Gradient Institute Ltd.
Level 2 Merewether Building H04
Cnr City Rd & Butlin Ave
The University of Sydney NSW 2006
https://gradientinstitute.org

20 May 2022

Department of the Prime Minister and Cabinet
Digital Technology Taskforce
PO Box 6500
Canberra, ACT 2600

Dear Sir/Madam,

Response to Digital Technology Taskforce Issues Paper: Positioning Australia as a Leader in Digital Economy Regulation - Automated Decision Making and AI Regulation

Gradient Institute welcomes the initiative of PM&C to release this Issues Paper addressing a crucial matter for the future of Australia's economy, namely how Australia will regulate the digital economy and the use of AI and ADM in particular.

Gradient Institute has a focus on the scientific and technical aspects of Responsible AI. As such, we see our contribution to policy discussions as explaining how scientific and technical considerations can help evaluate the appropriateness of policy directions.

In this response, we address in order each of the questions posed in the Issues Paper.

1. **What are the most significant regulatory barriers to achieving the potential offered by AI and ADM? How can those barriers be overcome?**

We see as the main challenge not the fact that existing regulation is preventing AI/ADM from realising their positive potential, but the lack of systems, including regulatory systems, that prevent AI/ADM from realising their negative potential. AI/ADM are still immature technologies that, despite the positives, have already conclusively demonstrated great potential for causing harm.

When considering regulation aiming at increasing the adoption of AI/ADM, it is paramount to think preemptively about the risks and downsides of these technologies so as to avoid indiscriminate adoption, which may cause more harm than benefits.[1]

2. Are there specific examples of regulatory overlap or duplication that create a barrier to the adoption of AI or ADM? If so, how could that overlap or duplication be addressed?

There are cases when it isn't sufficiently clear how the law applies to AI/ADM. Anti-discrimination law is a salient case in point. We routinely observe organisations struggling with the lack of clarity about how to interpret anti-discrimination law in the case of automated decisions.

For example, discrimination on the basis of attributes protected under the law (such as race, age, disability or gender) is illegal. How can an organisation instruct their AIs not to discriminate in this sense? Perhaps by not providing the AI with the corresponding data flags for race, gender, etc? In the absence of clear guidance, that's what many organisations do. However this practice is known to be unsound and can even increase harm to the group it aims to protect.[2] Organisations are currently in the dark about how they should appropriately address this problem due to lack of clarity of how to encode compliance with anti-discrimination law into an AI.

What can be done? There is a need for lawmakers and technical experts to collaborate in order to establish guidance and any necessary legislative reform that is both sound according to the objectives of anti-discrimination law and sufficiently specific so as to be implementable in code. This is likely to be a taxing and lengthy enterprise, but it needs to be prioritised if we want automated decisions to adhere to similar anti-discrimination standards as human decisions.

3. What specific regulatory changes could the Commonwealth implement to promote increased adoption of AI and ADM? What are the costs and benefits (in general terms) of any suggested policy change?

---

[1] Several books for a wide readership are now available which report the many ways AI and ADM can and have created harm. For instance: O'Neil, C. (2017). Weapons of math destruction. Penguin Books; Eubanks, V. (2018). Automating inequality: how high-tech tools profile, police, and punish the poor. First edition. New York, NY: St. Martin's Press; Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press; Reich, R., Sahami, M., Weinstein, J.M. (2021). System Error: Where Big Tech Went Wrong and How We Can Reboot. HarperCollins Publishers.

[2] The issue stems from conflating the meaning people ascribe to "direct discrimination on the basis of X" and the meaning machines ascribe to "presence of data flag X". These statements are not equivalent, for several reasons. One is that other variables can stand in as proxies for X (e.g. if X is gender, occupation can be a proxy), thus not entirely removing information about X. Another is that not all judgement based on X is discriminatory (e.g. it can be legitimate to use gender information for the purposes of marketing mentrual pads). Yet another is that whereas the former statement has scope for interpretation, the latter has a unique and precise mathematical interpretation. For more information see https://medium.com/gradient-institute/ignorance-isnt-bliss-6d133ee00f51

As a preamble to answering this, we wish to highlight the perhaps self-evident fact that the main lever to increase adoption of AI and ADM is not regulation, but investment. Powerful levers to increase AI adoption are (i) more investment in cultural change towards research and innovation (ii) more investment in AI research (iii) more investment in entrepreneurial and VC ecosystems (iv) more investment in targeted education programs.

In terms of regulatory levers, at a general rather than specific level, we do believe the most effective approach to increase appropriate adoption of AI and ADM is *to encourage a mature risk management mindset* in public, private and non-profit sectors. Among other things, this means educating them on the reasons and motivations to manage AI/ADM risk, as well as working with expert organisations to provide these sectors the tools and resources required to identify and mitigate risks, and offer remedy where things go wrong.

A risk management mindset is crucial because the costs and benefits of using AI/ADM can be greatly uncertain. This is due to the novelty of the technology, its intrinsic complexity, and its extrinsic complexity reflected on the multitude of ways in which it can interact with the outside world. The realised costs and benefits will be greatly dependent on the specific context in which the technology is used, which demands risk management tools that can be applied in a bespoke manner.

We emphasise the need for "mature" risk management because the *dynamics* of the use of AI/ADM, as well as of any regulatory response, should be accounted for (it often isn't).

For instance, "tail risks" will play an important role when considering uncertainty arising from the use of AI/ADM.[3] In the same way as the risks of pandemics were underestimated before COVID, and risks of nuclear war were underestimated before the recent ongoing conflict in Europe, risks of catastrophic harms arising from the use of AI systems will likely be underestimated before something very bad happens – unless deliberate effort is made to manage tail risks. This strongly suggests that any specific use of AI/ADM should be assessed with respect to whether it constitutes a tail risk, and if so ensure that measures are taken to control the worst-case scenario (ensuring downside certainty), even if it's at the risk of potentially reducing the best-case scenario (letting go of upside certainty). Trade downside uncertainty for upside uncertainty – not the opposite.

[3] Tail risks (https://en.wikipedia.org/wiki/Tail_risk) can emerge in hyper-connected complex systems where "bad things" or "errors" *multiply* instead of just adding up. This can lead to an exponential increase in risk, which can grow faster than our ability to respond appropriately. This is the case for infectious diseases, which can rapidly cause a pandemic, as well as a nuclear attack, which can rapidly lead to all-out nuclear exchange. AI/ADM, depending on the context of its use, could also fall into this category of risks.

4. Are there specific examples where regulations have limited opportunities to innovate through the adoption of AI or ADM?

See answers to Q2 and Q3.

5. Are there opportunities to make regulation more technology neutral, so that it will more apply more appropriately to AI, ADM and future changes to technology?

This question suggests another: if we want to make regulation more "technology neutral", doesn't that imply that perhaps technology itself may not be quite the right target of regulation after all?[4]

Technological development moves faster than regulation. This is one reason why having a fast evolving technology as the core target of regulation is a fragile idea. It is more robust if regulation targets the *terms according to which technological artefacts interact with the world*.[5] It is also more pragmatic, since it's those terms that determine the impact the technology has on the world – which is what matters after all.[6]

For instance, regulation will most certainly be required to determine the terms according to which self-driving cars will be allowed to operate on the roads, but setting precise rules on how to collect the training data used by their machine learning algorithms is both ineffective and unnecessary to achieve that goal.[7]

Or consider the use of ADM platforms for matching job seekers to job opportunities. Regulation is required to determine the terms according to which the matchings should be allowed to happen. For instance, anti-discrimination law establishes certain *fairness* terms according to which certain decisions impacting people are allowed to be made – which would apply in this case. As we've seen in the answer to Q2, this is technically challenging to quantify in terms of ADM decisions, but it's a legal requirement and the legislation should evolve to clarify what accounts for illegal discrimination in the context of ADM. It would instead be counterproductive (and both ineffective

[4] Bennett Moses, Lyria, How to Think About Law, Regulation and Technology: Problems with 'Technology' as a Regulatory Target (2013). (2013) 5(1) Law, Innovation and Technology 1-20, UNSW Law Research Paper No. 2014-30, Available at SSRN: https://ssrn.com/abstract=2464750

[5] There may be cases where the development of a certain technology may be in and of itself unacceptable, regardless of the context of its use. This may arise for example when some feasible use of the technology entails material existential or catastrophic risks.

[6] More rigorously, this conclusion follows from the intuitive mathematical concept of conditional independence as applied to causal reasoning, which in non-technical language says that if the only path from X to Z is through Y, and if we can fix Y, any residual change in X doesn't affect Z. See https://en.wikipedia.org/wiki/Conditional_independence

[7] Ineffective because it's the company developing the vehicle that has the best knowledge of how to engineer its systems so as to achieve set constraints on how the vehicle should operate on the roads; unnecessary because of the previous footnote.

and unnecessary) to focus the efforts of regulation on which types of "matching algorithms" could be used, for instance whether based on deep learning or not.[8]

One powerful heuristic to design durable artefacts can be derived from the mathematically substantiated "Lindy effect",[9] which says that, for anything without an expiry date, life expectancy *grows* with age.[10] This could be applied to regulation: When designing a regulatory intervention aiming to be effective over the next N years, pretend to be N years into the past and try to imagine which intervention, if designed then, would still be effective today. If you design this intervention now, there is a good chance it will be effective N years from now.

The reason this is a good rule of thumb is that it doesn't allow you to consider anything that happened in the recent past and therefore may be too ephemeral ("the new gets replaced by the newer, the old doesn't get replaced"). For instance, regulation focusing on compliance with rules depending on specific aspects of recent technologies would likely soon become inapplicable as these aspects change. On the other hand, regulation focusing on what types of harms the technology should in principle avoid creating (e.g. manipulation, unfair treatment, lack of access to recourse, carbon emissions, viral contagion, etc.) would be much less fragile as the nature of harms isn't so dependent on the specifics of the technologies potentially causing them.

6. Are there actions that regulators could be taking to facilitate the adoption of AI and ADM?

See answer to Q3.

7. Is there a need for new regulation or guidance to minimise existing and emerging risks of adopting AI and ADM?

Yes. However, noting the above, it will be more helpful if new regulation/guidance doesn't focus on the technology itself but on the terms of its interaction with the world. For instance, regulation and guidance focused on preventing or reducing specific types of potential *harms* caused by AI/ADM are likely to be more actionable and effective than restrictions on the internal workings of the technology. On this matter, we refer to this recent report by Gradient Institute on management of AI risk and governance: https://gradientinstitute.org/de-risking-automated-decisions/

[8] For the same reasons as those in the previous example and footnote.
[9] https://en.wikipedia.org/wiki/Lindy_effect
[10] For instance: ideas, books, etc. The older an idea has been around for, the longer it's expected to endure; the longer a book has been in print, the longer it can be expected to be in print. The opposite is true for "perishable" things, or things *with* an expiry date (e.g. people: life expectancy reduces with age).

8. Would increased automation of decision making have adverse implications for vulnerable groups? How could any adverse implications be ameliorated?

Increased use of ADM systems has the *potential* to cause greater harm to individuals in "vulnerable circumstances" than to other individuals – it won't necessarily do so, and that very much depends on the extent to which there are failures of legitimacy, design and implementation of such systems.

This potential for harm is, at present, often realised due to such failures being the norm. For instance, since there is less data on minorities than on the general population, an off-the-shelf application of AI without due consideration to its impacts (which is still the norm) will likely use data based on its accessibility, leading to more mistakes on minority groups.[11] Another example is that "negative treatments", such as debt collection, are correlated with various forms of vulnerability, and thus any errors due to poor design and implementation may differentially harm the most vulnerable (as was the case with Robodebt[12]).

However, use of ADM also has the potential to benefit minorities. When designing an ADM system, we can make deliberate choices about how to distribute burdens and benefits across different segments of the population, based on the data telling us who is whom. If we wish to afford certain benefits to a specific minority, we can explicitly design the system to do so, sometimes even without the need of any material sacrifice to the majority, or other minority groups.[13]

In summary, what determines whether minorities suffer adverse consequences from the use of ADM fundamentally comes down to matters of legitimacy, design and implementation/execution. All these can be addressed in practice. Gradient Institute's report we referred to above points to practices and tools to ameliorate legitimacy, design and execution risks of ADM.[14]

9. Are there specific circumstances in which AI or ADM are not appropriate?

In order to sensibly determine whether AI/ADM are or are not appropriate in a certain circumstance, we require decision *criteria* that speak to "appropriateness". Gut feelings, intuitions or moral injunctions without reasonable justifications can work, but can also fail, and without a good theory we can't know when and by how much they can fail, which is a significant risk.[15]

---

[11] With less data to learn from, the AI will tend to make more errors.

[12] *Prygodicz v Commonwealth of Australia (No 2)* [2021] FCA 634

[13] This blog post by Gradient Institute explains in detail how this can be achieved in certain cases: https://ambiata.com/blog/2021-03-22-nba-for-social-good/

[14] https://gradientinstitute.org/de-risking-automated-decisions/

[15] We may intuit that some types of decisions should never be made "by machines", and yet our moral intuitions can be less trustworthy than we take them to be. For instance, some people will argue that on matters of life and death, a machine should not have the final word. But what if the potential victim prefers the machine's decision? And what if they have a strong argument supporting that preference? Consider the following thought experiment. Ten expert

In a recent report, Gradient Institute has outlined some criteria to help guide practitioners to assess the appropriateness of using AI/ADM when the application context is known.[16]

Within a given application context, we must compare AI/ADM with the feasible and reasonable alternatives. Could humans do the job? If yes, how well could they do it? What happens if the job isn't done? Some type of cost-benefit thinking is helpful, while acknowledging uncertainty in determining the costs and benefits in each circumstance, and managing downside risk. Other moral considerations are also appropriate, such as legitimacy: what do the people potentially affected think? In asking them, have they been provided an accurate account of the best knowledge of the likely material consequences of each choice? (e.g. using AI, not using AI, ignoring the problem, etc).

10. Are there international policy measures, legal frameworks or proposals on AI or ADM that should be considered for adoption in Australia? Is consistency or interoperability with foreign approaches desirable?

We don't have specific views on consistency and interoperability. With regards to references for adoption, we believe there are positive examples to be learned from and negative examples to avoid.

Specifically, let's consider the major new legal regulatory framework for AI: the European proposal for an "Artificial Intelligence Act".[17] In general terms, we wholeheartedly agree with one aspect of this proposal, namely the notion that a *risk management approach* is required – as should be clear from our answers to previous questions.

On the other hand, as should also be clear from previous answers, we are not in favour of the approach taken by the EU of framing the regulatory intervention as one that broadly targets "AI" itself as a technology.

---

surgeons are asked to give their own opinion about whether a patient should undergo a life or death surgery (let's assume the goal asked by the patient is to maximise their expected lifespan). They all do so, and some advise to proceed with the surgery while others don't. An AI is also given the opportunity to give its own "opinion". It says its opinion will be based on a weighted majority vote of the opinions of the experts in which each surgeon's opinion is weighted by a proxy of their competence in similar cases: the observed lifespan of their previous patients with similar risk factors who were subject to the same kind of decision in the past; the opinion with highest total weight is then chosen. In this situation, which decision would the patient prefer, the AI's or the decision of any other doctor? This example not only shows that our moral intuitions about what AI should be allowed to do may be less robust than we think, but also that we can't confuse "using machines" with "not using people".

[16] https://gradientinstitute.org/de-risking-automated-decisions/ Appendix A.

[17] Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act), Brussels, 25.11.2020 COM(2020) 767 Final, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

We welcome the opportunity to discuss these issues further with PM&C through the consultation. Through our work helping Australian organisations ensure that their AI systems are operating responsibly, we are continually learning about what makes for good governance of AI systems and this knowledge can be used to help Australia develop good regulation that supports the appropriate and broad adoption of AI/ADM.

Tiberio Caetano
Chief Scientist

William Simpson-Young
Chief Executive