2019). 'The global landscape of AI ethics guidelines'. *Nature Machine Intelligence*, September 2019:389-399.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, Jennifer Wortman Vaughan (2019). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. CHI 2020, April 25–30, 2020, Honolulu, HI, USA.

Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

Jacob Metcalf, Emanuel Moss and danah boyd, 'Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics' (2019) 82(2) *Social Research: An International Quarterly*, 82:2, 449-476.

Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences', (2019) 267 *Artificial Intelligence* 1-38

Ming, Y., Cao, S., Zhang, R., Li, Z., Chen, Y., Song, Y., and Qu, H. (2017). Understanding hidden memories of recurrent neural networks. In 2017 IEEE Conference on Visual Analytics Science and Technology (VAST), pages 13{24. IEEE.

OECD (2019). *Artificial Intelligence in Society (Summary in English)* (2019), 1 <https://read.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society/summary/english_9f3159b8-en#page1>.

Andrew Selbst and Solon Barocas, 'The Intuitive Appeal of Explainable Machines' (2018) 87 *Fordham Law Review* 1085-1139.

Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, Anna Goldenberg (2019), 'What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use', *Proceedings of Machine Learning Research* 1-21, 2019.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579.