Gradient Institute

"Artificial Intelligence: Australia's Ethics Framework" CSIRO's Data61 discussion paper

Gradient Institute's submission to the public consultation

Tiberio Caetano and Bill Simpson-Young

(on behalf of the Gradient Institute team)

Executive Summary

This document constitutes Gradient Institute's[1] submission to the public consultation[2] on the discussion paper "Artificial Intelligence: Australia's Ethics Framework (A Discussion Paper)", developed by CSIRO's Data61 and released by the Department of Industry, Innovation and Science on 5 April 2019. We enthusiastically welcome the Australian Government's initiative to start a public discussion on Ethical AI. We congratulate the Department of Industry, Innovation and Science, as well as CSIRO's Data61, and all individuals who have contributed to the work that culminated with the release of the Discussion Paper.

There is urgency to develop an Ethics Framework for AI.

In the space of decades, AI systems have migrated from science fiction, to the lab, to the real world. The AI technology humans are building is already steering the lives of billions of people in ways unknown to them, and the sophistication and reach of this influence are growing very rapidly. Data-driven algorithms are deciding who gets insurance, who gets a loan, and who gets a job. Parole and sentencing risk scores, social media feeds, web search results, traffic routes, advertising, job recruitment and online dating recommendations are all consequential decisions that are already algorithmically personalised today. Powered by algorithms and data, AI systems are increasingly helping decide what happens in the lives of billions of people.

The ubiquity of AI systems could be a great thing if we *knew* their influence was for the good of all. But we know that to be false. Numerous media headlines in recent years have illuminated that AI systems can make all types of unethical decisions, and in particular cause even greater harm to already vulnerable people. If we, as a society, want to fix this, we need to understand the causes. First, most AI systems currently in operation have not been designed, developed and deployed so as to assure their decisions are ethical and avoid harm. In contrast to humans, who are incentivised to uphold norms of behaviour that are at least perceived to be sufficiently ethical so as to afford social acceptance, current algorithms have no such default ethical restraints - any restraints have to be *designed into them*. Second, AI systems are based on learning from data, and learning can only be based on what *did happen* in the past, not on what *should* have happened but didn't. Therefore, if we blindly apply the learnings of a machine learning algorithm to decide what *should* be done into the future, we are effectively building a system to perpetuate the past. Often this will create a positive reinforcement cycle that will not only perpetuate existing inequalities, but further amplify them.

[1] https://gradientinstitute.org/
[2] https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/

The status quo does not align with the goal of creating fair and just outcomes for the good of all. Instead, it gives us an imperative to embed ethical considerations into the design, development and deployment of AI systems, in order to best ensure such systems operate ethically.

In this submission, we provide our current views on some of the actions we believe need to be taken by the Australian Government to address the challenge of making AI systems operate ethically. Although we do so in the context of the Discussion Paper, the primary intention of our response is to direct attention to what we think are some essential considerations the Australian Government needs to embrace to successfully address the challenge.

Gradient Institute's expertise lies eminently on the *technical* considerations pertinent to Ethical AI systems. As such, this submission focuses in large part on recommendations about the importance of getting the technical components right to ensure the *intent* of realising Ethical AI is materialised into the *actions* coming out of an AI system.

Recommendations

The Australian Government should address the issues around Ethical AI in the following ways:

- **It should clearly state a set of *Ethical Principles* that aim to encapsulate, at the most fundamental level, what is necessary and ideally sufficient to guide ethical action for Australia as a country.**

  The development of clear *Ethical Principles* will address the question, 'What are the ultimate values that an Ethics Framework for Australia should promote?'. For example: **wellbeing, fairness and autonomy** could be a tentative set of values that capture the imperatives of increasing wellbeing and reducing harm, while distributing goods and any harms fairly and at the same time respecting the autonomy (e.g. freedom of choice) of individuals.

- **It should ensure that in the list of proposed *Ethical Principles* there are not only concepts addressing wellbeing and fairness, but also *autonomy*.**

  The concept of *autonomy* is missing from the set of principles proposed in the Discussion Paper. This is related to the value of individual liberty, freedom and choice, which is core to liberal democracies. An AI system that generated great and fairly distributed wellbeing but didn't allow individuals any freedom of choice of whether or how to interact with the technology would potentially satisfy all the proposed principles. In addition, there needs to be a recognition that the extent to which people can exercise their autonomy when dealing with government is fundamentally different than when dealing with industry.

- **It should recognise, acknowledge and embrace the fact that *there are fundamental trade-offs* when making ethical decisions**

  Ethics is complex largely because of *fundamental trade-offs*. If we knew how to take actions that equally helped everyone in everything that mattered to them, Ethics would be much simpler. Making trade-offs is complex. It means thinking about the nature of the harm suffered by different groups, and perhaps identifying certain harms as unacceptable, while others as acceptable but with some attempt to mitigate or compensate for the harm suffered. This has nothing to do with algorithms, data or AI, but is simply a fact of the world. There needs to be a recognition that algorithms and AI systems, in particular, will be also subject to such fundamental trade-offs.

- **It should establish *Ethical Infrastructure* that supports the realisation of the *Ethical Principles*.**

  The development of *Ethical Infrastructure* must be undertaken as a separate exercise to developing *Ethical Principles.* The development of *Ethical Infrastructure* should take into account technology-specific considerations (such as those AI-related), as well as legal, governance, economic, environmental and any other societal considerations.

  In this document we aim to provide a *technical* perspective on key requirements for *Ethical Infrastructure* as far as AI systems are concerned, since that's where our expertise lies. In particular we emphasise the need for:

  (i) the mathematical quantification of ethical concepts,
  (ii) the use of a scientific approach when designing AI systems, and
  (iii) the use of engineering principles and methodologies to design, build, deploy and validate AI systems.

  Importantly, *Ethical Principles* and *Ethical Infrastructure* (including any AI-related components) should be clearly disentangled and demarcated. The former is about what values we want as a society and the latter is about how those values can be realised by systems within our society. This is not only motivated by conceptual clarity, but also by pragmatic reasons since changes in *Ethical Infrastructure* should not imply changes in *Ethical Principles*.

- **It should recognise, acknowledge and embrace the ideal of *precise and accurate quantification of ethical objectives and trade-offs***

The imperatives of precision and accuracy should underpin the development of policy, regulation and legislation. There must be an acknowledgement of the necessity to **quantify ethics with mathematical precision** so it can be encoded into machines, which only accept mathematical instructions. In addition, there must be an emphasis on the importance of *accuracy*, which is different from precision[3]. For instance, the more accurately the *objectives* for an AI system are specified, the smaller the chances of unintended consequences.

What is perhaps not so much emphasized is that *fundamental trade-offs* must also be precisely and accurately quantified. Having to live with fundamental trade-offs is hard enough. We should not accept having to live with trade-offs that apparently exist but in reality are illusions. Decisions informed by false trade-offs often cause unnecessary harm. For instance, if a government agency makes the executive decision not to collect gender information from its clients with a view to avoid unintentional discrimination at the expense of a loss in overall accuracy, it is effectively making a false trade-off. Gender discrimination not only can still happen after explicitly removing gender information, but can be even amplified *because* this "protected attribute" was removed.[4] Had gender data been proactively considered in an effort to avoid discrimination on the basis of this protected attribute, the government agency could have had a stronger chance of achieving its goal. False trade-offs made out of ignorance are commonplace, and technical competence is paramount to avoid them. This means it is imperative to characterise ethical trade-offs as accurately as possible. Any inaccuracies will give room to making false trade-offs, which may give rise to unnecessary harm.

- **It should acknowledge and embrace the ideal that bespoke scientific analysis that resists a prescriptive or rule-based approach is a *requirement* for realising ethical AI systems.**

  The truth is that we, as a society, don't yet know how to build ethical AI. We need to recognise that truth if we want to make real progress. At present, most aspects in the design and development of ethical AI systems for decision-making require bespoke scientific analysis and investigation. When building systems that make consequential decisions about people, there must be an emphasis to **uphold a scientific approach** in their design, development and deployment, as opposed to applying "off-the-shelf" tools for decision-making, lest we face unintended consequences. A scientific approach, among other things, recognises that:

[3] Precision relates to the granularity or resolution with which we represent a certain quantity, whereas accuracy relates to the extent to which the representation faithfully encodes the truth.
[4] This approach is called "fairness through unawareness" and is known by the algorithmic fairness research community to be a flawed strategy to pursue fairness. See this blog post from Gradient Institute for a detailed case study: https://gradientinstitute.org/blog/2/

(i) *uncertainty* is always present and needs to be acknowledged and incorporated quantitatively into decision-making processes,
(ii) understanding of *causality*, not correlation, is what informs which decisions are likely to cause which consequences.

It is crucial to draw significant attention to these points because the current success of AI systems is overwhelmingly due to their capacity to detect *correlations*, rather than accurately representing uncertainty or causality. In order to effectively deal with these factors, we cannot rely only on existing machine learning technology in order to design AI systems. We need to fully subscribe to a bespoke scientific approach which requires significant human involvement. The recent "Robo-debt" episode is a good example of a failure to do so: Centrelink didn't take proper care in *quantifying the uncertainty* associated with an averaging procedure that determined potential debt recovery amounts for welfare recipients. This resulted in large-scale automated issuing of debt recovery notices that often greatly exceeded the true amount owed.[5]

- **It should acknowledge and embrace the ideal that treating AI systems as *engineering artefacts* is a *requirement* for realising ethical AI systems.**

  There should be a strong emphasis on the imperative to ***properly engineer*** **AI systems** according to methodologies and practices of software and systems engineering, and in addition to recognise the need for a new engineering discipline that caters for the data-driven reality of software systems driven by data. This is required to afford the *technical assurance* that the systems will be fair, secure, and respect privacy and autonomy while being performant.

- **It should acknowledge and embrace the ideal that effective systems of accountability, legal and otherwise, are a *requirement* for realising ethical AI systems - in particular the need for a legal treatment of discrimination.**

  Although **accountability** is listed as a principle in the Discussion Paper, there isn't much discussion on the imperative of creating effective systems of organisational, legal and societal accountability. This is one of the most urgent and complex matters in the debate around Ethical AI, and needs to be addressed by the Australian Government immediately and effectively.

[5] The New Digital Future for Welfare: Debts without Legal Proof or Moral Authority? Terry Carney. UNSW Law Journal Forum, pp 1-16, 2018

In particular, there must be significant attention devoted to the **legal treatment of discrimination,** including the concepts of direct and indirect discrimination, which are pertinent to ethical AI. There is a need to deeply scrutinise the existing anti-discrimination legislation and identify if, where and how it may fall short with regards to the new reality imposed by automated decision systems. It is imperative to ensure that anti-discrimination legislation be suitable in the context of decisions made by AI systems.

- **It should acknowledge the different regulatory regimes between industry and government when addressing the issue of AI systems.**

  These two types of organisations are subject to fundamentally different incentives and constraints of operation. Also, the nature of the impact of Government and Industry interventions is different, in particular the impact on human autonomy, given the fundamental differences in the extent to which autonomy can be exercised in each of these sectors. Arguably, government has additional responsibilities. In the legal system, it is often said that government should be a "model litigant"[6] - that is, it should not just observe the law and the rules, but it should go above and beyond to act appropriately and not rely on loopholes to frustrate the goals of the law. In the same way, government should be perhaps the 'model deployer of technology' - it should not just aim to 'get away with what it can' or accept minimum viable protections for affected people, but uphold the true spirit of ethical principles.

What we need to realise Ethical AI

Here we provide a conceptual framework that reflects our thinking in order to contextualise the above recommendations.

- **Ethical AI is guided by humans**.

  There is no need to say that Ethical AI *must* be guided by humans. It simply *is* guided by humans. There is no one else in charge. Humans choose both the *values* they uphold and the *actions* they take. Building Ethical AI is a human enterprise. From writing mathematical equations that quantify ethical objectives to be programmed into AI systems to enacting new

[6] Model Litigant Obligations: What are They and How are They Enforced? Eugene Wheelahan, Federal Court Ethics Seminar Series.
https://www.fedcourt.gov.au/digital-law-library/seminars/ethics-seminar-series/20160315-eugene-wheelahan

anti-discrimination regulation catering for automated decision systems, *it's humans who are in charge*.

We need to be wary of anthropomorphising AI. AI is software ultimately produced by people, by leveraging data collected by people. AI has no agency of its own. It has no internal values, no intentionality, no purpose - we do. The reason why it sometimes *seems* like AI systems have agency is because humans have become really good at building systems that *mimic* certain human capabilities, which we exercise with agency. But this an illusion - AI systems have no agency.

Humans decide what AI systems can and can't do. Humans decide what goals to program into an AI system. Humans decide which legislation to enact and which other systems to build to surround AI for legitimacy and accountability. We just need to pursue the *knowledge* of how to do all of this effectively.

- **Ethical AI should be driven by a "technical triad": Mathematics, Science and Engineering**.

Mathematics. AI systems are computer systems and as such they speak only one language: mathematics. Or, more precisely, they work by maximising objectives that must be specified mathematically. This means if we want to make AI systems operate according to a set of defined ethical principles, we must *quantify ethics* with mathematical precision. For instance, a computer doesn't understand the instructions "be unbiased" or "be fair". We are actually forced to specify what we mean with ultimate levels of mathematical precision, such as "change prices so as to maximise total profitability conditional on the profit margin on males and females differing by no more than 10%". When we conduct this exercise transparently we illuminate the trade-offs that we make through the quantification of ethics - it doesn't take long for us to realise that some of the 'ethical principles' we would like to enforce end up being mathematically incompatible, and force us to prioritise one principle over another.[7] As such, quantifying the system's ethical intent or objectives is not only a requirement for ensuring AI systems operate ethically, but is also a way to make transparent for humans what is possible and what isn't, thus helping them to refine their own ethical thinking while operating such systems.

Science. In order to create AI systems that operate ethically we need to understand the cause and effect relationships between different design choices made in the design of these systems

[7] Inherent Trade-Offs in the Fair Determination of Risk Scores. J. Kleinberg, S. Mullainathan, M. Raghavan. https://arxiv.org/abs/1609.05807

and the ethical valence of the outcomes they produce. This is because we may commit to an action with positive intent, but that action may cause harm that we did not intend. The only known reliable strategy to infer cause and effect relationships are the methods of *science*. Rigorous measurement, observation, controlled experimentation, explicit modelling and quantification of uncertainty are some of the hallmarks of the scientific approach. In order to accurately assess the uncertainties and risks associated with a deployment and maximise our confidence in the outcomes of AI systems, a commitment to a scientific approach is paramount.

Engineering. The reason why people trust an airplane with their lives is not because they understand the inner workings of the machine, but simply because they believe that airplanes just *work*. The same is true for bridges - drivers without a degree in civil engineering are still happy to cross them.[8] Humans take for granted other engineering marvels such as electricity on demand and safe online credit card transactions. Engineering is successful precisely when we don't notice its presence, because successful engineering means that things simply work. It is imperative that AI systems are *engineered competently*. As a consequence of mathematising ethical ideals and understanding from the science of ethical AI, we will be able to develop mechanisms of *assuring* the degree to which AI systems will be fair, secure, and respect privacy while being performant. Mathematics, Science and Engineering form a core "technical triad" to support the creation of ethical AI systems.

- **Ethical AI needs *Ethical Infrastructure***

As mentioned in the recommendations, *Ethical Infrastructure* must exist to support the realisation of the ethical principles. This infrastructure should have two sub-components: *Intent Infrastructure* and *Action Infrastructure*.

Intent Infrastructure. AI systems are deployed by organisations, and if the *intent* of such organisations is unethical, the objectives programmed into these systems will be unethical, which will result in unethical actions. Due to the natural forces driving human nature and markets, organisations require *incentives* to ground their actions in ethical intent. For example, a key approach to generate incentives to behave ethically is through *effective systems of accountability*. Accountability systems can be both regulatory and non-regulatory (such as board oversight and employee, public and customer pressure).

To ensure that AI systems operate ethically, it is therefore crucial that Government and society more broadly work towards the emergence of an *Intent Infrastructure*, the purpose of which is

[8] It wasn't always like that. During the Victorian era several bridges collapsed because people didn't know how to build and deploy safe bridges. Likewise the security of airplanes increased significantly since their first deployments.

to drive organisations to pursue ethical goals. When doing so, it is crucial to acknowledge the fundamental differences between industrial and governmental organisations - particularly with regard to their legal obligations. *Intent Infrastructure must be technically informed*. For instance, when developing new anti-discrimination legislation that aims to take into account algorithmic decisions, it is crucial to ensure that the natural language phrasing of a new law doesn't create loopholes that become only evident under mathematical scrutiny. If such loopholes are produced, organisations who deeply understand the technology and its underlying mathematics will be potentially able to identify and exploit them to find new paths towards achieving unethical outcomes. One way to avoid this is to ensure that when creating new legislation a proper consultation process takes place with subject matter experts.

Action Infrastructure. Ethical intent does not guarantee ethical action. *Action Infrastructure* needs to be built between the world of ethical principles and the tangible actions and artefacts that an organisation utilises to realise value. The purpose of this infrastructure must be to minimise the translation error between ethical intent (e.g. the values and principles an organisation forms as part of its purpose) and ethical execution, such as that which is driven by computer code the organisation is responsible for. Action Infrastructure is what is needed to prevent the occurrence of "unintended consequences" in the deployment of AI systems.

Action Infrastructure will require non-technical and technical components.

On the non-technical side, we must ask: how could we utilise organisational governance models in an age of AI systems to actualise the concepts and goals agreed in the organisation's ethical principles? For instance, the question of minimising the principal-agent problem whilst still giving developers and model users the flexibility to deliver competitively and effectively. Potentially the use of risk and compliance management could effectively assure action against corporate policy goals, and flow down mitigation to the level of individual activity, allowing autonomy within an envelope whilst protecting boundaries of permissible actions. *Action Infrastructure* should also address questions of organisational provenance. What happens when key designers and executives responsible for an AI system leave an organisation? Can the organisation still justify the use of the system, its settings, or even understand what the system is trying to achieve?

On the technical side, interactive visual interfaces will need to be designed and developed to help humans clearly visualise the trade-offs between promoting different versions of objective mathematical measures of ethical goals. This will help people use familiar forms of communication such as visual and written data, to accurately understand the hard-to-grasp mathematical representations of ethical concepts that are present in the underlying code, the algorithms and data used, and the models created. Supporting such interfaces in the backend

there must be properly engineered systems implementing the mathematics and science used to quantitatively represent ethical reasoning and ideals.

Response to "questions for consideration"

The discussion paper released by the Department of Industry, Innovation and Science includes seven questions for consideration. This section provides Gradient Institute's answers to these questions.

1. *Are the principles put forward in the discussion paper the right ones? Is anything missing?*

The set of 'principles' proposed entangles fundamental ethical principles and methods to realise them. These elements should be clearly demarcated not only for conceptual clarity, but also for very practical reasons such as avoiding that any methodological changes necessarily impact the fundamental principles.

Some principles may be better framed as requirements, or desirable properties (such as what has been done in the European ethics guidelines for trustworthy AI[9]). Finally, the discussion offered doesn't acknowledge, let alone emphasize, perhaps the most important and difficult part of developing ethical AI systems: the process of *deciding* and *quantifying* the objectives and *making trade-offs* between competing objectives.

There is also an important redundancy (related to principles 1 and 2 - see "additional comments" below) and at least one important omission: a principle of **autonomy**.

Additional comments:

- The principles outlined don't include an acknowledgement of *uncertainty*, i.e., the fact that we never have complete knowledge about the world and any attempts to empirically assess notions of "harm" or "fairness" will be subject to error.[10]
- There are limitations in Principle 1, 'generate net benefits' and Principle 2, 'do no harm'. Assuming we are talking about ethically consequential actions, then Principle 2 logically implies Principle 1 (since doing no harm to anyone implies the only actions allowed are those that only cause good, and those always have positive net value, so their total is also positive). An important limitation of Principle 2 is that it may encourage an exceedingly conservative attitude towards causing good (for instance, the deployment of a system that improves the

[9] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[10] In general, there is a lack of recognition of epistemic limitations when defining or making inferences about ethically relevant concepts.

lives of thousands of people at the cost of causing minimum harm to a single person would be discouraged by this principle). For those reasons, we believe these two principles need to be combined and revised. We believe phrasings like "promoting **wellbeing** and demoting harm" would capture the likely intent behind Principles 1 and 2. The principle of "Prevention of harm" from the European guidelines is an improvement to the current language, however it doesn't capture the promotion of wellbeing or the remediation of harm, so we think it's not sufficient. Also, in cases where preventing harm upfront isn't possible, then compensation or mitigation steps to ensure the problem is addressed may be necessary.

- We recommend that Principle 3, 'regulatory and legal compliance' should not be a principle, but a requirement, since it is simply stating that the law must be obeyed.
- We consider that Principle 4, 'privacy protection' is addressing an important concern. The treatment of this topic can be significantly improved, and we recommend drawing from the approach outlined in the Salinger Privacy response[11]. We see privacy as a value in the service of more fundamental principles, such as **wellbeing, fairness and autonomy**. Privacy could instead be phrased as a desired property.
- We agree that Principle 5, '**fairness**' should be a fundamental principle. The description of this Principle could however be improved. *"This requires particular attention to ensure the "training data" is free from bias or characteristics which may cause the algorithm to behave unfairly."*[12] This statement is somewhat misleading. Removing bias from training data (pre-processing) is not the only (or generally the most effective) way to ensure algorithms satisfy fairness principles. The statement is also somewhat imprecise. There are different ways to interpret the word *bias*. For instance, if we interpret "bias" as a form of "unfairness", then every data set, which is inherently biased, will be considered unfair. Also, reference to "training data" must be further qualified. In the standard *supervised learning* approach, there is a conceptual distinction between training and "test" data. However, more realistic settings of deployment of AI systems involve feedback loops, as characterised by *reinforcement learning*, and such a distinction is no longer meaningful.
- We disagree with Principle 6, 'Transparency and Explainability' because we see these as potential mechanisms for affording the *assurance* that the system is operating in accordance with certain requirements. We emphasise that there are alternative ways to seek assurance. Rarely do we request to see the precise technical specifications of the technology that we interact with: the people who trust an airplane with their lives don't demand the machine's inner workings to be transparent or explainable to them before boarding. We recognise that transparency and explainability can be very desirable and even required properties in many cases, but we reject the idea that they ought to constitute undisputable principles.
- We recommend that more thought be given to Principles 7 and 8, 'contestability' and 'accountability': These concepts may be viewed as requirements or desirable properties rather

[11] https://www.salingerprivacy.com.au/2019/04/27/ai-ethics/
[12] *Discussion Paper,* pg.6.

than principles. For instance, contestability seems to be a way to exercise the principle of **autonomy**. It is important to recognise that notions of contestability, recourse and redress qualify merely as *opportunities* that can be given to people so they can exercise their autonomy for their own benefit. Some people have less knowledge, resources or time than others and most certainly there will be significant differences in the extent to which people leverage those opportunities.

In our view, there is a major principle missing: **autonomy**. This relates to individual liberty, freedom of choice, and self-determination.

Some recent publications have postulated fundamental ethical principles for AI, such as the *European Guidelines for Trustworthy AI* [13], and for technology more generally, such as the *Ethical by Design* paper by The Ethics Centre[14]. In both, a principle akin to autonomy is also present.

In general terms, we believe that, together, the concepts of **wellbeing, fairness** and **autonomy** capture fundamental ethical ideals broadly shared by liberal democracies. Both the **prevention of violations** and **remediation for violations** of these ideals should be sought. Coming up with quantifiable measures of these ideals and exposing to human oversight the controllable trade-offs between such measures are imperatives.

2. *Do the principles put forward in the discussion paper sufficiently reflect the values of the Australian public?*

The question rests on the assumption that we have already quantified Australian values, and that in itself illuminates the core issue of the need for *quantification* of Ethics.

To properly answer this question a significantly broader set of opinions, beyond those from readers of this discussion paper, should be heard. Also, we don't necessarily have all the same values and philosophies as some of the other countries from which the case studies in the Discussion Paper are drawn. Some of the countries discussed in the Discussion Paper would draw the balance differently, for example, on the importance of freedom of expression (related to autonomy).

One idea may be to discuss Australian values directly. For instance, the paper mentions the celebrated Australian motto “fair go”. How could we make the notion of a “fair go” more precise? Is “fair go”

[13] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[14] https://ethics.org.au/ethical-by-design/

aligned with equality of opportunity, equality of outcome, or perhaps other notions of fair treatment? Different notions of fairness abound and many are mutually incompatible.[15]

3. *As an organisation, if you designed or implemented an AI system based on these principles, would this meet the needs of your customers and/or suppliers? What other principles might be required to meet the needs of your customers and/or suppliers?*

No. They would still have to set the trade-offs and objectives for the AI systems, and this problem isn't mentioned in the document. Also since the principle of autonomy isn't mentioned, there is a risk of implementing systems that do not respect freedom of choice.

4. *Would the proposed tools enable you or your organisation to implement the core principles for ethical AI?*

Section 7.2 contains a number of issues:

- *'It has been written with the goal of creating a toolkit of practical and implementable methods (such as developing best practice guidelines or providing education and training) that can be used to support core ethical principles designed to assist both AI developers and Australia as a whole'*
    - We would like to focus on the risk management toolkit offered (7.2), **which has important issues**. Some comments:
        - If we want a system to be able to do anything about risks, it is imperative that we (1) *Quantify* risks, and the related uncertainties, and (2) Find out how to *explicitly trade them off*. This section proposes a risk assessment approach that does not attend to these requirements.
        - Table 3 proposes actions to mitigate risk *for a given level of risk*, but the methodology proposed to *assess* the level of risk in the first place (table 2) is a simple qualitative template as opposed to a comprehensive and detailed risk assessment tool.
        - **In the lack of a reliable and accurate risk assessment tool, it is <u>risky</u> to exclude mitigation actions from "low" or "medium" levels of risk, simply because such estimates of risk may be imprecise in the first place. A conservative approach here is crucial.**
        - The procedure of assessing risk though expected cost (multiplying probability by cost, which is what is done here) can be appropriate for assessing aggregate risk, but falls short for assessing the risk of harm on minorities. The expected risk can be low over the entire population, but very high for a small fraction of the population. This fundamental issue is not discussed.

[15] Inherent Trade-Offs in the Fair Determination of Risk Scores. J. Kleinberg, S. Mullainathan, M. Raghavan. https://arxiv.org/abs/1609.05807

- Technical comment: First table: risk is actually decomposable into ***probability*** and ***cost***. So the column headings should relate to costs, not risk, and row headings should refer to likelihood of cost being materialised, not likelihood of risk.

5. *What other tools or support mechanisms would you need to be able to implement principles for ethical AI?*

To implement ethical principles within an organisation, designers must have access to a comprehensive **design process guide** that sits between high level principles and a technical "best practice" handbook. Below we briefly outline a number of elements that we think should be included in such a design process:

1. **Identify what matters:**

   Senior decision-makers, designers, domain experts, legal experts and other relevant stakeholders should elicit all the considerations that matter from the point of view of evaluating the performance of the organisation deploying the AI system (including ethical considerations, but also others such as business considerations). This is the time to start considering what could "unintended consequences" look like.

2. **Measure what matters:**

   Ensure that all these performance considerations can be measured and quantified, obtaining *metrics* for everything that matters. Quantification should include the *uncertainty* associated with the quantities being measured (we can't always measure things with absolute precision).

3. **Build a decision-making system that influences what matters:**

   Build a *configurable* decision-making system designed to influence the performance metrics, including minimising the likelihood of unintended consequences.

4. **Estimate impact of different configurations of the system on what matters**:

   Use scientific approaches to estimate the causal effect that different configurations of the AI system have on the performance metrics chosen.

5. **Expose trade-offs to humans:**

   Expose through a human-machine interface how different *configurations* of the system are likely to *impact* the different performance metrics, as well as the explicit trade-offs between the

various different performance objectives and unintended consequences. *This is the key component that will enable human decision makers responsible for the deployment of the system to make an informed decision about if, when and how it should be deployed - subject to successful testing.*

6. **Test and Scrutinise:**

   Test the system to catch unexpected behaviour, design errors, and to allow senior decision makers to adapt how factors involved in trade-offs are weighted.

7. **Iterate:**

   Iterate the design process to re-examine the specification and measurement of goals, the identification and measurement of unintended consequences, the repertoire of configurations for the AI system, the causal estimates, the exploration of trade-offs, and the possibility of performance improvements through design changes or data collection.

The question of *when and at which scale to deploy* a system is a particularly difficult one. In many cases it is likely that several iterations of the above design process should be conducted in a simulated environment first, and only then gradually rolled-out to an increasing number of people throughout additional iterations.

The following are important measures to be considered during the different stages in the design process:

- *Action measurement.* The types of AI systems we consider all interact with the world by taking actions. Designers need to be intimately aware of the (intended and unintended) effects of these actions. Experiments may need to be run before an AI system is deployed to measure the effects of potential actions. While the system is running, the effects of actions should be closely scrutinised.
- *Objective and trade-off design.* By using an AI system we have to be particularly precise about its objectives and constraints. For instance, maximising overall accuracy almost always comes at the cost of higher error rates on minority groups, which may not be acceptable in certain circumstances. We must explicitly consider these trade-offs in the design of the system at the *outset*. This requires senior executives for contextual awareness, as well as the people who implement and engineer AI systems.
- *Continual iteration and improvement.* It is unreasonable to assume that an AI decision system will be free from harmful impact at the outset, even if it has been designed to the best of our ability. Hence, we need to slowly scale, iterate design decisions, and improve these systems between their first deployment to their full scale operation. Any legislation aiming to minimise risks of unethical deployment must also consider error tolerance as a potential strategy to incentivise improvements to occur in a transparent manner.

- *Privacy and fairness tradeoff.* To properly measure fairness and implement fair outcomes we need access to potentially sensitive attributes of individuals, including protected attributes such as race and gender. By identifying and acknowledging these attributes within an AI system, they can be used to proactively prevent unfair bias. This however, potentially comes at the cost of the individual's privacy. "Fairness through unawareness" (not using these sensitive attributes in prediction or monitoring) is a flawed approach[16], and so we need to be prepared to face potential trade-offs between the value of an individual's privacy and the cost of potentially unfair outcomes to that individual. This doesn't necessarily mean that private information can't be managed using best-practice data privacy techniques or that compliance with privacy laws may be threatened if we want to avoid discrimination. It just means that there is a potential tension between privacy and fairness that must be recognised and technical work needs to be done to precisely characterise any real trade-offs that may exist.

6. *Are there already best-practice models that you know of in related fields that can serve as a template to follow in the practical application of ethical AI?*

The approach of following a template would most likely be a mistake in this instance. This is a new field, the knowledge so far formed is fluid, and in a year from now we will probably have a much improved picture of what the right questions to ask are. Templates tend to promote compliance and demote deep thinking and learning, which are absolutely vital when the foundations of a new field are still being erected. This suggests something more dynamic than a template. Perhaps a "digital handbook" in the form of a frequently updated web page, like a wiki, would be a good format.

We draw attention to the point 5 of the Toolkit here: industry standards. Standards are necessary and can be very effective, but there are risks that need acknowledgement, such as promoting groupthink and divergence across sectors without reason (e.g. why should insurance marketing be different than FMCG marketing?) need to be recognised and dealt with. We also question whether certification is a good idea for such an indeterminate thing as ethical applications of AI. To what degree would certification cause people to disengage their minds or be unjustifiably confident when tackling new challenges? Also, AI is at its core software, which is a very general-purpose tool. Certification of software isn't something that has really worked well, so perhaps it is unlikely that this will work for AI systems.

7. *Are there additional ethical issues related to AI that have not been raised in the discussion paper? What are they and why are they important?*

[16] https://gradientinstitute.org/blog/2/

See our response to question (4), as these are missing from both the discussion of ethical AI and the toolkit. Furthermore, there is not much discussion on the legal treatment of both direct and indirect discrimination (e.g. the potential conflicts between them), which is pertinent to this issue.

Comments on the Discussion Paper

General comments

- Much of the paper is concerned with case studies, example applications in specific domains, and ideals rather than actionable principles. Instead this should be material used to motivate or support the proposed principles. Chapter 2-6 could be appendix material, or come after the core material so the presented framework could be described with reference to them.
- Core principles: see answer to question for consideration (1) above.
- Page 7: What is meant *precisely* by the term **black box**? This paragraph conflates a few potential meanings of black box. Is this referring to proprietary/secret implementations, or complex architectures (e.g. deep nets)?
- The discussion around discrimination and AI on page 18 needs clarification and expansion. The statement *"AI systems are vulnerable to discriminatory outcomes"* implies that these outcomes in the cases discussed are a consequence of the AI system. In reality, the implementation of AI often *exposes* the discriminatory outcomes that already existed within the decision making context.

Section 3

- While there is mention of biased datasets in this section, there is no mention of experimental design - which is a crucial consideration when gathering data.[17]

Section 4

- This section is too general, and is therefore difficult to translate into action. Some of the ideas here could be simplified into more fundamental considerations, and ordered more appropriately to mirror the design process of implementing an ethical AI system. Big blind spots in this section are causation and uncertainty - which are given no consideration.
- 4.1: This section is quite vague. It would have been a good opportunity to mention one of the most serious issues with deploying AI systems: positive feedback loops that reinforce the status quo and amplify inequalities. There is no mention of it. This could be re-framed as human oversight, continual monitoring and development based on the system's impacts. Also there should be a section *before* this section on estimating the potential causal impact of these systems on society before they are implemented.

[17] Chen, I., Johansson, F.D. and Sontag, D., 2018. Why Is My Classifier Discriminatory? In Advances in Neural Information Processing Systems (pp. 3539-3550). https://arxiv.org/abs/1805.12002

- 4.2: Again this section is imprecise (Black box issues and transparency). It conflates the issues of transparency, reliability/validity, and accountability. What action can one take as a result of this section? It is unclear.
- 4.3: Automation bias -- again this section needs to be made more actionable. Organisations need to decide what is and what is not appropriate to automate, and if they have decided to automate something, suitable validation, testing, monitoring and safeguards are required. Also, the example chosen in this instance (Enbridge pipeline leak) is clearly more nuanced than the conclusion drawn from it may suggest.

Section 5

- 5.2 opening. The following is incorrect. *"Indirect discrimination occurs when data variables that are highly correlated with discriminatory variables are included in a model"*. It is only true when these variables have been used to cause harm (indirectly) to a demographic. They could also (a) have no effect or (b) be used to compensate for disadvantage. In addition, indirect discrimination can occur when no individual variable in the model is highly correlated with a discriminatory variable - e.g. the variables in the model may predict the sensitive attribute in combination but not individually. A more suitable statement would be: indirect discrimination can occur when a sensitive attribute is correlated with the outcome of interest, and can also be predicted (better than at random) from variables included in the model.
- 5.2. *"This also prompts another ethical question for consideration: beyond racial discrimination, should location-based discrimination be permissible or is this still discrimination?"*. This sounds like an arbitrary suggestion. There are many variables that are sometimes closely correlated with a protected attribute. Adding all of them to the list of sensitive attributes and developing corresponding anti-discrimination legislation is clearly infeasible. If this was the standard, we would need discrimination acts preventing discrimination on the basis of people's video game preferences, eating habits, etc. Instead, we need to focus on clarifying how indirect discrimination should be interpreted in the context of algorithmic decision making, and how we manage the conflict between indirect and direct discrimination legislation. This is currently an open problem. For instance, it can be argued that the concept of *causality* seems to be key in reasoning about discrimination. If the location of a vehicle is a known causal factor on the risk of accidents, we may think it's permissible to price motor insurance policies based on location. However if a protected attribute (e.g. race) is correlated with location, indirect discrimination may arise as a result.

Section 6

- 6.4: why is only gender diversity the focus of this section? And why is the issue of pay brought into this? It is unclear what is the link between diversity and the pay gap this section is making? What about other disadvantaged groups?

Section 7

- Core principles - see response to questions for consideration (question 1).
- 7.1: “Putting principles into practice” - The goals of this section are unclear.
- 7.1.2 Describes software to calculate fairness measures. It is wrong to suggest these are “AIs” themselves, and also that the danger of using them is they may have the same flaws they are trying to assess. The paper describes them as a solution, but in fact they are only a component of a solution because they do not help determine the relative importance of violations to different notions of fairness.
- 7.1.5 Education, training and standards: In setting standards for data scientists, care would need to be taken to ensure that any compulsory or industry standard accreditation demonstrably improves the quality of data science and analysis in Australia, and is accessible and affordable to obtain. The (regulatory capture) risk that an accreditation process is co-opted to prevent people from entering the sector and limit competition needs to be recognised.
- 7.2: **This section has important issues**. See answer to question 5.

Conclusion

We welcome the Australian Government’s initiative to start a public debate on Ethical AI. There is urgency to develop an AI Ethics Framework for Australia. We believe the Discussion Paper is a good start for this conversation which will stimulate community engagement and gather valuable feedback on key issues in this space.

We at the Gradient Institute believe Australia needs *Ethical Principles* as well as an *Ethical Infrastructure* to realise them. The set of *Ethical Principles* needs to encapsulate the ideals of promoting wellbeing, fairness and autonomy. *Ethical Infrastructure* is multi-faceted, involving organisational, technical, legal and broader societal considerations. *Ethical Infrastructure* must (i) incentivise organisations to develop *ethical intent* while remaining competitive, and (ii) ensuring the organisation’s ethical intent translates into *ethical action*. The latter can be seen as the challenge of “unintended consequences”: ethical intent does not guarantee ethical action. The reality of AI systems adds colour to each of these challenges: both require significant *technical considerations*.

To drive and measure *ethical intent*, proper societal, institutional and legal accountability mechanisms need to be developed. It is crucial that such mechanisms, when developed, are properly *technically informed* by considerations of how AI systems actually operate. For instance, consider *legal* mechanisms to define and enforce accountability with a view to promote *ethical intent*. Because legislation is written in ambiguous natural language and AI systems operate with precise mathematical instructions, there is an increased propensity for the emergence of regulatory loopholes that can be

exploited by organisations to subvert the authority of the state. It is therefore critical that technical subject matter experts are involved in the development of policy and regulation for AI.

Technical considerations play a significant role in ensuring organisations accurately translate ethical intent into *ethical action*. First, ethical concepts must be made mathematically precise since actions of AI systems can only be driven by mathematical instructions. This requires technical mathematical knowledge in addition to ethical knowledge. Second, a scientific approach for ethical action is required because in order to design AI systems that operate ethically we need to understand the *cause and effect relationships* between design choices for AI systems and the ethical valence of their decisions. Finally, an engineering approach for ethical action is required to afford the *technical assurance* that the systems will be fair, secure, and respect privacy and autonomy while being performant.

Gradient Institute is committed to helping Australia make concrete progress towards making ethical AI a reality. We hope the release of this Discussion Paper and the ensuing public consultation will, in time, be seen as a pivotal moment in the history of the attempts to make the country of the 'fair go' even fairer, freer, and more prosperous.

Contact

info@gradientinstitute.org

Acknowledgements

We thank Linda Przhedetsky, Kimberlee Weatherall, Chris Dolman, Matthew Newman, Ross Buckley and Dimitri Semenovich who were consulted and provided valuable input on earlier drafts of this submission. Linda Przhedetsky provided detailed feedback on all aspects of the submission.

