Beyond Asimov's Three Laws: A New Ethical Framework for AI Developers

Paul Kallenbach, Vanessa Mellis, Siegfried Clarke at MinterEllison

In 1942, science fiction author Isaac Asimov introduced the world to *The Three Laws of Robotics* in his short story *Runaround*. The simple rules to not injure, to obey and to protect have become a touchstone for science fiction authors ever since. However, regrettably for real world developers incorporating artificial intelligence and machine learning (**AI**) into their systems, these modest commands offer very little guidance on the real world ethical considerations which they face.

Given the particular breadth of applications for AI, it is also difficult to point to any single law or set of laws that are relevant for Australian AI developers. In any given project, a myriad of laws may be relevant. Depending on the project, this might include a combination of government specific legislation (e.g. the *Social Security (Administration) Act 1999*), human rights obligations, anti-discrimination legislation, data sharing legislation and the *Privacy Act 1988* (Cth), none of which has been developed with AI technologies in mind.

Acknowledging these challenges, the Australian government has recently joined jurisdictions around the world and started work on a framework to assist decision makers and developers to create and deploy AI driven technologies responsibly. While this article focuses on Australia's efforts, more information about developments abroad is discussed elsewhere in this edition of CLB in our article *The Ethics of Artificial Intelligence: laws from around the world.*

AHRC White Paper: Artificial Intelligence: governance and leadership

Following an inquiry, in January 2019 the Australian Human Rights Commission released a White Paper entitled ***'Artificial Intelligence: governance and leadership'*** which highlighted a number of key ethical concerns linked to AI:

- **Human dignity and life and apportionment of responsibility:** The effect of AI informed decision making and systems on everyday life is unprecedented and raises myriad questions regarding how we, as a society, should apportion responsibility and accountability when things go wrong.
- **Fairness and non-discrimination:** AI can be a powerful tool for identifying trends, bias and discrimination in decision making. However, using AI-informed decision making runs the risk of perpetuating existing trends, biases and discrimination. If the algorithm is trained on data that has trended towards favouring a certain demographic, gender or ethnicity, then the algorithm may continue to make decisions that follow those ingrained biases. There is currently no legal framework that implements safeguards at the design, modelling and execution phases of technological development.
- **Data, privacy and personal autonomy:** The personal data that individuals provide in return for services has become a highly valuable commodity. Private organisations hold large amounts of data containing personal information which can be analysed and on-sold to advertisers seeking new markets.

Data61 Ethics Framework

Following the White Paper, the Department of Industry, Innovation and Science published the ***Artificial Intelligence: Australia's Ethics Framework*** (**Framework**) authored by CSIRO's Data 61 to discuss how to best harness the benefits of AI technology, while limiting the risks which accompany it.

Given the pace and breadth of development in AI, government, industry and developers will each need to play their role in addressing ethics. The Framework is intended to create a dialogue and serve as a starting point to help guide decision making and to engender trust.

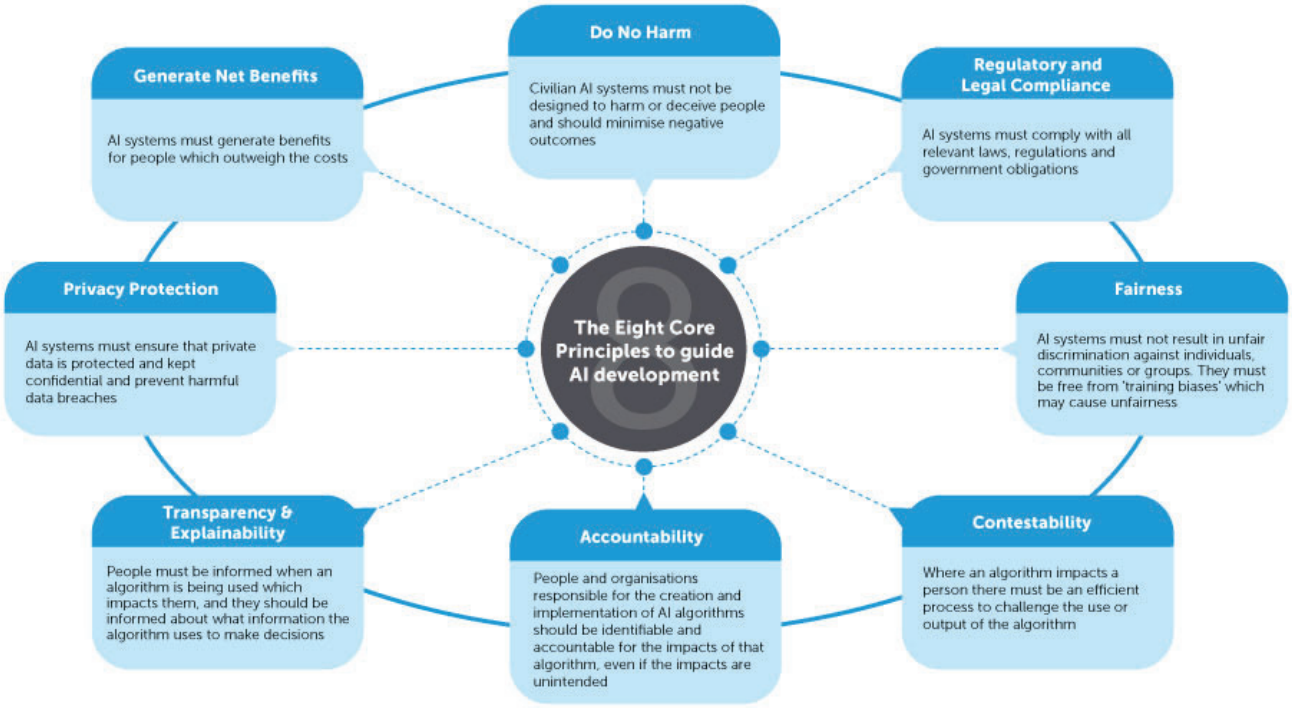
In developing the Framework, CSIRO's Data 61 formed the view that to truly unlock the potential of AI the public will need to have trust in AI applications. One mechanism to achieve this is aligning AI application development with ethical and inclusive values from the outset. It is not about rewriting laws or ethical standards but updating them so they can be applied in the context of new AI technologies.

The Framework draws on a number of complementary works and case studies from around the world to identify eight core principles which it suggests should guide AI development. See facing page:

The Ethical Development Toolkit

In addition to the eight core principles the Framework also proposes a 'toolkit' to help individuals, teams and organisations practically apply these core principles to their work. The Framework emphasises that there is no 'one-size fits all' solution when it comes to addressing ethics in AI, and it notes that the issues are challenging and not likely to remain static over time. As a starting point it extols those responsible for AI systems to ask

- what is the purpose of this system?
- which principles will guide the ethical use and deployment of the system?; and



- how would the requirements of meeting those principles be assessed?
- The Framework suggests a number of tools which might go to assist stakeholders in understanding how their systems express and incorporate these core principles. The tools include:
- **Impact assessments:** auditable assessments of potential direct and indirect impact of AI which address the potential negative impacts on individuals, communities and groups and mitigation procedures.
- **Internal or external review:** undertaken either by specialist professionals, groups or even in some cases other software to report on how the system is operating and whether it is adhering to ethical principles and applicable laws.
- **Risk assessments:** in particular with respect to assessing as a threshold matter whether certain uses of AI require additional assessment or review.
- **Best practice guidelines:** in particular to provide a flexible, accessible cross-industry guide for developers to implement that is adjusted as both technology and experience develop over time.
- **Industry standards:** in particular in the form of certification or standards which can be used as a short hand to assess off-the-shelf solutions. At this stage there is no agreed standard for AI systems or data science generally; however, both Standards Australia and the International Standards Organisation are working to develop technical and ethical standards in this space.
- **Collaboration:** programs that incentivise ‹ethical by design› AI drawing together industry and academia and groups from different backgrounds, combatting demographic bias and ensuring robust parallel development of ethical standards and technology both in theory and practice.
- **Monitoring and improvement mechanisms** which regularly review the outcomes of the system for accuracy, fairness and suitability – including whether the original goals of the algorithm remain relevant.
- **Recourse mechanisms** to create a path for appeals and human review of potentially erroneous automated decisions.
- **Consultation:** public or specialist consultation to provide an opportunity for ethical issues to be discussed by key stakeholders, including (as relevant) academics, industry and the public. In particular, the Framework notes the value of consultation in understanding the full breadth of ideas, concerns and solutions regarding ethical development of AI systems.

As developments in AI advance, governments and private actors seem acutely aware of the serious ethical considerations at play but are loath to miss out on the opportunities which AI technology presents. It is unlikely that anything as elegant as Asimov's Three Laws will ever be feasible as a rule of law; however, legislation will no doubt evolve to keep pace as the sector consolidates. With the community consultation window for the Framework having closed on 31 May, and the return of the incumbent government, we can expect developments in this space to be ongoing. In the meantime, governments abroad and major players in the tech sector are forming their own frameworks and best practices to address the ethical risks which advancements in AI pose.