Society on Social Implications of Technology (SSIT) Australia response to the Discussion Paper on Artificial Intelligence: Australia's Ethics Framework

May 2019

Authors (in alphabetical order):

Greg Adamson, University of Melbourne
Simone Alesich, Independent researcher
Morgan M. Broman, Queensland University of Technology & Griffith University
Michael Guihot, Queensland University of Technology
Samuli Haataja, Griffith University
Aurélie Jacquet, Lawyer
Michael Rigby, University of Melbourne
Marcus Wigan, IEEE, Australian Computer Society, University of Melbourne

Contents

1. Introduction

The following submission has been prepared by members of the IEEE Society on Social Implications of Technology (SSIT) in Australia. IEEE is the world's largest technology professional association, with 420,000 members in 161 countries. Founded in 1972, SSIT is the unit within IEEE which addresses the relationship between technology and society. SSIT Australia was formed in 2005 and has members in all states. SSIT Australia members have actively contributed to the IEEE's Ethically Aligned Design First Edition report ('IEEE EAD')[1], referred to in the Discussion Paper on Artificial Intelligence: Australia's Ethics Framework ('Discussion Paper').

SSIT Australia welcomes this opportunity to participate in the development of Australia's Ethics Framework. Overall, we see the Discussion Paper as a positive contribution and a useful starting point to stimulate further and wider discussions on the topic. In this submission we highlight a range of concerns with the Discussion Paper and also propose a number of improvements.

2. Ethics and regulation

2.1 Ethics not a substitute for regulation

The IEEE Ethically Aligned Design program has engaged more than 2,000 experts around the world over the past three years. While this is a significant effort, IEEE makes no claim that an ethical approach to the design and management of AI technology can replace the obligations of countries to develop laws to protect human rights and other existing principles as new technologies emerge: neither does the development of ethical principles relieve institutions of the obligation to enforce existing laws, such as laws on rights to privacy. For technologists, ethical behaviour such as found in our Codes of Ethics is a professional obligation on us to create and maintain technologies which provide benefit to the world, and thereby validate expectations and privileges (such as being considered experts) which societies bestow on our profession. We hope that ethical principles such as product safety will become incorporated into law over time, but ethical behaviour remains a separate domain to regulatory practice.

2.2 A need to expand the coverage of professional ethics

Additionally, it is clear that there is a real need to expand on the coverage of **professional** ethics for all computing fields as well as targeting AI and ML specifically. It is equally clear that governance of the intersection of IT, computing and society is underdeveloped and that this is a major omission in the Discussion Paper.

The ability of IT professionals of all kinds to create computing capacities that can be misapplied and abused is growing swiftly, and the current formal ethical frameworks, designed as they are primarily for reputational and professional protection, are simply no longer sufficient. Risk assessment is no longer from the viewpoint of organisations, but is moving into the hands of the professionals themselves. Examples such as former CIA employee Edward Snowden are moving quickly from

[1] Available here: https://ethicsinaction.ieee.org/

outliers to essential components of the entire social system. It follows that for any of the positive outcomes from growing AI and ML applications to occur, we must encourage and protect whistle-blowers at the origin of the algorithms that they create-and certainly at the application stage of what they develop.[2]

While the Discussion Paper covers some of the issues involved in training data, it does not include sufficient coverage of the various forms of information ownership and responsibility that the law already requires - and certainly not far enough to encompass the General Data Protection Regulation ('GDPR').

3. Concerns surrounding Core Principles

3.1 Distinction between law and ethics

We believe the main shortcoming of the Core Principles (and the Discussion Paper) is that many of the ethical principles simply involve re-stating the law (opposed to focusing on the ethical dimension which goes beyond simply compliance with the law). Ethical principles should not simply require a person/organisation to follow the law.

3.2 Human values

On page 14 of the Discussion Paper, it provides that technology should be aligned with 'human values'. However, human values are only obliquely referenced in the Core Principles. These values must be set out and defined more explicitly.

Also, the first principle ('Generate net benefits') does not ensure that the technology is human-centred, trustworthy or benefits society. Most ethical frameworks, such as the EU Ethics Guidelines for Trustworthy AI, and the Ethics Advisory Group to the European Data Protection Supervisor (EDPS) recognised the necessity to make the technology human centric. The definition of 'benefits' is also flawed as the implication is that such benefits are solely fiscal.

Furthermore, to ensure that technology is human centric, most ethics initiatives consider how to promote the wellbeing of individuals and society. For example, wellbeing is a key principle of the IEEE EAD, and pursuant to this principle "A/IS creators are required to adopt increased human well-being as a primary success criterion for development".

3.3 Australian values, fairness, and privacy

The Discussion Paper also does not adequately address Australian values. These are limited to the notion of 'a fair go' which is only vaguely linked to fairness. The notion of fairness in the Discussion Paper is also quite vaguely defined.

[2] See, for example, Wigan, M.R.(2015) BigData - Can Virtue Ethics Play a Role? at https://works.bepress.com/mwigan/29/

By referring to the principles such as fairness, do no harm, and privacy there is a possibility to go beyond the current limitation of the law and develop an ethical perspective. For example, privacy rights in Australia are much lower than in Europe so the Discussion Paper offers an opportunity to develop privacy in this context and go beyond the current limitation of Australian privacy laws.

The data on which ML is based is not considered personal data in Australia, but the expansion of the new underpinning global standard of the EU GDPR which affects any party dealing with the EU- suggests that a recommendation to incorporate the GDPR into Australian practice would be timely and appropriate, as it is not solely a technical but also an ethical development.

3.4 Environment

While the Discussion Paper mentions the potential utility of AI in addressing environmental issues,[3] the environment's well-being could be given more primacy in the core principles themselves. For example, the EU Ethics Guidelines for Trustworthy AI specifically include 'Societal and environmental well-being' as a principle requiring it is ensured that the broader impact of AI systems on the natural environment and other living beings is taken into consideration.

Once again the fiscal valuation of 'wellbeing' is taken as read – and this can no longer be left if it is to balance ethical considerations and costs in the utilitarian ethical mould implied by the Discussion Paper.

4. Other concerns

4.1 Understanding of technology

Two principles included in the IEEE EAD which reflect the concerns of technologists are:

- 'Effectiveness … creators and operators shall provide evidence of the effectiveness and fitness for purpose' and;
- 'Competence … creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.' These principles underlie the concern that technologists have for the effectiveness of technology.

In contrast, the Australian Ethics Framework is far more confident in the functioning of technologies. This comes up particularly in discussion of "'black boxes' in which the inner workings of an AI are shrouded in secrecy".[4] This suggests that the problem of a black box is deliberate secrecy such as resulting from intellectual property. A black box in the AI context emerged (in the late 1940s) from an understanding of the unknowability of processes, particularly when analogue or feedback functionality is involved.

[3] See, for example, pages 8 and 49 of the Discussion Paper.
[4] Discussion Paper, p 7.

It would be worthwhile for the document to address the unknowability of some current technology processes, and approach ethics and regulation from the assumption that in all but trivial cases humans (and machines) can only give likely reasons for the behaviour of AI (eg listing factors considered or steps taken), not predictive mechanical descriptions of how an AI will work. Asking an AI to explain why it undertook a particular function also risks encouraging the AI to game the problem (or "optimise" it, to use engineering language). This comes up again on page 11 of the Discussion Paper in the discussion of input data and outcomes, which is a good summary of the black box concept.

The question of whether processes can be understood or not is ambiguously addressed throughout the document.

- "very few people would have the expertise to understand" (page 11, para 4)
- "difficult for most people to understand." (page 21, para 7)
- "the explanations may still be far too complex". (page 34, para 5)

These all still suggest that at least some humans have the capacity to understand how machines are doing what they are doing, which is a poorly supported assertion.

For example, consider use cases where black boxes are deployed for the provision of essential services in a community. Degrees of transparency are required at different levels to ensure that a system's levels of control, which may comprise multiple black boxes, are fully known and can be suitably manageable by specific roles. In the design of such a system biases may exist in the underlying algorithms used for services matching and the datasets used for training the ML/AI during the initial design phase. In the case of mobility as a service (MaaS) for the matching of vehicles for on-demand transportation, biases may be deliberate (eg nudging users to a certain options over others) and or unintended with effects that are far reaching. Further, without knowing the relationships between algorithms and data, social implications can be expanded to economic and environmental dimensions.

Once again, introducing the personal data concepts in the GPDR would make it clear that there is now an emergent property right to data **about** people, an issue that will be strenuously denied by many, but is now inescapable if AI and ML is not to fail the growing severity of tests of social licence. Treating AI and ML separately from the issues of big data generally avoids the emergent major problem of the Private Data Commons vanishing entirely, as is clearly the likely outcome of Google and Facebook's operations. It is difficult to address AI Ethics[5] without considering these unexpected and little discussed aspects of the data that drives ML.

4.2 Technology having agency

At various points in the Discussion Paper the language used lends itself to suggest that technology has agency:

[5] See Wigan, M.R ; Roger Clarke, R.A 'Big Data's Big Unintended Consequences' (2013) IEEE Computer 46(6)46-53 at https://ieeexplore.ieee.org/document/6527249/

- "Consider the system, and design a suitable framework for keeping it transparent *and accountable*." (Emphasis added) (page 1, para 4)
- "what do we need to know about this algorithm and keep it accountable…" (page 34, para 6)

Elsewhere in the Discussion Paper it is, however, noted that:

- "an AI system has no moral authority, it cannot be held accountable in a judicial sense for its decisions and judgements. As such, a human must be accountable for the consequences of decisions made by the AI." (page 36, para 3).

This is correct and it is important to emphasise that AI systems do not have moral agency and avoid using language suggesting otherwise.

Additionally, AI systems do not have legal personality. While the question and extent of an entity's legal personality is a matter to be determined for each nation-state within their domestic legal systems, there is currently no practical need for AI systems to have legal personality. However (and regardless of the exact legal characterisation of AI systems), as noted in the IEEE EAD it is pivotal that there are systems for registration and record keeping to ensure it is always possible to determine legal responsibility for a particular AI system.[6]

The discussion around legal personality, as presented by the 2016 EU study around 'European Civil Law Rules in Robotics'[7] and used as a basis for the 2019 EU 'Ethics Guidelines for Trustworthy AI', focuses strongly on an autonomous entity combining a robotic (physical) body with an integrated artificial intelligence (mind).[8] The 2016 study concludes that assigning a legal personality to a robot is not an answer from the legal perspective as it does not necessarily align with *current* civil liability law.[9] The 2019 report does not utilize the term "legal personality" at all, but neither does it use the term "*moral authority*". However, there is nothing presented in any document that proves that the question about "legal personality" as an issue will not reoccur as the autonomy of the AI expands. In the Discussion Paper the responsibility for an AI's (automated) decision (page 36, para 4) is raised utilizing four (4) questions. These questions are closely related to liability and therefore to attached legal rights and obligations. The key points presented in the Discussion Paper (page 37) ends with a good question – "Ask: Who is responsible for the decisions made by the system?"

The EU's approach to this is to establish rules for "*trustworthy AI*" containing steps to ensure, as far as possible, the trustworthiness of all actors and processes involved as "*…part of the system's socio-technical context throughout its entire life cycle.*"[10] The Discussion Paper mentions the three components identified and presented by the EU paper (page 20, para 9) – *lawful, ethical and robust* - but goes no further. While the EU draft proposal may not be the final answer, the three components presented do provide a good starting point that Australia could utilize for its own development of an ethical-legal answer to their own question.

---

[6] IEEE EAD, p 30.
[7] Nathalie Nevejans, 'European Civil Law Rules in Robotics' (2016) available at http://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2016)571379
[8] Nevejans (2016) pp 8-12.
[9] Nevejans (2016) p 14.
[10] High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (2019) available at https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai pp 5-7.

4.3 Assumption of technology beneficence

The Discussion Paper provides that "… where the slow pace of regulatory adaptation has hindered the development of potentially life-saving AI technologies." (page 15, para 5). If the claim is that privacy can be traded off for better health outcomes, this should be dealt with in all its complexity, not by assuming that privacy is disposable.

Additionally, the recommendation in the Discussion Paper for "helping the workers transition smoothly and proactively into new jobs and new careers." (page 54, section 6.3) is based on the assumption that as old jobs disappear, new jobs will be created, which is under debate. For example, the replacement of horse and cart with motor vehicles did not eliminate drivers, while autonomous vehicles do. Rather than assuming that things will be okay, it would be useful to call for the creation of measures of technology impact on jobs, as proposed in the AI Now 2016 report.

4.4 Technology and marginalisation

A key issue regarding ethics and AI is the impact of AI on vulnerable or marginalised groups. Where a group or groups are already marginalised, an increase in their marginal status can be considered a key ethical issue, one that is relevant to the core principles for AI regarding net-benefit and fairness.

Marginalised groups include:

- Women
- People from sex and gender diverse groups (including LGBTIQ)
- Low income elderly people
- People with disability (including physical, cognitive and mental health)
- Indigenous people
- People from culturally diverse backgrounds
- People experiencing intergenerational poverty

Marginalised groups usually do worse in technological advancement. This is because:

1. Most developers are not from marginalised groups and are not aware of their needs and preferences.
2. Many technologies are deployed outside the sociocultural context in which they are designed and developed
3. Technology is often developed by private companies motivated by profit, not social benefit
4. Technology can replicate social bias. This is particularly the case with algorithms and machine learning, where past data is used to 'teach' the AI. Data, and the organisation of data (including metadata) can contain social bias. This is demonstrated by automated decisions in recruitment and sentencing that adversely impacts marginalised groups.
5. Technology development is directed at replacing roles or interacting with people who are less valued by society. For example, developing robots to care for the elderly or assist autistic children reflects social bias regarding the undervalued role of caregivers, and the lack of priority placed on the elderly and children with disability. These marginalised groups become 'problems' to be 'solved' through AI.

6. Technology is seen to be neutral or unbiased by those interacting with it, making it more able to replicate and reinforce bias without human awareness or questioning.

An exception that is often cited is the mobile phone, which has been used innovatively by groups that it was not necessarily designed for. For example, it has increased connectivity and access to services by people living in the global south, and by indigenous people. However, this technology was not designed for this purpose, therefore the benefits for non-target groups are incidental. Another example that is cited is the benefit of smart phones for people with hearing impairment, allowing them to interact through visual means. Again, this is an incidental benefit. While these examples may be championed as ways in which technology benefits marginal groups, this is a misleading argument. One benefit for a group does not lead to inclusion.

Technology with a specific social benefit has much greater value for marginalised groups than mainstream technology. An example of a technology with social benefit is the website created by Infoxchange to connect people in insecure housing with key social services. A target group for this organisation is women experiencing family violence. Anecdotal evidence suggests that this website, codesigned with women who have experienced family violence, could reduce the time that women spend in violent situations by several years. This is an example of substantial social benefit for a marginalised group.

Like all technology, AI is biased. For example, science fiction, twentieth century political systems and major events have had substantial impacts on technological development in the United States, along with American values of individualism and consumerism. When these products are exported to other countries, the values and biases are exported along with the technology.

Fairness means not just combating discrimination, but proactively working to identify marginalised groups and find ways to create inclusion. This leads to a more level playing field, noting that groups start from different points depending on their backgrounds. Methods can include consultation and codesign, as well as actively countering bias.

Ways to address AI and ethics for marginalised groups:

1. Assume at the outset that all technology is biased. Work to identify what the bias is and how to counter it (either through adjusting the technology, or through regulation) (part of risk assessments).
2. Encourage and support companies and social enterprise that develop technology with social benefits. While these will still have bias, they are more likely to benefit marginalised groups.
3. Actively seek to consult and design with marginalised groups when developing new technology and when designing laws to regulate it (collaboration).
4. Question the development of technology where there is no clear social benefit (this could also be part of risk assessments).
5. Independently assess all new technologies for their social benefit / harm and bias before they are released to market (impact assessments). Seek to include marginalised groups in impact assessments.

6. Aim for universal design. Rather than seeing marginalised groups as 'problems' to be solved through tailored solutions, aim to create universally applicable technologies that address the needs and interests of all rather than a narrow few.

Areas of addictive behaviour that disproportionately affect marginalised groups need to be considered for additional regulation, for example, the use of AI in the gambling industry. Governments must consider whether it is ethical for gambling companies to use AI to predict the behaviour of customers through physical movement / surveillance, machine interaction or brand loyalty, and the impact on principles such as do no harm.

4.5 Need for consensus

Australia is also among 42 countries that have recently adhered to a new values-based Recommendation on Artificial Intelligence (Recommendation) put forward by the Organisation for Economic Co-operation and Development (OECD).[11]

The Recommendation promotes five values-based principles for the responsible development of trustworthy AI (Principles), and five complementary strategies for developing national policy and international cooperation (Strategies). Given the Recommendation comes from the OECD, it treads the line between promoting economic improvement and innovation and fostering fundamental values and trust in the development of AI. The five Principles encourage (1) inclusive growth, sustainable development and well-being, (2) human-centred values and fairness, (3) transparency and explainability, (4) robustness, security and safety, and (5) accountability. Australia's Ethics Framework should consider these principles.

The OECD Recommendation is the latest of a number of projects and guidelines from governments and other bodies around the world seeking to instil an ethical approach to developing AI. These include the Institute of Electrical and Electronics Engineers (IEEE) (discussed above), the French Data Protection Authority, the Hong Kong Office of the Privacy Commissioner, and the European Commission.

Strategy five in the OECD Recommendation seeks to encourage 'international co-operation for trustworthy AI'. While the Discussion Paper seeks to foster Australia's Ethics Framework, it should also work towards achieving some form of consensus among these various proposals. Developments in technology, particularly in AI, have ensured that we are no longer isolated from the world. Indeed, the bulk of AI development occurs in other parts of the world. Only by finding common ground might we one day be able to develop a more inclusive global ethical approach to AI development.

[11] See http://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm