Electronic Frontiers
AUSTRALIA

W www.efa.org.au
E email@efa.org.au
@efa_oz

31 May 2019

Artificial Intelligence
Strategic Policy Division
Department of Industry, Innovation and Science
GPO Box 2013
CANBERRA ACT 2601

By eLodgement and Email: artificial.intelligence@industry.gov.au

Dear Minister,

RE: Artificial Intelligence: Australia's Ethics Framework

Electronic Frontiers Australia (EFA) appreciates the opportunity to provide this submission in relation to the AI Ethical Framework Consultation ("**the Discussion Paper**"). EFA's submission is contained in the following pages.

Artificial Intelligence ("**AI**") creates significant opportunities and issues for the Australian and global community. EFA commends the existence of the Discussion Paper; however, significant ongoing consultation must occur with the recommendations contained within submissions being taken seriously. A profound junction of human rights and technology exist within the ambit of Artificial Intelligence and these crossroads must be carefully and meaningfully examined.

About EFA

Established in January 1994, EFA is a national, membership-based non-profit organisation representing Internet users concerned with digital freedoms and rights. EFA is independent of government and commerce, and is funded by membership subscriptions and donations from individuals and organisations with an altruistic interest in promoting civil liberties in the digital context. EFA members and supporters come from all parts of Australia and from diverse backgrounds.

Our major objectives are to protect and promote the civil liberties of users of digital communications systems (such as the Internet) and of those affected by their use and to educate the community at large about the social, political and civil liberties issues involved in the use of digital communications systems.

Yours sincerely,

[redacted]

Angus Murray
Chair of the Policy Committee
Electronic Frontiers Australia

Introduction

At the outset, EFA respectfully agrees with the analogy of an ethical framework for a party proposed by Anna Johnston of Salinger Privacy[1]. The AI Framework must be understood as the beginning to a broader, more detailed and informed series of consultations regarding the totality of the impact of AI.

In this context, EFA considers that the most important aspect of an AI Framework begins with the introduction of an enforceable Federal human rights legislative framework. This has been a consistent recommendation across numerous recent submissions (including the *Data Sharing and Release* consultation, the Human Rights Commissioner's Human Rights & Technology Issues Paper[2] and the subsequent White Paper on Artificial Intelligence[3]).

The introduction of a source for a comprehensive human rights framework would provide a safeguard for Australians, create market certainty regarding the operation of new and emerging technology and bring Australia into line with other western democratic societies. Once this legislative step has occurred, EFA considers that an AI Framework ought to be underpinned by the following non-exhaustive ethical principles:

- Privacy by design;
- Transparency;
- Trustworthiness;
- An informed society is fundamental;
- Decision making metrics need to be human reviewable; and
- Sustainability.

It is important that these ethical principles are non-exhaustive as the scope and application of what AI *could* do in the future is not reasonably predictable. In our opinion, in addition to an ethical and legal framework within which AI may operate, it is also important to ensure that AI remains trustworthy. In this regard, EFA adopts and integrates the European Commission's Ethical Guidelines for Trustworthy AI released in December 2018[4] namely being as follows:

- Human agency and oversight;
- Technical robustness and safety;
- Privacy and data governance;
- Transparency;
- Diversity, non-discrimination and fairness;
- Societal and environmental well-being; and
- Accountability[5].

[1] See: https://www.salingerprivacy.com.au/2019/04/27/ai-ethics/.

[2] https://tech.humanrights.gov.au/sites/default/files/inline-files/29%20-%20Australian%20Privacy%20Foundation_2.pdf.

[3] https://tech.humanrights.gov.au/sites/default/files/inline-files/26%20-%20Electronic%20Frontiers%20Australia.pdf.

[4] See: https://ec.europa.eu/futurium/en/ai-alliance-consultation.

[5] See also: European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)): "*the guiding ethical framework should be based on the principles of beneficence, non-maleficence, autonomy and justice, on the principles and values enshrined in Article 2 of the Treaty on European Union and in the Charter of Fundamental Rights, such as human dignity, equality, justice and equity, non-discrimination, informed consent, private and family life and data protection, as well as on other underlying principles and values of the Union law, such as non-stigmatisation, transparency, autonomy, individual responsibility and social responsibility, and on existing ethical practices and codes*".

Although the ethical principles and trustworthiness principles seemingly overlap, EFA considers that a multi-layered ethical framework is more likely to capture emerging applications of AI than an exhaustive proclamation of standardised ethical principles. Furthermore, each principle contained within any AI framework to be proposed or developed must underpin the introduction of enforceable human rights legislation at the federal level.

On the foundation of this introduction, EFA has responded to each of the questions posed in the Discussion Paper in the following pages before concluding with general comments regarding the current state of the AI Ethical Framework Discussion Paper.

EFA thanks its Policy Committee for their work on this submission - https://www.efa.org.au/our-work/policy-team/.

Responses to Questions

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

In the interest of providing a comprehensive response to this question, we have addressed each of the principles proposed in the Discussion Paper separately before addressing what is missing.

1. *Generates net-benefits: the AI system must generate benefits for people that are greater than the costs.*

   This is an excellent example of why Australia must introduce an enforceable human rights legislative framework. Whilst this principle is seemingly well-intended, it is difficult to accept a "greater good" basis for AI. This basis is misguided for the following reasons:

   a. It lacks an understanding that AI is a rapidly evolving technology and the consequence of actions taken today will impact future generations in a variety of ways that are not yet foreseeable. The "greater good" today could, in fact, be to the greatest detriment tomorrow.

   b. It fails to clearly establish "benefits for people". It is beneficial, for example, to reduce the cost associated with radiology, however, that same or similar process may generate benefits for insurance companies.

   c. The cost(s) may not be immediate or tangible. Whether this is in the context of national security or anti-spyware. The cost may become a sophisticated surveillance network that cannot be removed from digital infrastructure in the future.

2. *Do no harm: civilian AI systems must not be designed to harm or deceive people and should be implemented in ways that minimise any negative outcomes.*

   EFA agrees with this principle; however, it is relevant to note that the Discussion Paper expressly excludes military operations. The concept of deceit also requires further analysis as AI has the potential to alter human behaviour in non-traditional means.

Furthermore, the issue of winners and losers deserves some attention in the context of 'net benefit' and 'do no harm'. It is part of the nature of AI systems that they are developed to produce acceptably good results most of the time and that it is understood that not all decisions will be accurate, correct or appropriate. The AI has no understanding of 'fairness'. There are predictable groups of humans who are likely to be more difficult for autonomous systems to serve: those with language difficulties, access difficulties, the young, the elderly and so on. Social justice demands that any system employing AI should address the equity of access of all Australians.

3. *Regulatory and legal compliance: the AI system must comply with all relevant international, Australian local, state/territory and federal government obligations, regulations and laws.*

   EFA agrees with this principle; however, reiterates that an enforceable human rights framework must be introduced for this principle to be meaningful.

4. *Privacy protection: any system, including AI systems, must ensure people's private data is protected and kept confidential plus prevent data breaches which could cause reputational, psychological, financial, professional or other types of harm.*

   EFA agrees with this principle; however, as framed, it seems to oversimplify an extremely complex interplay of issue social, legal and architectural issues.

5. *Fairness: the development or use of the AI system must not result in unfair discrimination against individuals, communities or groups. This requires particular attention to ensure the "training data" is free from bias or characteristics which may cause the algorithm to behave unfairly.*

   This principle engages issues of privacy and complex data integrity concepts. EFA agrees that the development or use of AI systems must not result in unfair discrimination against individuals, communities or groups. However, this principle is unsustainable without an underlying principle regarding the responsible person(s), manner and means by which training data is freed from bias.

6. *Transparency & Explainability: people must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions.*

   EFA agrees with this principle and further recommends that the decision making processes be human-reviewable or at least capable of extrapolation to human reviewable logic.

7. *Contestability: when an algorithm impacts a person there must be an efficient process to allow that person to challenge the use of the algorithm.*

   EFA repeats the need for a human rights legislative framework that contains enforcement provisions. This principle cannot operate in isolation to a principle-based legal solution.

8. *Accountability: people and organisations responsible for the creation and implementation of AI algorithms should be identifiable and accountable for the impacts of that algorithm, even if the impacts are unintended.*

   EFA agrees with this principle; however, it is useful to note the work being undertaken in Europe in relation to Civil Rights for Robotics. In 2017, it was recommended as follows:

   "*51. Asks the Commission to submit, on the basis of Article 114 TFEU, a proposal for a legislative instrument on legal questions related to the development and use of robotics and AI foreseeable in the next 10 to 15 years, combined with non-legislative instruments such as guidelines and codes of conduct as referred to in recommendations set out in the Annex;*

   52. Considers that, whatever legal solution it applies to the civil liability for damage caused by robots in cases other than those of damage to property, the future legislative instrument should in no way restrict the type or the extent of the damages which may be recovered, nor should it limit the forms of compensation which may be offered to the aggrieved party, on the sole grounds that damage is caused by a non-human agent;

   53. Considers that the future legislative instrument should be based on an in-depth evaluation by the Commission determining whether the strict liability or the risk management approach should be applied;

   54. Notes at the same time that strict liability requires only proof that damage has occurred and the establishment of a causal link between the harmful functioning of the robot and the damage suffered by the injured party;

   55. Notes that the risk management approach does not focus on the person "who acted negligently" as individually liable but on the person who is able, under certain circumstances, to minimise risks and deal with negative impacts;

   56. Considers that, in principle, once the parties bearing the ultimate responsibility have been identified, their liability should be proportional to the actual level of instructions given to the robot and of its degree of autonomy, so that the greater a robot's learning capability or autonomy, and the longer a robot's training, the greater the responsibility of its trainer should be; notes, in particular, that skills resulting from "training" given to a robot should be not confused with skills depending strictly on its self-learning abilities when seeking to identify the person to whom the robot's harmful behaviour is actually attributable; notes that at least at the present stage the responsibility must lie with a human and not a robot;

   57. Points out that a possible solution to the complexity of allocating responsibility for damage caused by increasingly autonomous robots could be an obligatory insurance scheme, as is already the case, for instance, with cars; notes, nevertheless, that unlike the insurance system for road traffic, where the insurance covers human acts and failures, an insurance system for robotics should take into account all potential responsibilities in the chain;

> *58. Considers that, as is the case with the insurance of motor vehicles, such an insurance system could be supplemented by a fund in order to ensure that reparation can be made for damage in cases where no insurance cover exists; calls on the insurance industry to develop new products and types of offers that are in line with the advances in robotics".*

In our view, the principles form an overly simplified summary of what AI might engage and not what or where issues could occur. In this regard, we consider that the following non-exhaustive additional principles must also be involved:

1. Sustainability

   There is an environmental aspect of all digital technology. A principle of Ethical AI ought to understand the underlying physical cost associated with complex computational processes.

2. Author knowledge sharing

   AI (and works created by AI) is arguably capable of intellectual property protection as well as protection via trade secrets. A principle that addresses the "property" of AI is required to ensure a comprehensive understanding of the obligations and communal benefit of the creators of AI and the creations of AI in turn.

3. Safety

   AI ought to be subject to testing and peer review before implementation.

4. Retrievability

   AI ought to be developed and/or implemented in a manner that allows removal at any point in the future so as to ensure that society does not become systemically reliant on AI.

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

Firstly, this question must be understood in the context of the extremely rapid progression of technology, significant complexities of AI and the inherent difficulties that arise from asking whether the abovementioned principles (noting the issues raised with same) "sufficiently reflect the values of the Australian public". This question is apt to mislead.

For example, to understand the principles requires a technological and mathematical understanding of AI. It is incongruous to ask a member of the public whether they agree that the above principles are sufficient insofar as that person considers they apply to a maximum entropy reinforcement learning framework without a campaign on increasing informed decision making, as we recommend.

Nonetheless, the values of the Australian public are often characterised by the idea of the 'Fair Go'. While most Australians are not directly engaged with discussions of the ethics of AI, they may never-the-less recognise that the 'Fair Go' concept would encompass straight forward dealing between parties (transparency and fairness), honesty (transparency and legal compliance) and respect for the boundaries and responsibilities of a relationship between provider and customer (privacy and accountability). The Australian 'Fair Go' would also demand that all decisions, especially AI informed decisions, are reviewable and corrected when in error.

Specifically Australian concepts of the 'Fair Go' would abhor:

- Long, verbose and legalistic terms of use being used to defend abhorrent business practices;
- The involvement of third parties in transactions to the detriment of the customer;
- The use of personal information contributed for one purpose for some unrelated alternate purpose; and
- Lack of accountability and responsibility for actions affecting others including those as a result of AI aided decision making.

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?

Different sectors will need to comply with different regulatory frameworks and legislation. The principles as outlined in the Discussion Paper are not sufficient to ensure that those requirements are met. In our view, it is overly simplified to pose the question in its current form. It is important to understand that AI has broad and evolving applications. It can be used to play chess[6], to identify wildlife[7], within the legal profession[8] or in malware[9].

The health Sector may be a useful example - the Royal College of Radiologists recently held a consultation for its members and published 'Ethical Principles for AI in Medicine' which proposed 8 principles[10] including:

1. Safety;
2. Avoidance of Bias;
3. Transparency and Explainability;
4. Privacy and protection of Data;
5. Decision-making on Diagnosis and Treatment
6. Liability for decisions made;
7. Application of Human values; and
8. Governance.

The Nuffield Council on Bioethics has also published a Briefing Note[11] on this issue and specifically dealt with the ethical issues of AI in healthcare and raised the important issue of the effects AI may have on people's sense of dignity and isolation in care situations and the potential for AI to be used for malicious purposes.

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

[6] https://deepmind.com/research/alphago/.

[7] https://www.nationalgeographic.com.au/animals/how-artificial-intelligence-is-changing-wildlife-research.aspx.

[8] Murray, Angus and Owen, Daniel. Legal forecast: Building judge Hercules: Blending the science of analytics with the art of law [online]. Proctor, The, Vol. 37, No. 8, Sep 2017: 34-35. Availability: <https://search.informit.com.au/documentSummary;dn=083412570358424;res=IELHSS> ISSN: 1321-8794.

[9] https://www.2-spyware.com/security-experts-create-deeplocker-the-ai-based-malware.

[10] See: https://www.ranzcr.com/college/document-library/ranzcr-ethical-principles-for-ai-in-medicine-consultation.

[11] http://nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research.pdf.

We respectfully consider this question to be premature. The principles ought to be properly deliberated and resolved before engaging in discussion regarding AI tools.

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

We respectfully consider this question to be premature. The principles ought to be properly deliberated and resolved before engaging in discussion regarding AI tools.

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 health Sector may be a useful example - the Royal College of Radiologists recently held a consultation for its members and published 'Ethical Principles for AI in Medicine' which proposed 8 principles[12] including:

1. Safety;
2. Avoidance of Bias;
3. Transparency and Explainability;
4. Privacy and protection of Data;
5. Decision-making on Diagnosis and Treatment
6. Liability for decisions made;
7. Application of Human values; and
8. Governance.

We additionally repeat the work being undertaken in Europe in relation to Civil Rights for Robotics and the EU Ethical Guidelines for Ethical AI are a useful model to regard.

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?

We repeat the contents of our submissions to the Human Rights and Technology Issue Paper[13] and AI Whitepaper[14].

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

In our submission, the principles for ethical AI must be established before a discussion regarding the proposed tools can meaningfully occur.

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

In our submission, the principles for ethical AI must be established before a discussion regarding the proposed tools can meaningfully occur.

[12] See: https://www.ranzcr.com/college/document-library/ranzcr-ethical-principles-for-ai-in-medicine-consultation.

[13] https://tech.humanrights.gov.au/sites/default/files/inline-files/29%20-%20Australian%20Privacy%20Foundation_2.pdf.

[14] https://tech.humanrights.gov.au/sites/default/files/inline-files/26%20-%20Electronic%20Frontiers%20Australia.pdf.

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?

In our submission, the principles for ethical AI must be established before a discussion regarding the proposed tools can meaningfully occur.

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?

In our view, the principles form an overly simplified summary of what AI might engage and not what or where issues could occur. In this regard, we consider that the following non-exhaustive additional principles must also be involved:

1. Sustainability

   There is an environmental aspect of all digital technology. A principle of Ethical AI ought to understand the underlying physical cost associated with complex computational processes.

2. Author knowledge sharing

   AI (and works created by AI) is arguably capable of intellectual property protection as well as protection via trade secrets. A principle that addresses the "property" of AI is required to ensure a comprehensive understanding of the obligations and communal benefit of the creators of AI and the creations of AI in turn.

3. Safety

   AI ought to be subject to testing and peer review before implementation.

4. Retrievability

   AI ought to be developed and/or implemented in a manner that allows removal at any point in the future so as to ensure that society does not become systemically reliant on AI.

Furthermore, one of the more difficult attributes of large data sets being used in AI is that it is difficult to trace the lineage of the data from source to use and so this may not be possible in some cases. This lineage issue also affects other areas such as privacy where PII is present in decision making but it is not possible to identify where the PII came from.

In the circumstance of discrimination based on gender/physical features, this is already happening with services like Amazon's Mechanical Turk where humans are involved in determining the training datasets using subjective measures about images of other humans[15].

[15] See: https://pursuit.unimelb.edu.au/articles/holding-a-black-mirror-up-to-artificial-intelligence; https://pursuit.unimelb.edu.au/articles/why-does-artificial-intelligence-discriminate.

Further Comments

We thank you for the Discussion Paper and the opportunity to make comments regarding same. As expressed in this submission, we consider that there is further work to be done to ensure that this process is comprehensive, informed and based on informed deliberation of the underlying principles. We appreciate that there is likely good reason for the non-inclusion of military applications of AI in the Discussion Paper; however, this should not be treated as a separate issue and ethical principles must be universally sound.

It is our view that a large portion of the underlying premise for the Discussion Paper could readily be resolved by the introduction of an enforceable human rights framework at the Federal level and we reiterate the importance of this step being promoted via the outcome to this consultation process.

Separately, we encourage an ongoing consultation to define and clearly frame the relevant principles for AI ethics to ensure that they are comprehensive, well considered and sufficiently layered to be adaptable to emerging applications of AI.

We trust that this submission will be carefully considered and that our responses to the Discussion Paper are useful.

Please do not hesitate to contact Angus Murray, Chair of the EFA Policy Committee should you require any further comment in relation to these submissions or the ongoing consolation process generally.