Will we ever agree to just one set of rules on the ethical development of artificial intelligence?

May 29 2019, by Michael Guihot

Credit: Pavel Danilyuk from Pexels

Australia is among 42 countries that last week signed up to a new set of policy guidelines for the development of artificial intelligence (AI)

systems.

Yet Australia has its own draft guidelines for ethics in AI out for public consultation, and a number of other countries and industry bodies have developed their own AI guidelines.

So why do we need so many guidelines, and are any of them enforceable?

The new principles

The latest set of policy guidelines is the Recommendation on Artificial Intelligence from the Organisation for Economic Co-operation and Development (OECD).

It promotes five principles for the responsible development of trustworthy AI. It also includes five complementary strategies for developing national policy and international cooperation.

Given this 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 AI 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
5. accountability.

These recommendations are broad and do not carry the force of laws or even rules. Instead they seek to encourage member countries to

incorporate these values or ethics in the development of AI.

But what do we mean by AI?

It is hard to make specific recommendations in relation to AI. That is partly because AI is not one thing with a single application that poses singular risks or threats.

Instead, AI has become a blanket term to refer to a vast number of different systems. Each is typically designed to collect and process data using computing technology, adapt to change, and act rationally to achieve its objectives, ultimately without human intervention.

Those objectives could be as different as translating language, identifying faces, or even playing chess.

The type of AI that is exceptionally good at completing these objectives is often referred to as narrow AI. A good example is a chess-playing AI. This is specifically designed to play chess—and is extremely good at it—but completely useless at other tasks.

On the other hand is general AI. This is AI that it is said will replace human intelligence in most if not all tasks. This is still a long way off but remains the ultimate goal of some AI developers.

Yet it is this idea of general AI that drives many of the fears and misconceptions that surround AI.

Many many guidelines

Responding to these fears and a number of very real problems with narrow AI, the OECD recommendations are the latest of a number of

projects and guidelines from governments and other bodies around the world that seek to instil an ethical approach to developing AI.

These include initiatives by the Institute of Electrical and Electronics Engineers, the French data protection authority, the Hong Kong Office of the Privacy Commissioner and the European Commission.

The Australian government funded CSIRO's Data61 to develop an AI ethics framework, which is now open for public feedback, and the Australian Council of Learned Academies is yet to publish its report on the future of AI in Australia.

The Australian Human Rights Commission, together with the World Economic Forum, is also reviewing and reporting on the impact of AI on human rights.

The aim of these initiatives is to encourage or to nudge ethical development of AI. But this presupposes unethical behaviour. What is the mischief in AI?

Unethical AI

One study identified three broad potential malicious uses of AI. These target:

- digital security (for example, through cyber-attacks)
- physical security (for example, attacks using drones or hacking)
- political security (for example, if AI is used for mass surveillance, persuasion and deception).

One area of concern is evolving in China, where several regions are developing a social credit system linked to mass surveillance using AI technologies.

The system can identify a person breaching social norms (such as jaywalking, consorting with criminals, or misusing social media) and debit social credit points from the individual.

When a credit score is reduced, that person's freedoms (such as the freedom to travel or borrow money) are restricted. While this is not yet a nationwide system, reports indicate this could be the ultimate aim.

Added to these deliberate misuses of AI are several unintentional side effects of poorly constructed or implemented narrow AI. These include bias and discrimination and the erosion of trust.

Building consensus on AI

Societies differ on what is ethical. Even people within societies differ on what they regard as ethical behaviour. So how can there ever be a global consensus on the ethical development of AI?

Given the very broad scope of AI development, any policies in relation to ethical AI cannot yet be more specific until we can identify shared norms of ethical behaviour that might form the basis of some agreed global rules.

By developing and expressing the values, rights and norms that we consider to be important now in the form of the reports and guidelines outlined above, we are working toward building trust among nations.

Common themes are emerging in the various guidelines, such as the need for AI that considers human rights, security, safety, transparency, trustworthiness and accountability, so we may yet be on the way to some global consensus.

Commons license. Read the original article.

Provided by The Conversation