Explainer: what is artificial intelligence?

AI has jumped from sci-fi movie plots into mainstream news headlines in just a couple of years, according to this *ABC News* report by Margot O'Neill

And the headlines are often contradictory. AI is either a technological leap into greater prosperity or mass unemployment; it will either be our most valuable servant or terrifying master.

But what is AI, how does it work, and what are the benefits and the concerns?

WHAT IS ARTIFICIAL INTELLIGENCE?

AI is a computer system that can do tasks that humans need intelligence to do.

"An intelligent computer system could be as simple as a program that plays chess or as complex as a driverless car," Mary-Anne Williams, professor of social robotics at the University of Technology, Sydney, said.

A driverless car, for example, relies on multiple sensors to understand where it is and what's around it. These include speed, location, direction and 360-degree vision. Based on those inputs, among others, the "intelligent" computer system controls the car by deciding, like a human would, when to turn the steering and when to accelerate or brake.

Then there's machine learning, a subset of AI, which involves teaching computer programs to learn by finding patterns in data. The more data, the more the computer system improves.

"Whether it's recognising objects, identifying people in photos, reading lung scans or transcribing spoken mandarin, if we pick a narrow task like that [and] we give it enough data, the computer learns to do it as well as, if not better, than us," University of New South Wales professor of artificial intelligence Toby Walsh said.

AI doesn't have to sleep or make the same mistake twice. It can also access vast troves of digital data in seconds. Our brains cannot.

An intelligent computer system could be as simple as a program that plays chess or as complex as a driverless car.

DO I ALREADY USE AI?

Yes, probably every day. AI is in your smart phone; it's there every time you ask a question of iPhone's Siri or Amazon's Alexa. It's in your satellite navigation system and instant translation apps.

AI algorithms recognise your speech, provide search results, help sort your emails and recommend what you should buy, watch or read.

"AI is the new electricity," according to Andrew Ng, former chief scientist at Baidu, one of the leading Chinese web services companies. AI will increasingly be all around you from your phone to your TV, car and home appliances.

WHY ARE WE TALKING ABOUT IT NOW?

Four factors have now converged to push AI beyond games and into our everyday lives and workplaces:

- Computer processing power is doubling every two years (known as Moore's Law)
- The amount of data being generated is doubling every year (AI algorithms are hungry for data)
- Recently, the amount of AI funding has also been doubling every two years
- There is now 50 years of established AI research, giving us better and better algorithms.

The term artificial intelligence was first coined in 1956 by US computer scientist John McCarthy. Until recently, the public mostly heard about AI in Hollywood movies like *The Terminator* or whenever it defeated a human in a competition.

In 1997, IBM's Deep Blue computer beat Russian chess master Garry Kasparov. In 2011, IBM's supercomputer Watson beat human players on the US game show *Jeopardy!*. Last year, Google's AlphaGo beat Go master Lee Sedol.

"We now have the computing power, the data, the algorithms and a lot of people working on the problems," Professor Walsh said.

Machine learning, a subset of AI, involves teaching computer programs to learn by finding patterns in data. The more data, the more the computer system improves.

CAN AI HELP US?

AI promises spectacular benefits for humanity, including better and more precise medical diagnosis and treatment; relieving the drudgery and danger of repetitive and dehumanising jobs; and super-charging decision making and problem solving.

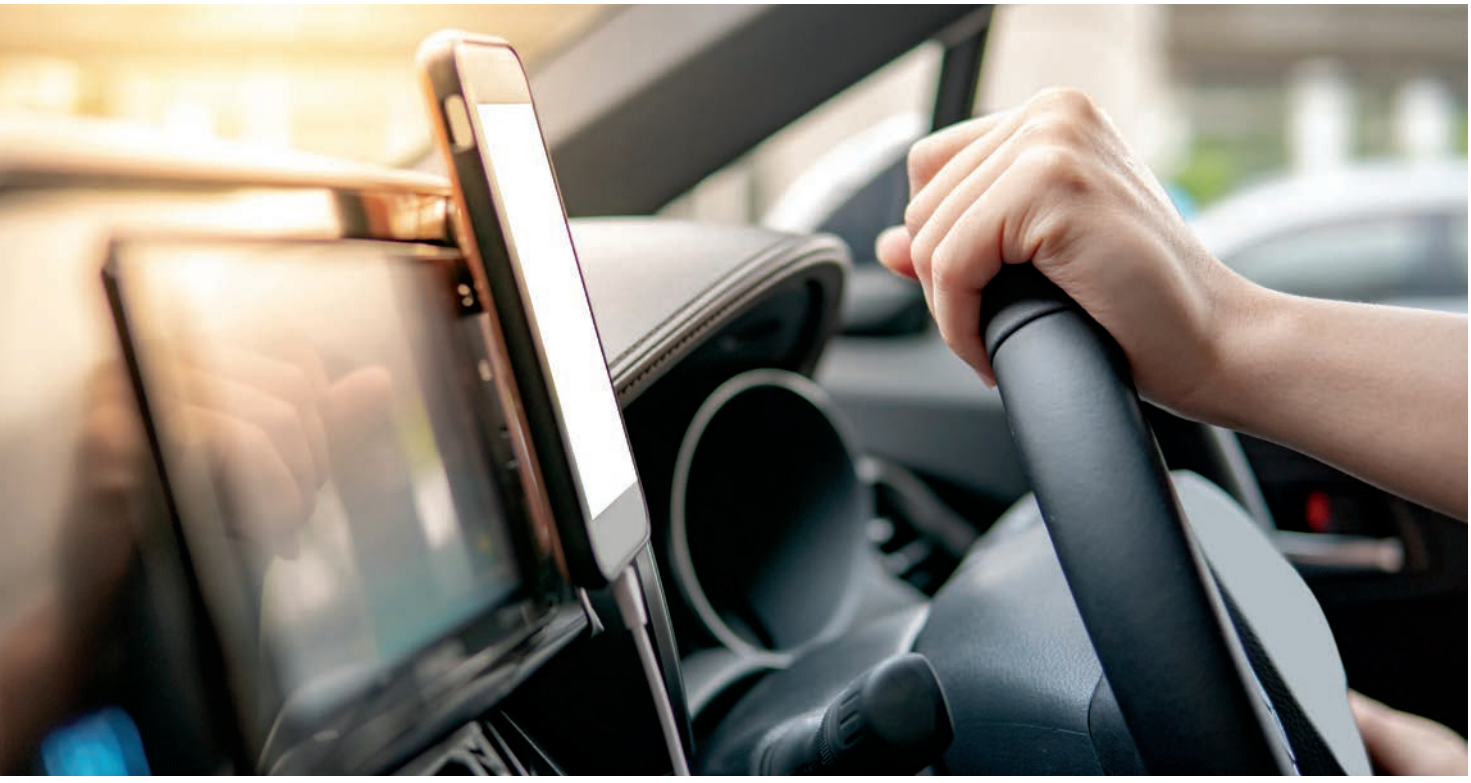
"Driverless cars could save many, many lives because 95 per cent of accidents are due to human error," Professor Walsh said.

"Many of the problems that are stressing our planet today will be tackled through having better decision making with computers that access and analyse vast troves of data," he said.

BUT CAN IT ALSO HURT US?

There are a range of concerns:

- That the AI and robotics revolution might create mass unemployment inside a generation
- That AI will further undermine privacy and democracy through greater mass surveillance by governments and companies
- That we will be more easily manipulated by personalised algorithms creating fake news



ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) is a broad term used to describe a collection of technologies that can solve problems and perform tasks to achieve defined objectives without explicit human guidance.

Central to AI are automation and machine learning that underpin applications such as natural language processing (Apple Siri or Amazon Alexa), computer vision (Tesla Autopilot), and optimisation and decision support (Google Maps). AI has the potential to automate repetitive or dangerous tasks, increase productivity and allow the development of innovative consumer products. It is forecast to add trillions of dollars to the global economy in the coming decades.

Examples include:

- Using advanced data analytics techniques to diagnose diseases at earlier and more treatable stages
- Using automated machines for hauling and drilling on mine sites, increasing productivity and reducing risks to workers
- Enabling greater use of smart forms that can tailor legal information to assist individuals to draft a will or settle financial arrangements following a break-up
- Tailoring content on entertainment platforms to meet user preferences.

There is considerable effort, both in Australia and internationally, focused on ensuring that AI is applied ethically and delivers broad societal benefits.

Department of Industry, Innovation and Science. *Understanding emerging technologies.* Retrieved from www.industry.gov.au on 9 July 2019.

- That algorithms will be biased but will be used to decide important issues in our lives such as insurance claims, job applications, loan applications and even judicial sentencing.

THAT ALL SOUNDS BAD. SO, WILL IT OVERTAKE HUMANITY?

Experts are famously split on this. Prominent tech entrepreneurs and scientists such as Elon Musk and Stephen Hawking, among others, warn that AI could reach and quickly surpass humans, transforming into super-intelligence that would render us the second most intelligent species on the planet.

Musk has compared it to "summoning the demon". Scientists call it 'singularity', "where machines improve themselves almost without end," Professor Walsh said.

Facebook's Mark Zuckerberg accuses Musk of being alarmist. Professor Walsh says we don't yet even fully understand all the facets of human intelligence and there may be limits to how far AI can develop.

He's surveyed 300 of his AI colleagues around the world and most believe if AI can reach human level intelligence, it is at least 50 to 100 years away.

If it happens, humanity will likely have already solved most of the problems about whether the machines' values are aligned with ours. "I'm not so worried about that," he says.

AND WHO CONTROLS IT?

The recent push into AI came from big US tech companies such as Google, Facebook, Amazon, Microsoft and Apple. And the US military. What could go wrong?

There's growing concern that these companies are too big and control too much data, which trains the AI algorithms.

China has now also joined the race with plans to dominate the world in AI development by 2030.

There's presently very little national or international regulation around how AI is developed. The Big Tech companies have begun discussing the need for guiding principles to ensure AI is only used for public good.

> AI promises spectacular benefits for humanity, including better and more precise medical diagnosis and treatment; relieving the drudgery and danger of repetitive and dehumanising jobs; and super-charging decision making and problem solving.

"One of those is what is the point of AI? It has to be to augment people, to support people, not replace them," Microsoft Australia national technology officer James Kavanagh says.

"Secondly, it has to be democratised. It can't be in the hands of a small number of technology companies.

"Thirdly, it has to be built on foundations of trust. We need to be able to understand any biases in algorithms and how they make decisions."

O'Neill, M (7 August 2017). 'Explainer: What is artificial intelligence?', *ABC News*. Retrieved from www.abc.net.au/news on 9 July 2019.

ARTIFICIAL INTELLIGENCE: 10 WAYS SOCIETY WILL CHANGE BY 2050

Leading Australian artificial intelligence scientist Professor Toby Walsh is warning that we are "sleepwalking" into an AI future in which billions of machines and computers will be able to think.

Professor Walsh, from the University of New South Wales, is calling for a national discussion about whether society needs to adopt clear boundaries and guidelines around how AI is developed and how it's used in our lives.

In his book *It's Alive: Artificial Intelligence from The Logic Piano to Killer Robots*, he has highlighted key questions in a series of predictions that describe how our future could be far better or far worse because of AI.

Here's how he thinks society might change by 2050, thanks to artificial intelligence.

1. YOU ARE BANNED FROM DRIVING

Humans drive drunk, tired and distracted and cause 95 per cent of accidents. The roads will be much safer without human drivers and most likely far less congested, as networked vehicles service passengers 24/7.

Street parking and most car parks will disappear, transport will be cheaper and groups such the elderly and disabled will have greater personal mobility.

Most people won't bother buying cars and will lose driving skills. And autonomous vehicles will arrive quickly – within 15-20 years.

"By 2050, the year 2000 will look as quaintly old-fashioned as the horse-drawn era of 1900 did to people in 1950," Professor Walsh said.

2. YOU SEE THE DOCTOR DAILY

Your personal 24/7 AI doctor will know your gene sequence and vulnerabilities to particular diseases. It will continually monitor your blood pressure, sugar levels, sleep and exercise. It will process data from your toilet, which will automatically analyse your urine and stools.

Your future version of a smartphone or fitness watch will regularly take selfies to identify melanomas and eye disease. It will record your voice for signs of a cold, dementia or a stroke. It will call for help if you faint. It will also be a trillion-dollar global business.

"Our personal AI physician will have our life history, it will know far more about medicine than any single doctor, and it will stay on top of all the emerging medical literature," Professor Walsh said.

3. MARILYN MONROE IS BACK IN THE MOVIES

Avatars will be programmed to act and talk like anyone we choose in interactive movies, including ourselves or celebrities from recent history. Where the story goes depends on what you do or say.

Hollywood and the computer games industry will merge and immerse us in hyper-real worlds. But there will be increasing concern about the seductive nature of these unreal, alternate worlds. There may be an underclass of addicts who spend every waking moment in them. And some who behave in distasteful or illegal ways.

"This problem will likely trouble our society greatly," Professor Walsh said.

"There will be calls that behaviours which are illegal in the real world should be made illegal or impossible in the virtual."

4. A COMPUTER HIRES AND FIRES YOU

That's just the beginning. AI systems will also increasingly take over managing how you work: scheduling your activities, approving holidays, monitoring and rewarding your performance.

"By 2050, the year 2000 will look as quaintly old-fashioned as the horse-drawn era of 1900 did to people in 1950."

But should we hand over decisions like hiring and especially firing to a computer?

"We will have to learn when to say to computers: 'Sorry, I can't let you do that.' It's not enough for a machine to do a task better than a human. There are some decisions we simply should not allow machines to make."

5. YOU TALK TO ROOMS

You will walk into a room and say "lights on" and "who won the football?" and one of the many AI devices in your house will recognise your voice and understand you well enough to know which football code you follow.

A few people will resist and determinedly follow a disconnected 20th century life. But most of us will take advantage of having just about everything in our lives connected: fridges, toasters, baths, door locks, windows, bicycles and pot plants.

AI will operate through the so-called Internet of Things using conversation instead of typing.

"Our privacy, diversity and democracy will be challenged," Professor Walsh said.

"[Government] intelligence [agencies] can't wait for every room to be listening to us. Marketers, too, would love all this data about our everyday lives.

"So, the next time you get asked to check your privacy settings, think long and hard about what you may be giving up."

6. A ROBOT ROBS A BANK

Cyber-crime to date has been relatively low-tech with phishing and malware attacks. But AI will surpass human hackers – and the only defence will be another AI program.

Warfare is also moving into cyberspace. But these technologies will also quickly find their way into the civilian sphere. One of the challenges will be that many advances in AI used to defend systems will be quickly turned around to attack systems.

"The supposed hacking by Russians in order to influence the 2016 US presidential election demonstrates the impact that such cyber attacks can have," Professor Walsh said.

"Banks [and other companies and governments] will have no choice but to invest more and more in sophisticated AI systems to defend themselves from attack."

7. WORLD SOCCER CHAMPIONS LOSE TO A ROBOT TEAM

Robots will have superior ball skills, including unfailing accuracy in passes and penalties. They will know precisely where all players are at all times and will know how to interpret that information because their AI system learned strategic play from watching every World Cup match ever recorded.

The human team will be soundly defeated. Even fans of the robots will call for the humans to be given a break. That's why most sporting teams will stay human. But AI will change football and most other games with managers and players using AI to train and play better.

"Data scientists will be some of the best paid members of football [and other sporting] clubs," Professor Walsh said.

"Scouts will hang out at [top universities] to recruit young computer scientists."

8. GHOST SHIPS, PLANES AND TRAINS CROSS THE GLOBE

The oceans, skies and railroads of the planet will be filled with autonomous ships, planes and trains transporting cargo without any people on board, as driverless car technology spreads to other industries. It will improve safety and efficiency. And children will no longer grow up wanting to be train drivers.

"Planes carrying people will probably continue to be piloted by humans," Professor Walsh said.

"But after several decades of safe flights by cargo planes, the debate will begin whether humans should still be airline pilots."

9. TV NEWS IS MADE WITHOUT HUMANS

Nearly every part of this prediction is already here – it's just that no one has yet pulled all the pieces together. Computers now write simple sport and financial stories but as technology improves, AI will write more complex stories. Avatars and chatbots will play the role of presenters filmed by robotic cameras. And the

news you watch will be narrowcast, or tailored to your personal preferences.

"There will be ongoing debate about the biases of algorithms, especially when humans take no part in deciding what news we see," Professor Walsh said.

"Our viewpoints are shaped by the lens through which we look at the world. Will algorithms challenge us enough? Will they understand lies and deception? Will they care about what we care about?"

10. HUMANS LIVE ON AFTER DEATH

It will be common to leave behind an AI chatbot that will talk like you, know the story of your life and comfort your family when you die.

Some people might give their chatbot the task of reading their will; settling old scores; or relieving grief through humour.

Digital doubles will also appear in place of the living. Celebrities will use bots to create social media; many of us will similarly use them to manage our diaries.

This digital outsourcing will fuel a lively debate.

Professor Walsh asked: "What redress do you have against an AI bot that pretends to be you? Do you have a right to know if you're interacting with a computer rather than a real person? Should AI bots be prohibited from political discourse? Who can switch off your bot after you die? Do bots have freedom of speech? It will be an interesting future."

Walsh, T (7 August 2017). 'Artificial intelligence: Professor Toby Walsh on 10 ways society will change by 2050', *ABC News*. Retrieved from www.abc.net.au/news on 9 July 2019.

YOUR QUESTIONS ANSWERED ON ARTIFICIAL INTELLIGENCE

Artificial intelligence and robotics have enjoyed a resurgence of interest, and there is renewed optimism about their place in our future. But what do they mean for us? Following are answers to questions put to experts about AI and robotics, first published by The Conversation

Q1. How plausible is human-like artificial intelligence?

A. Toby Walsh, Professor of AI

It is 100% plausible that we'll have human-like artificial intelligence. I say this even though the human brain is the most complex system in the universe that we know of. There's nothing approaching the complexity of the brain's billions of neurons and trillions of connections. But there are also no physical laws we know of that would prevent us reproducing or exceeding its capabilities.

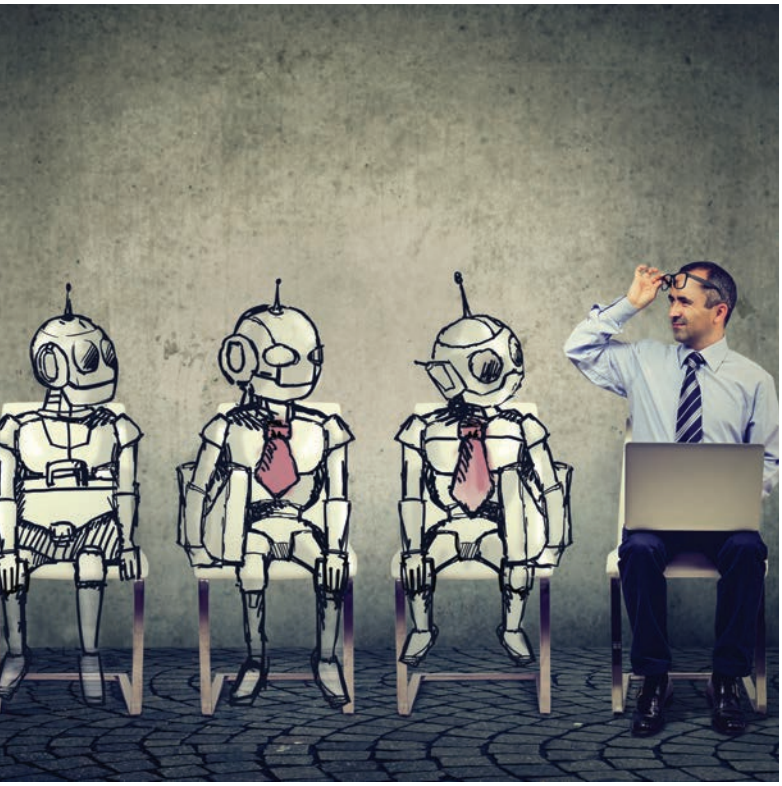
A. Kevin Korb, Reader in Computer Science

Popular fiction AI from Issac Asimov to Steven Spielberg is plausible. What the question doesn't address is: when will it be plausible?

Most AI researchers (including me) see little or no evidence of it coming anytime soon. Progress on the major AI challenges is slow, if real.

What I find less plausible than the AI in fiction is the emotional and moral lives of robots. They seem to be either unrealistically empty, such as the emotionless Data in *Star Trek*, or unrealistically human-identical or superior, such as the AI in Spike Jonze's *Her*.

All three – emotion, ethics and intelligence – travel together, and are not genuinely possible in some form without the others, but fiction writers tend to treat them as separate. Plato's Socrates made a similar mistake.



A. Gary Lea, Researcher in Artificial Intelligence Regulation

AI is not impossible, but the real issue is: "how like is like?" The answer probably lies in applied tests: the Turing test was already (arguably) passed in 2014 but there is also the coffee test (can an embodied AI walk into an unfamiliar house and make a cup of coffee?), the college degree test and the job test.

If AI systems could progressively pass all of those tests (plus whatever else the psychologists might think of), then we would be getting very close. Perhaps the ultimate challenge would be whether a suitably embodied AI could live among us as J. Average and go undetected for five years or so before declaring itself.

Q2. Automation is already replacing many jobs. Is it time to make laws to protect some of these industries?

A. Jonathan Roberts, Professor of Robotics

Researchers at the University of Oxford published a now well-cited paper in 2013 that ranked jobs in order of how feasible it was to computerise or automate them. They found that nearly half of jobs in the USA could be at risk from computerisation within 20 years.

This research was followed in 2014 by the viral video hit, *Humans Need Not Apply*, which argued that many jobs will be replaced by robots or automated systems and that employment would be a major issue for humans in the future.

Of course, it is difficult to predict what will happen, as the reasons for replacing people with machines are not simply based around available technology. The major factor is actually the business case and the social attitudes and behaviour of people in particular markets.

A. Rob Sparrow, Professor of Philosophy

Advances in computing and robotic technologies are undoubtedly going to lead to the replacement of many jobs currently done by humans. I'm not convinced that we should be making laws to protect particular industries though. Rather, I think we should be doing two things.

First, we should be making sure that people are assured of a good standard of living and an opportunity to pursue meaningful projects even in a world in which many more jobs are being done by machines. After all, the idea that, in the future, machines would work so that human beings didn't have to toil used to

be a common theme in utopian thought.

When we accept that machines putting people out of work is bad, what we are really accepting is the idea that whether ordinary people have an income and access to activities that can give their lives meaning should be up to the wealthy, who may choose to employ them or not. Instead, we should be looking to redistribute the wealth generated by machines in order to reduce the need for people to work without thereby reducing the opportunities available to them to be doing things that they care about and gain value from.

Second, we should be protecting vulnerable people in our society from being treated worse by machines than they would be treated by human beings. With my mother, Linda Sparrow, I have argued that introducing robots into the aged care setting will most likely result in older people receiving a worse standard of treatment than they already do in the aged care sector. Prisoners and children are also groups who are vulnerable to suffering at the hands of robots introduced without their consent.

A. Toby Walsh, Professor of AI

There are some big changes about to happen. The #1 job in the US today is truck driver. In 30 years time, most trucks will be autonomous.

How we cope with this change is a question not for technologists like myself but for society as a whole. History would suggest that protectionism is unlikely to work. We would, for instance, need every country in the world to sign up.

But there are other ways we can adjust to this brave new world. My vote would be to ensure we have an educated workforce that can adapt to the new jobs that technology create.

We need people to enter the workforce with skills for jobs that will exist in a couple of decades time when the technologies for these jobs have been invented.

We need to ensure that everyone benefits from the rising tide of technology, not just the owners of the robots. Perhaps we can all work less and share the economic benefits of automation? This is likely to require fundamental changes to our taxation and welfare system informed by the ideas of people like the economist Thomas Piketty.

A. Kevin Korb, Reader in Computer Science

Industrial protection and restriction are the wrong way to go. I'd rather we develop our technology so as to help solve some of our very real problems. That's bound to bring with it economic dislocation, so a caring society will accommodate those who lose out because of it.

But there's no reason we can't address that with improving technology as long as we keep the oligarchs under control. And if we educate people for flexibility rather than to fit into a particular job, intelligent people will be able to cope with the dislocation.

A. Jai Galliot, Defence Analyst

The standard argument is that workers displaced by

automation go on to find more meaningful work. However, this does not hold in all cases.

Think about someone who signed up with the air force to fly jets. These pilots may have spent their whole social, physical and psychological lives preparing or maintaining readiness to defend their nation and its people.

For service personnel, there are few higher-value jobs than serving one's nation through rendering active military service on the battlefield, so this assurance of finding alternative and meaningful work in a more passive role is likely to be of little consolation to a displaced soldier.

Thinking beyond the military, we need to be concerned that the Foundation for Young Australians indicates that as many as 60% of today's young people are being trained for jobs that will soon be transformed due to automation.

The sad fact of the matter is that one robot can replace many workers. The future of developed economies therefore depends on youth adapting to globalised and/or shared jobs that are increasingly complemented by automation within what will inevitably be an innovation and knowledge economy.

Q3. Where will AI be in five to ten years?

A. Toby Walsh, Professor of AI

AI will become the operating system of all our connected devices. Apps like Siri and Cortana will morph into the way we interact with the connected world.

AI will be the way we interact with our smartphones, cars, fridges, central heating system and front door. We will be living in an always-on world.

A. *Jonathan Roberts, Professor of Robotics*

It is likely that in the next five to ten years we will see machine learning systems interact with us in the form of robots. The next large technology hurdle that must be overcome in robotics is to give them the power of sight.

This is a grand challenge and one that has filled the research careers of many thousands of robotics researchers over the past four or five decades. There is a growing feeling in the robotics community that machine learning using large datasets will finally crack some of the problems in enabling a robot to actually see.

Four universities have recently teamed up in Australia in an ARC funded Centre of Excellence in Robotic Vision. Their mission is to solve many of the problems that prevent robots seeing.

Q4. Should we be concerned about military and other armed robots?

A. *Rob Sparrow, Professor of Philosophy*

The last thing humanity needs now is for many of its most talented engineers and roboticists to be working on machines for killing people.

Robotic weapons will greatly lower the threshold of conflict. They will make it easier for governments to start wars because they will hold out the illusion of being able to fight without taking any casualties. They will increase the risk of accidental war because militaries will deploy unmanned systems in high-threat environments, where it would be too risky to place a human being, such as just outside a potential enemy's airspace or deep sea ports.

In these circumstances, robots may even start wars without any human being having the chance to veto the decision. The use of autonomous robots to kill people threatens to further erode respect for human life.

It was for these reasons that, with several colleagues overseas, I co-founded the International Committee for Robot Arms Control, which has in turn supported the Campaign to Stop Killer Robots.

A. *Toby Walsh, Professor of AI*

"Killer robots" are the next revolution in warfare, after gunpowder and nuclear bombs. If we act now, we can perhaps get a ban in place and prevent an arms race to develop better and better killer robots.

A ban won't uninvent the technology. It's much the same technology that will go, for instance, into our autonomous cars. And autonomous cars will prevent the 1,000 or so deaths on the roads of Australia each year.

But a ban will associate enough stigma with the technology that arms companies won't sell them, that arms companies won't develop them to be better and better at killing humans. This has worked with a number of other weapon types in the past like blinding lasers. If we don't put a ban in place, you can be sure that terrorists and rogue nations will use killer robots against us.

For those who argue that killer robots are already covered by existing humanitarian law, I profoundly disagree. We cannot correctly engineer them today not to cause excessive collateral damage. And in the future, when we can, there is little stopping them being hacked and made to behave unethically. Even used lawfully, they will be weapons of terror.

You can learn more about these issues by watching my TEDx talk on this topic, *How can you stop killer robots?*

A. *Sean Welsh, Researcher in Robot Ethics*

We should be concerned about military robots. How-

ever, we should not be under the illusion that there is no existing legislation that regulates weaponised robots.

There is no specific law that bans murdering with piano wire. There is simply a general law against murder. We do not need to ban piano wire to stop murders. Similarly, existing laws already forbid the use of any weapons to commit murder in peacetime and to cause unlawful deaths in wartime.

There is no need to ban autonomous weapons as a result of fears that they may be used unlawfully any more than there is a need to ban autonomous cars for fear they might be used illegally (as car bombs). The use of any weapon that is indiscriminate, disproportionate and causes unnecessary suffering is already unlawful under international humanitarian law.

Some advocate that autonomous weapons should be put in the same category as biological and chemical weapons. However, the main reason for bans on chemical and biological weapons is that they are inherently indiscriminate (cannot tell friend from foe from civilian) and cause unnecessary suffering (slow painful deaths). They have no humanitarian positives.

By contrast, there is no suggestion that "killer robots" (even in the examples given by opponents) will necessarily be indiscriminate or cause painful deaths. The increased precision and accuracy of robotic weapons systems compared to human operated ones is a key point in their favour.

If correctly engineered, they would be less likely to cause collateral damage to innocents than human-operated weapons. Indeed robot weapons might be engineered so as to be more likely to capture rather than kill. Autonomous weapons do have potential humanitarian positives.

Q5. How plausible is super-intelligent AI?

A. David Dowe, Associate Professor in Machine Learning and Artificial Intelligence

We can look at the progress made at various tasks once said to be impossible for machines to do, and see them one by one gradually being achieved. For example: beating the human world chess champion (1997); winning at *Jeopardy!* (2011); driverless vehicles, which are now somewhat standard on mining sites; automated translation, etc.

And, insofar as intelligence test problems are a measure of intelligence, I've recently looked at how computers are performing on these tests.

A. Rob Sparrow, Professor of Philosophy

If there can be artificial intelligence then there can be super-intelligent artificial intelligences. There doesn't seem to be any reason why entities other than human beings could not be intelligent. Nor does there seem to be any reason to think that highest human IQ represents the upper limit on intelligence.

If there is any danger of human beings creating such machines in the near future, we should be very scared. Think about how human beings treat rats. Why should machines that were as many times more intelligent than us, as we are more intelligent than rats, treat us any better?

Q6. Given what little we know about our own minds, can we expect to intentionally create artificial consciousness?

A. Kevin Korb, Reader in Computer Science

As a believer in functionalism, I believe it is possible to create artificial consciousness. It doesn't follow that we can "expect" to do it, but only that we might.

John Searle's arguments against the possibility of artificial consciousness seem to confuse functional realisability with computational realisability. That is, it may well be (logically) impossible to "compute" consciousness, but that doesn't mean that an embedded, functional computer cannot be conscious.

A. Rob Sparrow, Professor of Philosophy

A number of engineers, computer scientists, and science fiction authors argue that we are on the verge of creating artificial consciousness. They usually proceed by estimating the number of neurons in the human brain and pointing out that we will soon be able to build computers with a similar number of logic gates.

If you ask a psychologist or a psychiatrist, whose job it is to actually "fix" minds, I think you will likely get a very different answer. After all, the state-of-the-art treatment for severe depression still consists of shocking the brain with electricity, which looks remarkably like trying to fix a stalled car by pouring petrol over the top of the engine. So I'm sceptical that we understand enough about the mind to design one.

Q7. How do cyborgs differ (technically or conceptually) from AI?

A. Katina Michael, Associate Professor in Information Systems

A cyborg is a human-machine combination. By definition, a cyborg is any human who adds parts, or enhances his or her abilities by using technology. As we have advanced our technological capabilities, we have discovered that we can merge technology onto and into the human body for prosthesis and/or amplification. Thus, technology is no longer an extension of us, but "becomes" a part of us if we opt into that design.

In contrast, artificial intelligence is the capability of a computer system to learn from its experiences and simulate human intelligence in decision-making. A cyborg usually begins as a human and may undergo a transformational process, whereas artificial intelligence is imbued into a computer system itself predominantly in the form of software.

Some researchers have claimed that a cyborg can also begin in a humanoid robot and incorporate the living tissue of a human or other organism. Regardless, whether it is a human-to-machine or machine-to-organism coalescence, when AI is applied via silicon microchips or nanotechnology embedded into prosthetic forms like a dependent limb, a vital organ, or a replacement/

additional sensory input, a human or piece of machinery is said to be a cyborg.

There are already early experiments with such cybernetics. In 1998 Professor Kevin Warwick named his first experiment Cyborg 1.0, surgically implanting a silicon chip transponder into his forearm. In 2002 in project Cyborg 2.0, Warwick had a one hundred electrode array surgically implanted into the median nerve fibres of his left arm.

Ultimately we need to be extremely careful that any artificial intelligence we invite into our bodies does not submerge the human consciousness and, in doing so, rule over it.

Q8. Are you generally optimistic or pessimistic about future of artificial intelligence and its benefits for humanity?

A. Toby Walsh, Professor of AI

I am both optimistic and pessimistic. AI is one of humankind's truly revolutionary endeavours. It will transform our economies, our society and our position in the centre of this world. If we get this right, the world will be a much better place. We'll all be healthier, wealthier and happier.

Of course, as with any technology, there are also bad paths we might end up following instead of the good ones. And unfortunately, humankind has a track record of late of following the bad paths.

We know global warming is coming but we seem unable not to follow this path. We know that terrorism is fracturing the world but we seem unable to prevent this. AI will also challenge our society in deep and fundamental ways. It will, for instance, completely change the nature of work. Science fiction will soon be science fact.

A. Rob Sparrow, Professor of Philosophy

I am generally pessimistic about the long-term impact of artificial intelligence research on humanity. I don't want to deny that artificial intelligence has many benefits to offer, especially in supporting human beings to make better decisions and to pursue scientific goals that are currently beyond our reach. Investigating how brains work by trying to build machines that can do what they do is an interesting and worthwhile project in its own right.

However, there is a real danger the systems that AI researchers come up with will mainly be used to further enrich the wealthy and to entrench the power of the powerful. I also think there is a risk that the prospect of AI will allow people to delude themselves that we don't need to do something about climate change now. It may also distract them from the fact that we already know what to do, but we lack the political will to do it.

Finally, even though I don't think we've currently got much of a clue of how this might happen, if engineers do eventually succeed in creating genuine AIs that are smarter than we are, this might well be a species-level extinction threat.

A. Jonathan Roberts, Professor of Robotics

I am generally optimistic about the long-term future of AI to humanity. I think that AI has the potential to radically change humanity and hence, if you don't like change, you are not going to like the future.

I think that AI will revolutionise health care, especially diagnosis, and will enable the customisation of medicine to the individual. It is very possible that AI GPs and robot doctors will share their knowledge as they acquire it, creating a super doctor that will have access to all the medical data of the world.

I am also optimistic because humans tend to recognise when technology is having major negative consequences, and we eventually deal with it. Humans are in control and will naturally try and use technology to make a better world.

A. Kevin Korb, Reader in Computer Science

I'm pessimistic about the medium-term future of humanity. I think climate change and attendant dislocations, wars etc. may well massively disrupt science and technology. In that case progress on AI may stop.

If that doesn't happen, then I think progress will continue and we'll achieve AI in the long-term. Along the way, AI research will produce spin-offs that help economy and society, so I think as long as it exists AI tech will be important.

A. Gary Lea, Researcher in Artificial Intelligence Regulation

I suspect the long-term future for AI will turn out to be the usual mixed bag: some good, some bad. If scientists and engineers think sensibly about safety and public welfare when making their research, design and build choices (and provided there are suitable regulatory frameworks in place as a backstop), I think we should be okay.

So, on balance, I am cautiously optimistic on this front – but there are many other long-term existential risks for humanity.

Toby Walsh is Professor of AI, Research Group Leader, Optimisation Research Group, Data61. David Dowe is Associate Professor, Clayton School of Information Technology, Monash University. Gary Lea is Visiting Researcher in Artificial Intelligence Regulation, Australian National University. Jai Galliott is Research Fellow in Indo-Pacific Defence, UNSW. Jonathan Roberts is Professor in Robotics, Queensland University of Technology. Katina Michael is Associate Professor, School of Information Systems and Technology, University of Wollongong. Kevin Korb is Reader in Computer Science, Monash University. Robert Sparrow is Professor, Department of Philosophy; Adjunct Professor, Centre for Human Bioethics, Monash University. Sean Welsh is Doctoral Candidate in Robot Ethics, University of Canterbury.

THE CONVERSATION

Walsh, T, Dowe, D, Lea, G, Galliott, J, Roberts, J, Michael, K, Korb, K, Sparrow, R, and Welsh, S (16 November 2015). *Your questions answered on artificial intelligence.* Retrieved from http://theconversation.com on 9 July 2019.

HISTORY OF ARTIFICIAL INTELLIGENCE

A BRIEF HISTORY OF AI FROM THE QUEENSLAND BRAIN INSTITUTE

Modern AI began in the 1950s with the view to solving complex mathematical problems and creating 'thinking machines'.

From the outset, there were two competing approaches. One used formal rules to manipulate symbols, a logic-based approach not at all based on biology. This became known as 'good old-fashioned artificial intelligence': GOFAI. The other camp took inspiration from how the brain works and created 'artificial neural networks' loosely inspired by our brains. These still had to be trained using certain procedures to solve problems.

In the first 20 years, GOFAI was the more successful approach, leading to much hype and significant government funding. But in real-world settings GOFAI didn't achieve its outcomes. Artificial neural networks also struggled, and in the 1970s funding dried up, research slowed and the AI community shrank.

In the 1980s, improvements were made in both the rules-based GOFAI systems and biologically-inspired neural networks. Previously difficult problems became achievable and AI seemed promising once again. However, the hope and hype exceeded reality, and by the 1990s AI research again diminished.

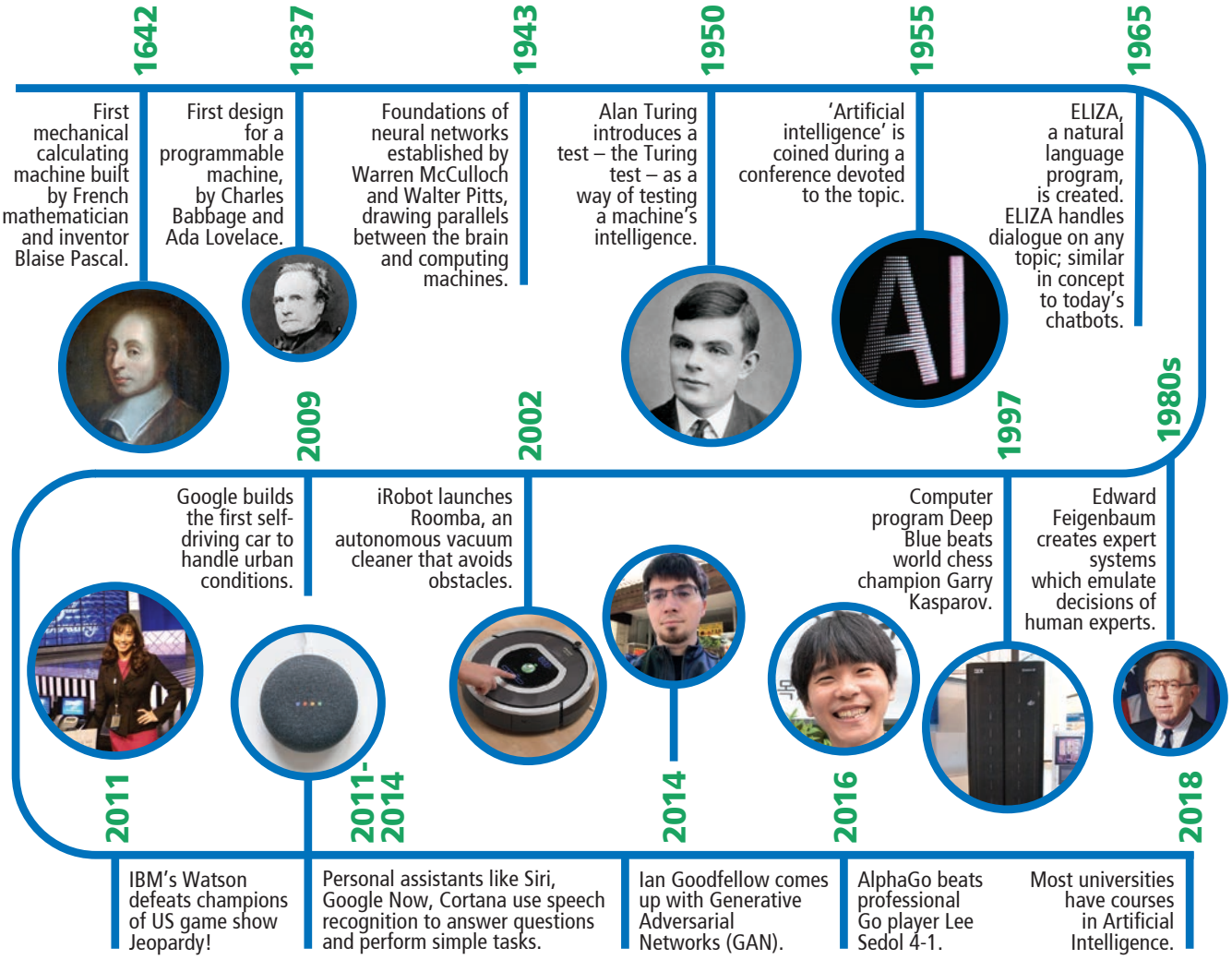
The latest surge of interest comes off the back of the power of deep learning, a type of biologically-inspired neural network that harnesses the huge amounts of data now available, and the massive computational power and speed of today's computers.

With enormous data sets, modern AI neural networks can often exceed human performance in many tasks, including pattern recognition and playing games like Go, previously very difficult for AI. Importantly, these systems can learn from experience, unlike GOFAI.

AI's ubiquity might now appear like it's not far off reaching human-level intelligence. But AI needs massive amounts of data to learn, unlike our brains, which can learn from a single experience.

Some researchers argue that for AI to advance further, more needs to be understood about the basic principles of how our brains function, and the kinds of biological shortcuts our brains take to complete tasks.

Queensland Brain Institute (30 January 2019). *History of Artificial Intelligence.* Retrieved from http://qbi.uq.edu.au on 9 July 2019.



ARTIFICIAL INTELLIGENCE: AUSTRALIA'S ETHICS FRAMEWORK

EXECUTIVE SUMMARY FROM A DISCUSSION PAPER BY THE CSIRO AND DATA61

The ethics of artificial intelligence are of growing importance

Artificial intelligence (AI) is changing societies and economies around the world. Data61 analysis reveals that over the past few years, 14 countries and international organisations have announced AU$86 billion for AI programs. Some of these technologies are powerful, which means they have considerable potential for both improved ethical outcomes as well as ethical risks. This report identifies key principles and measures that can be used to achieve the best possible results from AI, while keeping the wellbeing of Australians as the top priority.

Countries worldwide are developing solutions

Recent advances in AI-enabled technologies have prompted a wave of responses across the globe, as nations attempt to tackle emerging ethical issues (*Figure 1*).

Germany has delved into the ethics of automated vehicles, rolling out the most comprehensive government-led ethical guidance on their development available[1]. New York has put in place an automated decisions task force, to review key systems used by government agencies for accountability and fairness[2]. The UK has a number of government advisory bodies, notably the Centre for Data Ethics and Innovation[3]. The European Union has explicitly highlighted ethical AI development as a source of competitive advantage[4].

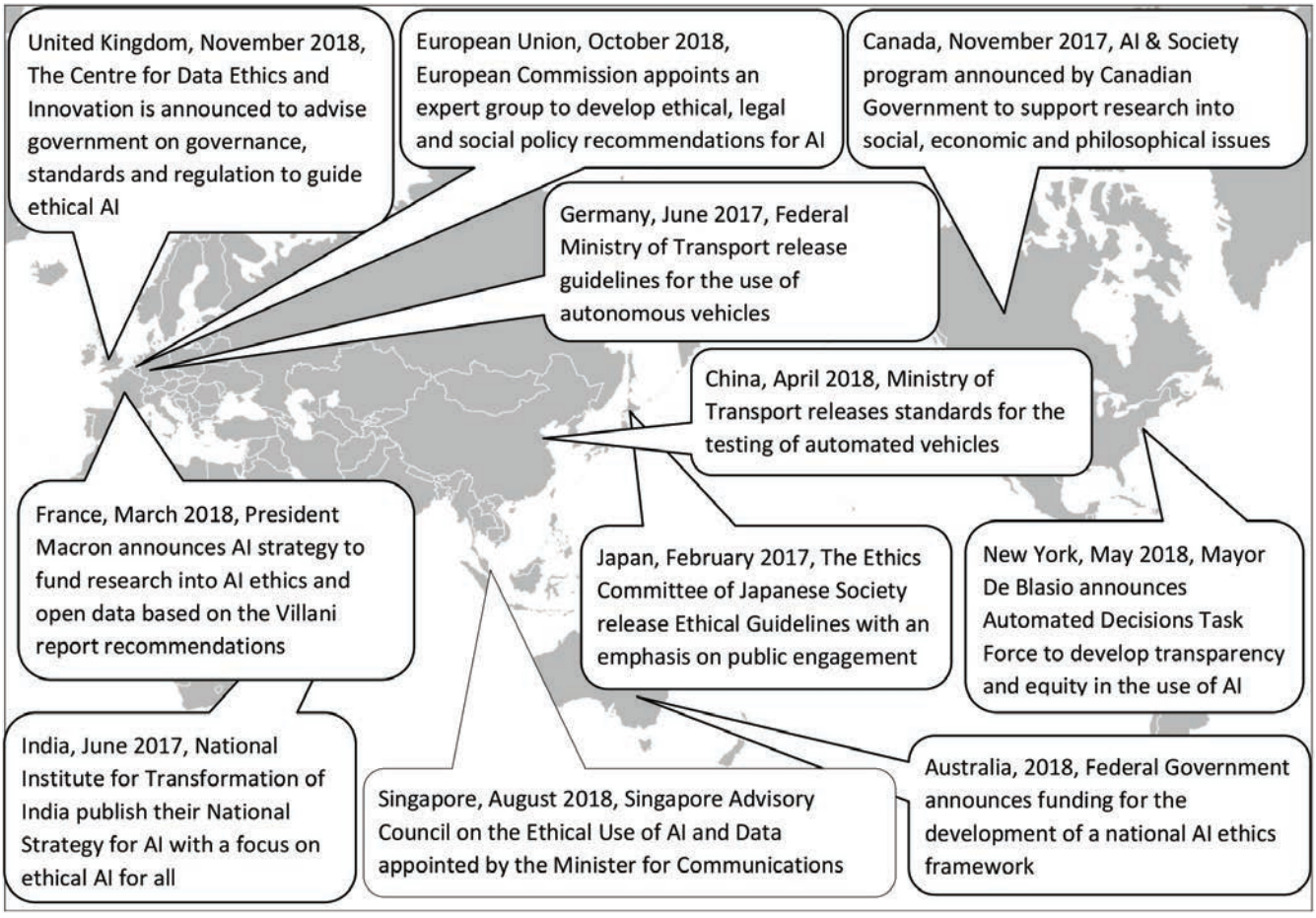
An approach based on case studies

This report examines key issues through exploring a series of case studies and trends that have prompted ethical debate in Australia and worldwide (*see Figure 2*).

Artificial intelligence (AI) holds enormous potential to improve society

While a "general AI" that replicates human intelligence is seen as an unlikely prospect in the coming few decades, there are numerous "narrow AI" technologies which are already incredibly sophisticated at handling specific tasks[3]. Medical AI technologies and autonomous vehicles are just a few high profile examples of AI that have potential to save lives and transform society.

FIGURE 1: MAP OF RECENT DEVELOPMENTS IN ARTIFICIAL INTELLIGENCE ETHICS WORLDWIDE



Data sources: Pan-Canadian AI Strategy[5], Australian Federal Budget 2018-19[6], German Ministry of Transport and Digital Infrastructure[1], National Institute for Transformation of India[7], The Villani Report[8], Reuters[9], Japanese Society for Artificial Intelligence[10], European Commission[11], UK Parliament[12], Singapore Government[13], China's State Council[14], New York City Hall[2].

FIGURE 2: TABLE OF KEY ISSUES EXAMINED IN CHAPTERS, CASE STUDIES AND RELEVANT PRINCIPLES

	Examples of case studies	Most relevant principles
Data governance and AI	**Identifying de-identified data** In 2016, a dataset that included de-identified health information was uploaded to data.gov.au. It was expected that the data would be a useful tool for medical research and policy development. Unfortunately, it was discovered that in combination with other publicly available information, researchers were able to personally identify individuals from the data source. Quick action was taken to remove the dataset from data.gov.au.	**Privacy protection** **Fairness**
Automated decisions	**Houston teachers fired by automated system** An AI was used by the Houston school district to assess teacher performance and in some cases fire them. There was little transparency regarding the way that the AI was operating. The use of this AI was challenged in court by the teacher's union, as the system was proprietary software and its inner workings were hidden. The case was settled and the district stopped using it[15].	**Fairness** **Transparency and explainability** **Contestability** **Accountability**
Predicting human behaviour	**The COMPAS sentencing tool** COMPAS is a tool used in the US to give recommendations to judges about whether prospective parolee will re-offend. There is extensive debate over the accuracy of the system and whether it is fair to African Americans. Investigations by a non-profit outlet have indicated that incorrect predictions unfairly categorise black Americans as a higher risk. The system is proprietary software[16-19].	**Do no harm** **Regulatory and legal compliance** **Privacy protection** **Fairness** **Transparency and explainability**

Data sources: Office of the Australian Information Commissioner[20], US Senate Community Affairs Committee Secretariat[15], ProPublica[16,18,19], Northpointe[17].

The benefits come with risks

Automated decision systems can limit issues associated with human bias, but only if due care is focused on the data used by those systems and the ways they assess what is fair or safe. Automated vehicles could save thousands of lives by limiting accidents caused by human error, but as Germany's Transport Ministry has highlighted in its ethics framework for AVs, they require regulation to ensure safety[1].

Existing ethics in context, not reinvented

Philosophers, academics, political leaders and ethicists have spent centuries developing ethical concepts, culminating in the human-rights based framework used in international and Australian law. Australia is a party to seven core human rights agreements which have shaped our laws[21].

An ethics framework for AI is not about rewriting these laws or ethical standards, it is about updating them to ensure that existing laws and ethical principles can be applied in the context of new AI technologies.

Data is at the core of AI

The recent advances in key AI capabilities such as deep

CORE PRINCIPLES FOR AI

1. **Generates net benefits.** The AI system must generate benefits for people that are greater than the costs.
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.
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.
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.
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.
6. **Transparency and 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.
7. **Contestability.** When an algorithm impacts a person there must be an efficient process to allow that person to challenge the use or output of the algorithm.
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.

learning have been made possible by vast troves of data. This data has to be collected and used, which means issues related to AI are closely intertwined with those that relate to privacy and data.

The nature of the data used also shapes the results of any decision or prediction made by an AI, opening the door to discrimination when inappropriate or inaccurate datasets are used. There are also key requirements of Australia's *Privacy Act* which will be difficult to navigate in the AI age[22].

Predictions about people have added ethical layers

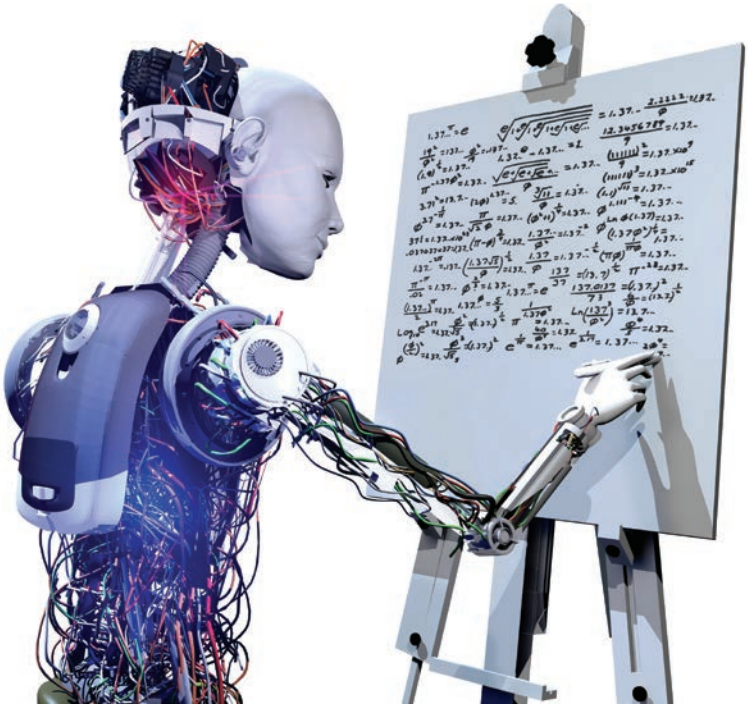
Around the world, AI is making all kinds of predictions about people, ranging from potential health issues through to the probability that they will end up reappearing in court[16]. When it comes to medicine, this can provide enormous benefits for healthcare. When it comes to human behaviour, however, it's a challenging philosophical question with a wide range of viewpoints[23].

There are benefits, to be sure, but risks as well in creating self-fulfilling prophecies[24]. The heart of big data is all about risk and probabilities, which humans struggle to accurately assess.

AI for a fairer go

Australia's colloquial motto is a "fair go" for all. Ensuring fairness across the many different groups in Australian society will be challenging, but this cuts right to the heart of ethical AI.

There are different ideas of what a "fair go" means. Algorithms can't necessarily treat every person exactly the same either; they should operate according to similar principles in similar situations. But while like goes with like, justice sometimes demands that different situations be treated differently. When developers need to codify fairness into AI algorithms, there are various challenges in managing often inevitable trade-offs and sometimes there's no "right" choice because what is considered optimal may be disputed.



When the stakes are high, it's imperative to have a human decision-maker accountable for automated decisions – Australian laws already mandate it to a degree in some circumstances[25].

Transparency is key, but not a panacea

Transparency and AI is a complex issue. The ultimate goal of transparency measures are to achieve accountability, but the inner workings of some AI technologies defy easy explanation. Even in these cases, it is still possible to keep the developers and users of algorithms accountable[26].

An analogy can be drawn with people: an explanation of brain chemistry when making a decision doesn't necessarily help you understand how that decision was made – an explanation of that person's priorities is much more helpful.

There are also complex issues relating to commercial secrecy as well as the fact that making the inner workings of AI open to the public would leave them susceptible to being gamed[26].

Black boxes pose risks

On the other hand, AI "black boxes" in which the inner workings of an AI are shrouded in secrecy are not acceptable when public interest is at stake. Pathways forward involve a variety of measures for different situations, ranging from explainable AI technologies[27], testing, regulation that requires transparency in the key priorities and fairness measures used in an AI system, through to measures enabling external review and monitoring[26].

People should always be aware when a decision that affects them has been made by an AI, as difficulties with automated decisions by government departments have already been before Australian courts[28].

Justifying decisions

The transparency debate is one component feeding into another debate: justifiability. Can the designers of a machine justify what their AI is doing? How do we know what it is doing?

An independent, normative framework can serve to inform the development of AI, as well as justify or revise the decisions made by AI. This document is part of that conversation.

Privacy measures need to keep up with new AI capabilities

For decades, society has had rules about how fingerprints are collected and used. With new AI-enabled facial recognition, gait and iris scanning technologies, biometric information goes well beyond fingerprints in many respects[29].

Incidents like the Cambridge Analytica scandal demonstrate how far-reaching privacy breaches can be in the modern age, and AI technologies have the potential to impact this in significant ways. We may need to further explore what privacy means in a digital world.

Keeping the bigger picture in focus

Discussions on the ethics of autonomous vehicles tend to focus on issues like the "trolley problem" where the vehicle is given a choice of who to save in a life-or-death situation. Swerve to the right and hit an elderly person, stay straight and hit a child, or swerve to the left and kill the passengers?

These are important questions worth examining[30], but if widespread adoption of autonomous vehicles can improve safety and cut down on the hundreds of lives lost on Australian roads every year, then there is a risk that lives could be lost if relatively far-fetched scenarios dominate the discussion and delay testing and implementation. The values programmed into autonomous vehicles are important, though they need to be considered alongside potential costs of inaction.

AI will reduce the need for some skills and increase the demand for others

Disruption in the job market is a constant. However, AI may fuel the pace of change. There will be challenges in ensuring equality of opportunity and inclusiveness[31].

An ethical approach to AI development requires helping people who are negatively impacted by automation transition their careers. This could involve training, reskilling and new career pathways. Improved information on risks and opportunities can help workers take proactive action. Incentives can be used to encourage the right type of training at the right times. Overall, acting early improves the chances of avoiding job loss or ongoing unemployment.

AI can help with intractable problems

Long-standing health and environmental issues are in need of novel solutions, and AI may be able to help.

Australia's vast natural environment is in need of new tools to aid in its preservation, some of which are already being implemented[32]. People with serious disabilities or health problems are able to participate more in society thanks to AI-enabled technologies[33].

International coordination is crucial

Developing standards for electrical and industrial products requires international coordination to make devices safe and functional across borders[34]. Many AI technologies used in Australia won't be made here. There are already plenty of off-the-shelf foreign AI products being used[35].

Regulations can induce foreign developers to work to Australian standards to a point, but there are limits. International coordination with partners overseas, including the International Standards Organisation (ISO),will be necessary to ensure AI products and software meet the required standards.

A TOOLKIT FOR ETHICAL AI

1. **Impact assessments:** Auditable assessments of the potential direct and indirect impacts of AI, which address the potential negative impacts on individuals, communities and groups, along with mitigation procedures.
2. **Internal or external review:** The use of specialised professionals or groups to review the AI and/or use of AI systems to ensure that they adhere to ethical principles and Australian policies and legislation.
3. **Risk assessments:** The use of risk assessments to classify the level of risk associated with the development and/or use of AI.
4. **Best practice guidelines:** The development of accessible cross industry best practice principles to help guide developers and AI users on gold standard practices.
5. **Industry standards:** The provision of educational guides, training programs and potentially certification to help implement ethical standards in AI use and development
6. **Collaboration:** Programs that promote and incentivise collaboration between industry and academia in the development of 'ethical by design' AI, along with demographic diversity in AI development.
7. **Mechanisms for monitoring and improvement:** Regular monitoring of AI for accuracy, fairness and suitability for the task at hand. This should also involve consideration of whether the original goals of the algorithm are still relevant.
8. **Recourse mechanisms:** Avenues for appeal when an automated decision or the use of an algorithm negatively affects a member of the public.
9. **Consultation:** The use of public or specialist consultation to give the opportunity for the ethical issues of an AI to be discussed by key stakeholders.

Implementing ethical AI

AI is a broad set of technologies with a range of legal and ethical implications. There is no one-size-fits all solution to these emerging issues.

There are, however, tools which can be used to assess risk and ensure compliance and oversight. The most appropriate tools can be selected for each individual circumstance.

Best practice based on ethical principles

The development of best practice guidelines can help industry and society achieve better outcomes. This requires the identification of values, ethical principles and concepts that can serve as their basis.

ABOUT THIS REPORT

This report covers civilian applications of AI. Military applications are out of scope. This report also acknowledges research into AI ethics occurring as part of a project by the Australian Human Rights Commission[36], as well as work being undertaken by the recently established Gradient Institute.

This work complements research being conducted by the Australian Council of Learned Academies (ACOLA) and builds upon the Robotics Roadmap for Australia by the Australian Centre for Robotic Vision.

From a research perspective, this framework sits alongside existing standards, such as the National Health and Medical Research Council (NHMRC) Australian Code for the Responsible Conduct of Research and the NHMRC's National Statement on Ethical Conduct in Human Research.

REFERENCES

1. Ethics Commission Federal Ministry of Transport and Digital Infrastructure. 2017. *Automated and connected driving. German Government*. Germany.
2. New York City Hall. 2018. *Mayor de Blasio announces first-in-nation Task Force to examine automated decision systems used by the city*. Mayor's Office.
3. House of Lords Select Committee on Artificial Intelligence. 2018. *AI in the UK: Ready, willing and able?* UK Parliament. United Kingdom.
4. European Commission. 2018. *Artificial Intelligence for Europe*.
5. Canadian Institute for Advanced Research. 2017. *Pan-Canadian artificial intelligence strategy*.
6. Australian Government. 2018. *2018-2019 Budget Overview*. Canberra.
7. National Institute for Transforming India. 2018. *National strategy for artificial intelligence*.
8. Villani C. 2018. *For a Meaningful Artificial Intelligence*.
9. Rosemain M and Rose M. *France to spend $1.8 billion on AI to compete with U.S., China*. Reuters. 29 March 2018.
10. Japanese Society for Artificial Intelligence. 2017. *The Japanese society for artificial intelligence ethical guidelines*. Japan.
11. European Commission. 2018. *High-level expert group on artificial intelligence*.
12. UK Parliament. 2018. *World first Centre for Data Ethics and Innovation: Government statement*.
13. Infocomm Media Development Authority. 2018. *Composition of the Advisory Council on the Ethical Use of Artificial Intelligence and Data*. Minister for Communications and Information. Singapore.
14. The State Council: The People's Republic of China. 2018. *Guidelines to ensure safe self-driving vehicle tests. The State Council: The People's Republic of China.*
15. Senate Community Affairs Committee Secretariat. 2017. *Design, scope, cost-benefit analysis, contracts awarded and implementation associated with the Better Management of the Social Welfare System initiative*. Parliament of Australia.
16. Angwin J, Larson J, Mattu S et al. 2016. Machine bias risk assessments in criminal sentencing. *ProPublica*.
17. Dieterich W, Mendoza C, Brennan T. 2016. COMPAS risk scales: Demonstrating accuracy and predictive parity. Northpointe Suite.
18. Angwin J, Larson J. 2016. ProPublica responds to company's critique of machine bias story. *ProPublica*.
19. Angwin J, Larson J. 2016. Bias in criminal risk scores is mathematically inevitable, researchers say. *ProPublica*.
20. Office of the Australian Information Commissioner. 2018. *Publication of Medicare Benefits Schedule and Pharmaceutical Benefits Schedule data: Commissioner initiated investigation report*. Australian Government.
21. Australian Government. *International human rights system*. Attorney-General's Department.
22. Australian Government. 1988. *Privacy Act 1988, No. 119, 1988 as amended*. Australian Government. Canberra.
23. Corbett-Davies S, Pierson E, Feller A et al. 2016. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post*.
24. Moses L B, Chan J. 2016. Algorithmic prediction in policing: assumptions, evaluation, and accountability. *Policing and Society*, 28(7): 806-822.
25. Australian Government. 1999. *Social Security (Administration) Act 1999: No 191, 1999*. Federal Register of Legislation.
26. Reisman D, Schultz J, Crawford K et al. 2018. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*.
27. Kleinman Z. *IBM launches tool aimed at detecting AI bias*.
28. Khadem N. Tax office computer says yes, Federal Court says no. *ABC*. 8 October 2018.
29. Human Rights Law Centre. 2018. *The dangers of unregulated biometric use: Submission to the Inquiry into the Identity-matching Services Bill 2018 and the Australian Passports Amendment (Identity-matching Services) Bill 2018*. Human Rights Law Centre. Australia.
30. Awad E, Dsouza S, Kim R et al. 2018. The Moral Machine experiment. *Nature*, 563(7729): 59-64.
31. Berg A, Buffie E, Zanna L-F. 2018. *Should we fear the robot revolution? (The correct answer is yes)*. International Monetary Fund.
32. CSIRO. 2015. *Robots to ResQu our rainforests*. CSIRO.
33. Ray S. 2018. Data guru living with ALS modernizes industries by typing with his eyes. *Microsoft News*.
34. International Electrotechnical Commission. *Functional Safety and IEC 61508*.
35. Redrup Y. 2018. Google to make AI accessible to all businesses with Cloud AutoML. *Australian Financial Review*.
36. Australian Human Rights Commission. 2018. *Human rights and technology issues paper*. Sydney.

Dawson, D, Schleiger, E, Horton, J, McLaughlin, J, Robinson, C, Quezada, G, Scowcroft, J, and Hajkowicz, S. *Artificial Intelligence: Australia's Ethics Framework*, Executive summary, pp. 4-9, Data61/CSIRO, Australia. Retrieved from http://consult.industry.gov.au on 9 July 2019.

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

EVERYONE HAS THEIR OWN IDEA ON THE ETHICAL USE OF AI, BUT CAN WE GET A GLOBAL CONSENSUS? BY MICHAEL GUIHOT

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 wellbeing
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?

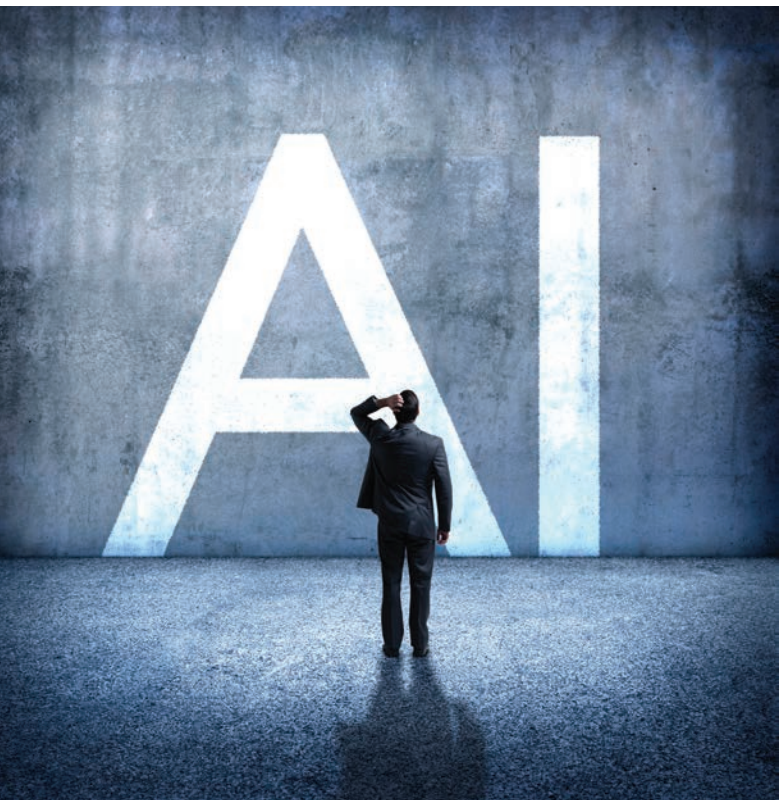
It is hard to make specific recommendations in relation to AI. That is partly because AI is not one thing with a single application that poses singular risks or threats.

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 ultim-



ate 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.

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.

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.

DISCLOSURE STATEMENT

Michael Guihot does not work for, consult, own shares in or receive funding from any company or organisation that would benefit from this article, and has disclosed no relevant affiliations beyond his academic appointment.

Michael Guihot is Senior Lecturer in Law, Queensland University of Technology.

THE CONVERSATION

Guihot, M (29 May 2019). *Will we ever agree to just one set of rules on the ethical development of artificial intelligence?* Retrieved from http://theconversation.com on 9 July 2019.

HOW DO WE MAKE ARTIFICIAL INTELLIGENCE MORE HUMANE?

How do we build trust in the development and use of exciting new technologies, while also addressing the possible threats to universal human rights?
By Edward Santow and Nicholas Davis

We've all had something like it happen: one minute you're searching for a present suitable for a two-year-old; the next, ads for nappies and prams are on every site you visit. It's unsettling. No one feels comfortable about bots following us surreptitiously as we roam around the web, when companies use what they learn from our online behaviour to promote products and services in creepy ways.

But could concerns around privacy and informed consent – though undeniably important – be distracting us from what we should be really worried about?

The exploitation of personal information for marketing purposes is a real problem. But the more serious risk is that our personal information can be used against us – not just to advertise a product we don't want, but to discriminate against us on the basis of our age, race, gender or some other characteristic we can't control.

PRECISION PREJUDICE

For example, if you have darker skin, facial-recognition technology is dramatically less accurate than if you have a light complexion. As this technology is progressively rolled out across law enforcement, in border security and even in delivering financial services, the risk that you'll be unfairly disadvantaged increases depending on your ethnicity.

Similarly, there are examples of artificial intelligence (AI) operating to prevent women or older people seeing certain online employment opportunities.

Not only does this violate the human rights of anyone negatively affected, but it also undermines community trust in AI more broadly. A collapse in community trust in AI would be disastrous, because AI has the potential to be an enormous boon – not just for national economies, but also in making communities more inclusive.

For every instance of AI causing harm, there's also an uplifting counter-example. This could be anything from AI-powered smartphone applications allowing blind people to "see" the world around them, to huge strides in precision medicine.

The challenge, therefore, is to build enduring trust in the development and use of a tremendously exciting set of technologies, so that citizens and organisations around the world can take advantage of the opportunities while addressing the threats to universal human rights.

This might sound eminently sensible to you. Unfortunately, this challenge is made harder by a damaging but pervasive myth.

RIGHTING THE WRONGS

Take Australia as an example. A common counterpoint to the idea that common sense norms and rules should apply to the development or implementation of AI relates to the fact that other countries are less likely to do the same.

AI, is developing at breakneck speed ... in almost every country around the world, laws are slow to adapt.

In which case, the argument goes, if Australia is to compete globally in developing AI products, Australian researchers and companies must not be fettered by human rights concerns, because other countries certainly aren't.

China, for example, is investing heavily in AI technology such as facial recognition to support its "social credit score" system, which involves conducting precise and determinative surveillance of its citizens. In the context of a global AI arms race, it is argued, Australia can't compete with one arm tied behind its back.

This argument is dangerous and misguided. Australia's liberal democratic values are one of its core strengths. The Australian Human Rights Commission's consultation on human rights and technology has shown that,

as Australians learn more about AI, there's a growing demand that AI only be used in ways that respect their human rights.

Consumers in liberal democracies want the benefits of AI, through self-driving cars, better healthcare and super-powerful computers. However, they won't accept a trade-off that involves mass surveillance, the exclusion of entire groups and a rise in discrimination.

This suggests that embedding human-rights protection in AI as it's developed isn't just morally right – it's also smart. If Australia can become known for developing AI that gets the balance right, it can gain a competitive advantage.

After all, consumers in liberal democracies want the benefits of AI, through self-driving cars, better health-care and super-powerful computers. However, they won't accept a trade-off that involves mass surveillance, the exclusion of entire groups and a rise in discrimination.

So, what's the solution? We know that technology, and especially AI, is developing at breakneck speed. We also know that, in almost every country around the world, laws are slow to adapt.

This puts greater pressure on institutions in countries such as Australia to smooth AI's rough edges in ways that allow us to harness the opportunities without allowing vulnerable members of our community to be crushed.

AI LEADERSHIP

Luckily, there is a way forward. And it might just be in Australia that real progress is made.

Several influential voices have already called for an Australian organisation to lead on AI. The World Economic Forum and the Australian Human Rights Commission have formed a partnership to consider this idea. These two bodies have invited leading decision-makers in government, industry and academia to meet at the University of Technology Sydney (UTS) to consider how we will tackle this AI leadership challenge.

Based on the consultation we have conducted to date, some of the key issues that should be considered include the following.

First, we should clearly articulate the values that should underpin AI. In Australia, these should be quintessentially Australian values such as equality or the fair go.

Second, there has been some support among stakeholders for a specialised organisation – either a new or existing one – to take a central role in assessing technologies and formulating laws, guidelines, accountability and capacity-building strategies in AI. This should be a national organisation with close connections with all stakeholders.

Third, this organisation should work closely with industry, government and the community to support the development of AI technologies that respect human rights.

The World Economic Forum and the Australian Human Rights Commission are consulting on these issues right now and have produced a white paper inviting comments, focused on the Australian context. But this is an issue facing all countries around the world, and we welcome your input in this process.

Edward Santow is a Human Rights Commissioner and leads the Australian Human Rights Commission's work on technology and human rights.

Nicholas Davis is Head of Society and Innovation, Member of Executive Committee, World Economic Forum.

Santow, E, and Davis, N (5 March 2019). *How do we make artificial intelligence more humane?* Retrieved from www.humanrights.gov.au on 9 July 2019.

PROTECTING HUMAN RIGHTS IN THE CONTEXT OF AI

A CHAPTER EXTRACT FROM A WHITE PAPER PRODUCED BY THE AUSTRALIAN HUMAN RIGHTS COMMISSION AND THE WORLD ECONOMIC FORUM

International and Australian human rights law requires that individuals be treated without discrimination.[17] Governments must uphold human rights, while businesses have a responsibility to respect human rights in all their operations.[18]

The most effective way of complying with these obligations in the context of AI is to ensure that it is designed and used responsibly, by protecting privacy, fairness, equality and other human rights.

This is, of course, not solely an Australian challenge. Public and private sector organisations globally are exploring ways to understand and manage the impact of bias in AI and ML (machine learning). The interconnected nature of the global economy, combined with the fact that such explorations in innovative governance are at an early stage in all jurisdictions, means that Australian institutions have an opportunity to lead in developing new structures, policies and relationships that can help address these important issues on behalf of all Australians.

Accordingly, a responsible innovation framework could accommodate imperatives that sometimes sit in tension, by anticipating and addressing the potential harms of AI, so that it can be deployed in a way that is safe and beneficial for Australia.

There are three primary reasons for Australia to be concerned about bias and discrimination in AI systems:

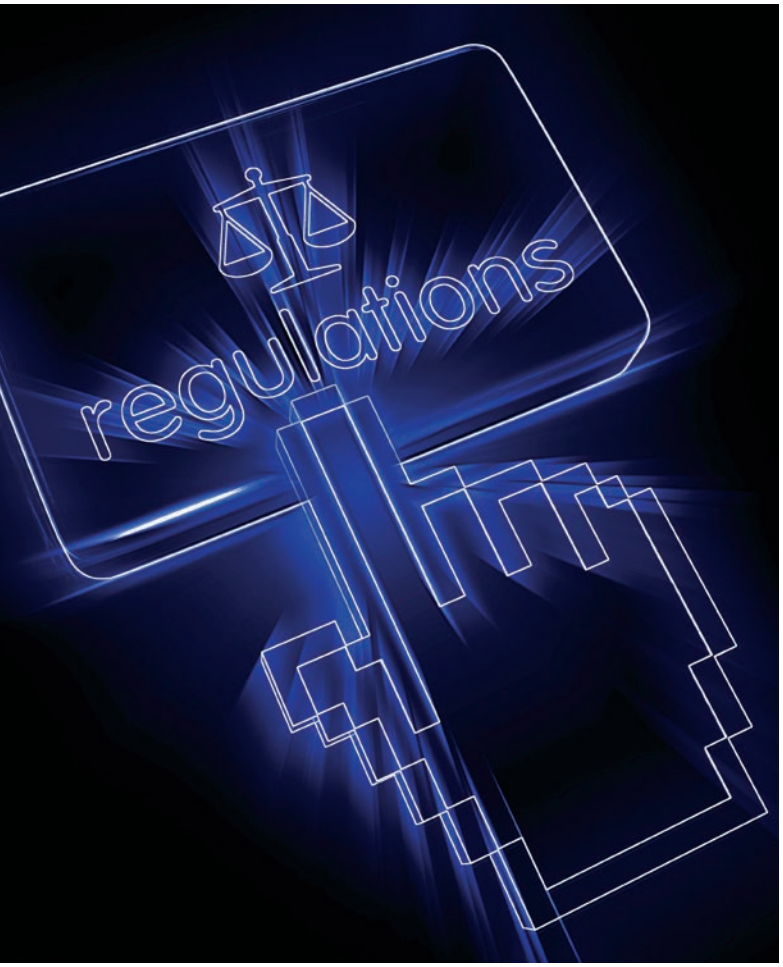
1. Automated decision-making systems will be applied more often by both the private and public sectors

Case study: Artificial intelligence and the risk of discrimination

Bias and discrimination in technology have entered the public consciousness along with our increasing reliance on and understanding of AI and ML. AI systems can discriminate and operate unfairly for many reasons.

For example:

- AI is designed by human beings who possess inherent biases and is often trained with data that reflects the imperfect world that we live in.[19]
- Training AI systems with data that is not representative, or using data that reflects bias or prejudice (for example, sexism or racism), can lead to an AI-supported decision that is unfair, unjust, unlawful or otherwise wrong.[20]
- AI's algorithms can include discriminatory variables (for example, including a variable for private school attendance in a loan application algorithm) that results in further discrimination.[21]
- Where users do not understand AI's limitations, especially if they assume AI's predictions to be more accurate and precise (and thus more authoritative) than those made by people, this can result in unfairness.[22]
- AI can be deployed in an inappropriate context (for example, deploying a model in a different cultural context from that in which it was originally trained).[23]
- Personal data is the 'fuel' for AI.[24] It can be at risk when deployed in ML models, as hackers can often threaten individual privacy by reverse-engineering algorithms, which could allow access to the personal data the algorithm is trained on.[25]



and at a greater scale across a wide variety of essential services, from decisions in health care to financial services. Discrimination in these decisions is both more likely and of greater consequence for groups that are already vulnerable.[26]

2. It is difficult to know the decision-making process adopted in an AI system, because ML tends to involve opaque proprietary algorithms.[27] Without understanding this process, it is hard to discern whether, when or how such systems are discriminating against a group or individual. This fundamentally challenges the concept of procedural fairness in administrative decision-making.
3. Hasty implementation of AI puts at risk its benefits by undermining public trust in new technologies. Public trust in Australian businesses, the government, media and civil society has fallen rapidly in the last decade to record lows.[28] If this trust is further eroded by the emergence of widespread discrimination through the deployment of AI systems, it may slow adoption in ways that prevent Australians from harnessing the many positive impacts of AI and ML.

The first set of consultation questions (*see box below*) focuses on understanding your sense of the challenge itself and the general approach that you feel the government should take in this area.

CONSULTATION QUESTIONS

1. What should be the main goals of government regulation in the area of artificial intelligence?
2. Considering how artificial intelligence is currently regulated and influenced in Australia:
   (a) What existing bodies play an important role in this area?
   (b) What are the gaps in the current regulatory system?

ENDNOTES

17. Australian Human Rights Commission, *Legislation*, www.humanrights.gov.au/our-work/legal/legislation.
18. Australian Human Rights Commission, *Business and Human Rights*, www.humanrights.gov.au/employers/business-and-human-rights.
19. Microsoft, *The Future Computed: Artificial Intelligence and its Role in Society* (Microsoft, 2018) 58.
20. *Ibid* 58-59.
21. Nicholas Davis, World Economic Forum, Ch. 7: The Future Relationship Between Technology and Inequality, in *How Unequal? Insights on Inequality* (Committee for Economic Development of Australia, April 2018).
22. Microsoft, *The Future Computed: Artificial Intelligence and its Role in Society* (Microsoft, 2018) 59.
23. Nicholas Davis, World Economic Forum, Ch. 7: The Future Relationship Between Technology and Inequality, in *How Unequal? Insights on Inequality* (Committee for Economic Development of Australia, April 2018).
24. Christian Ehl, 'Data – the Fuel for Artificial Intelligence' on *Medium* (14 January 2018), https://medium.com/@cehl/data-the-fuel-for-artificial-intelligence-ed90bf141372; see also Bernard Marr, 'Why AI Would Be Nothing Without Big Data,' *Forbes* (online), 9 June 2017, www.forbes.com/sites/bernardmarr/2017/06/09/why-ai-would-be-nothing-without-big-data/#6fe8a7a94f6d.
25. Justin Sherman, 'AI Innovation: Security and Privacy Challenges' on *Medium* (22 January 2018), www.medium.com/swlh/ai-innovation-security-and-privacy-challenges-84c0200b1bae.
26. Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018).
27. World Economic Forum, *How to Prevent Discriminatory Outcomes in Machine Learning* (2018), www3.weforum.org/docs/wef_40065_white_paper_how_to_prevent_discriminatory_outcomes_in_machine_learning.pdf.
28. Edelman, *Edelman Trust Barometer 2018*, www.slideshare.net/.

Australian Human Rights Commission and World Economic Forum (White paper January 2019). *Artificial Intelligence: governance and leadership*, Chapter 3, pp. 9-10. Retrieved from http://tech.humanrights.gov.au on 9 July 2019.