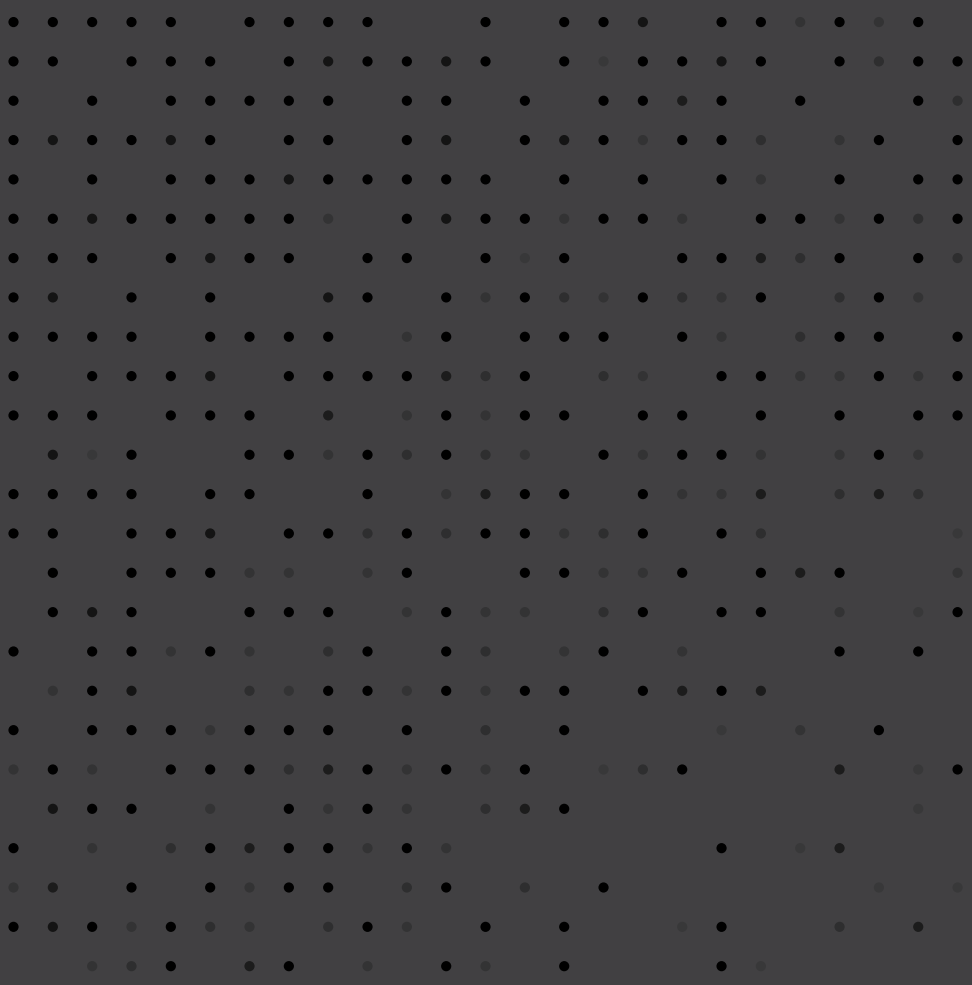
CLOSER TO THE MACHINE

Technical, social, and legal aspects of AI



OVIC

Authorised by the Victorian Information Commissioner

Published by the Office of the Victorian Information Commissioner
PO Box 24274
Melbourne, Victoria, 3001

t: 1300 006 842
e: enquiries@ovic.vic.gov.au
w: ovic.vic.gov.au

ISBN 978-0-6486723-0-2 Digital (online resource)
ISBN 978-0-6486723-1-9 Print (paperback)
ISBN 978-0-6486723-2-6 E-Book (electronic book text)

Date of publication
August 2019

Authors
Professor Toby Walsh - *Understanding AI*
Ms Katie Miller - *A matter of perspective: Discrimination, bias and inequality in AI*
Dr Jake Goldenfein - *Algorithmic transparency and decision-making accountability: Thoughts for buying machine learning algorithms*
Distinguished Professor Fang Chen and Dr Jianlong Zhou - *AI in the public interest*
Dr Richard Nock - *Algorithms, neural networks and other machine learning techniques*
Associate Professor Benjamin Rubinstein - *Data security and AI*
Emeritus Professor Margaret Jackson - *Regulating AI*

Editors
Cliff Bertram
Asher Gibson
Adriana Nugent

Design, layout & typography
Dean Bardell-Williams

FOREWORD

Artificial intelligence, or AI, has become a ubiquitous part of our lives. Hardly a day goes by without hearing or reading about AI and the impacts it is having on society.

Up until now, industry has led the charge in developing and implementing AI technologies to help achieve commercial goals. However, the public sector is increasingly turning to AI technologies to carry out its functions, develop and inform policy, and deliver services to its citizens.

How governments and regulators respond to technological and social developments in AI will have a large and lasting impact on our society. We need to encourage worthwhile technological innovation, but we need to do so with our eyes open. This requires us to be alert to the far-reaching effects AI can have. We all have a role to play in determining what the society in which we want to live looks like.

This is why we have developed this book. Its primary purpose is to increase the Victorian public sector's understanding of AI technologies that have the potential to impact our lives, and to assist those who implement AI systems to appreciate the technical, social, and legal aspects. But we also hope the book will be of interest to any member of the community who wishes to explore the ramifications of such a transformative technology and to participate in the debate around its adoption.

We must not be complacent about the potential effects of AI in public administration. As the Victorian public sector's uptake of AI technologies increases, we must design our systems in a way that is cognisant of the considerations discussed in this book.

In developing this publication, we drew on the extensive expertise of eight AI experts who authored the chapters of this book. I would like to express my thanks to each of them for their contributions, and for making these topics accessible, engaging and thought-provoking.

Sven Bluemmel
Information Commissioner
August 2019

CONTENTS

INTRODUCTION

There's an oft-quoted line from the Spielberg movie *Jurassic Park* in which Dr Ian Malcolm, the fictional character portrayed by actor Jeff Goldblum, says *"Your scientists were so preoccupied with whether or not they could, they didn't stop to think if they should".*

While the quote has been widely used in discussions of Artificial Intelligence (**AI**), it's less true of AI than it is of hypothetical re-imagined dinosaurs. AI offers potential benefits, as several of the chapters in this book illustrate. And it offers many benefits in the here and now, for example by way of the automatic braking systems we now take for granted in our cars, and the voice recognition used by tens of millions of people to conduct daily tasks. AI is here already. We can't stop to think about whether or not we should build it, but we might want to ask exactly what it is we're building, and think about how best to incorporate it into our lives.

If we're to avoid the unintended consequences of poorly planned or executed AI, we need to consider modifying the fictional Dr Malcolm's proposition:

> *"We should be concerned with whether or not we are solving the right problems, for the right people."*

This book contains seven chapters, exploring the technical, social and legal aspects of artificial intelligence. While each chapter looks at different aspects to be considered in developing and implementing AI, there are some common themes. One is that despite enormous progress in AI, we are only just beginning to see the potential benefits and problems it may bring. Another is that the decision-making and risk analysis frameworks for implementing AI into our modern lives may need some adjustment, but not at the expense of human rights and privacy.

UNDERSTANDING AI – provides a brief overview of the history of AI, from Aristotle, to Alan Turing, to what AI looks like today. It covers what AI is and is not, what it can and cannot do, and what it will and will not be able to do in the future.

A MATTER OF PERSPECTIVE: DISCRIMINATION, BIAS AND INEQUALITY IN AI – looks at discrimination, bias and inequality in AI, how these concepts are understood differently by law and technology, and how they might be addressed. The chapter also explores human rights and how they are impacted by AI.

ALGORITHMIC TRANSPARENCY AND DECISION-MAKING ACCOUNTABILITY: THOUGHTS FOR BUYING MACHINE LEARNING ALGORITHMS – focuses on how AI is increasingly used to make decisions, the ramifications that has on transparency and accountability, and how we could tackle those issues.

AI IN THE PUBLIC INTEREST – covers many of the ways in which AI can, and is, being used in the public interest, from making strategic decisions to maintaining the Sydney Harbour Bridge. It discusses how AI can both compromise and protect privacy, and highlights some of the barriers that AI will encounter in the future.

ALGORITHMS, NEURAL NETWORKS AND OTHER MACHINE LEARNING TECHNIQUES – dives into how AI actually works; how computers can perform tasks without being told how. It covers the different ways that computers have been 'taught' over the years, ramping up to the neural networks that power the modern world, and the problems that have come with them.

DATA SECURITY AND AI – takes a closer look at the security and privacy challenges that come with using AI outside of 'the lab', and the arms race of developing ways to use AI to protect and exploit personal information.

REGULATING AI – explores how AI is already being regulated, and how we could update old laws or create new ones to respond to AI.

KEY TERMS

Artificial intelligence, by its nature is a technical subject. The authors have strived to explain many of the terms and concepts in this book in a non-technical manner, however it is useful for you to have an overview of the key terms and concepts before you begin reading.

Artificial intelligence (AI)
The ability for a computer to do something that requires intelligence, such as learning or problem solving.

Machine learning (ML)
A subset of AI, machine learning is the ability for a computer to perform tasks without being given explicit instructions how, instead 'learning' how to preform those tasks by finding patterns and making inferences.

Black box
An AI system where the data inputted is known, and the decisions made from that data are known, but the way in which the data was used to make the decisions is not understood by humans.

Explainable AI (XAI) or ***white box*** or ***glass box***
An AI system where the way in which the system makes decisions is understood by humans.

Model
The 'intelligent' part of a machine learning system that learns how to perform tasks by making predictions or decisions.

Training
The process used to create a model.

Training data
The set of data used in the training process.

Instance or ***observation*** or ***sample*** or ***entity*** or ***case*** or ***record*** or ***pattern*** or ***row***
A single item in a larger set of data, for example, a single person in a spreadsheet of employees.

Feature* or *attribute
A property of a set of data. For example, the features of a spreadsheet of employees might be names, positions, salaries, phone numbers and addresses.

Classifier
A machine learning model that can classify information. For example, a classifier may sort pictures of bananas into groups of 'unripe', 'ripe' or 'overripe'.

Supervised learning
The process of training a model using training data that is labelled. For example, training a classifier to tell the difference between apples and oranges using training data made up of pictures labelled 'an apple' or 'an orange'.

Unsupervised learning
The process of training a model using training data that is unlabelled. For example, training an AI system to tell the difference between different kinds of vegetables using training data made up of unsorted and unlabelled pictures of vegetables.

Semi-supervised learning
The process of training a model where the training data is made up of both labelled and unlabelled data. Semi-supervised learning is often done by manually labelling a relatively small part of a large unlabelled data set.

Reinforcement learning
The process of training a model by using trial and error, where the system receives rewards for performing well and punishments for performing poorly.

Online machine learning
The process of performing machine learning where the learning is ongoing using real-time data, as opposed to being trained once with a fixed set of data.

Transfer learning
The ability to use an existing model for a new purpose, such as getting a chess playing AI to play checkers.

Neural network (NN) or ***artificial neural network (ANN)***
A common way to perform machine learning, neural networks are made up of sets of algorithms (called 'neurons'), each of which helps perform a very small part of a larger task. Neurons have connections (sometimes called 'edges') to other neurons with varying strengths (called 'weights').

Deep neural network (DNN)
A neural network that has its neurons organised into multiple 'layers', where the results of the first layer feed into the second layer and so on.

Deep learning (DL)
The process of performing machine learning using deep neural networks.

Backpropagation
A common way for neural networks to learn, especially when using supervised learning. After a neural network experiments with a new approach to performing a task, backpropagation is how the network evaluates how successful that approach was, and then adjusts the weights of connections throughout the network accordingly.

Convolutional neural network (CNN or ***ConvNet)***
A kind of deep neural network that looks at the spatial relationships of information, such as where pixels are located in a picture. Convolutional neural networks are commonly used for AI systems involving images, such as identifying objects that are in photographs.

Recurrent neural network (RNN)
A kind of deep neural network that looks at the sequence of information, such as the order of words in a sentence. Recurrent neural networks are commonly used for AI systems involving language, such as transcribing spoken words to text.

Generative adversarial network (GAN)
A kind of machine learning that pits two different neural networks against each other. For example, the first neural network (called a 'generative network') might try to create 'fake' pictures of human faces, while the second neural network (called a 'discriminative network') tries to guess if the faces are real or not. Backpropagation is done for both neural networks; the first network can learn how to make more convincing faces, and the second network can learn how to better spot artificial faces.

Autoencoder
A type of neural network that can learn using unsupervised learning. Autoencoders take an instance, break it down ('encode') into a representation of the features of the instance, then reconstruct ('decode') the instance using the representation. Autoencoders learn using backpropagation; they measure how 'successful' an attempt was by comparing the original instance to the reconstructed instance.

UNDERSTANDING AI

Toby Walsh

We all know what artificial intelligence (**AI**) is. Hollywood has given us plenty of examples. Artificial intelligence is the destructive T-800 robot played by Arnold Schwarzenegger in the *Terminator* movies. It is HAL 9000, the sentient and murdering talking computer in *2001: A Space Odyssey*. It is Ava, the female humanoid robot in *Ex Machina* that deceives humans to escape into the wild. And it is the Tyrell Corporation Nexus-6 replicant in *Blade Runner*, running away from would-be replicant Harrison Ford to save itself from being 'retired'.

In reality, artificial intelligence is none of these conscious robots. We cannot yet build machines that match the intelligence of a three year old. We can, however, program computers to do narrow focused tasks that require intelligence for humans to solve.

What is AI?

If artificial intelligence is not the stuff of Hollywood movies, what is it? Oddly enough, AI is already very much part of our lives. However, much of it is hidden from sight. For example, every time you ask Siri or Google Home a question, you are using a number of different types of artificial intelligence: the speech recognition software that converts your speech into a natural language question, the natural language processing algorithms that convert this question into a search query, the search algorithms that answer this query, and the ranking algorithms that predict the most 'useful' search results.

As a second example, if you're lucky enough to own a Tesla, you can sit in the driving seat whilst the car drives itself autonomously along the highway using a whole host of AI algorithms that sense the road and environment, plan a course of action, and drive the car to where you want to go. Artificial intelligence is also the machine learning algorithms predicting which criminals will reoffend, who will default on their loans, and shortlisting your CV for your next job.

For those of us working in the field, the fact that AI often falls out of sight in this way is gratifying evidence of its success. Ultimately, AI will be a pervasive and invaluable technology, like electricity, that invisibly permeates our lives. Almost every device in

our lives today uses electricity. It is an essential and largely unseen component of our homes, our cars, our farms, our factories, and our shops. It brings energy and data to almost everything we do. If electricity disappeared, the world would quickly grind to a halt. In a similar way, AI will shortly become an essential and mostly invisible component to our lives. It is already providing the smartness in our smart phones, but soon it will be powering the intelligence in our self-flying cars, smart cities, and intelligent factories.

Unlike many scientific endeavours, artificial intelligence has an official birth year. It started in 1956 when one of the founding fathers, John McCarthy proposed the name. He used the term to describe the topic of a famous meeting, the Dartmouth Conference held over the summer of 1956 that kicked off the field. There's arguably much wrong with the name that John McCarthy chose. 'Intelligence' is itself a poorly defined concept. And putting the adjective 'artificial' in front of anything opens you up to countless jokes about 'natural intelligence' and 'artificial stupidity'. But for better or worse, the name artificial intelligence has stuck.

The history of artificial intelligence goes much further back than 1956 when the name was coined. Indeed, it goes back before even the invention of the computer. Humans have been thinking about machines that might think, and how we might model thinking, for thousands of years. Like many stories, there is no clear beginning to humanity's quest to build machines that think. The story is, however, intimately connected to the story around the invention of logic.

One possible beginning is the 3rd century BC, when Aristotle founded the field of formal logic. Without logic, we would not have the modern digital computer. The computer is a practical implementation of logic. And logic has often and continues to be seen as a model for thinking. It is a means to make precise how we reason and form arguments. Moving forwards in time, the history of AI takes in many other great thinkers besides Aristotle, such as Ramon Llull, Gottfried Leibnitz, Charles Babbage, Ada Lovelace and George Boole, all of whom dreamed of mechanising thought.

One figure that stands out in the complex and surprisingly long history of artificial intelligence is the mathematician and code breaker, Alan Turing. Despite a tragic and early death in 1954, Turing played a pivotal role in the invention of the digital computer. He also wrote what is often considered the first scientific paper about artificial intelligence. In 1950, before the field had even been named, Turing wrote a paper titled 'Computing Machinery and Intelligence' for the journal MIND.[1] The paper asked the question of how we would know when AI had succeeded. When could we say that a machine thinks?

Turing's answer to this question is now known as the Turing Test. It is also called the Imitation Game, as it asks if a computer can imitate a human being. Turing suggested that if you interrogate a human and a computer remotely and cannot tell them apart then perhaps you might as well consider that the computer 'thinks' like a human. This functional test is reflected today in an equally functional definition of artificial intelligence: AI is getting computers to do tasks that, when humans do them, we think they require thinking. Driving a car. Translating English into German. Proving a mathematical theorem. Playing the ancient Chinese game of Go. Reading an x-ray. Diagnosing skin cancer. Composing a song. Painting some abstract art. Or coming up with a joke. These are all tasks we think require thinking. And it may surprise you to hear that computers can already do all of these tasks.

AI is many things

A common misconception is that AI is a single thing. Just like our intelligence is a collection of different skills, AI today is a collection of different technologies, such as machine learning, natural language processing, and speech recognition. As many of the recent advances in AI have been in the area of machine learning, artificial intelligence is often mistakenly conflated with machine learning. However, just as humans do more than learn how to solve tasks, artificial intelligence is more than just machine learning. In my 2017 book, *It's Alive!: Artificial Intelligence from the Logic Piano to Killer Robots*,[2] I introduce the four tribes of AI who are working on different aspects of building thinking machines: the learners, the reasoners, the roboticists and the linguists. Of course, the intellectual landscape of AI is much more complex than this quartet of tribes, but this decomposition is a good place to start in understanding what AI is.

The first tribe working on artificial intelligence is the tribe of *the learners*. The learners are interested in getting computers to learn to do intelligent tasks. Much of our human intelligence is learnt. We are born without language; without knowledge of what is good to eat; without an ability to walk, talk, or add up numbers; without knowledge of the sun and the moon; and without an understanding of Newton's laws of physics. But we learn all these things and more. One way, therefore, to build a thinking machine is to build a computer that can learn, just like humans do.

Giving computers the ability to learn also solves the problem of having to codify all of the knowledge we have acquired as we grow up; knowledge that is essential to operating in the real world. It is a long and painful task to itemise to a robot all the common sense knowledge it might need, such as the sky is blue, shadows are not objects, objects do not disappear when they go out of sight, and so on. Within the learners, a very successful group of late are those working with neural networks, and in particular those working with deep learning – neural networks with many layers.

This group borrows ideas from neuroscience to build learning mechanisms loosely modelled on those used in our brains. They construct 'neural networks' with thousands of abstract neurons, and millions of connections that are trained on examples of the concepts to be learnt. If you show the network thousands of images of cats and dogs, and adjust the weights in these connections, the network can be trained to distinguish between the two.

The second tribe working on building artificial intelligence is the tribe of *the reasoners*. This group explores how to equip machines with explicit rules of thought. Machines can reason over knowledge that either is explicitly encoded up front, or is learnt from interacting with the real world. Hence, the reasoners may depend on the tribe of the learners to prepare their way. Human reasoning is far more complex than the simple 0 and 1 logic of computers. We need to cope with incomplete knowledge, with inconsistent knowledge, with uncertainty, even with knowledge about knowledge. The reasoners therefore try to develop formal models of reasoning that can cope with partial information, with contradictory information, with probabilistic information, and with information about information itself (so called meta-information).

The third tribe working in artificial intelligence is the tribe of *the roboticists*. Human intelligence is a complex phenomenon. It arises in part from our interactions with the real world. The roboticists build machines that act in the real world, that can reason about their actions, and that can learn like we do from these interactions. The roboticists therefore also overlap with the tribes of the learners and the reasoners. Of course, robots need to sense the world in which they act, so a part of this tribe works on computer vision – giving computers the ability to perceive the state of the world. Vision not only helps us navigate in the real world but is an important part of our ability to learn about that world. Much of what we have learnt came from what we have seen.

The fourth tribe working on building a thinking machine is the tribe of *the linguists*. Language is an important part of human thought. For machines to think, they must therefore understand and manipulate natural language. The tribe of the linguists develop computer programs that can parse written text, that can understand and answer questions, and that can translate between two languages. We also use language in speech. Therefore, a part of this tribe also works on speech recognition – getting computers to understand audio input.

What AI can and can't do today

Artificial intelligence is almost certainly at the peak of inflated expectations in its hype cycle and will likely descend shortly into a trough of disillusionment as reality fails to match expectations. If you added up everything written in the newspapers about the

progress being made, or believed the more optimistic surveys, you might suspect that computers would shortly match humans in their intelligence. The reality is that whilst we have made good progress in getting machines to solve narrow problems, we have made almost no progress on building more general intelligence that can tackle a wide range of problems. AI systems are surprisingly brittle. If you change the problem, even slightly, even the smartest AI systems tend to break catastrophically.

A lot of the hype about artificial intelligence today is due to the remarkable progress being made in the area of deep learning, especially for perceptive tasks like seeing or hearing the world. For example, Baidu's Deep-Speech 2 system is now competitive with humans at transcribing speech into text. It would be, however, wrong to conclude that machine learning will solve AI, and that with a few more refinements, it will get us to human level intelligence.

One limitation of deep learning, and indeed of almost all machine learning techniques, is the amount of data that is needed. Often hundreds of thousands, or even millions, of training examples are needed to train a system to reach human level performance. Whilst many enterprises are collecting large sets of data, there are nevertheless many domains where data is hard to collect or is simply not available. In robotics, the laws of physics may limit how quickly we can collect data. The robot cannot move faster than its motors allow. We may also have to be careful not to break the robot. There are other domains where we simply cannot have a lot of data. We might want to predict success rates for heart-lung transplants, but the number of such operations worldwide is numbered in the hundreds. We cannot have thousands, let alone millions of training examples. Humans are, by comparison, very fast learners. For instance, we can learn from a single example, and can generalise easily to new situations.

There are several other limitations of deep learning. First, it is largely a black box and is unable to explain itself in very helpful ways. Second, it cannot guarantee certain behaviours. For example, we can break many computer vision systems by changing a single pixel.[3] Third, we often need to do more than make predictions. We might need to also make decisions based on these predictions. For example, given some predicted demands, these are the best products to manufacture in the next quarter. Or, given this predicted traffic, these are the best routes for our truck fleet to deliver to the shops tomorrow.

It would be impossible to discuss recent advances in artificial intelligence without mentioning the role that big data has played. Many enterprises are leveraging big data sets to build practical applications using machine learning. Banks are using big data and machine learning to detect credit card fraud. Online stores like Amazon are using big data and machine learning to tune their product recommendations. And scientists have identified promising new drugs using machine learning applied to large data sets.[4]

In general, machine learning helps us classify, cluster and make predictions about data. It is impossible to list all the applications, but I will mention a few to illustrate the breadth of the field. Machine learning is being successfully used to detect malware; to predict hospital admissions; to check legal contracts for errors; to prevent money laundering; to identify birds from their tweets; to predict gene function; to discover new materials; to mark essays; to identify the best crops to plant; and somewhat controversially, to predict crime and schedule police patrols. Indeed, it might be easier to list the areas where machine learning is *not* being used, except it is almost impossible to think of an area where machine learning isn't being used currently.

There are several areas in which machine learning techniques are challenged. One such area mentioned earlier is explanation. Unlike humans, many machine learning algorithms are unable to explain how or why they came up with their answers. Another area is in learning from limited amounts of data, as well as from noisy data. Machine learning has a long way to go to match human performance in such settings. A third challenge is learning across problems, also known as 'transfer learning'. Humans can apply their expertise in one domain to get up to speed quickly in another. Once you are good at playing squash, you will likely be reasonable at playing tennis. By comparison, machine learning algorithms have to start from scratch when dealing with new tasks.

Another area in which machine learning remains challenged is in what is called 'unsupervised learning'. Many of the recent advances in machine learning have been in supervised learning, where we have training data that is correctly labelled. The data is labelled with the concept to be learnt. This email is 'spam'. These emails are 'genuine'. This web traffic is 'suspicious'. These web queries are 'legitimate'. But in many application domains, we don't have such labelled data. When we learnt to see the world as babies, the world didn't helpfully come labelled 'book', 'table', 'chair', etc. We would like computers to learn from unlabelled data like we can.

Despite all these limitations, I am not concerned that the field will go through another 'AI Winter', as it did in the 1990s following the expert systems boom of the 1980s. Even if we made no more technical progress on building AI, which I doubt, we can now solve a wide range of useful problems. Just rolling out AI to new areas in which it has yet to be applied would be of considerable practical value. A report by PricewaterhouseCoopers in 2017 estimated that AI will add over $15 trillion to the world's gross domestic product (**GDP**) in inflation adjusted terms by 2030.[5] In some countries, like China, it might help grow GDP by a quarter. In Australia, it was estimated to add over 10% to our GDP. This is perhaps half of all the economic growth we can expect in the next decade.

What AI can and can't do tomorrow

In the future, will we run into limits that prevent us from building even smarter artificial intelligence?

There are many machines we would like to build that we will likely never engineer. For example, we're unlikely ever to have time machines to take us back to the past, or perpetual motion machines that run without limit. Perhaps artificial intelligence will never match human level intelligence. Many other scientific fields have discovered fundamental limits, both practical and theoretical. Indeed, computing is already running into quantum limits in trying to shrink transistors further and squeeze more onto silicon chips. Perhaps there are also practical and theoretical limits that will defeat the goal of building machines that match or even exceed human level intelligence.

One argument against AI, discussed by Alan Turing himself, is the argument of disability. This is the argument that computers may act somewhat intelligently but they will never do some particular activity. They will never be wrong. Or never fall in love. Or never learn from experiences. Or never invent a joke. Or never appreciate the music of Beethoven. The list goes on. Unfortunately, people rarely back up such arguments with any evidence that machines cannot do these tasks. It is merely that they have not seen a machine do this task yet.

Some of these arguments are very easy to dismiss. There are many documented cases of computers doing something new. Making a new type of opening move in Go. Writing a poem. Composing music in the style of Bach. The list goes on. There are also many examples of computers learning from experiences. AlphaGo learnt to play Go by playing against itself. Google Translate learnt how to translate sentences from thousands of UN transcripts. Computer vision systems have learnt to recognise skin cancer better than human eyes.

One of the most popular and important arguments against AI is one of the oldest and is due to Ada Lovelace, often considered to be the first computer programmer. Ada Lovelace worked alongside Charles Babbage in the 19th century when he was trying (and failing) to build a mechanical computer. Ada Lovelace suggested that computers only do what we know how to do and in particular cannot be creative. There are many responses to this objection. One is that computers have been creative many times already, writing poems, composing music, inventing new mathematics and painting paintings. Another response to Ada Lovelace's objection is humans are limited by the same deterministic laws as computers. Are we not merely biological machines? How then can we be creative if machines aren't? A third response to Ada Lovelace's objection is that AI systems often behave in ways that we do not expect. Perhaps creativity can be found within these unexpected moments.

Another limit besides creativity is that machines may never be conscious. In 1951, Geoffrey Jefferson eloquently put this argument forwards as follows:

> *Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain – that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants.*

Of course, consciousness is itself a difficult problem to explain in biological systems. Indeed, artificial intelligence may throw some light on human consciousness. It is not clear if computers will ever develop some sort of consciousness, or if it is a uniquely biological phenomenon. We might prefer that machines do not gain consciousness. For once machines are conscious, we may have ethical obligations towards how they are treated. Is it reasonable to turn them off? Do they suffer? In any case, since we understand so little today about consciousness, it is not at all clear that it is necessarily a limit on artificial intelligence.

We may also run into various tacit limits in trying to build more intelligent machines. A lot of the 'intelligent' activities we do are ones that we cannot explain to anyone else. Or even to ourselves. One example of tacit limits is facial recognition. You know your mother's face, and you can recognise it out of a million, or indeed a billion others. Yet you are not conscious about your knowledge of her face. You would probably struggle to describe the precise arrangement of her eyes, nose, and mouth. Instead, you recognise her face as a whole unconsciously. There are many other examples. Riding a bicycle. Shooting a hoop. Even deciding on a good move in the game of Go. We can read about them in books. But you have to do them, to learn them.

Easy tasks for humans are often hard to get computers to do. And vice versa. This is known as Moravec's Paradox, after a famous roboticist Hans Moravec who identified it in the 1980s. Other well-known AI researchers like Rodney Brooks and Marvin Minsky made similar observations around this time. Indeed, in 1994 the cognitive scientist Steven Pinker has claimed that this is one of the most important ideas discovered in AI research so far:

> *The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard. The mental abilities of a four-year-old that we take for granted – recognizing a face, lifting a pencil, walking across a room, answering a question – in fact solve some of the hardest engineering problems ever conceived. Do not be fooled by the assembly-line robots in the automobile*

commercials; all they do is weld and spray-paint, tasks that do not require these clumsy Mr. Magoos to see or hold or place anything. And if you want to stump an artificial intelligence system, ask it questions like, Which is bigger, Chicago or a breadbox? Do zebras wear underwear? Is the floor likely to rise up and bite you? If Susan goes to the store, does her head go with her? Most fears of automation are misplaced. As the new generation of intelligent devices appears, it will be the stock analysts and petrochemical engineers and parole board members who are in danger of being replaced by machines. The gardeners, receptionists, and cooks are secure in their jobs for decades to come.

Moravec's paradox means that many jobs will be difficult for AI to replace. Equally there will be many new jobs that AI creates. In addition, the jobs at risk are not perhaps those that you suspect. The construction worker is relatively safe thanks to Moravec's paradox. But he or she has become a cyborg of sorts, with fork lifts, cranes, drills and power tools that amplify their productivity manyfold. There is perhaps no real paradox to Moravec's paradox. Our brains encode billions of years of evolution. They have been fine-tuned over millions of years. It is the higher-level cognitive tasks like playing Go, reading X-rays, or rostering staff that are easier to get computers to do.

How long have we got?

In the view of many experts in AI, myself included, we still have a substantial way to go to build artificial intelligence that matches human intelligence. The AI we can build today solves narrow problems. Nothing matches the breadth and depth of abilities of humans. We can write a computer program to play Chess, but it would have no hope at a game of chance like Poker. And it certainly couldn't translate Chinese into English. Or identify pneumonia in chest x-rays. Humans have a remarkable ability to adapt to new circumstances. It is likely we will need significant technical breakthroughs to build AI with such general purpose capabilities.

In 2012, Vincent Muller and Nick Bostrom of the University of Oxford surveyed a number of AI researchers about when high-level machine intelligence would be achieved. In particular, they asked when we might build a machine that could carry out most jobs at least as well as an average human. As there is significant uncertainty when this might happen, they asked for an estimate of when this was 50 percent likely. The median of these estimates was the year 2040.

I conducted a more recent survey in January 2017. I asked over three hundred of my colleagues, researchers working in artificial intelligence, to give their best estimate of the time it would take to tackle the obstacles between us and high-level machine intelligence. And to put their answers in perspective, I also asked nearly five hundred

non-experts for their opinion. I was expecting there might be some mismatch between the predictions of the experts and the non-experts. I was right. The median prediction of the experts was 2062. This compares to a prediction by the non-experts of 2039, over two decades earlier. The non-experts were a little more optimistic than Ray Kurzweil, futurist and director of engineering at Google, who predicts computers passing humans around 2045.

Why are experts so much less optimistic than non-experts? One of the perception problems faced by AI is that people see systems playing complex games like Chess and Go and, reasoning that these games require lots of intelligence, imbue these systems with all the other intellectual abilities that we humans have. In the case of human Chess and Go players, this is a reasonable assumption. A good Go player is likely to be an intelligent person. But this is not the case with computers. A good Go program isn't necessarily able even to play Chess. And there is a very long distance to get from playing Go to doing many of the other tasks humans can do that require intelligence.

What you can take from these surveys is that we are still a considerable distance, scientifically, from building AI that matches human intelligence. It is not something that is likely to be achieved in the next decade. Equally, it is not something that many experts think will take a thousand years. It may take fifty to a hundred years, so it is entirely conceivable that it will happen in the lifetime of our children. And if we are lucky, it might even happen in our own lifetimes. It is therefore a reasonable moment to consider the impact that AI will have on our lives, and how best to prepare for its arrival. Equally, you don't have to lose sleep over the machines surpassing us in the very near future.

How does AI work?

As I mentioned earlier, machine learning has fuelled many of the recent spectacular advances seen in artificial intelligence. It powered Google's AlphaGo to beat the best Go players on the planet. It is also the secret sauce behind Google Translate. And machine learning is behind the success of many other AI programs that can beat humans at tasks, like diagnosing skin cancer or playing Poker.

One common reaction to the idea of machine learning is that computers only do what you program them to do. On a simple level, this is correct. Computers are entirely deterministic. They follow the instructions in their computer code. They do not deviate. They cannot deviate. But on a deeper level, computers can do things they weren't explicitly programmed to do. They can learn new programs. They can even be creative. Just like us, they learn to do new things from their experiences. AlphaGo wasn't programmed to play the ancient Chinese game of Go better than a world champion.

No one sat down and worked out how to program playing Go like an expert. AlphaGo learnt to play Go well by playing against itself millions of times.

The reason that it got better than humans was that it played more games of Go than a human could in a lifetime of playing Go. If you played Go for the whole of your life, from the moment you woke up every morning to the moment you fell asleep, you wouldn't have played even a fraction of the number of games of Go that AlphaGo did in order to beat the world champion. And in learning to play Go well, it even became a little creative. It played moves that Go masters never expected, opening up new possibilities in how Go is played.

The claim that computers cannot be creative is an often repeated but flawed argument against the possibility of artificial intelligence. AlphaGo isn't the only world champion that is a computer. Computers are now better than humans in a wide variety of games including Backgammon, Poker, Scrabble, and Chess. Whenever someone tells me that computers can only do what they have been programmed to do, I list all the games where computers are already world champions. In almost every case, these computer programs were programmed by players of intermediate ability, and the program became world champions by *learning* to play better than us.

How does AI break?

Humans are robust decision makers. We can adapt easily to new circumstances. And our performance degrades gracefully when the problem changes slightly. This is far from the case with artificial intelligence today. We can change a single pixel and a computer vision system will classify a cat as a dog. Or more worryingly, a stop sign as a go sign. Even the impressive Alpha Zero program which learnt to play world class Chess and Go in a matter of a few hours has no idea how to play a game of chance like Poker. And it certainly has no idea how to translate from English into German. Or to interpret an x-ray.

When I talk to people about artificial intelligence, they often focus on the word 'intelligence'. This is perhaps not surprising. Intelligence is what lets humans dominate the planet, for better or worse. And intelligence is what AI is trying to build. But I also remind people to think about the word 'artificial'. It might be a very different, a very *artificial*, intelligence to the natural intelligence that we have.

A good analogy is flight. Artificial flight that humans invented is quite different to the natural flight that evolution found. We came at the problem of flight from a completely different angle to nature. We use a fixed wing and a powerful engine. Nature uses a wing that flaps. Both natural and artificial flight depend on the same theory of

aerodynamics. But they are different solutions to the problem. In a similar way, artificial intelligence may look very different to human intelligence. For example, it currently breaks in quite different ways to natural intelligence.

One important feature of AI today is that it is much more statistical than human decision making. If you ask Google Translate to convert “He is a nurse” into Turkish, and then translate the result back into English, you get “She is a nurse”. On the other hand, if you ask Google Translate to convert “She is an engineer” into Turkish, and translate the result back into English, you get “He is an engineer”. Turkish is a gender-neutral language, so both he and she get translated into the same word. But when translating back into English, Google Translate has some old-fashioned prejudices about nurses and engineers. If we’re not careful, AI will perpetuate many of the biases like sexism and racism that we’ve been trying to overcome for decades.

The reason for this gender bias is that Google Translate, like many machine learning algorithms, is based on statistics. And these statistics are generated by training on a corpus of text containing such gender biases. They thus reflect a bias that exists today in written text. But it is a bias that most of us wouldn’t want baked into our society. And even though AI systems are only just starting to enter the mainstream, many other examples of algorithmic bias have already been identified.

Technology companies have amplified the problem. They have promoted a myth that algorithms don’t have the unconscious biases of humans. They have suggested that algorithms simply and blindly serve up the best result. This lie has let them avoid taking responsibility. Humans are, of course, terrible decision makers. Behavioural economics is full of examples of their biases, and evidence that people often behave irrationally. But we can build machines that are just as biased as humans if we are not careful. In fact, algorithms are in some ways more problematic than humans. Unlike humans, many algorithms are unable to explain how they make their decisions. By comparison, humans can be asked to explain why they made a particular decision. But with most AI today, we simply have to accept the answer it gives.

One reason algorithms make biased decisions is that they are trained on biased data. We can train a machine learning program to predict who to hire. However, this is not trained on data about which people are actually the best to hire. We don’t know who is best to hire. Some people were not hired so we don’t know how well they would have performed in a particular job. We only know how the people who were actually hired performed. The training data may therefore have racial, gender and other biases which are reflected in its predictions.

There are also examples where algorithms have been intentionally designed to be biased. In 2012, it was discovered that Orbitz was offering Mac users more expensive

hotels than those using Windows. In particular, they were more likely to offer an expensive room to a Mac user, and a cheap room to a Windows user. Orbitz defended themselves by arguing that they were serving the needs of their customers, as Mac users spend around 30% more per night than Windows users. Orbitz claimed not to offer the same room at different prices to different users, but such dynamic pricing is the logical next step. Dynamic pricing may not sound fair, but it is legal in most countries. By finding features like operating systems that expose our different sensitivity to price, online retailers are likely to increase their profits. AI therefore throws up important issues about equitability and fairness.

In addition to fairness, there is a vital need for transparency in AI systems. We want decisions taken by machines not just to be fair but to be seen to be fair. This is a major challenge for AI systems today. Popular approaches like deep learning cannot explain how they made decisions in any meaningful way. Their decisions are often the product of being trained on more data than a human could look at in a lifetime.

Humans, of course, are also not very transparent. We are very good at 'inventing' explanations for our decisions. But there is a fundamental difference between humans making decisions and computers making decisions. We can hold humans to account for their decisions. If my decision results in harm being caused, I will face the financial or even legal consequences. Computers cannot be held to account in a similar way. It is thus more important that machines be able to explain their decision making. Transparency will help to bring trust to systems. If a medical app recommends you need some dangerous treatment, most of us would prefer a transparent system that can explain what is wrong with us, and why this is the best course of action. Transparency will also help correct systems when they make mistakes. There are, of course, places where transparency might be a luxury. The control software to a nuclear reactor might not need to explain why it is shutting down. We might accept the inconvenience of losing power temporarily, compared to the risk of a melt down.

There are many other challenging problems in building AI systems that we can trust. It is unlikely that there is a simple or single solution to building trustworthy systems. However, we can learn from other areas. We literally trust doctors with our lives. We have built a medical system in which we can do so, safe in the knowledge that doctors who harm their patients will be struck off, and medicines that don't work will not be approved. We perhaps need similar systems in place to ensure we can trust AI. This may require governments to regulate, industry bodies to set standards, as well as citizens to be better educated. We don't expect consumers buying a new washing machine to know much about water conservation other than to look for the star rating. Perhaps we need similar mechanisms so consumers can trust AI?

A final challenge regarding AI systems is accountability. Machines have no sentience. They do not suffer. They cannot be punished. It seems unlikely then that we can hold them accountable in any meaningful way for their actions. For example, when an autonomous car kills an innocent person, it is very unclear who we can hold responsible. This is an especial challenge with autonomous weapons where the design is actually to kill. Large numbers of AI researchers have called to regulate such weapons given the significant legal, technical and moral problems of allowing machines to decide who to kill.

Conclusion

It should be clear by now that AI offers significant promise to transform our society. The potential benefits of AI cover almost every aspect of our lives, including agriculture, banking, construction, health care, housing, education, entertainment, finance, government, law, manufacturing, mining, retail and transportation. Indeed, it is hard to think of an area that it will not touch. And the benefits are not purely economic. Artificial intelligence also offers major opportunities to improve our societal and environmental well-being. It can, for example, be used to make buildings and transportation more efficient, helping us conserve the planet's limited resources, and tackle wicked problems facing the world, like climate change.

Alongside these benefits, artificial intelligence also presents significant potential risks, some of which are global. These risks include the displacement of jobs, an increase in inequality within and between countries, the transformation of war, the corrosion of political discourse, and the erosion of privacy and other human rights. Indeed, we can already see worrying trends in many of these areas. Further development of AI should therefore be directed to evolve society in a direction that improves prosperity, reduces inequity, improves political engagement, and enhances the rights of all citizens.

Given these opportunities and risks, we need to ensure that AI is developed for the common good and that no one is left behind. The protection of human rights and fairness must be built in from the start. This will ensure that AI benefits all parts of society. Meaningful dialogue between civil society, industry, academia and government will be needed to decide the kind of society we want for future generations. The public will need to be actively engaged in this dialogue as it is *their* future which is being decided.

One area of significant importance is inclusivity. Artificial intelligence offers many opportunities to make society more inclusive. Those with difficulties hearing can use AI to hear. Those with difficulties seeing can use AI to see. AI can also help those with learning difficulties to learn. AI thus offers the possibility to improve the lives of people with disabilities, as well as many groups experiencing disadvantage.

To ensure the benefits of AI are shared amongst all these parts of society, we will likely need regulation. This may involve legislation to ensure the technology is built with accessibility and equity at its centre. We may need to extend legal concepts such as liability to decisions made by AI, as well as develop ethical standards for AI systems. Regulatory systems developed to ensure the safe and inclusive deployment of AI must increase public trust in these technologies and limit adverse outcomes. We already see a growing mistrust in AI which threatens the ability of society to take advantage of these technologies. There is an especial need to consider the human rights implications of AI based technologies, especially regarding areas such as privacy, discrimination, bias and transparency.

In the last few years, many ethical frameworks have been proposed by government, industry and civil society to deal with the challenges that AI poses. However, it is also becoming clear that it is often basic human rights like privacy that are under threat. We may not need too many new laws but simply to apply more vigorously existing ones. We may also wish to follow other counties like the UK in having an independently led AI body that brings stakeholders together from government, academia and the public and private sectors. This body could provide a critical mass of skills and leadership to develop AI technologies ethically. If we get it right, we can hope that our grandchildren will inherit a better, fairer and more prosperous AI-powered world.

Biography

Toby Walsh is Scientia Professor of Artificial Intelligence at the University of New South Wales and Data61, guest professor at TU Berlin and adjunct professor at QUT. He was named by The Australian newspaper as one of the "rock stars" of Australia's digital revolution. He is a Fellow of the Australia Academy of Science and recipient of the NSW Premier's Prize for Excellence in Engineering and ICT. He appears regularly on TV and radio, and has authored two books on AI for a general audience, the most recent entitled 2062: The World that AI Made. His research is funded by the European Research Council under the Horizon 2020 Programme via ERC Advanced Grant AMPLify 670077.

A MATTER OF PERSPECTIVE:

Discrimination, bias and inequality in AI

Katie Miller

An Aboriginal woman is refused housing.
A man with a disability is refused a contract for a phone service.
A woman is denied employment as a pilot.

Each of these examples involve discrimination. But discrimination by whom? Can you distinguish the case where the discrimination is caused by a human and the case where the discrimination is caused by artificial intelligence (**AI**)?

This is the challenge presented in an age when some decisions are made by humans, some are made by AI, and some are made by a combination of AI and humans. For the person refused housing, a phone service or employment, the experience is the same – but the ability to understand what has happened and obtain a remedy may be very different if the discrimination is attributable to, or contributed by, an AI system.

The questions posed above are, of course, trick questions. Each of the examples given have resulted from discrimination by humans in the past.[6] Each of the examples given could, or already have, resulted from the use of AI systems. Each of the examples given can result from direct discrimination on the basis of race, gender or parental responsibilities, or indirectly through discrimination on the basis of criminal record, employment status or mental health status. One of the challenges presented by AI systems is that we increasingly do not know why decisions are made or how traditionally protected attributes factor into AI decisions, recommendations and advice.

If we are to preserve the policy intentions of our discrimination, equal opportunity and human rights laws, we need to understand how discrimination arises in AI systems, how design in AI systems can mitigate such discrimination, and whether our existing

laws are adequate to address discrimination in AI. This chapter endeavours to provide this understanding. In doing so, it focuses on narrow, but advanced, forms of artificial intelligence, such as natural language processing, facial recognition and cognitive neural networks.

Are we speaking the same language?

The challenges of discrimination, bias and equality in AI involves the intersection of multiple domains of law, sociology and technology, each with their own experts and language. In order to have a shared understanding of the issues and possible solutions, we must first ensure that we are speaking the same language. In particular, we need to know what we mean by 'discrimination' and 'bias'. While the same words may be used across domains, they can have different meanings and connotations within different domains.

Discrimination

In everyday speech, to 'discriminate' is to "note or observe a difference; distinguish",[7] and 'discrimination' is "the process of differentiating between persons or things possessing different properties".[8] Understood in this sense, an AI system is a discriminating machine. The ability to discriminate, quickly and over large data sets, is one of AI's greatest strengths and a large part of the reason for its adoption and incorporation into so much of our daily lives. For example, AI assistants such as Cortana, Siri and Alexa rely on natural language processing, speech recognition and deep learning algorithms that can differentiate between words (or the sounds we use to represent words) and the contexts in which they are used.[9]

In a legal sense, 'discrimination' involves treating, or proposing to treat, someone unfavourably because of a personal characteristic protected by law.[10] For example, refusing a person a job because of their gender, racial background, disability or sexual orientation constitutes an unlawful form of discrimination.

Considerations of discrimination in AI involve questions about the types of discrimination that are acceptable, desirable and intended. Discriminating between cancer cells may be acceptable, desirable and intended.[11] Discriminating against women may not be acceptable or desirable – and, depending on the design of the AI system, such discrimination may not be intended.[12] The challenge for designers, users and subjects of AI is that cancer cells are clearly something we want to discriminate against – they are objectively 'bad'. People are more complex – and so too the questions of when it is acceptable or desirable to discriminate against them.

The legal concepts of direct and indirect discrimination are both important and helpful to the discussion about discrimination in AI systems. Understanding these concepts can assist in designing AI systems that are lawful because they comply with discrimination laws. More fundamentally, they are helpful in ascertaining if the discrimination undertaken by an AI system is acceptable or desirable.

Direct discrimination occurs when a person is treated less favourably *because of* an attribute that is protected by law, such as race, gender, religious belief or (dis) ability. Laws prohibiting direct discrimination are based on the idea that a person's protected attribute must be an irrelevant consideration when dealing with that person.[13] Discrimination is prohibited not just on the actual protected attribute, such as a person's gender, but also characteristics that are stereotypically attributed to persons of the protected group. Such imputed characteristics can include the susceptibility of married women to the influence of their spouses, and the clothing and grooming preferences of persons of particular sexual orientations.[14]

Indirect discrimination is directed towards activities that are "fair in form but discriminatory in outcome".[15] It requires consideration of how an ostensibly neutral action affects people with one or more protected attributes. For example, preferring to employ people who can attend work at 8am may indirectly discriminate against parents who have child care responsibilities for young or school-aged children. Indirect discrimination is not automatically unlawful; it requires consideration of the reasonableness of that requirement.[16]

This distinction between direct and indirect discrimination finds an analogy in algorithms and mathematics, where a distinction is made between direct and indirect variables. Direct variables are specific characteristics that the algorithm is programmed to recognise and consider; indirect variables, or 'proxies', are statistical correlations between one attribute, such as a postcode, and another attribute, which may or may not be protected, such as race or social class.[17]

If and when an AI system is challenged for breaching discrimination laws, there will be complex and novel arguments about whether an AI system is engaged in direct discrimination and whether it can be said that an AI system involves a "requirement condition or practice" constituting indirect discrimination.[18] This chapter does not engage with these complex arguments, which will no doubt depend on the particular circumstances of the AI system and discriminatory effect alleged. Instead, it is sufficient for now to distinguish between the AI system coded to rely or use directly protected attributes such as gender, race, (dis)ability; and the ostensibly neutral AI systems that operate in a discriminatory way.

Bias

Understanding the meaning of 'bias' is arguably more straight forward. While used in different ways between law and technology, the mechanism is generally accepted across domains. 'Bias' refers to a predisposition, prejudgment or distortion. In law, this often refers to a prejudice, inclination or prejudgment of a question.[19] In technology and mathematics, 'bias' may refer to a "systemic distortion of a statistical result due to a factor not allowed for in its derivation".[20] Across the domains of law and technology, bias implies that some parts of the picture are being preferenced, and others ignored.

Bias is generally recognised as a problem to be managed and something that can affect the integrity and quality of the final result – whether it be a decision by a government official or the ability of an AI system to recognise and match a face accurately. We strive for unbiased AI systems because we implicitly understand and accept that a biased decision is less desirable than an unbiased decision.

Yet there is a tension in our desire for unbiased AI systems, because every AI system has some inherent bias. Any AI system is limited, in the sense that it is merely a model or representation of a real-world situation.[21] In designing and implementing the model, choices are made about what to include or exclude. Just as we are becoming increasingly aware of the inherent, unconscious biases that all humans have,[22] so too must we be open to the presence of inherent bias in AI systems and models.

Equality

Related notions of 'fairness' and 'equality' are much more complicated – and always have been. There are different formulations and understandings of both 'fairness' and 'equality' across societies, cultures, and socioeconomic divides.[23] Our understanding of what is fair or equal can change depending on whether we consider it from our position as an individual or as between groups; and when we consider the extent to which we can control or influence particular outcomes.[24]

Both law and technology offer responses to the philosophical questions of 'what is fair?' and 'what is equal?', which are informed by and applied within their respective domains. For example, in administrative law, a distinction is drawn between substantive fairness – the fairness of a decision or outcome, and procedural fairness – the fairness of the manner in which a decision is made. The latter attracts remedies in administrative law, whereas the former does not.[25] In discrimination law, a distinction is drawn between equality of outcome and equality of opportunity; some advocate that laws should be directed towards equality of outcome whereas others argue that equality of opportunity is sufficient.[26]

Within the domain of artificial intelligence, 'fairness' can refer to notions of parity between data sets, classifiers and outcomes.[27] Each of these definitions is workable within the respective domain. However, they necessarily represent particular perspectives, which ignore many facets of fairness and equality that may be provided by other perspectives, such as cultural or philosophical perspectives. The difference in language, which in turn is based on a difference of conceptual understanding, makes any discussion about whether an AI system is 'fair' or produces 'equal' results challenging. For present purposes, it is sufficient to note that real differences in language exist, and encourage AI users and designers to be transparent about how they define fairness and equality when using such terms.

The promise of AI

AI is not the only machine that can discriminate. The human brain has also evolved to be a highly effective discriminating machine, filtering irrelevant information and creating mental heuristics to discriminate quickly between friend and foe; in and out groups. While useful to our early survival in ensuring we could quickly identify a threat and respond to it, these heuristics also inform the stereotypes and quick judgments that underpin or lead to discrimination.[28]

We are increasingly aware of the role that unconscious bias and human fallibility play in human decision making and discrimination. Humans segregate people on the basis of medical conditions,[29] refuse to make adjustments for people who use different languages,[30] and treat women employees unfairly when they are pregnant.[31] Humans are less inclined to grant parole immediately before lunch than after,[32] are more likely to perceive male candidates as competent and worthy of higher salaries,[33] and more likely to prefer names that don't "sound foreign".[34] We suffer from the halo effect and confirmation bias – and can become more discriminatory, the more objective we think we are.[35]

Mental heuristics assist us in rendering complex information simple, distinguishing between relevant and irrelevant information and to make decisions quickly. Yet these are also the conditions in which an AI system will excel. An AI system can process large amounts of data quickly, differentiating between the relevant and irrelevant and rendering the complex simple.

AI systems may therefore present opportunities to replace or support fallible human decision makers with objective, rational technology systems, thereby reducing the risk of discrimination. AI systems have been used to increase diversity in employment practices,[36] select members for company boards,[37] as well as increase access to financial services for traditionally under-represented consumer cohorts.[38]

Replacing humans with AI systems enables data to be sorted and differentiated at high speed and with access to a greater range of data and evidence than humans could process efficiently. AI can be programmed to exclude certain matters from the data set or algorithm more cheaply and effectively than training humans not to have regard to such matters or enforcing laws that prohibit them from doing so. AI does not get tired, hungry or distracted by the challenges of their personal lives.

Even where humans remain in control, AI can inform, support and assist our decisions and functions. AI provides insights and identifies patterns that humans could not effectively or efficiently do alone. AI replaces the gut instincts and reliance on personal experience that often inform human efforts to predict risk.

AI can also promote less discriminatory policy by analysing big data to identify trends, norms and outliers in our social practices and policies. Regulators could use AI systems to understand and monitor how different groups within society are treated in the provision of goods and services and to identify outliers who may be discriminating directly or indirectly. For example, analysing payroll tax, employment data, and data about maternity leave payments, could reveal insights about which employers have employees end their employment in close proximity to pregnancy or birth, which may in turn invite consideration of whether the employer's policies and practices are directly or indirectly discriminating against women on the basis of pregnancy or status as a parent.

One of the harms inherent with discrimination is that it treats an individual according to the characteristics of the group to which they belong, or are assumed to belong, rather than treating the individual on their own merits.[39] For decades, both policy makers and businesses have sought to decrease their reliance on broad-brush generalisations and potentially discriminatory stereotypes by increasing their understanding of individuals within groups through surveys, statistics and customer research.[40] AI systems carry the allure and promise of perfecting targeting and matching, thereby allowing governments and businesses to meet the individual's personal needs – whether it be in the context of social services or music preferences.

Improved targeting of individuals could result in people who are eligible for social security payments being identified and paid their entitlements – "no more and no less".[41] It can also improve risk-based approaches to the use of state power, reducing the bureaucratic burden on those who are not a risk while focusing resources on those identified as a risk.[42] This can reduce the time the average traveller spends at an airport by focusing on persons identified as being a higher risk of infringing immigration requirements.[43] Implemented well, such systems carry the promise of reducing friction – for both government and citizen – in the administration of government schemes and functions.

The dark side of AI

AI carries the promise of decreased discrimination and enhanced efficiency. But is that promise always realised?

The Australian Human Rights Commission and World Economic Forum have identified several ways in which AI is susceptible to discriminating and operating unfairly:[44]

- AI is designed by human beings who possess inherent biases and is often trained with data that reflects the imperfect world in which we live.
- 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.
- 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.
- 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.
- 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.
- Personal data is the 'fuel' for AI. It can be at risk when deployed in machine learning 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.

As AI systems become more prevalent around the world, more examples of discrimination in AI are being discovered. Some of these are outlined below.

- Amazon's experimental hiring tool, which used AI to review resumes and give a job applicant's resume a score from one to five stars. The experiment was discontinued after Amazon realised the tool was biased towards men. The data used to train the AI tool was 10 years' worth of resumes submitted to Amazon, most of which came from men because men still represent the vast majority of employees in the technology industry.[45]
- Centrelink's 'Robodebt' algorithm is more likely to raise a disputed or unfair debt for a person with inconsistent income and work hours because it relied on averaging annual income reported to the Australian Taxation Office across 26 fortnights and using that average as evidence of an overpayment in a given fortnight.[46]

- Companies advertising housing on Facebook were permitted to exclude persons of particular races from seeing the advertisements.[47]
- A US Immigration algorithm used to assist immigration officials decide whether to detain or release a person pending deportation was modified to remove the system recommendation of 'release' – resulting in the only possible answer being 'detain'.[48]
- Online targeted advertising through popular search engines and email services can change depending on whether the names used in association with those searches and email services are associated with particular racial backgrounds.[49]
- AI systems used to predict a person's risk of recidivism underestimates the recidivism risk of white people, while overestimating the recidivism risk of black people – and ultimately are no more accurate than random human decision makers.[50]
- Voice recognition software appears to respond more accurately to male voices than female voices, as well as people with certain accents and first languages.[51]

We readily recognise discrimination in an AI system when we can identify that a person has been unfairly denied a service, opportunity or resource or, alternatively, has been unfairly targeted for scrutiny, investigation or suspicion. This form of harm is described as 'allocative harm'.[52] It is the type of harm that most aligns with our understanding of direct discrimination – a person denied or targeted for something based on an attribute or characteristic that is unrelated to the outcome. The bail and sentencing algorithms referenced above are recognised as being unfair and discriminatory because the outcomes produced are more influenced by a person's racial background than the crimes they have committed or are alleged to have committed. The Amazon resume algorithm is recognised as unfair because it gives too much weight to being male without any evidence that men develop better software than women.

Discrimination in AI systems can also produce an additional, systematic harm, known as 'representational harm', which involves the reproducing and application of harmful stereotypes.[53] Targeted online advertising algorithms are harmful not just because they result in people of colour being targeted for some services – such as exploitative loan and debt recovery services – or missing opportunities for other services – such as housing or certain types of jobs – but because they reinforce stereotypes about people of colour. The AI systems in turn absorb the lessons of these reinforced stereotypes, producing even more discrimination in the AI system. This phenomenon was seen in Microsoft's chatbot, Tay, which, without the intention of its designers, learnt to be racist. Designed to learn from the tweets it was sent, Tay was vulnerable to learning (and did learn) to be racist after it was sent tweets containing intentionally racist and offensive content.[54]

The interaction between human and AI systems can contribute to representational harms becoming allocative harms by converting, over time, an ostensibly neutral and objective AI system into one with discriminatory effect. For example, an AI system may identify (or 'flag') a child at risk of abuse using non-discriminatory characteristics. A human acts on this flag to conduct an investigation into the circumstances of the child and their family. The investigation produces more data about the family, which is fed back into the AI system – enhancing the specificity of the algorithm in respect of the investigated family and families with similar characteristics. As a result, more children in similar situations are flagged – not necessarily because they are more at risk of abuse, but because there is more data about children like them in the AI system. The effect is discriminatory, without the design necessarily being so.[55]

Like all discrimination, representational harm is not the sole province of AI. Representational harm can also be seen in the discriminatory practices of humans and cultures of "deep-seated, pervasive prejudice that lingers".[56] Discrimination laws have traditionally not addressed or dealt with representational harm, preferring instead to focus on individual acts, harms and "overt, explicit and formal inequality".[57] This preference for the overt and tangible discrimination is reflected in discussions about reducing discrimination in AI, which focus on the reduction of allocative harm, rather than representational harm.[58]

Detecting discrimination in AI systems

Identifying that discrimination has occurred has always been a difficult task. In the words of Justice Kirby of the High Court of Australia, "human motivation is complex", "[d]iscriminatory conduct can rarely be ascribed to a single 'reason' or 'ground' and much discrimination occurs unconsciously, thoughtlessly or ignorantly".[59] A person alleging discrimination is at a disadvantage, because the information relevant to whether discrimination has occurred or not is held by the alleged discriminator – who may be under no obligation to explain their decision and may not even be aware of the full reasons for their decision.[60]

These problems are present and compounded in the case of AI systems. AI systems allow fragmentation of responsibility for any particular decision or action. The person designing the AI system is likely to be different to the person using the AI system. When the user of the AI system is asked for an explanation, the most common response will be "because the computer said so".[61]

Furthermore, technologists have focused much of their developments and research efforts on refining the outputs of AI systems, rather than explaining why those outputs are produced. The problem of 'explainability' has attracted more research attention in recent years, but researchers are playing catch up.[62]

For a person who has been the subject of AI discrimination, the following barriers exist to understanding what has occurred:

1. The individual affected may not realise that an AI system has been used in making the decision or taking the action that affects them.
2. The user of the AI system may not be obliged to provide an explanation. This is generally the case where the decision or action is taken in a commercial setting. While there is a limited obligation to provide reasons in respect of some government decisions,[i] that obligation may not extend to assistance provided by an AI system, for example, where a risk assessment is provided to a human who ultimately makes the decision.
3. The AI system may be a 'black box', in that it may not be capable of producing an explanation. Increasingly advanced forms of artificial neural networks are producing outcomes based on correlations and patterns that are unseen to the human eye. From the perspective of data analytics, this is one of the key strengths of advanced neural networks. From an explainability perspective, it is a significant barrier.
4. The designer of the AI system may resist disclosing the AI system's reasoning process in order to maintain commercial and competitive advantages and secrecy. While many AI systems are considered 'black boxes', so too are their creators.[63]
5. While the AI system may be able to produce an audit trail of the factors considered, such a trail may not extend to identifying why those factors have been marked as relevant to the decision or recommendation made by the AI system.[64] Given that the power of an AI system is to identify and learn patterns drawn from large datasets over time, the answer to a question about the relevance of a particular feature may lie in millions of algorithmic cycles.

Technology can assist in helping disparate individuals understand that they share the experience of adverse action from an AI system. Social media in particular is an effective way of individuals sharing their experiences and 'connecting the dots' to understand that they have been subject to an algorithm.[65] For example, our collective understanding of the technology system utilised by Centrelink to automatically raise debts (colloquially described as 'Robodebt') owes, in large part, to social media. The issue first came to mainstream attention after affected individuals posted on social media that they had received debt notices they did not understand. Social media helped similarly affected individuals to realise that they were subject to an automated

[i] For example, s 8 of the *Administrative Law Act 1978* (Vic) obliges some public officials to provide a statement of reasons in respect of certain decisions.

system, rather than traditional, individualised human decision making. Social media facilitated the spread of understanding of how the system operated and tools and techniques to challenge it. Social media connected affected individuals with experts and advisors and created a community of sufficient size to attract attention from the media, politicians and oversight bodies.

Acting on this shared experience can be difficult due to the problem of explainability. In Europe, Article 22 of the General Data Protection Regulation attempts to remedy this by introducing rights in respect of automated decision making, requiring safeguards that may include a right to obtain an explanation of a decision reached and to receive meaningful information about the logic involved in automated decision-making.[66] Such protections have not been introduced into Australian law.

Addressing discrimination

Although some anti-discrimination bodies have the power to consider systemic discrimination, Australia's discrimination laws are still based on a model of individuals making complaints about specific incidents of discrimination.[67] Given the barriers to individuals identifying that they have been discriminated against by an AI system, or understanding how that discrimination has occurred, the model of individual complaint is ill-equipped to effectively address discrimination in AI systems.

Technologists are aware of the problems posed by discrimination, bias and inequality in AI systems. The pursuit of 'fairness' in AI systems is increasingly a field of research. Researchers are posing different algorithms, mathematical techniques and definitions of 'fairness' to counter biases in datasets or discriminatory outcomes.[68]

It is unlikely that the problem of discrimination in AI systems will be solved by mathematics alone. Humans remain the designers, trainers and operators of AI systems – they are made in our image and reflect our own imperfect biases and prejudices. For mathematical solutions to fairness to be effective, they must be developed within "a framework that accounts for social and political contexts and histories". Without such a framework, mathematical 'solutions' may "serve to paper over deeper problems in ways that ultimately increase harm or ignore justice".[69]

The benefits of AI systems – such as increased efficiency and insights – largely accrue to the operators of AI systems who implement them within existing business or government practices. Just as a company or government agency would be responsible for any discriminatory actions of their staff, policies or procedures, so too should they be responsible for any discriminatory actions or decisions made by the AI systems they implement. This imperative is made stronger by the increasing understanding of the risks of discrimination in AI systems. If a body knowingly imports such risk into their

services and practices, it is not fair to then outsource the costs of those risks occurring to customers, citizens and the public.

For Victorian public authorities, the obligation to consider the potentially discriminatory effects of an AI system is even clearer. *The Charter of Human Rights and Responsibilities Act 2006* (Vic) (**Charter**) obliges public authorities to consider and act compatibly with human rights. One of those rights is the right to recognition and equality before the law, which includes the right to equal and effective protection against discrimination.

Operators of AI systems should review and monitor their AI systems for indicators of discrimination, bias and inequality. This requires consideration across the "full stack supply chain" of an AI system, encompassing the "origins and use of training data, test data, models, application program interfaces, and other infrastructural components over a product life cycle".[70]

In particular, an operator of an AI system should:

- ask the vendor of the AI system about how it has been programmed and trained to counter potentially discriminatory actions;
- consider the effects of choices made in designing the model expressed in the AI system;
- consider and test for potential discrimination embedded in any training data;
- supervise machine learning to detect early if the AI system is learning to be discriminatory in process or outcome; and
- regularly test the outcomes of AI systems to identify if they are producing unequal results that may reflect discrimination and/or bias.

Discovering that your AI system is discriminatory after it has been implemented costs time, money, and public trust and confidence. Remedying or retraining a discriminatory AI system is rarely a matter of 'tweaking' and often requires abandoning an existing system and developing a new one.[71] Taking a proactive approach by involving affected communities can assist designers of AI systems to better understand the problem the AI system is trying to solve, the data on which the AI system is to be trained, and the effects of the AI system.[72]

Calibrating AI with human rights

Just as a mathematical understanding of fairness is inadequate on its own to address discrimination in AI, the lens of discrimination alone is inadequate to understanding equality.

Laws against discrimination ensure that people are not treated unequally on the basis of protected attributes. However, laws prohibiting discrimination alone are insufficient to ensure that everyone is treated equally. As discussed above, concepts like 'equality' and 'fairness' are understood differently depending on our social, cultural, philosophical and ideological standpoints. In a liberal democracy, we accept some forms of inequality, such as those based on income, while prohibiting others, such as those based on gender.[73] Even when discrimination is prohibited, inequality can remain due to the complex interaction of social, economic and legal practices. Pay discrimination on the basis of gender has been illegal since the 1970s, yet a gender pay gap remains.[74] Understanding 'equality' requires us to understand how much and which types of inequality we accept and are prepared to justify.

Within the legal context, attempts have been made to reframe the contest about what constitutes 'equality', by understanding equality as fundamental to human dignity and protected to a minimum standard, when rights and freedoms that are recognised as common to all humans and integral to human dignity, are respected. In Victoria, those rights and freedoms find expression in the Charter.[ii]

AI systems engage human rights directly and indirectly; that is, both through the direct operation of AI systems, and indirectly by affecting the ability and confidence of persons to exercise their human rights.

The right to equality before the law

Given the preceding discussion about discrimination in AI systems, it is clear that AI systems engage the right to recognition and equality before the law, which expressly includes the right to protection from discrimination.

The right to equality before the law "prohibits treatment based on distinctions between persons which are arbitrary, in the sense of lacking objective justification, in the application and administration of the law".[75] In developing AI systems to apply and administer the law, there is a real question about whether distinctions drawn by AI based on group characteristics, correlations and imperfect data can be considered as having 'objective justification' and therefore not arbitrary.

The right to equal protection of the law without discrimination, and equal and effective protection against discrimination, requires equality in both the content and outcome of the law. It is not sufficient to ensure that the processes or opportunities are

[ii] Similar Acts are found in the ACT and Queensland: see the *Human Rights Act 2004* (ACT) and *Human Rights Act 2019* (Qld).

equal; the outcome of the law in action must also be similar across different types of people. As such, both the processes and outcomes produced by an AI system must be non-discriminatory. The right to equal protection of the law protects against the possibility that discriminatory outcomes of an AI system – such as disproportionately misidentifying people of colour, or targeting for audit or inspection women of child-bearing age – cannot be ignored or excused on the basis that the code was not designed to be discriminatory.

The right to privacy

The right to privacy is most obviously engaged by AI. It is a concept that 'defies precise definition'.[76] With each new era of technology comes a new concept of privacy. The existing paradigm that underpins most privacy legislation in Australia is 'information privacy', which is closely related to the concept of 'data protection' and was developed when computerised databases and ecommerce were the new technologies challenging our sense of privacy. The growth of cloud computing and algorithmic search functions has prompted debates about our right to be forgotten.[77] AI, combined with drones, CCTV networks and digital tracking, provokes debates about spatial privacy, the right to be left alone, and the right to obscurity.[78]

Implicit in these concepts about privacy is a recognition that privacy has traditionally been protected by practical limitations. It was once very expensive to undertake widespread surveillance and it was therefore unnecessary to express legal protections in great detail. Advances in technology, including AI, have made surveillance cheap, and access to it universal and easy. Our laws have not adapted to recognise this reality and our debates about the privacy impacts of AI continue to be characterised by a legal standard developed for a less invasive form of technology.

Under the Charter, the right to privacy is expressed as a right to not have your privacy, family, home or correspondence unlawfully or arbitrarily interfered with; and to not have your reputation unlawfully attacked.[79] Arbitrary interferences with privacy may include interferences that are "capricious and not based on any identifiable criterion or criteria"[80] and interferences that are "unreasonable in the sense of not being proportionate to a legitimate aim sought".[iii]

AI systems involve potentially arbitrary interferences with privacy. For many computer scientists, the answer to poor AI outcomes is often 'more data'. Yet the more data

[iii] The question of whether 'arbitrarily' should be given its ordinary English meaning or a meaning informed by human rights jurisprudence is an unresolved question. See Pound, A. & Evans, K. (2019). Annotated Victorian Charter of Rights (2nd ed), Thomson Reuters (Professional) Australia Limited, p. 11.

collected, the more the data collection approaches capriciousness and becomes disproportionate to the end sought. As AI advances, it may be capable of taking more actions that cannot be explained by the humans programming the system. As the criteria on which actions are taken become less identifiable, the more AI systems start to act in ways that seem capricious and unreasonable.

Freedoms of association, expression and movement

AI systems may also indirectly engage some of the 'freedom' rights, such as freedom of movement, freedom of expression and freedom of association. While an algorithm per se does not stop a person moving about Victoria, expressing their views verbally or in print, or joining groups, AI systems do facilitate the surveillance of such activities. AI systems regularly use data from the devices we carry with us daily, especially our mobile phones and tablets. AI systems are fed and can produce data about where we are and with whom. We are already familiar with advertising targeted to particular mobile devices in particular locations. It is just as possible that such technology could be used in ways that deter people from exercising their freedoms.

When the machine knows us better than we know ourselves – knows when we are pregnant,[81] seeking mental health services, or associating with people who may be seen as undesirable – we are each compelled to take evasive action on a daily basis, never knowing when our actions may be captured by the AI system and rendered meaningful. The feeling of being watched may alone be sufficient to deter people – especially those from marginalised or minority groups – from feeling truly free to exercise their human rights. People with characteristics flagged as 'risky' and subjected to increased monitoring may find themselves needing to "do more to prove and justify themselves simply because they 'look' like past transgressors".[82]

Some European authorities have found that storage of personal information on registers as a means of surveillance constituted an unjustified interference with the rights of freedom of assembly and expression, even where it had not been established that the person's exercise of their rights had in fact been hindered.[83] In Victoria and the UK, courts and tribunals have been less willing to accept that collection and storage alone will constitute an unjustified interference without evidence that there has, in fact, been a 'chilling effect' and persons in fact deterred from exercising their freedoms.[84]

Conclusion

AI systems are discriminating, data hoarding machines that provide the means for surveillance. While AI systems have the potential to both help or harm humans, they may not be not neutral. At every stage of their development and use, AI systems may discriminate in ways that adversely and unfairly affect humans. The humans inviting AI systems into our workplaces, economy and justice system should be aware of these risks and manage them proactively.

So too should our lawmakers. While discrimination by AI systems may be unlawful under existing discrimination laws, establishing the case will raise novel and complex arguments about definitions and attributes. Would an AI system be considered as a 'condition or requirement' for the purposes of indirect discrimination? Would an AI system that was not programmed to consider protected attributes nevertheless be considered to have directly discriminated if it learnt discriminatory patterns through machine learning? The benefits of AI systems will accrue to their operators, while the costs of testing whether our discrimination laws are adequate in the face of AI systems will likely fall to the people discriminated against by the AI system.

Biography

Katie Miller is a lawyer and public servant exploring how innovation and technology are changing legal practice and public service. Katie is currently a Deputy Commissioner of the Independent Broad-based Anti-corruption Commission and is undertaking a PhD under the supervision of Professor Matthew Groves on whether administrative law is fit for the technology age. Katie is a Law Institute of Victoria Accredited Specialist in Administrative Law, and an occasional guest host of ABC 774's Writs and Cures.

ALGORITHMIC TRANSPARENCY AND DECISION-MAKING ACCOUNTABILITY:

Thoughts for buying machine learning algorithms

Jake Goldenfein

More than 20 years ago, Helen Nissenbaum published 'Accountability in a computerized society', asking questions about how to think about accountability and responsibility in the unique environments of software production.[85] The problems of accountability she described then have been amplified further as computing technologies have become more complex and less intelligible. Nevertheless, these technical systems are increasingly adopted by governments, who at the same time import their accountability deficits into their regulatory function. This has prompted a great deal of research into ways to achieve algorithmic accountability and transparency in automated decision-making systems – especially those used in public governance.

Good accountability for the implementation of decision-making systems is far from simple, with multiple overlapping institutional, technical, and political considerations. In automated decision systems, the accountability typically associated with a public official (the human decision-maker) is displaced into the actions and decisions of those creating system specifications, software engineers, and any other party involved in producing the decision-making system. Questions of accountability therefore need to be considered prior to the deployment of a system, rather than after any decision is made. That means the ideal time to address questions of accountability and transparency in automated systems is during their procurement and development.

Thinking about accountability and transparency at that early stage requires knowing how automated decision systems actually work, knowing what to ask for when procuring such systems, and building the capacity to evaluate whether systems do what vendors say they do. In order to assist with thinking about accountability and transparency in procurement, this chapter describes some contemporary transparency and accountability mechanisms for automated decision systems, while at the same time offering some warnings about their susceptibility to industrial or corporate co-option that risks undermining their utility.

What algorithms?

Questions of accountability in AI systems have been studied for a long time, especially in the context of symbolic AI or 'expert systems'. That type of automated decision-making technology uses a 'rule base' wherein a certain body of knowledge – typically a decision-making process as specified by legislation, regulation, or policy – is represented symbolically, often through 'if-then' rules. Systems that encode legislative or regulatory rules for decision automation in this way make up the majority of automated decision-making system used by governments. In the Australian context, there has been *some* policy guidance on the deployment of automated decision-making in administrative government for expert systems. [86]

While questions of accountability and transparency with respect to those technologies are a long way from being solved, the focus of this chapter is a different type of automated decision-making technology that is increasingly the focus of scientific research and government procurement – machine learning. These systems do not simply automate rules, but instead use large amounts of data and statistical pattern-matching to generate predictions, classifications, scores, and decisions. In the machine learning domain, there is an ongoing and lively debate about what is necessary for transparency and accountability. Whereas expert systems require a direct translation of existing rules into a programming language, machine learning systems analyse the past (in the form of data), cluster that data according to certain 'features', and then generate a rule that correlates subsequent input data with a classification, without a human directly programming how those classifications are made. Not having a professional that can explain why they encoded a rule in a particular way, as with expert systems, makes accountability more complex.

Jenna Burrell has described the three forms opacity in machine learning systems that present the greatest challenges for accountability and transparency. These are:

- corporate concealment or trade secrets that may ultimately constitute some type of knowing deception;

- the reality that few people are sufficiently knowledgeable with relevant programming languages and machine learning systems; and
- that the complexity and dimensionality (high number of data points per sample) of the statistical processes used in machine learning means decisions are not consistent with human-scale reasoning.[87]

Each of these issues make algorithmic accountability and transparency – and ensuring that the values of public governance are embedded in decision-making systems – very difficult. They also make the landscape of government procurement for these systems particularly fraught. In particular, opaque automated decision-making systems risk undermining the already limited mechanisms that afford access into, and oversight of, government decision-making processes.

There are multiple ways that machine learning systems might be implemented by governments, and different applications require very different accountability and transparency approaches. In some contexts, machine learning outputs may enact a fully automated decision. Alternatively, such a system may only inform or assist a subsequent decision made by a human. There are also examples of algorithms being used to optimise public services, like public school bus routes and schedules in the city of Boston,[88] and thus not for individual decision-making, but rather allocating broader infrastructural services. Each of these types of implementation requires different levels of consultation with stakeholders, different types of transparency, and different levels of institutional accountability. Accountability and transparency of automated decision-making systems requires thinking beyond the technical 'decision', and including the broader decision system of technology, people, and institution. To that end, the focus on accountability here is not limited to technical artefacts but must also extend into the relationship of those technical artefacts to decision processes more generally, including processes of development and implementation.

Dealing with these issues at the design and procurement stage is a critical way to make up for the problematic belief that a 'human-in-the-loop' is an adequate solution for accountability. As the studies into 'automation bias' (the phenomena whereby humans uncritically defer to the outputs of technological systems) show,[89] a 'human-in-the-loop' is rarely going to be a sufficient solution to the accountability deficit these systems generate. In fact, in many cases, the idea of a 'human-in-the-loop' is a red-herring, especially if that human does not sufficiently understand, and cannot meaningfully explain, the basis on which a system has come to a determination. Rather, the time and place for instilling public values like accountability and transparency is in the design and development of technological systems, rather than after-the-fact regulation and review.

Various accountability and transparency mechanisms are discussed below, including 'human-in-the-loop' approaches, institutional transparency tools, fairness in machine learning, and explainable automated decisions. The next section, however, first offers a fuller example of the use of statistical prediction systems in a specific context.

Automation in justice and policing applications

Automation is used across multiple domains of government, each with their particular idiosyncrasies and regulatory environments. This section briefly describes the uptake of automated systems in the particularly fraught domain of risk assessments for policing and criminal justice.

It should be noted that beyond criminal justice applications, machine learning systems are being deployed around the world in, for instance, determining access to welfare benefits, public housing, educational resources, as well as for immigration decisions, and in many other fields. Many of the issues that have been brought up in the context of risk assessments are also critical issues in those other domains, however. That said, often the issues are very different. This overview of risk assessments here is thus not intended to be a comprehensive outline of all the problems these systems generate. Rather, these examples are described because they clearly demonstrate some of the risks associated with the use of automation and statistical approaches in government decision-making, and the issues associated with private industry developing software for public governance.

Although criminal intelligence and security tools are sometimes very sophisticated, the majority of automated systems in policing and criminal justice applications are relatively simple prediction systems. While intelligence systems might analyse data generated from a wide variety of sources,[90] risk assessments typically use a limited, and human curated dataset. Machine learning approaches look for statistical likelihoods of, for instance, re-offending, within that curated data to generate risk scores for decision subjects. This is a way of automating the 'actuarial approaches' that have been dominant in criminology, penology, and scientific governance for some time. Bernard Harcourt describes the types of questions commonly used in parole decisions, with categories such as criminal history, education and employment, financial position, family and marital status, accommodation, leisure and recreational interests, companions, alcohol and drug status, emotional and personal characteristics, and attitude and orientation.[91] Contemporary automated systems used for determinations about bail, parole, or sentencing use relatively similar data. With machine learning however, which particular data points (or features) are used in the calculation, and what weight (or significance) they have to the determination becomes less clear.

Risk assessment systems can be traced back to the increasing use of statistics in criminology and penology from the end of the 19th century for the sake of individualising criminal sanctions and imprisonment. Statistical approaches supposedly allowed for greater control of individual behaviour and shifted the role of incarceration to a method of general crime prevention. As the statistical work developed, it was used to analyse whether there was a social benefit in focusing penological and policing resources on specific offenders that supposedly committed the majority of crime. The ideological premise of that approach was the ability to predict dangerousness through the probabilistic relationship between certain characteristics or behavioural traits and criminality. Through the middle of the 20th century, the RAND Corporation in the United States (**US**) pushed this statistical work on risk assessment tools further under the moniker of 'evidence-based practices', and many of the privately developed products on the market grew out of that intellectual environment. Risk assessments for bail, parole, and sentencing are now extremely common in numerous jurisdictions, especially in the US.

Another application of automated decision-making becoming more widely used is 'predictive policing'.[92] Typically, predictive policing focuses on questions of 'when' or 'where' crime is likely to occur, but more and more, also 'who' is likely to be involved. Private vendors such as Palantir, a company launched shortly after the September 11 attacks, have already built systems trialled in US cities like New Orleans, Los Angeles, and Chicago. The data inputs for the New Orleans system included criminal databases looking at ballistics, gang, probation and parole data, jail telephone records, as well as central case management histories and a repository of 'field interview cards', for the goal of identifying potential future offenders and victims of crime. Field interview cards were collected from suspects and non-suspects when interviewed by police officers and greatly expanded the scope of data input for Palantir's systems. *The Verge* has also reported that in 2016, the Danish National Police and intelligence services also signed an 84-month contract with Palantir for a similar predictive technology package to identify potential terrorists.[93] That system, however, also uses law enforcement data taken from automated number plate readers and CCTV video, which would be analysed by computer vision systems.

In Australia, the New South Wales (**NSW**) Police have been using an algorithmic risk assessment and predictive policing system called the 'Suspect Target Management Plan' (**STMP**) that generates lists of suspects for police targeting. Investigation into the system identified that 44% of individuals targeted were Indigenous.[94] When questioned by the Legislative Assembly, the Minister for NSW Police refused to release any details on the procurement process or system itself.[95] It is unclear if the NSW Police force even know what 'features' are used in that risk assessment software, or whether it was developed in-house or purchased as a product. Access to this material was refused by a NSW Tribunal, deciding there was an overriding public interest against disclosure

because of the system's use in intelligence work.[96] Apparently, a senior police officer oversees the validity of targeting on all occasions (as a human-in-the-loop), although it is unclear what role they actually play or how the system contributes to decision-making.

There are also several 'transparency challenges' for risk assessment tools, as well as other automated decision-making systems generally.[97] When it comes to risk assessments and predictive policing specifically, there is already a growing literature that critiques many aspects of those systems, including the selection of data, their efficacy,[98] their detrimental social impact,[99] the problematic mismatch between the data used and the data that matters,[100] the discriminatory over-policing, and the over-incarceration of particular groups. But there are also transparency issues associated with these systems that generalise more readily to other machine learning applications. For instance, a critical problem is corporate reluctance to reveal which specific 'features' are relevant in automated calculations, or what weight those features are given. A 'feature' is a data point that supposedly measures some meaningful aspect of the phenomenon being observed.

Claims to trade secrecy here are particularly dangerous, and officials should refuse to work with vendors who are not willing to make their system sufficiently transparent for appropriate auditing and review. In the US at least, the courts seem content to protect those trade secrets claims, despite the clear necessity of auditing those systems for legality and desirability.

In the high profile *Loomis* case, in which the applicant argued that the proprietary nature of the COMPAS risk assessment tool prevented challenging its scientific validity, the court refused to provide access to the vendor's trade secrets for the sake of fairness auditing, despite acknowledging that meant it could not evaluate how features such as gender were used in the assessment.[101] Even if courts are willing to protect algorithms as intellectual property however, the issue can be avoided by good procurement practices that ensure sufficient transparency from industry, and that trade secrets and copyright claims do not trump the values of good governance.

Scholars have also criticised the power of tech vendors in this domain, arguing it can constitute a form of undue influence.[102] This dynamic sometimes means that governments have limited input into the design and specification of these products, including which datasets are, and are not, appropriate to include. Vendors may also refuse to share the data collected during the operation of their systems. Such *de facto* privatisation of public data must be resisted at every turn.

Certain risk assessment and predictive policing systems, have thus become emblematic of machine learning procurement processes gone wrong, and a demonstration of the problematic political economy shaping this environment.

Public sector agencies should recognise their power as market actors in these procurement contexts and use their positions to ensure automated decision-making systems perform adequately and appropriately, and are subject to proper governance and oversight. The procurement of these systems actually offers some space for ethical imagination with respect to the values they ought to be servicing, rather than merely eliminating the worst possible outcomes.

As noted previously, risk assessments are not the only applications of machine learning in government, and indeed some risk assessments do not use machine learning at all. There is a wealth of data available to government from various sources however, and this data can be used to train decision-making systems in multiple areas. Many of the issues described above in one domain can offer some useful lessons for thinking about accountability and transparency irrespective of application. To that end, a great deal of research into achieving transparency and accountability has emerged, which has generated a variety of approaches. These different accountability tools, and their limitations, are described in the following sections.

Human-in-the-loop

In Australia, there no explicit legal algorithmic accountability regime. The only (partially) regulated dimension of automated decision-making is the necessity of a 'human-in-the-loop'. However, even this issue is governed by antiquated policy, produced more than a decade ago, and in the context of different decision-making technologies.[103]

That guidance suggests that full automation is permissible depending on the level of discretion provided by the legislation, for example, if the provision suggests the decision-maker 'must' or 'may' make a decision. Government lawyers also appear influenced by the principle from the 1943 case of *Carltona Ltd v Commissioner of Works* concerning the capacity to delegate decision-making powers,[104] which has been interpreted to mean that as long as authorised by law, automated decision-making systems have few constraints.[105] But this approach ignores critical concerns about the quality and validity of technical systems, how they are integrated into decision-making systems, and who ought to be involved in the process.

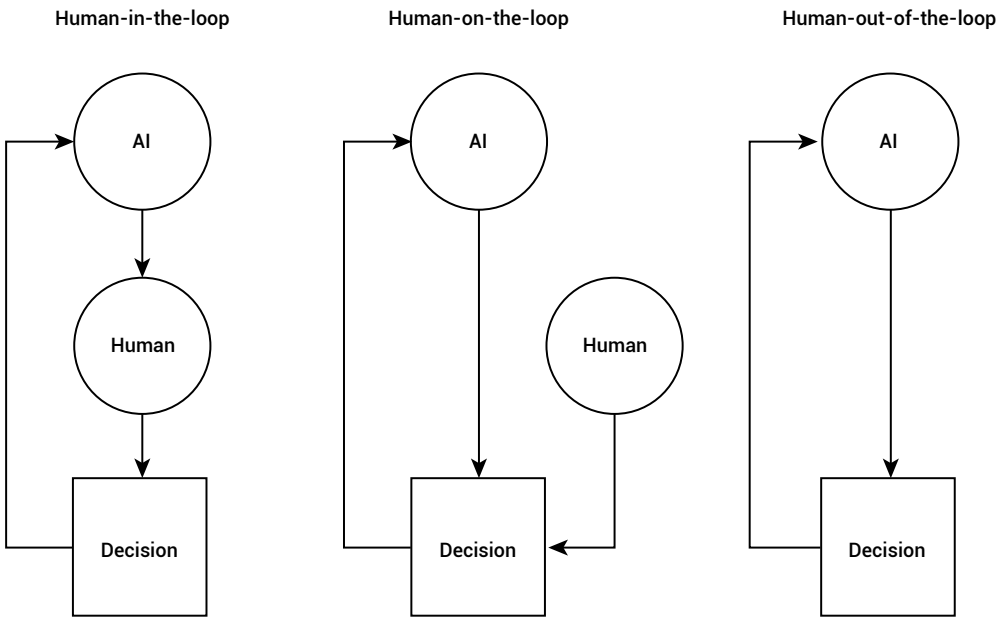


Figure 1

Three different kinds of human involvement with AI: human-in-the-loop – where an AI system provides information to a human in order for them to make a decision; human-on-the-loop – where a human supervises an AI system making a decision; and human-out-of-the-loop – where an AI system makes a decision without any human involvement.

High-level political discourse also seems to suggest that a human-in-the-loop approach is adequate. For instance, in 2018, Michael Pezzullo, Secretary of Australia's Department of Home Affairs – the department seeking to build Australia's interconnected facial recognition system, and also the official in charge of Australia's intelligence and security agencies and immigration decisions – made the claim that "No robot or artificial intelligence system should ever take away someone's right, privilege or entitlement in a way that can't ultimately be linked back to an accountable human decision-maker".[106] He called this the 'Golden Rule', which was then endorsed by Australia's Chief Scientist, Alan Finkel.[107] But this regulatory approach may be a distraction from the need for stricter rules regarding auditing or certification of decision systems, as well as the more important but labour-intensive process of scrutinising when, why, and how automated systems ought to be implemented.

European data protection law has, since at least 1995, included limitations on automated decision-making,[108] offering the right to obtain a human decision-maker in certain cases. This has been replicated in Article 22 of the General Data Protection Regulation (**GDPR**). These rights have been interpreted as requiring a 'human-in-the-

loop', as they afford individuals a right to obtain human intervention, express their view, and *contest* a decision. Even though such provisions have been law for some time, there is growing awareness that they might not provide effective oversight of decision-making systems, and there is no record of these Articles having ever been invoked or litigated. One reason for that may be that those laws are problematically vague.

To begin with, there are conceptual difficulties in identifying whether a decision is fully automated or not. Does a human decision-maker anywhere in the relevant information processing chain mean the decision is not made solely by an automated system? [109] That a human decision-maker *anywhere* in the process might remove an automated system from the purview of the law is a worrying limitation.

Some have argued that the EU text, "decision based solely on automated processing", should still include decisions made with some human participation,[110] although, in a decision on credit scores, the German Federal Court of Justice suggested any human participation in the decision would mean the law did not apply.[111] What 'degree' of human intervention might be permissible therefore needs resolution. Further, it is also unclear what a contest made to a human overseer would offer. What different information could a person present to a decision-maker to challenge an automated decision?

The human-in-the-loop paradigm may make sense in certain very high-level applications, but it is hardly the structural solution to algorithmic accountability some might wish it to be. As technologies develop, the speed and scale of automated decisions may ultimately undermine the capacity for human supervision.

The focus on a human-in-the-loop also risks taking attention away from the more important task of ensuring that systems are properly designed in the first place. That is why other provisions in the GDPR, such as the specific access rights dealing with the logic of decision-making systems (what is sometimes called a 'right to explanation'), as well as requirements for certification, impact assessments and auditing, may be more relevant for transparency and accountability. However, these are not legally required in the Australian context.

To be meaningful, the human-in-the-loop approach needs supplementation. When thinking through the design and specification of a system intended to inform a decision-maker, it is imperative that the human decision-maker knows how the system works, and understands the limitations of its outputs. Scholars like Mireille Hildebrandt have suggested 'agonistic' machine learning technologies be developed, wherein the system would demonstrate how each act of computation relies on a particular system of measurements, representations, and analytics.[112] Ultimately, it requires ensuring that the results are presented to the decision-maker in a manner conducive to scepticism. This might mean including specifications about dashboard design and data

visualisation strategies when a system is being procured, as well as adequate training of human decision-makers for interpreting the outputs of opaque systems. Because machine learning is effectively premised on the capacity to make decisions using large amounts of data rather than with domain knowledge, a human-in-the-loop is only meaningful if experienced human decision-makers retain their capacity to express that domain knowledge, rather than simply rubber stamping an automated system.

Institutional transparency and public values

There are many dimensions to algorithmic 'transparency', but in the context of institutional actors, it requires clarity in the procurement, implementation and technical mechanisms associated with automated decision-making systems. This type of transparency is useful for keeping track of the impacts of decision systems over time, and achieving some public disclosure on their purpose, reach, policies, and techniques. Freedom of information laws may appear relevant in this context. But while freedom of information may be available, and indeed may yield some useful documentation, it does not necessarily contribute to meaningful accountability. Good practice requires that when these systems are used in ways that affect people's lives, there is sufficient consultation and review, such that accountability and transparency are built into the implementation process. The public should not have to use freedom of information against departments and agencies to find out how they are governed by automated technical systems.

Researchers that have attempted to go through freedom of information processes, in the US at least, have achieved relatively little in terms of exposing either the contract terms between governments and private software providers, or any meaningful information about how that software actually works.[113] Researchers working in this area have also demonstrated that trade secrets remain a critical obstacle when governments use systems produced by private software providers, as that forms the basis of an exemption from freedom of information. Even the preamble to the GDPR acknowledges that transparency and access rights must be balanced against the rights of those who build the technologies, including copyright or trade secrets. [114]

Rather than thinking about whether or not information will have to be released under freedom of information laws, a better approach for good governance is to think about what information is meaningful to release, to whom, and when. Simply making *all* the information about an algorithmic process available may be of limited utility. To that end, some have proposed a 'how', 'what', 'why' model for thinking about the different types of disclosures necessary for regulating automated decision-making systems.[115] The 'what' question would ask the reasons for a specific outcome – for example, for a given input, what led to the output? The 'how' question is the set of rules that govern

the decision process and may involve exposure of its logic, including the particular algorithms and system design. The 'why' question looks to the ultimate goals of the system, the assumptions made in its implementation, data selection, and compliance with legal standards.

Some researchers have argued for another approach to algorithmic accountability and compliance that would not require transparency, calling instead, for what Joshua Kroll and his collaborators call 'procedural regularity'.[116] These systems would enable the subjects of automated decisions to know that the procedure applied to them was the same procedure applied to everyone else, that the same policy is used for each decision, that those decisions are reproducible, and that the decision policy was specified before the particular subjects of the decision were known. Proponents of that approach cite the whole toolbox of computer science tools used for testing and verification of software, and mechanisms like zero-knowledge proofs (a computational way of proving that certain information exists without revealing that information), that can prove the existence of certain features without revealing the whole system's operation. That approach, however, is primarily targeted at clearly egregious or abusive uses of those technologies, not necessarily for ensuring accountability in routine ordinary operation.

An approach gaining more traction, accordingly, is the embedding of structural compliance mechanisms in the form of auditing, certification, and impact assessments. As noted previously, the GDPR includes requirements for data protection impact assessments (Article 24), codes of conduct (Article 40), and certification (Article 42) in certain situations.

In the US, some have argued for 'public agency accountability' that involves self-assessments with respect to fairness, justice, and bias.[117] Others frame these structural requirements in terms of 'technological due process' or as meaning the introduction of an 'FDA for algorithms'.[118] Danielle Keats Citron advocated for this approach as early as 2008, noting then that automated computer systems were becoming *primary* decision-makers in administrative decisions.[119] She argued that a system of technological due process was necessary to bolster the procedural safeguards being undermined by automation, and considers the 'due process' clauses of the US Constitution (and other US federal and state laws) as an appropriate mechanism. This position was then extended in collaboration with Frank Pasquale to the use of scoring algorithms by the private sector,[120] arguing that 'technological due process' – "procedures ensuring that predictive algorithms live up to some standard of review and revision to ensure their fairness and accuracy" – is the proper path to accountability. Whatever the specific approach, the intention is to bring public administrative law and constitutional-type accountability to automated decision-making systems.

An example of technological due process in practice is New York City's statutory taskforce on algorithmic accountability, which includes a fact-finding group of researchers to evaluate the city's use of automated decision-making systems, but without exposing their technical details to the public,[121] and without interfering with the trade secrets of the companies that build those technologies.[122] The purview is systems used by the city for automated decision-making in policing and criminal justice, welfare entitlements, public housing, education, and wherever else. It is unclear whether disclosure to this taskforce has yet had any meaningful regulatory impact, or in fact whether the taskforce has even received any meaningful disclosures, indicating it might be a cynical political exercise.

Nonetheless, similar approaches are being suggested for the private sector, for instance, with the introduction of the US *Algorithmic Accountability Act* that would require automated decision and data protection impact assessments for 'highly sensitive' machine learning systems, produced by large companies, for bias and discrimination. Those compliance obligations would apply to any system affecting consumers' legal rights, any system involving large amounts of sensitive data, or involving systematic monitoring of a publicly accessible physical space. This approach effectively legislates for 'fairness' requirements, discussed further below. However, this approach would require companies to perform impact assessments on their own decision-making systems, rather than instituting a government agency to perform audits and certification.

Compliance through self-regulation may have various benefits, but it also threatens the legitimacy of any accountability exercise. While the *Algorithmic Accountability Act* would encourage the participation of external third parties, independent experts, and auditors if reasonably possible, it does not create the bureaucratic compliance infrastructure that would be more useful, and that the GDPR, for instance, requires. Another problem is that the companies building automated decision-making systems are sometimes the same companies that build the auditing and fairness tools for evaluating those decision-making systems. To that end, when thinking through the regulation of these systems, instilling public values is a challenge that cannot be left to industry self-regulation.

While total transparency may not be the most desirable outcome, instituting public mechanisms for certification, auditing, and evaluation of automated decision-making systems produced by private companies for public governance is important. Alternatively, if building those systems in-house, making the process as transparent to the public as possible in order to facilitate auditing would dramatically improve the implementation of those tools.

Fairness

'Transparency' and human supervision may be elements of algorithmic accountability, but what exactly are we looking for when these systems are opened up, or when a decision-maker is asked to account for their decision? As scandal after scandal associated with machine learning reveals, the issue is often bias or discrimination. 'Fairness' in machine learning has thus emerged as a sub-discipline of computer science, focused on exposing and limiting bias in algorithmic calculation. The fairness in machine learning paradigm grew predominantly out of the US legal environment's prohibitions on discrimination, although prohibitions on the use of sensitive data types in automated profiling in the GDPR mean 'fairness' is also becoming a global paradigm. [123]

'Fairness' frames the harm of automated decision-making as discriminatory or otherwise unfair evaluation of an individual in an algorithmic decision-making or classification system. Unfair discrimination finds its way into automated systems in multiple ways,[124] and improving the outcomes of automated decision systems, for instance, by removing the impact or influence of sensitive or protected data (any category you do not want influencing decision outputs, such as race, sexuality, or political position), is critically important.

But what exactly 'fairness' requires, or means, is complex. Indeed, both the NSW Police STMP program, and the US example of the COMPAS risk analysis tool discussed in the *Loomis* case, demonstrated bias against particular racial groups in their risk scores. Those tools, and the opacity of their operation are highly problematic. But the discussion amongst researchers that followed *Loomis*, or more precisely, followed the investigation by *Pro Publica* into the COMPAS tool,[125] was interesting for how it also highlighted the multiple possible interpretations of 'fairness' in statistical applications, the impossibility of embedding multiple ideas of fairness concurrently, and that a system's fairness optimisation will necessarily reflect a political agenda.

Fairness has to be optimised towards one outcome or another. For example, this might be approached through 'predictive parity' (ratio of true positives to those labelled high-risk generally), or 'error rate balance' (the distribution of false positives and negatives for specific groups).[126] Commentators rightly point out that the former approach, which Northpointe (the vendor of COMPAS) argued indicated their system was fair, may reflect the world view of those who have studied actuarial and preventative penology and policing, whereas the latter (which Pro Publica argued would be necessary for fairness) may reflect a social justice paradigm of idealised outcomes. But these goals cannot be implemented concurrently.[127]

This raises difficult problems in the procurement and specification of automated decision systems. Even if an official recognises the necessity of ensuring such

systems do not operate in a biased or discriminatory manner, how such a system should be optimised is unclear. Simply specifying 'fair' or 'non-discriminatory' machine learning when designing a system is insufficient. To that end, 'fairness' research has now revealed an entire catalogue of mechanisms for identifying (and sometimes correcting) bias and discrimination. There are now somewhere between 15 and 25 plausible definitions of 'fairness' being used, each optimising for different things. Those definitions can generally be grouped into three different categories – 'statistical', 'similarity', and 'causal' based reasoning methods.[128]

'Statistical' measures describe fairness metrics that can be calculated from observational data, such as how a range of features is distributed across a population. Statistical fairness measures attempt to equalise the distribution of features across that population, but they often mask unfairness towards smaller minority or sub-groups within the population that are not identified by the feature being equalised for. These measures are therefore only fair for an average member of a protected group.

'Similarity' based approaches analyse the similarity of treatment between two individuals. Similarity is measured by a 'distance metric' or 'statistical distance' between the distribution of outcomes for those two individuals. But this is difficult to translate into a meaningful notion of fairness because the relationship between that 'distance' and some tangible fairness criterion is highly abstract.

Finally, 'causal reasoning' fairness systems take unobservable variables into account in order to understand the influence of different attributes on each other. The most common example is 'counterfactual fairness', which involves changing the value of any one protected attribute while keeping non-causally-dependent attributes constant to examine any influence on the outcome. A system will be counterfactually fair if it generates the same outcome in both the 'real world' and a 'counterfactual world' where the individual belonged to a different demographic group.[129] But this does not address how protected features typically represent structural inequalities that will be expressed through the entirety of an individual's data.

Most critically, work has not been done to analyse which of these definitions is appropriate or useful in different types of decision-making application. This is now becoming the next frontier of work in algorithmic accountability, and rigorous attention is urgently needed to improve the procurement process.

These fairness approaches are necessary because the math demonstrates that simply removing potentially discriminatory data points will not produce fairer outcomes. In fact, eliminating specific data categories may be counter-productive, as the biases associated with those social categories linger on elsewhere in the data. That is, when the explicitly discriminatory data points are removed, the effects of those protected

categories on the system are replicated in other non-protected categories of data. Eventually, tracking the effect of discriminatory features through an entire dataset becomes an intractable mathematical problem.

In the end, no purely technical approach can solve every problem of machine learning discrimination, and all 'fairness' approaches become political because they privilege some stakeholders and marginalise others.

Another risk when attempting to make machine learning fairer is not thinking about those political questions and simply purchasing off-the-shelf corporate fairness products. Fairness tools are being created, standardised, and marketed by large tech firms like Google and IBM as 'fairness solutions'. However, it is not always clear what type of fairness these tools measure. Nonetheless, this has not prevented those companies from expressing how their systems have 'solved' the fairness problem.[130]

Off-the-shelf products generate other risks too – they may not always be suitable for an application for which they were not specifically designed, despite vendors suggesting otherwise. Investigation into the COMPAS tool revealed that although it was being used to predict likelihood of re-offence at sentencing, it was actually designed for use at the pre-trial phase. It is therefore important to note that validation of a machine learning tool in one context does not necessarily translate to another. Even validation in one geographic context does not mean validity when used amongst a population situated in another geographic area. Different contexts mean different data and different outcomes with different impacts on people's lives.

Unquestionably, these companies have an interest in removing discrimination from their machine learning products, as 'fairer' machine learning is more legitimate machine learning, and eliminating improper discrimination leads to better decision-making. But there is a risk that the valuable work of the academic community investigating fairness is being co-opted by industrial interests to justify the proliferation of these systems, without properly attending to the issues they create. Officials involved in specifying these systems should be cautious that any off-the-shelf fairness solution may not address the issues that are relevant to that particular application.

Beyond 'fair' calculation, some scholars have turned their attention to other issues for building fair systems. For instance, selecting relevant and appropriate data sets to train a system is critical. Often the data used to train a system does not sufficiently measure the real-world effect that the system is trying to regulate. The calculation mistakes associated with the 'Robodebt' scandal were consequences of the data input for the system (information about income earned over a year period) not reflecting what the system was actually trying to measure (money earned through employment while also receiving benefits).[131] To prevent these issues, predictive systems ought to be open to proper scrutiny prior to their deployment.

Explainability

In much the same way that administrative law requires 'reasons' in administrative decision-making, the capacity to understand why an automated system has reached a particular decision is important for algorithmic accountability and transparency. However, the illegibility of machine learning makes this a challenge. While no jurisdiction has explicitly legislated this requirement, some have read a 'right to explanation' into Article 15(1)(h) of the GDPR, even though the term is not actually used in that provision. In fact, there is rigorous debate over what that provision truly affords, how useful it may be, and whether explanation of machine learning is even possible. The law provides that in the case of automated decision-making and profiling, individuals should have access to "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such [automated] processing for the data subject [subject of the decision]." What constitutes 'logic' here is disputed,[132] as is what type of disclosure about that logic would be 'meaningful'.

The debate focuses on whether the law requires explanation of system functionality – "the logic, significance, envisaged consequences and general functionality of an automated decision-making system, eg the system's requirements specification, decision trees, pre-defined models, criteria, and classification structure" – or an explanation of specific decisions – "the rationale, reasons, and individual circumstances of a specific automated decision, eg the weightings of features, machine-defined case-specific decision rules, information about reference or profile groups".[133] Authors like Sandra Wachter, Brent Mittelstadt and Luciano Floridi adopt the former position, arguing that the GDPR does not grant a right to explanation of the logic, significance, and consequences of a specific decision after it has been made. They claim the law only requires a general explanation of system functionality.

On the other hand, authors like Andrew Selbst and Julia Powles argue that when you read various provisions of the GDPR together, they do require explanation of specific decisions, because 'meaningfulness' is tied to the right to contest a decision in Article 22(3).[134] This approach frames the right to explanation as giving ammunition for a contest or appeal rather than simply justifying how a system works. That latter interpretation would greatly enhance the utility and significance of the law.

If a 'right to explanation' of specific decisions does exist in the GDPR, or is legislated in another jurisdiction, there are still questions as to what that explanation should communicate. Does explanation mean: disclosures about the specification and design of an algorithm; the system's explicit purpose; the features and weightings the system uses; the kind of outputs it generates and how they contribute to a decision; what level of human intervention remains or is possible; whether the system has been validated, certified or audited, and in what context; whether the system uses a fairness model and

what type of model; or anything else? In this context an approach growing in popularity is to build machine learning systems capable of explaining themselves. The computer science community is taking on board the possibility that machine learning-based profiling may need to be explainable to be legally acceptable.[135] There is therefore a growing effort in computer science to design automated decision-making systems that are in some way explainable. This field is called 'explainable artificial intelligence' or XAI.

XAI entered the mainstream agenda after a 2016 US Defense Advanced Research Projects Agency (**DARPA**) grant solicitation funded multiple research laboratories to work on the problem that the development and effectiveness of machine learning technologies will ultimately be limited by their opacity.[136] The broader policy goals behind XAI are thus to improve machine learning systems, to satisfy any emerging legal compliance, and to enhance public trust in the use of those systems.[137] There are a growing number of XAI models and a great deal of competing theorisations of the value, function, and proper expression of computational 'explanation'. But what constitutes explanation is as yet unclear in the computer science community, let alone in the broader community that has long contemplated that question, or in the communities that would put XAI systems to work in socially operative systems.

Some XAI projects address explanation by attempting to produce more comprehensible or intelligible decisions by communicating to users simplified approximations of what a system is doing or what the model does. Some argue that simplified approximations are insufficient, and that explaining the line of reasoning a system engages is also necessary. Others have proposed 'what', 'why', and 'how' models (again), noting that reasons are not necessarily explanations unless they give insight into the mechanisms at play in any particular decision.[138] Some projects focus on specific domains like images, and describe explanation as "presenting textual or visual artefacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction".[139] A prominent example is the Local Interpretable Model-agnostic Explanations system (or LIME), capable of describing which elements of an image pushed a classifier to make a particular prediction by identifying elements of the image connected with those predictions. Gaining particular popularity however, are 'counterfactual' approaches that account for how a model has behaved in a particular context or application, rather than how it functions.

Finale Doshi-Velez and others have described, for the sake of counterfactual XAI, what they understand as the three core elements of 'explanation': What were the main factors in a decision? Would changing a certain factor have changed the decision? And why did two similar-looking cases get different decisions, or vice versa? [140] For them, the goal of explanation is to produce after-the-fact analysis of decision output rather than exposing how a model actually works.[141] In other words, they are primarily 'justification' rather than 'introspection' methods.

To that end, counterfactual approaches may not be the best approach to accountability in governmental decision-making because they do not demonstrate how a particular classifier has interpreted and dealt with a particular rule – that is, the rule a contesting party might argue the decision-maker has failed to abide by in their decision. Instead, counterfactual approaches only probe how a decision-maker might deal with different factual scenarios, not how the system interprets the world or the relationship between data and rule or policy. Accordingly, while a counterfactual approach may be useful for justifying a decision, it may not offer meaningful explainability for contesting automated decisions. In other words, if the goal of explanation is to give ammunition for an appeal, a piece of software explaining how the system would have computed other input data may not be enough to build a persuasive argument that a decision was made improperly.

Ultimately, each of these systems includes inherent trade-offs, perform better or worse in certain situations, provide one type of information at the expense of another, or end up being arbitrary when pushed too far.[142] When computer scientists generate 'interpretable models', it is important to recognise that 'interpretable' in that sense is a purely mathematical and quantified concept.[143] As Mittelstadt and his co-authors rightly note, XAI is more akin to scientific modelling than explanation-giving.[144] Indeed, research in this field is running up against the reality that the legal conception of explanation may not be what technical systems can foreseeably provide.[145]

Without arguing for what might be the most appropriate XAI approach, it is suggested here that XAI actually risks becoming a harmful approach to accountability. XAI has the potential to entrench problematic automated decision-making by narrowing the types of reasons that are given for decisions, therefore narrowing the grounds for contesting them. Being subjected to automated decisions without understanding how or why that decision was made may be problematic; but receiving automated explanations that do not provide a premise on which to base an appeal or contest – and simply justify the decision – might be worse.

There is therefore a risk that what 'explanation' means for law – a subject long discussed in legal theory – may ultimately be reduced to what a computational system is capable of explaining about itself, or what the entities that build and commercialise machine learning systems, or the institutions that deploy them, may prefer to constitute an explanation. In other words, XAI risks ceding to data science the epistemological terrain of what constitutes explanatory information.[146] If the function of explanation is similar to the provision of legal 'reasons', then this is a very precarious trajectory.[147] The simultaneous motivations behind XAI, of legal compliance and enhancing trust in a system, might then be contradictory, depending on how that notion of legal compliance is configured.

Building machine learning systems capable of giving a cogent and meaningful explanation for their output is a desirable goal, but it is critical that this development

in the technology is not directed to merely justifying its outputs. Explanation must be geared towards challenging decisions more than justifying them. It must be situated around exposing how an automated system may have used the wrong data; how the data used may not represent the totality of the data relevant to the question; how the system may have miscalculated or not understood the significance of that data; or how the rules, when applied to that data, might not produce the desired outcomes.

Conclusion

These critiques and warnings about transparency, fairness, and explainability all have a similar flavour. If an official specifies for a system to be transparent, fair, or explainable, it is important that they understand the limitations of such a specification. That means significant resources must be invested in developing the skills necessary to decide whether a machine learning system is useful and desirable, and how it might be made as accountable and transparent as possible. The ability to write good requests for proposals or tender documentation is critical. The capacity to closely review a system after implementation and consider its ongoing effectiveness and impacts is also important.

Political communities need to have discussions about their non-negotiables when dealing with vendors – that is, what absolutely must be included and excluded in a system and procurement contract. That might be something like retaining ownership of data and systematic reporting. It might mean involving universities and community organisations, or other stakeholders, to understand the impact of a decision system. This is because proper transparency and accountability means more than simply knowing what 'features' are important in the technical system. It also means inclusion of the people who are most likely to be affected by a machine learning system in the design and review process. Governments need to build these technical, political and organisational skills in simple deployments of machine learning systems in order to develop the ability to manage much larger projects, filled with massive-scale computation, associated with 'smart cities', that may be coming in the near future, with more massive impacts on people's lives.

Without those skills and transparent processes, all the governance standards in the world will not lead to good political outcomes, as too much is left to the private sector, where motives are not the same as those in government. In that context, transparency and accountability are at risk of losing their value. As Gloria Gonzalez Fuster argues with respect to algorithmic transparency in the GDPR:

> *In European data protection law... transparency is fundamentally not about a vague, utopic state of objective clarity, but about something else. It is not about letting data subjects sneak into the real life of their data and into the algorithms*

> *that move them, but about providing individuals with a certain narrative about all this processing; a narrative de facto constructed for data subjects on the basis of the interests of the data controllers, and adapted to fit a certain idea of the data subject's presumed needs and ability to discern.* [148]

Fuster describes how transparency might simply deliver to data subjects an account of what is being done to their personal data, tailored to a certain idea of what individuals might want to hear, and what they can perceive. The point is that transparency can become an instrument that distracts us or even actively undermines the capacity to meaningfully challenge or bring oversight to these decision-making processes.

In light of these problems, scholars like Frank Pasquale have begun to ask whether these computational forms of accountability adequately consider the question of 'accountability to whom'. [149] Yarden Katz similarly comments that "If AI runs society, then grievances with society's institutions can get reframed as questions of 'algorithmic accountability.' This move paves the way for AI experts and entrepreneurs to present themselves as the architects of society."[150] Without proper attention to these issues, accountability risks becoming part of the feedback mechanism that unthinkingly proliferates automated decision-making, without paying attention to its social desirability or political consequences.

Biography

Jake Goldenfein is a law and technology scholar interested in the emerging structures of governance in computational society. He completed a PhD at Melbourne Law School (University of Melbourne), has been a lecturer at Swinburne Law School since 2016, and is currently a postdoctoral research fellow at the Digital Life Initiative at Cornell Tech in New York City. Jake's recent work can be found in Law and Critique, Columbia Journal of Law and Arts, Internet Policy Review, and the University of New South Wales Law Journal.

AI IN THE PUBLIC INTEREST

Fang Chen & Jianlong Zhou

Like the steam engine in the first industrial revolution, electricity in the second industrial revolution, and electronics and information technology in the third industrial revolution, artificial intelligence (**AI**), which has powerful capabilities in prediction, automation, planning, targeting, and personalisation, is claimed to be the driving force of the next industrial revolution (Industry 4.0).[151] It is transforming our world and our society, affecting virtually every aspect of our modern lives.

AI enables the monitoring of climate change and natural disasters, enhances the management of public health and safety, can predict crop output in agriculture,[152] and automates administration of government services. AI might also be used to prevent human bias in criminal justice,[153] enable efficient fraud detection (such as in welfare, tax, trading), enhance the protection of national security, and in more esoteric fields, AI can help discover new galaxies,[154] and design new drugs.[155]

AI can provide benefits across a large range of public interests and deliver revolutionary change in both the efficiency and effectiveness of services. AI can enhance the quality of human and support better decision making, especially decision making in government.

However, there are many ethical, legal, social and cultural barriers to AI, and its applications are not without their costs. AI often requires an enormous amount of information to learn, and that information is often personal in nature. Privacy issues are becoming increasingly complex in the digital age, and the ethical implications of AI are among the community's top concerns with its proliferation. But while there are many privacy considerations with AI, not all are negative.

In some cases, AI can actually enhance privacy by reducing, if not removing, the need for personal information to be collected. This chapter will explore these issues, using examples from transport, infrastructure assets, energy and education.

Smart AI applications affect the quality of life

Many people assume that AI can enable computers to exhibit human-like cognition, and that AI is more efficient than humans in various tasks, for example, by having a higher accuracy, being faster and working 24 hours a day. Claims about the promise of AI are abundant and ever growing in relation to different areas of our lives. These diverse and ambitious claims have led to interest in AI in a wide range of industry sectors including retail, education, healthcare and others. According to surveys by McKinsey, the leading sectors in AI adoption today are mainly high tech and telecommunications, automotive and assembly, financial services, resources and utilities, media and entertainment, consumer packaged goods, followed by transportation and logistics.[156] Theoretically, at least, these uses should ultimately help to deliver better quality of life with manageable cost of living, a better environment, and easy access of transport for time saving.

McKinsey produced an analysis of more than 400 use cases of AI, representing $6 trillion in value across 19 industries and nine business functions, demonstrating the broad use and significant economic potential of AI.[157] In more than two-thirds of the listed use cases, AI was used to improve the performance beyond that provided by other analytics techniques. For example, AI is already helping financial institutions augment financial planning and investment strategy, and in some cases, AI powered diagnostics systems have proven to be more accurate than human doctors in diagnosing serious disease.

Using AI and data to support better decision making

AI may help humans by automating tasks that would take much human power or time to deal with. It can also find patterns that are usually difficult to catch by humans, allowing insights that might not otherwise be apparent. For example, AI has powerful capabilities in coordinating data delivery, extracting data trends, making predictions, quantifying uncertainty, checking data consistency, generating new data, and suggesting courses of action. Such capabilities can augment human intelligence dramatically in tasks and enable decision making processes to be done automatically. As a result, AI powered decision making could not only improve decision quality by reducing errors and biases that are common with human decisions, but it may also decrease the human workload involved in critical decision making. AI has the potential to make a revolutionary impact on the way in which humans make decisions.

AI and government decision making

The applications of AI in the public sector are already broad and are continuing to grow. Today, the sources of information accessible to government is massive, ranging from organisation data, program data, service data, health data, data created by Internet of Things devices, as well as many others. While any application of AI in the public sector must be balanced with careful governance, review and ethical considerations, there are several ways AI could be used to improve government operations:

- **To streamline or automate high frequency, high workload decisions:** AI can improve the quality of decisions and reduce the cost of services by automating time-consuming, manual bureaucratic processes. Typical examples of such decisions include making welfare payments and immigration decisions.
- **To make decisions in complex public sector problems:** AI may identify leading indicators that signal potential problems in public sector applications. For example, tax fraud is a serious problem that could be detected by AI, enabling government agencies and departments to make informed decisions about its enforcement and policy.
- **To make strategic decisions:** the ability to efficiently process large amounts of data from various sources for analyses and prediction allows AI to help government make high level strategic decisions for different areas of public policy. For example, AI could help government identify which skill sets might be required for a particular program, assisting in workforce planning. AI could also undertake predictive analytics for the requirements of infrastructure assets in a new suburb, helping government to make planning decisions.

There are four areas in particular in which AI has potential to enhance the work of government: transport, infrastructure management, energy, and education.

AI in transport

With growing urbanisation, more and more vehicles will be on our roads, resulting in significant transportation issues such as heavier congestion and an increase in serious accidents. These transport issues may cause economic and social loss. For example, road users in Sydney and Melbourne currently need to allow an average of 50% more time to complete their journeys during peak hours than during non-peak times.[158] AI could use data from different sources such as road cameras, mobile phones, road networks, and even social media to set up machine learning models, which could revolutionise different aspects of transportation.

AI is already being used in the prediction and detection of traffic accidents and congestion. Traffic simulation is used to simulate the effects of actions that could be conducted when traffic accidents or congestion occur. These solutions allow for real-time monitoring of transport networks and the identification of operational anomalies so that transport operators and travellers can make better decisions. For example, AI has been used (along with collating traffic data) to reduce congestion and improve the scheduling of public transport in some cities.[159]

In addition, AI can be used to predict the locations and frequencies of traffic accidents in order to suggest actions to improve safety. For example, pedestrians and cyclists are the most vulnerable road users to serious injuries in a traffic accident. If AI were used to predict common paths that pedestrians and cyclists take, it may help in decreasing instances of traffic accidents and injuries. With a predicted path of pedestrians or cyclists, transport signals – such as speed limit advisories and red/green lights at cross-roads – can be automatically adapted to control the movement speed and paths of vehicles on the road to reduce potential traffic accidents.

With appropriate inputs, AI could also be used to predict future transport requirements based on historical transportation volumes, population increases and other factors. This could help authorities optimise traffic networks and make informed plans for both traffic networks and city planning.

Autonomous vehicles have the potential to enhance mobility of people who cannot drive and dramatically change how we get from one place to another. Self-driving trucks and remote-controlled cargo ships could relieve drivers from intensive workloads and dangerous work conditions. Autonomous delivery trucks could change the way we receive goods, offering faster speeds through optimised paths, potentially bringing significant economic and environmental benefits if this is done effectively. AI could be used to sense the environment of autonomous vehicles while they are in operation and allow vehicles to automatically conduct various operations. These functions in self-driving cars, driverless buses, and driverless trains could improve traffic safety by automatically sensing risk around them and taking action to minimise the risk, resulting in an overall improvement of public safety. For example, computer vision and other AI technologies can be used to recognise objects surrounding operating vehicles; if a walking human is recognised in front of a vehicle, AI can decrease the speed of the vehicle or change lanes to avoid an accident.

These examples demonstrate the dramatic impact of AI in areas of public safety, administration of government services, and productivity.

AI and infrastructure management

Governments have significant challenges sustaining public infrastructure assets, such as water supply networks, roads, and bridges. Asset management strategies generally focus on the maintenance, replacement, and rehabilitation of assets in the later stages of their service life cycles. An effective implementation of infrastructure asset management can not only bring economic benefits to the government, but can also minimise potential failures of those assets.

Water supply networks constitute one of the most crucial and valuable urban assets. The combination of growing populations and aging pipe networks require water utilities to develop advanced risk management strategies in order to maintain their distribution systems in a financially viable way.[160] Especially in the case of critical water mains, risk should be defined based on the potential impact. The network size and location are key factors in determining this. For example, the failure of a single trunk line connecting distribution areas or a pipe under a major road typically brings severe consequences due to service interruptions and negative economic and social impacts, such as flooding and traffic disruption.[161] From an asset management perspective, there are two goals for water pipe management: to minimise unexpected pipe failures by prioritising timely renewals, and to avoid replacing a pipe too early before the end of its economic life.[162] AI can help to predict and identify high-risk pipes before failures. If used correctly, it is likely that repairs could be completed with minimal service interruption, water loss and negative social and economic impacts.

In addition to water pipes, AI can also be used for the predictive maintenance of other public infrastructure and equipment. For example, road maintenance is one of the key responsibilities of government. With appropriate privacy protections, we might use mobile network technologies and some of the sensors already present on modern vehicles to collect information on the condition of roads in real-time, using data from both passengers (such as from their mobile phones) and from vehicles. This may be a cost-effective alternative to using expensive special road inspection vehicles that are based on radar, high-definition cameras and LiDAR technologies. With the use of AI, information on the locations, sizes and types of road defects could be identified, analysed, prioritised and sent to road management authorities for appropriate action.

Traditionally, bridge inspection is conducted in person by professionals at predetermined time intervals (based on already developed risk assessment models). This is time-consuming and expensive. In contrast, different sensors (such as vibration sensors and displacement sensors) could be installed on various elements of a bridge to record any physical changes to the bridge. AI could analyse the data from these sensors, leading to a more reliable and continuous monitoring of the condition of the bridge in real-time. For example, the Sydney Harbour Bridge, which was completed in 1932, needs to be

inspected regularly. But finding faults along the 1,149-metre long deck and 134-metre high steel arch bridge visually is a difficult and time-consuming process. The Commonwealth Scientific and Industrial Research Organisation's (**CSIRO**) Data61 developed an AI system based on predictive analytics to continuously monitor the structural health of the bridge, and provide early warnings of problems before the bridge services are affected.[163] Around 2,400 sensors were installed on the bridge to collect information on its condition, allowing this intelligent monitoring system to work continuously.

AI in energy

The provision of adequate, reliable and affordable energy to meet future energy consumption needs is one of government's significant missions. AI can help to achieve such an ambitious objective.

With the significant increase in the use of renewable energy generation systems, electricity 'smart grids' are transitioning from creating intelligent energy distribution and flows within the existing grid structure, to using AI to restructure the grid by bringing in new, diverse and decentralised energy sources, such as solar and wind power.[164] Technologies such as batteries offer further opportunities for improvements, and electric vehicles provide increased demand requirements. This kind of future grid would be a complex network with both electrical generation and distribution assets. It could be expected to intelligently match supply and demand and operate automatically or semi-automatically. A future grid could also measure and predict the needs of individual customers, balancing them at different times or for different purposes, and then take appropriate actions throughout the network, delivering customised energy management solutions.

The adoption of AI and smart automation can aid the future grid from different perspectives:

- **AI for matching supply and demand:** The modern electricity ecosystem usually includes a mix of traditional energy, such as hydropower or thermal power; renewables, like solar and wind power; as well as energy storages. However, renewable energy sources are usually weather-dependent and therefore highly unpredictable, making it a challenge to match demand and supply. AI offers solutions to demand management problems by using predictive analytics to accurately estimate renewables to balance grids.[165]
- **AI for energy efficiency and reliability:** AI can help to improve the economic efficiency of energy. For example, AI can monitor and optimise the turbine parameters of wind power to increase its energy production. Turbines with less performance can be detected by monitoring and comparing the generation of other turbines in a wind farm. Parameters of the underperforming turbine, such

as the blade's direction relative to the wind direction, can then be optimised based on other turbines in the wind farm. AI can also automatically detect anomalies and faults in electrical grids by monitoring various parameters by way of smart meters in businesses and homes. Item sets of events that may be anomalous can be identified based on their patterns of appearance in the smart metre data stream.

- **AI and the consumer:** By monitoring the energy consumption behaviour of individuals and businesses, AI can offer solutions to tailor customers' energy consumption and reduce costs by giving suggestions on how and where customers can save energy. Customers can also benefit from AI in the supplier selection. Although market reform for improved competition would be needed to facilitate it, it might be possible for an AI to learn a customer's energy consumption and generation profile and then match the most suitable offers from the market and automatically switch suppliers.

AI could help provide adequate, reliable and affordable energy by using predictive analytics to accurately match demand and supply, optimise parameters of renewable energy for energy efficiency, detect anomalies automatically, and provide customised energy solutions for individual consumers. These functions combined could maximise the use of renewable energy and encourage the efficient use of energy, resulting in benefits for the environment.

AI in education

Education is, and will always be, a foundational part of humanity. Regardless of age, we are constantly developing new skills and understandings. AI technologies are well suited to achieving crucial education objectives, such as enhancing teaching efficiency and effectiveness, providing lifelong education for all, and developing personalised learning.

Conventional classroom teaching delivers one lesson to the entire class without considering individual differences in learning, which not only makes individuals frustrated when they cannot follow the pace of teaching, but also wastes the time of students who have already grasped the concepts. AI could improve adaptive learning and personalised teaching by identifying factors or indicators related to learning efficiency. Those factors or indicators are derived from students' behavioural, physiological information or even learning materials. For example, in mathematics, the cognitive load level of a student during learning, which is related to the difficulty level of mathematic questions, could be estimated through the examination of the student's behavioural features such as writing speed, orientation of the pen, and pressure of the pen tip.[166] When the cognitive load level of the student is too high or too low, the difficulty level of mathematic questions could be adjusted in order to keep the cognitive load to an appropriate level to maximise the learning performance of the

student. As a result, the learning could be adaptive based on students' responses, with a feedback loop for better learning performance.

AI can also complement the skills of classroom teachers. Teachers and AIs have complementary strengths and abilities. A teacher may have strengths in high level guidance and creativity, while AI may have strengths in analysing students' responses to learning materials. It is expected that AI can help fill needs gaps in learning and teaching that schools and teachers cannot provide.[167] For example, AI could provide personalised and streamlined teaching to students by analysing each student's responses and performance during learning, allowing teachers to focus on providing unique human capabilities of high-level guidance, high-order thinking and creativity.

The potential for AI to benefit the education sector also extends outside of the traditional classroom setting. Tutoring outside the classroom is often limited because teachers are not always available. However, students could receive additional support from AI tutors at any time, without being limited by locations (whether students are in urban areas or in remote rural areas). Most importantly, AI can also provide feedback to students on their success in the subject.

If implemented, AI will almost certainly shift the role of teachers in education. Because AI can help students with adaptive and personalised learning, and provide tutoring at any time, teachers will supplement AI-based learning, assisting students who are struggling as well as providing mentoring and hands-on coaching experiences for students – value-adding tasks that are uniquely suited to human beings.[168]

With the use of AI, students may not only have efficient learning based on their own capabilities but also benefit from lifelong learning. Most importantly, if made widely available, AI may also help achieve equity of learning for people, no matter their age, where they are from, or their profession. All these qualities of AI are of use for the advancement of society.

AI and privacy

In many cases, the use of AI will need to be balanced with other considerations, such as ethics and privacy. While we have so far demonstrated the far-reaching benefits that AI can have for the public sector and the community, governments are held to a high standard of accountability when it comes to the way they use personal information. The use of AI bears no exception, and privacy must be a key element in the design of AI systems.

There is no single or conclusive definition of privacy. It encompasses many connected but different ideas, including secrecy, confidentiality, freedom from surveillance, and having control over one's own personal information. Privacy is not a fixed concept – it can mean different things to different people, and individuals will experience privacy in varied ways. Often, one's past experiences will influence their relationship with privacy, and the extent to which they value this human right.[iv] Information privacy relates to an individual's ability to determine for themselves when, how, and for what purpose their personal information is handled by the organisations with whom they transact. Ultimately, it is about allowing people to maintain their individuality and autonomy.

Numerous countries, including Australia, have laws to protect information privacy,[v] and in many cases these laws are technology-neutral, meaning they apply to personal information regardless of the form it takes, or how it is collected, stored or used. When personal information is not protected, it can cause real harm, be that financial, reputational, or even physical. When privacy is invaded on large scales, it can lead to a 'chilling effect', deterring people from exercising their freedom of expression, stifling discourse that is necessary for democracies to function.

AI usually requires huge volumes of data in order to learn and make decisions. Because of such heavy demand for data – in many cases data containing personal information – privacy is one of the most important issues in the design and use of AI. Both AI technology developers and users need to consider how they can provide proactive protection to individuals in the face of AI technologies.

AI compromises privacy

The large amounts of data that enable AI to work can pose significant privacy risks, and these can be exacerbated if AI is permitted to perform tasks automatically without human intervention. Some examples are detailed below.

- **Data collection:** in a typical day, a person will get up in the morning, catch a bus or train to work, surf webpages on their devices, go to a restaurant to have a lunch, go to the supermarket to shop, and so on. All these activities generate data and can be recorded by mobile devices or other systems. People are mostly unaware of how much personal information their devices and systems generate, process, or share, and in many cases the consent mechanisms

[iv] The human right to privacy is enshrined in many instruments around the world, including the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights. In Victoria, a right to privacy is provided for by the *Charter of Human Rights and Responsibilities Act 2006*.

[v] Australia has a web of privacy laws, including the Commonwealth *Privacy Act 1988*, and state and territory-specific privacy laws (although not all states have privacy laws). In Victoria, the *Privacy and Data Protection Act 2014* promotes information privacy.

involved in the collection of the data are not well-developed. As more advanced technologies and devices are introduced, more personal information will be exposed to AI without the knowledge or agreement of individuals, increasing privacy concerns.

- **Identification and tracking:** AI has the capability to identify and track individuals across different data sources, which can result in shadow profiles of individuals. Even if personal information is de-identified, AI can relatively easily re-identify the data based on inferences from different data sources. For example, researchers have developed a way, based on convolutional neural networks, to identify and track individual animals by using the animal's movements, without facial recognition.[169] Gait analysis of humans is already in use in some limited circumstances. Such results could eventually be applied to public surveillance of humans by using an individual's movement, which creates significant privacy concerns.
- **Facial and speech identification:** face and voice are two typical signatures that we notice or hear to recognise someone. Successful business products based on face and voice recognition have been developed, and voice recognition in particular is already in widespread use. Many financial service providers are using facial biometrics for authenticating transactions, for example, some banks in China provide ATM services such as withdrawing cash by scanning the face.[170] Facial biometrics are also used for border entry and exit in many international airports, including Australian airports.[171] Such identification helps to improve the user experience from a consumer's perspective (such as easy authentication and fast check-in services) and the performance of authority management for public good. However, the technology also means that people's biometric data is held by the government or other authorities, with the potential that data may then be used for other purposes, leading to the erosion of privacy. Biometric identification applications such as these require very strict privacy management; individuals cannot change their faces, as they can with passwords.

AI enhances privacy

AI techniques pose a threat to privacy, as presented in the previous section. Just as one coin has two sides, AI can also enhance privacy through innovative methods, such as those described below.[172]

- **Reducing the need for training data:** AI requires large amounts of data, which can often be personal information, for training machine learning models. Fortunately, different techniques have been developed for generating synthetic data, which may reduce the need for training data associated with real people, decreasing the privacy-related risks. Using a generative adversarial network is one of the popular methods for generating synthetic data. This meets the needs

of having a large amount of data for the training of machine learning models, without the use of data containing real personal information.

- **Upholding data protection without reducing the basic dataset:** many AI models are trained using personal information, which can be sensitive. Ideally, to address privacy concerns, machine learning models should encode general patterns rather than facts about specific training examples. Other AI techniques can also help to overcome privacy concerns. For example, Google published a library named 'TensorFlow Privacy' for its TensorFlow machine learning framework, intended to make it easier for developers to train AI models with strong privacy guarantees.[173] It is based on the principle of differential privacy, a statistical technique that aims to maximise accuracy while balancing the users' privacy, and can prevent the memorisation of rare details.[174]
- **Enhancing knowledge share without centralised training data:** personal information is usually located on isolated 'islands' such as mobile phones, private cloud storage, private photo albums, and so on. While standard AI models are usually trained with centralised training data, which may cause privacy concerns, new machine learning techniques have been developed to relieve this problem. 'Federated' machine learning allows the training of models to be distributed among users, meaning that training is done directly on personal devices such as phones, so that personal information does not need to leave its 'island'.[175]
- **Avoiding the 'black box' issue:** AI is a 'black box' for general users, in that it accepts inputs and generates outputs, but does not disclose its internal working. Users do not know how the inputs are processed and how the outputs are generated. This is a challenge for both people who use AI systems and those whose data is used by systems. This can result in privacy concerns, as a common principle in privacy law is that organisations are expected to be transparent about how they use personal information. Explainable AI (**XAI**) is a new approach that tries to make the machine learning process understandable by giving explanations for AI outcomes. XAI attempts to explain how the training data is processed by a decision tree algorithm, to get outputs that let users understand that their data is processed in a way that does not – among other things – compromise privacy.

Barriers to AI

While we continuously find ourselves coming across appealing AI-based systems that seem to work (or have worked) surprisingly well in practical scenarios, AI is currently still facing significant barriers in user acceptance.

Human trust in AI

A major barrier to AI is the human trust in AI technologies and AI-based solutions.[176] In their paper 'Trust in automation: Designing for appropriate reliance,' John Lee and Katrina See of the University of Iowa defined trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability".[177] This definition shows that uncertainty is tightly coupled with trust. Uncertainty is a common phenomenon in AI technologies and machine learning, which can be found in every stage of learning, from input data and its pre-processing, to algorithm design, feature selection and model evaluation. In addition to uncertainty, the trust issues in AI are furthered by the black box nature of machine learning techniques, where users are unaware of what is going on inside a machine learning algorithm and how the prediction results are based on input data. Further investment is required to make AI explainable and transparent for trustworthy decisions driven by AI.

Data issues in AI

An abundance of high-quality data is critical for AI training systems. Poor data quality, such as sparse or missing data, is a key obstacle to the widespread adoption and high performance of AI. The saying 'garbage-in, garbage-out' has plagued analytics and decision making for generations, and this is especially an issue in AI. Data availability is also a big issue in some special areas for training AI with high accuracy; if limited data is available for model training, the overall AI model performance will be decreased, exacerbating risks of overfitting and low accuracy.

Government regulation of AI

With the continuous growth of AI uses, governments are increasingly expected to legislate or regulate the adoption of AI and use of data. The ethical implications of AI use in government is also a major focus of concern.[178] CSIRO's Data61 has developed a discussion paper to inform Australia's ethics framework for AI.[179] Similar frameworks have also been developed in other jurisdictions. For example, New York has reviewed key systems used by government agencies for accountability and fairness, and Germany has developed government-led advice on the ethics of automated vehicles. The regulation of AI is discussed in greater detail in the final chapter of this book.

Other barriers to AI

Aside from the issues met by AI as mentioned above, there are also other barriers to adoption and use of AI. The automation driven by AI has an important impact on both the future of jobs, and skills potentially required to design and implement AI. Employees are afraid of losing their jobs to automation, while employers are also worrying about the difficulty of finding people whose skills and capabilities are best matched to AI-driven work.

Additionally, the adoption of AI tends to be concentrated in relatively digitised industries that have access to massive amounts of data collected by their own infrastructures. Take Amazon, Alibaba, and Facebook, for example. Broader adoption of AI in different domains and especially in smaller firms could be important to drive improvements in product quality, performance, and markets. The barriers to smaller organisations adopting AI mainly lie in the lack of data availability, and a lack of appropriately skilled individuals working within these sectors.

Conclusion

The adoption of AI within government is still relatively low. Possible reasons for this could be ethical concerns such as fairness, transparency, explainability, accountability, and privacy, amongst the other possibilities noted throughout this chapter. For example, can AI be prevented from conscious or unconscious bias based on historical data? Can users accept decisions based on logic from black box deep-learning models?

While the strong capabilities of AI in prediction, automation, planning, targeting, and personalisation could deliver a revolutionary change in both the efficiency and effectiveness of government services, the adoption of AI within government is not assured and will need further work before it can succeed.

As we have shown, AI has powerful capabilities in prediction, automation, planning, targeting, and personalisation. But more attention needs to be paid to ethics and legal issues as well as social concerns related to AI. Government will play a central role in setting up ethical frameworks and policies to relieve these concerns and balance them against the public interest benefits of AI.

Biography

Professor Fang Chen is a thought leader in AI and data science. She has created many world-class AI innovations while working in Intel, Motorola, NICTA, CSIRO and now at the University of Technology Sydney. Professor Chen has also helped governments and industries utilise data and significantly increase productivity, safety and customer satisfaction. Through impactful successes, she has gained many recognitions such as the ITS (Intelligent Transport System) Australia National Award 2014, 2015 and 2018, and NSW iAwards 2017. She is the NSW Water Professional of the Year 2016, and National and NSW Research and Innovation Award recipient by the Australian Water Association. She received the Brian Shackle Award 2017 for 'the most outstanding contribution with international impact in the field of human interaction with computers and information technology. She is the winner of the Oscar Prize in Australian science – Australian Museum Eureka Prize 2018 for Excellence in Data Science. Professor Chen has 280 publications and 30 patents in eight countries. She is the Executive Director Data Science and Distinguished Professor, the University of Technology Sydney.

Dr Jianlong Zhou is a senior lecturer at the University of Technology Sydney. His research interests include human-centred AI, interactive behaviour analytics, human-computer interaction, machine learning, and visual analytics. He has extensive experience in data-driven multimodal cognitive load and trust measurement in AI-advised decision making. He leads interdisciplinary research on applying human behaviour analytics in trustworthy and transparent machine learning. He also works with industries in advanced data analytics for transforming data into actionable operations, particularly by incorporating human user aspects into machine learning to translate machine learning into impacts in real world applications.

ALGORITHMS, NEURAL NETWORKS AND OTHER MACHINE LEARNING TECHNIQUES

Richard Nock

It is hard to describe just how successful advances in machine learning have been over the past years, the field has reached a point where people refer to what is happening as a 'Cambrian explosion' of machine learning.[180] The geologic hyperbole of machine learning has been supported by a wide spectrum of specialists, from market analysts, to CEOs of major tech companies, and even high-profile machine learning researchers themselves.[181]

The metaphor is interesting for its implications: if we subscribe to it, then we imply that (i) there was a 'before', (ii) there is a reason for this Cambrian explosion and most importantly, (iii) there will be an 'after'. The earth's Cambrian explosion radically changed the planet forever. What should we expect for machine learning?

In this chapter, I will describe those three eras of machine learning in three parts, admitting the partially speculative nature of the third one.

The birth and motives for machine learning

Leslie Valiant, founder of modern supervised machine learning

The field of machine learning was born with computers as theorised by Alan Turing. The concept of a machine that could automate calculus was soon associated with the idea that it could be used to simulate intelligence.[182]

Statisticians developed the predictive power of data over decades, but it was only after Leslie Valiant in the 1980s, that different pieces could be assembled in a theory mixing both the computational machine and the mathematics of prediction.[183] Valiant's theory

was intuitive: a machine that learns would use an algorithm, a program, taking labelled observations as input and returning a *classifier*. This classifier would encode the way to predict the label of an observation.

How might we make the difference between good and bad classifiers? It seems reasonable to require that classification has to be accurate on the set of labelled observations it was trained from. Valiant's model adjusted this constraint in a more interesting direction, one dealing with *generalisation ability*: the classifier has to be accurate on the whole domain from which the training sample was sampled, with high probability.

The difference is subtle but fundamental: if the classifier we get predicts whether the profile of a job applicant (an observation) is a good one for an interview (the class), then we will want this classifier to be as accurate as possible on all applicants, not just the ones that we had in the database that was used to train the classifier. Because it seems unreasonable to require good generalisation systematically (our training sample may be poorly representative of the whole domain), we just require good generalisation with sufficient probability.

Historically, classifiers were simple: in one of his seminal works, Valiant was just considering simple sets of 'if-then' rules, remarking that humans tend to express their ideas using simple symbolic concepts: *if the polygon has three edges, then it is a triangle*.

Valiant's model made the assumption that the source of randomness in the data set being analysed does not change. This was reasonable at the time it was made, but it would have implications later when new methods of machine learning became available.

Valiant's model captured the essence of *supervised learning*: the training sample contains an observation whose label is given to the machine.

To explain this by example, let us elaborate on our introductory example above and look at machine learning in the context of a hypothetical recruitment process. Observations could be the description of first round job applicants to a company, which might have been collected by a standard questionnaire or populated from resumes: age; gender; marital status; postcode; activity; diplomas; past experience; current salary; and any other variable that could be easy to collect. Many of these observations would come from employees of the company, for which it therefore had work history and, in particular, a record as to whether this work history depicted a good fit for the job or not.

A supervised learning algorithm would then take this labelled dataset as input and output a classifier to decide whether the answers to the questionnaire describe an

applicant potentially of good profile for a first interview. Instead of a binary answer, we could also ask the machine to predict a number, say between 0 and 10, to represent in a more precise way, the goodness-of-fit of the candidate – 0 denoting a poor fit and 10 a perfect fit.

One might imagine that a system that would be good at classifying candidates for a first round of selection could potentially just replace a hiring panel for a second round of selection, because after all, the task would also be a supervised learning problem, the outcome of which would now be to make an offer or decline (and eventually quantify the offer). The input for this stage would be significantly more complex because it would consider candidates' feedback from the interview, not from their resumes as in the first step. Instead of asking basic questions about age, gender and the like, candidates might, for example, face Rorschach inkblot tests during their interview, for which they would have to give a description. They could be asked to draw a figure on a particular topic, draw a person standing in the rain, or answer technical questions about the job for which they are applying.

All this could easily be performed automatically; the candidate interacting with the machine using a simple device like a tablet. All the data stored would then be processed by a model more complex than the one in the first round of applications. The business has a history of hiring, and therefore a history of who was successful (or not) in their job inside the company. This process would represent *in fine* the exact same kind of supervised learning problem as the one used in the first round – predict whether a given profile is going to be successful in the job.

There is obviously a huge difference in the inputs to the model – Rorschach figures, drawings, and free-form texts are more complex in nature than a resume, which is (more often than not) subject to formatting designed to be immediately appealing to a department of human resources.

Two standard frameworks for machine learning: supervised and unsupervised

For the moment, let us just step back in the process to our first application round. Simple if-then rules were not necessarily the standard: at the end of the 20th century, decision trees were very popular, and are still popular today because they happen to be relatively simple for a machine to learn, and are easy to understand by humans. In the case of our interview example, a simple decision tree that could be used to decide to proceed further with an applicant is given in Figure 2. Interpreting the tree is very simple, and even transcribing it in sets of if-then rules is straightforward: in the case of Figure 2, the tree gives us three mutually exclusive rules, each of which proceeds from the root test of the tree on gender, to a leaf deciding the interview. For example,

reading from the top (root) of the tree, we get the rule: *If gender is male and education is at a lower level than PhD, then we do not proceed.*

Starting from this simple example, let us focus on the types of problems on which the whole field of machine learning has been created.

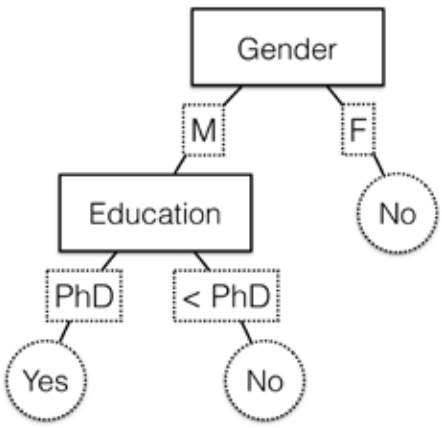


Figure 2

A simple decision tree to predict whether or not to interview a person, here based on two variables. Classification proceeds from the topmost test, which here questions the gender and then, if the applicant is male, questions his education. Essentially, only male candidates with a PhD would be recommended for interview by such a decision tree.

Supervised learning has always been an important component of machine learning – and is still a key component of the field. Another method is called unsupervised learning. In the kinds of cases for which *unsupervised* learning is utilised, we do not have labels, so the task is not so much to predict a class, but rather to organise the data according to patterns that the machine is left to find, giving it an objective that is in general very loose compared to supervised learning. One popular way to carry out unsupervised learning is to divide the data into a fixed number of clusters. To return to our interview example, the department of human resources of the company might just want to split a large set of resumes into a number of subsets matching the number of human resource employees who will be looking at the resumes; it would then make sense to ask the machine to make those subsets as homogeneous as possible so that each human employee really compares apples with apples, for whatever this notion might mean. In this example, the company might just leave it up to the machine to decide how to construct those homogenous subsets.

Beyond the standard frameworks

Supervised and unsupervised learning have been the foundation of the field of machine learning and they are still driving the field today. While both make sense as methods to be used in the example of hiring people, they were, even back in the 1980s and 1990s, not the only frameworks people were interested in. Early on it became apparent that a host of variations were necessary to capture the needs of many applications that were not fitting exactly into the supervised versus unsupervised picture.

One such important case related to supervised learning is *on-line* learning. In our hiring example, supervised learning is a *batch* operation; we can have a huge number of resumes and ask the machine to train a model that is going to be used over potentially a very long time. We might retrain a model after a number of new candidates get into the system to refresh it, make it fit to the current market and new profiles better, but it would clearly make little sense to retrain the model from scratch after *each* update to the database, after *each* resume has been submitted to the company.

This is exactly what matters in on-line learning: suppose our database consists of past history of a portfolio of goods alongside their returns over decades, for example using the Standard & Poor's 500 index. In this case, it would clearly be a terrible mistake to train a model to decide whether a stock is going to go up or not in a short horizon, and then leave it to decide allocations for a long period of time without any update to the model. In on-line learning, the model has to be updated after *each* update to the input: we update our portfolio or the predictions after *each* market update.

In the 1980s, we did not have the constraints imposed today by high-frequency trading, but the *framework* of on-line learning was already elaborated in the context of machine learning and under the scrutiny of researchers.

In the case of unsupervised learning, as applied to our recruiting example, we might imagine a further problem: that the company would like to do more than just organise its complete database of resumes. Maybe there is *that* candidate in the database, this person is different from all others, and their profile would be a perfect fit for an unusual kind of job. Isolating such an *outlier* is the purpose of *outlier detection*, which is arguably different from general purpose unsupervised learning. This refers to a popular set of techniques born in the 1980s and 1990s, named *anomaly detection*, because what we are looking for is the part of data that clearly departs from the mainstream sample, either denoting fraud (for example in credit card transactions, or votes), severe weather patterns (climate analysis), or intrusion in a network (hacking).

Reinforcement learning and the origin of 'machine learning'

On-line learning is an important model of learning because it puts the machine in an environment which is susceptible to feedback, to which it has to react, update its model, make it more accurate, and better fit to the objective.

It may be sufficient to deal with simple models of interactions as in our (over)simplified portfolio selection model; it is, however, way too simple if the machine is supposed to receive much more complex forms of interaction from the outside world, as would be the case of an autonomous robot wandering an office for its surveillance, to clean it, or to distribute mail to humans. When the machine is interacting with an environment and needs to figure out a complex policy, not just a simple model, to maximise rewards in interaction with the environment, the design of the machine learning algorithms belongs to another field, *reinforcement learning*. The robot may just start its task by knowing little of the best strategies available; we are going to ask the machine to *learn* those strategies. For example, in a hot-desk or flex-space organisation, the machine could have to learn to adapt to day-to-day changes of the floor plan occupancy for best cleaning, or optimal surveillance.

Interestingly, reinforcement learning did not meet with early fame in the robotic domain, but in a domain that inspired a whole field of artificial intelligence: board games. This domain is at intermediate complexity level, certainly not as simple as the database of our hiring company and not as complicated as for our office robot.

The case of board games is interesting because it sparked the very first allusion to a general definition of machine learning. In the late fifties, artificial intelligence pioneer Arthur Samuel wrote, in the abstract of his paper on making a program that learns to play Checkers, that the objective was: "a computer can be programmed so that it will learn to play a better game of Checkers than can be played by the person who wrote the program".[184]

Later, a broader definition emerged, which can be summarised as the ability of a computer to learn how to solve a given task from past experience. In his seminal paper, Samuel developed search algorithms that bypassed the combinatorial difficulty of the game by locally estimating a score function used to prune the search for the best moves,[vi] instead of trying to achieve the impossible task of computing all possible plays until the end of the game – a task that could only be completed in the 21st century after almost two decades of number crunching.[185]

[vi] This could be the number of pieces of the player left on the board after a limited series of rounds of play, or more complex functions as in Samuel's original article.

Samuel's approach was purely algorithmic: for a human, the difficulty of calculating winning options in a board game stems from the impossibility of calculating all possible combinations of plays in order to pick the best. However, the computer sees the complete state of the world in which it operates. Unlike a game like Poker, where the state of the game is partially hidden for each player, a board game operates on what is called *perfect information*. In this sense, it is a long way from the hiring company in our recruitment example, whose objective is to also come up with a model that is going to be accurate on *unseen* data, because in the recruiting case, the impossibility resides in the unavailability of the resume information of all possible candidates on the planet, as well as for their potential fit to the job at hand. If such a complete set of information were available, it would be much easier for the company to find the best hire, all the more as the maximal number of applicants to the job would still be billions of times smaller than the number of possible board positions in Checkers.

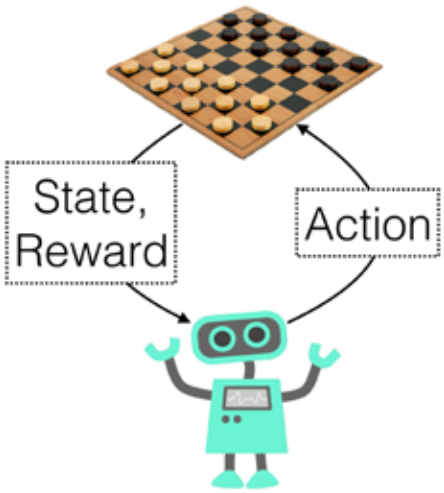


Figure 3

In reinforcement learning (simplified picture), the machine perceives the state of its environment and receives rewards from its own actions. The goal is to learn a policy, mapping states to actions in such a way that rewards are maximised through a sequence of interactions.

Through these examples of problems that early machine learning researchers have focused on, we can better understand the early preoccupations of formal learning models: manage the potentially huge number of possibilities and come up with a solution to the problem within a reasonable amount of time; in general, a model. This model is going to be as good as possible given the uncertainty coming from unseen data. To be more rigorous, we could make the convenient assumption that the data we have has been randomly sampled and that this source of randomness never changes; our outlier candidate will always be an outlier, and the reason why we have come to observe him is independent of the observation of any other candidate. The 20th century history of machine learning has been deeply influenced by this 'static' vision of learning, which is in the foundations of Valiant's model, a model that contributed to his winning the 'Nobel of computer science' in 2010, the ACM Turing Award.

One step further: Deep Blue

The (board)game between humans and the machine that started with Samuel's Checkers example became famous in a subsequent step that achieved spectacular results in learning in highly complex environments: *custom hardware*. IBM's Deep Blue was focused more on how to get the machines to operate on proper hardware than on improving the state of the art in algorithmic decision making. In Deep Blue, the 'machine learning' part was reduced to a core not so different from Samuel's search ideas, but the hardware was custom and pushed to its limits to implement the search in parallel and with much better efficiency, with the objective to beat the world champion of that time, Garry Kasparov. In the Deep Blue story, an official hallmark of modern machine learning was carved; it was not sufficient anymore for the machine to beat its programmer, as in Samuel's paper – the machine needed to display superhuman capacity in solving its problem. While it was clearly not Samuel's objective, a team of Canadian researchers in the 1980s took over the objective of making a machine the world champion of Checkers, and was later recognised as achieving a first in the genre.

Let us return however to consider reinforcement learning, to unveil one of its core challenges. Samuel pioneered some of the early techniques of storing the past and trying to generalise from this past to forecast the future possibilities for the game.[186] In the more general setting, even if just for a more complex game like the ones we have seen since the advent of personal video gaming systems, the machine needs to be in constant balance between two competing objectives: *explore* the environment or *exploit* its current strategy. In the former case, the machine gets to know its environment better, but may lose rewards by making suboptimal choices. In the latter case, the machine uses its current model to take an action that supposedly is going to give sufficient reward given its past actions, but it may miss the discovery of a particular feature of the environment that could have led to even greater rewards.

Very often, the game used to display this dilemma is Bandits (slot machines). Imagine we built a machine to play. The machine is in a casino, facing a set of different bandits, with an objective to earn the largest amount of money by repeatedly choosing a bandit to pull its arm. Exploration, in this example, is the ability to test different bandit machines and exploitation is the ability to stick to the machine that has given the largest amount of money *so far.*

Lightweight summary

Ignoring subsidiary issues like on-line learning or anomaly detection, there are common elements in dissimilar methods such as supervised learning, unsupervised learning, and reinforcement learning.

1. The inputs are of the same kind: data which encodes the knowledge of the past; the current state of the machine's environment; and eventually the rewards, mistakes or failure achieved by the machine.

2. Learning requires the machine to be fast in its computations and accurate in its decisions, whether they are classifying a person as hireable, a move as winning, or a candidate as having a specific profile.

3. More importantly, learning requires the machine to learn parameters about the world.[vii] More often than not, it consists of a *model*, which is just meant to be a representation of its current knowledge about the task at hand. This can be a set of numbers representing how worthwhile a move in Checkers might be (the higher, the better), or a decision tree capturing the essence of a good or bad hire. In all these cases, the numbers are not encoded by the person who writes the program but are fitted to the model by the machine. The decision tree is not given to the machine; the machine is tasked to find it.

4. There is obviously a catch in item 3 above. Leaving the machine to wander around without giving it a goal would surely result in something barely better than a random prediction, and we would end up with a potentially very expensive unbiased coin. In fact, in absolutely all these cases – all these examples, all these domains of machine learning – the programmer of the

[vii] Interestingly, some machine learning techniques are exceptionally lazy; they do not learn anything. In supervised learning, this is the case for one of the oldest 'algorithms' which would, for example, classify a candidate as good to hire by just looking at the closest known profile in the history database and attributing the same score to the unknown candidate as that of the known one in the database. Such a rule is called the nearest neighbour rule and was born in the early 1950s (see Fix, E. & Hodges, J. L. (1951). 'Discriminatory analysis, non-parametric discrimination', Report 4, Project 21-49-004). One might think that such a strategy is exceptionally poor if the dataset at hand is small — imagine our candidate database contains a single labelled observation: every new resume would just be classified in the same way. What is, however, totally counter-intuitive, is that this simple rule becomes extremely competitive as the dataset size grows, leaving us with the task to find a way to efficiently store and query this potentially huge database (hint: almost nobody would in fact do that.)

software or designer of the algorithm always starts with an *objective function* that encodes the quality of any potential solution to the problem, without ever explicitly giving the best one to the machine. There is no exception to this rule in machine learning; it is the goal of the machine to figure out how to get a good model, a good prediction, a good output with respect to this objective function. The design of this objective can be very intuitive and simple; we could just ask the machine that learns our decision tree to minimise the errors its learned tree makes. The objective function is then simply the error proportion on the training data. Our machine exploring bandit arms in its casino could be required to maximise the dollar amount of its total play. A subtler objective could be to require the machine strategy to come up close to the best possible strategy, since the dollar amount does not in fact reflect the difficulty of the task at hand in the machine's environment (maybe the bandits work purely randomly in one casino and are completely rigged in another one).

The missing piece of the machine learning framework

There is also a catch in item 4 above, but subtler: giving the machine an objective function is typically not enough to have a workable solution to our problem. In general, one has to give it the basics of how to make the best of the objective function, to determine how to *optimise* it. Consider the example of a child to whom we give a metal detector with the objective to find coins and other useful metals lost on a beach. The objective function is obviously a mix of fun and to maximise money, but the task would not begin without us explaining how the metal detector works and guiding the child on the best places where such target objects could be hidden and how to properly reach them, eventually concluding with some hints. The child would then be left with its own defined model of the beach, and progressively learn the best way to manipulate the detector, and eventually the best or worst places to find interesting metals.

It is the same for any learning algorithms: we would indicate to our algorithm to build a decision tree from scratch and make it grow until it properly fits the data.

The algorithmic and statistical part of machine learning was augmented by a third field of mathematics which would later prove instrumental in getting the best training algorithms even for very complex models: *optimisation*. Such techniques typically just give local strategies to the machine on how to make a better model from its current one, leaving it to the computational power of the machine to then build the complete model from the repeated application of this basic 'hint'.

Towards more complex models

In the 1980s a paper was published by David E. Rumelhart, Geoffrey Hinton and Ronald J. Williams. Titled 'Learning representations by back-propagating errors', it identified

useful methods of training models that mimic the neural networks in the brain.[187] It was recognised three decades later as foundational for the whole field of computer science through the ACM Turing Award in 2019.[188]

In the 1990s, there would have been another common element in all the examples above: the model learned was, in the worst case, relatively simple to understand, and based on data that was simple to represent. It is probably obvious by now for decision trees or simple if-then rules. It would also have been the case for Checkers — we just need to store an 8 x 8 array with each value specifying one of three possible values (empty, black or white). It would also have been the case for our hypothetical databases of resumes, each of which probably reduced to a list of important variables, such as gender, age and education, with specific values for each of them. While the calculations involved in modelling outcomes were often beyond the capability of people to do themselves, the outcomes were interpretable after they were derived – we could understand how the models were obtained.

During this period, other work pushed the boundaries of the field, analysing much more complex data, typically text, sound or images. In several notable examples, researchers wanted to teach computers how to recognise objects in images. This was computer vision, which became a focus for automation of classification. The state of the art proceeded in two steps, including – in the first step – the automatic extraction of features from the image, features that would then be used to train a classifier in pretty much the same way as for any other classification problem.

The top image in Figure 4 presents a very schematic view of the overall recipe. Researchers circumvented the complexity of the data by guiding the machine towards working on carefully engineered and simple features that could be extracted from the image. Such an approach may be fine when no other proposal exists on the table, but it contains a pitfall: engineered features inevitably contain human bias. We impose on the machine our own understanding of the domain at hand – for example, what part of an image we think makes an 'A' look like an 'A' – which can be highly suboptimal and force the machine to learn models in the subsequent stage that are not as good as they could be.

The question to be asked then is *whether it is possible to dispense with the human part* in the task at hand and let the machine figure out its own way to learn not just how to classify data, but also how to learn the key features of an image that best encode the class.

Neural networks

This more complex task was solved two decades ago using a model representation closer to the one we supposedly use at the analytical level in our brain: neural networks.[189] It probably sounds surprising today that neural networks could be so successful in the 20th century but then be followed by more than a decade of relative quiet; we shall see later why this eventually happened. The architecture of this early achiever is represented in the bottom image of Figure 4. Given the task of handwritten character recognition, the machine managed to learn a neural network achieving less than 1% error on testing, which is not just very good, but in fact allowed the technique to be used for substantial industrial deployment.[190]

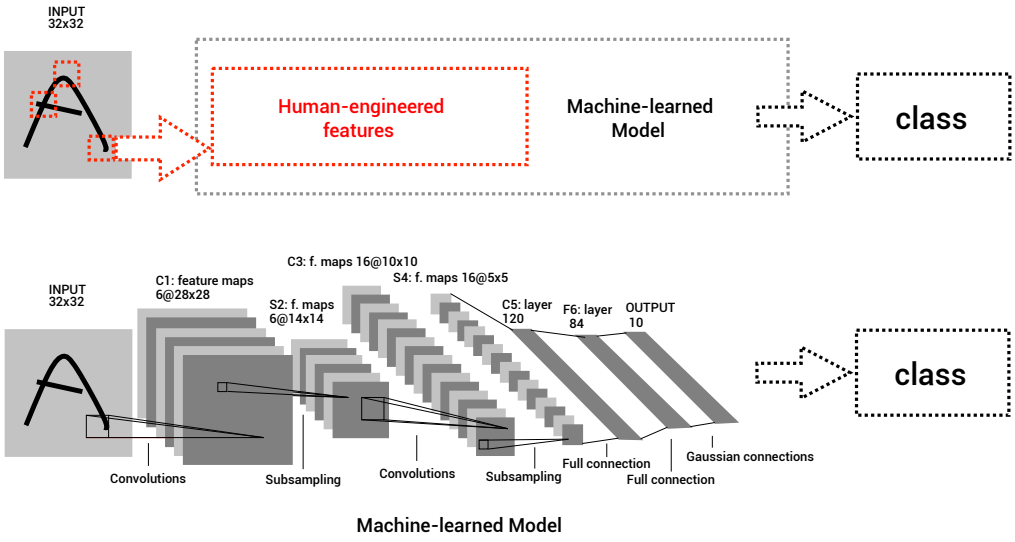


Figure 4

Top: Classifying an image had been historically done by a two-stage process, whose first step was to compute features from the raw image carefully engineered and optimised by humans (also called a feature extraction module). Learning a classifier was then based on these extracted features as input, rather than the raw image.

Bottom: LeNet5 was among the first attempts to get rid of this human bias in the process and let the machine decide by itself the best ways to learn a classifier directly from the image, using neural networks. Architecture taken from LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). 'Gradient-based learning applied to document recognition', Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278-2323.

The principle of a neural network is simple: it assembles simple basic functions, neurons, that are not much more complex than a local decision in our decision tree. Each neuron takes input from others and computes an output signal that aggregates all

inputs. Its output signal is then used as input for another neuron. This is an abstraction of the processing happening in our brain, but this local abstraction is simple and in fact not where the power of the whole network lies. The key to training a powerful neural network is its *architecture*, the global organisation of all neurons, typically in layers (seven in LeNet5, depicted in the bottom image of Figure 4). The layered design has this very intuitive notion that the machine is going to progressively learn an abstraction of the input features, towards new features that are good for the classification task at hand. In doing so, the machine is supposed to progressively bypass the step of human-engineered features by learning its own representation of the task. The power of the machine is essentially the ability to very carefully optimise this step, by considering a colossal number of possibilities in order to keep only the best one.

All that is left to the human is the design of the architecture, and then letting the machine learn the crux of the model – the weight of each connection from one neuron to another one. This very roughly approximates the way a human would learn, with the brain adjusting connections between neurons throughout learning. In LeNet5, the key part of the architecture is what is called *convolutions*, which requires some neurons to be receptive to only a small subset of the neurons in the previous layer, inspired by studies in the brain for vision. Such neural networks are called *convolutional neural networks*.

Applications using simply defined data flourish in the real world. In the 1990s *data mining* involved machine learning work prior to the progress of LeNet5. Perhaps the most prominent application targeted early by data mining was the general analysis of the shopping basket – requiring only a flat collection of transactions and therefore data represented in a much simpler manner than vision, speech or even text. Two decades later, convolutional neural networks would be recognised as a major landmark in machine learning. LeNet5 made it possible to analyse more complex data than just flat credit card transactions or simply defined resumes.

The bottleneck to scaling-up machine learning

It may come as a surprise that machine learning in the first decade of the 21st century was relatively quiet compared to today's activity. There is an explanation for this: nobody knew back then how to train neural networks substantially 'bigger' than LeNet5.

To grasp the importance of the challenge, consider that the brain analogy suggests that the source of the 'power' of a neural net lies in its ability to progressively learn and model abstractions of the features of the world in its *layered representation*. This is very natural: we would not characterise a bird by the local colour of its body parts but by higher-order features that can then be used to compare a bird with other animals, such as its feathers, wings, and beak. Once one realises that the source of such higher-level features comes from parts of the animal that are spatially related (one feather is

not split throughout the animal's body, but stands as a local description of the animal and is very useful for guessing that it is a bird), it does not take long to realise that this property also holds for other categories of complex data that humans process very well, such as texts in natural language, speech, and music.

In fact, the power of neural networks to carry out such higher understanding of natural language processing was also discovered in the 1990s.[191] It turns out that it also relied on a trick to capture, in the architecture, a specific property of data that we humans exploit to understand a text (or other kind of data for which this property holds, like music scores): the spatio-temporal dependencies that can be observed between words or sentences in a natural language written text. Expressed very roughly, the closer two words are in a text, the more likely they are to belong to the same grammatical or semantic unit.

Since we now understand why the architecture and its layered representation is key in neural networks – to model data that could be hard to model using, for example, simple if-then rules – we can return to our problem and can make it a bit more specific: how can we train not just bigger, but in fact *deeper* neural network architectures?

It took more than a decade to make a breakthrough that, by proposing the first scalable solution to this question, revolutionised computer science. It came with a new nickname: deep learning.

2012

If the analogy with the Cambrian explosion is appropriate, then 2012 is the year it all started, and it all started with a competition, but not (yet) with humans. Beginning in 2010, a large-scale image recognition competition was run using a now famous database, ImageNet.[192] The scale of the problem made it orders-of-magnitude more complex than the one solved by LeNet5: the dataset contained more than 1,000,000 images, with 1,000 different classes.

As in any competition, one would expect the top expert contenders to be really close to each other; such competitions – now popular in data science – happen to encourage new neat ideas to come forward and improve, even incrementally, the state of the art.

Things did not exactly happen this way for the ImageNet competition: in 2012, the winners delivered a model whose error almost divided by two the error of the runner up – while the previous year, the difference with the runner up was just a few percent. The competition was essentially a repeat of the LeNet5 achievement, but on a scale that virtually nobody could imagine: the runner up used human engineered features

(called SIFT) while the winner was, as with LeNet5, replacing the two-stage process with a single pipeline in which the machine crafted its own features while learning its deep neural network.

Getting such a big difference from the runner up took more than just one neat idea, especially considering that the final neural network had up to 60 million parameters and more than half a million neurons. In fact, it took two sets of new ideas to get there: a set of powerful new ideas on how to train a deep network, and the use of a hardware component that is now fundamental in training deep neural networks – Graphics Processing Units (instead of the classical Central Processing Unit of a computer). In other words, it took better algorithms *and* better hardware to get such results.

This breakthrough was experimental, but it reshaped the whole field of computer vision in the following years, to a point where many of the contributions of the leading computer vision conferences converged on the design of deep learning algorithms. The age of feature engineering as it had been done, and for the purpose it was designed for before 2012, was over.

What happened in computer vision was soon to happen in other fields and for similar reasons: text, natural language processing, speech, sound, video, network analysis – as in social networks. All these fields reimplemented the key feature of deep learning, which is essentially to give the machine the ability to learn its own features from raw complex data to solve the problem at hand, instead of relying on humans to 'pre-digest' those raw features into 'machine-readable', 'usable' ones. Returning to our recruitment example, if our hypothetical company wanted to design its second stage of interviews, including commenting on Rorschach inkblots, free-form drawing and text, it could utilise this new technology, and then eventually it could (in theory) rely on a machine for its analysis.

This was arguably the start of the deep learning revolution. From this starting point, deep neural networks not only started to be even deeper; they started to be used for more and more problems, soon reaching any number of sophisticated applications – autonomous driving, automatic translation, intelligent assistants, chatbots, and beyond – reaching whole scientific fields or industries including climate, health, finance, biosecurity, insurance, banking, entertainment, gaming, telecommunications, infrastructure, defence, social and political sciences, social networks, etc. This list cannot be exhaustive. To get an idea of where the applications are today, or what the applications could be tomorrow, keep in mind that wherever there is data, there is potential input for machine learning.

During the International Conference on Machine Learning that was held at Stanford University in 2000, conference chair Pat Langley made the joke that it was time to step from machine learning to machine *earning*, meaning that the field had to level up

its game for industrial rewards. This is certainly not a joke anymore, and this raises a number of issues today, regardless of what we take these earnings to be and whoever gets to enjoy them. A subtler problem is that any user of machine learning needs to be careful about the use of the technology itself and be warned that using the outputs of machine learning does not go without consequences, including highly unexpected ones, as we shall now see.

A new era for machine learning

Biased predictions and fairness

Let us step back for a moment: the reader might have already remarked that the picture of machine learning displayed so far – a field driven by a very strong technical backing to solve problems that matter – may in fact display weaknesses in the models it can learn.

If that is not the case, let us look back again to the decision tree in Figure 2. Another rule it yields is: *If gender is female then we do not proceed*. We conclude that if the machine gets to automatically process applications and reply to candidates for a first interview, then no female is going to show up at interview time, and no female is ever going to be hired as long as this model is used. If this decision tree were a real one, its impact would obviously pose a problem of fairness and discrimination. This example was crafted for the purpose of this chapter, but it turns out the problem described is real, and it in fact actually happened at a big tech company.[193]

Why this problem occurred is obviously the next question to ask, and the answer is simple: machine learning algorithms are not discriminatory on purpose, but they can be so good at learning that they manage to learn even the bias in their data, whether it discriminates against women, people of colour,[194] or against other qualities. Remember that one needs to give the machine an objective function to optimise the machine to learn that a particular model is good with respect to *this* function, *and one only gets what one wishes for*: can we blame a model for being unfair when in fact the source of unfairness may just come from the simple fact that the original bill of specifications for the machine learning algorithm did not include fairness in it?

In fact, this is not just about the goal assigned to the machine, but also about the freedom or constraints we give for the machine to learn in an environment which can rapidly escape any decent control. It took less than a day to transform a neutral chatbot learning from Twitter interactions into an absolute racist.[195] Such an event raises the question of accountability in a number of ways.

Why this is happening

At this point, it is useful to recall that the original bill of specifications for machine learning algorithms, as developed by Valiant, essentially contained the requirement of accuracy. This is just fine if the algorithm is supposed to learn a model to predict whether a board is winning or not in Checkers. This is just fine if the algorithm learns a model to predict whether a flower is from a given species. And this can be perfect if the algorithm predicts whether a plant has a specific disease. This is, however, not fine at all when we ask the model to predict whether a convicted person has a chance of reoffending given their past criminal records – and this is just one example. To understand the difference between the two categories of problems listed here, there needs to be an important metaphor put forward: Machine learning was born in the sterile room of computer science and mathematics.

To progressively reintroduce the Cambrian analogy, the Pre-Cambrian period for machine learning happened in the sterile room. Problems to be solved were just like formal models: simple in design, supported by simple assumptions that would make sense in a general purpose model, maybe naive in the belief that this would be sufficient to solve the biggest problems of the real world. For example, the problem of guessing flower species mentioned above was a popular one introduced in statistics during the 1930s.

In the Cambrian explosion period of machine learning, the whole field has been suddenly pushed out of the sterile chamber to expose its power to solve problems in the wilderness of the real world – its power, its weaknesses and the potential flaws in its deployment. It could have been possible to predict that deploying a chatbot that learns in an environment lacking sufficient control would result in unfortunate consequences. It is sometimes much less obvious to anticipate problems.

Subtle weaknesses and causality

If the discrimination problem in the example of the decision tree in Figure 2 can be easy to catch, some weaknesses can be subtler: the assumption that the source of randomness does not change in Valiant's model is mostly fine when we model games or predict plant diseases. It is absolutely not fine when it comes to health: suppose we have a model predicting whether or not to give a specific jab for a non-lethal condition. Once the riskiest population has been inoculated, if we keep on using the same model, we will just target the same people, whereas the source target of the disease might shift (as a function of weather, living conditions, development or just mutations). This is a case of what is called *distribution shift*.

Researchers are also investigating the extreme case of such shift which is done *on purpose*: train a model on a particular domain to predict a label, and then *transfer*

this model to work on a different domain. Such a *transfer learning* task is important because (i) it allows data scientists to solve several tasks with a single model and (ii) it is particularly useful when the information from labels is not available on the second task – which can happen when, for example, such information would be too costly to obtain.

Let us drill down into some other subtle consequence of applying machine learning, related to distribution shift, but not due to external factors as in our health example. This will explain another reason why some extra care and caution needs to be taken when using machine learning in highly sensitive applications, like the decision to hire people or decide on someone's chances to reoffend. Here another new component of post-Cambrian machine learning emerges: *causality*. Applying a model that is biased for a long time might serve to *reinforce* the hidden bias: women receiving fewer and fewer job offers from our decision tree will inevitably see their proportion grow in unemployment statistics, which will then reinforce any other subsequently trained model from current data into including even stronger bias against hiring women.

Explainability versus the rush for complexity

There are also some much subtler problems than those mentioned above, ones that were left hidden in the beginning of this chapter. We do not even need to apply our decision tree in Figure 2 to realise that the system only recommends men for interviews, and therefore realise after seeing a cohort of interviewed men that the system discriminates against women. It suffices to simply look at it to realise that the most influential variable, the one that appears in all if-then rules built from the decision tree, posits that gender is going to be the most influential feature in hiring people. This possibility, to guess that the model is going to be biased or unethical even before it is deployed, is no longer possible with deep neural networks.

A collateral event of the breakthrough in 2012 on the ImageNet competition was that it pushed for a race towards getting more and more complex models to solve problems: since the source of the breakthrough's result was believed to be its success in training more complex models, why not do the same strategy *systematically*: to get better results on another problem, one should just train more complex, deeper models. This brought about collateral damage of trading *interpretability* for more performance, which may be fine for the ImageNet competition (interpretability was not a requirement of the competition) but it will inevitably create problems if such models are applied in the public sphere, where rules and regulations would typically be developed to prevent this. Such is the framework of the European General Data Protection Regulation.[viii]

[viii] The General Data Protection Regulation, or GDPR, is explored further in other chapters..

Privacy

There are additional problems that do not appear in the first part of this chapter because they do not display a flaw or limit in the design of the early theories of machine learning. They appear because of the context in which machine learning is applied today (this could have been the case of our chatbot).

Consider another example: our hiring company happens to have competitors. Among those, it agrees to collude with one to share information related to their applicants, to learn a model developed from the union of their databases. Since it is trained over a bigger set of candidates, the model should be more accurate than if it were trained using just one of their databases. This is arguably a very strong motivation to share information. However, the companies require that the other (or any other external party) does *not* have access to their data in the clear. Such a constraint, that requires training a model using data that cannot be seen in the clear is called *federated learning*. It is usually addressed by a combination of machine learning and *cryptographic techniques*. Federated learning is also getting lots of attention because it addresses another concern against which early theories in machine learning were not challenged: *privacy*. We are witnessing the birth of marketplaces where data handlers do not share their data but instead share the 'hints' that help to train other peoples' algorithms.[ix] Such hints can be shared in exchange for remuneration and – if sufficient care is given – they should not unveil an individual's personal information.

However, it should be stressed that in the case of federated learning, the requirement to be privacy compliant usually comes with a significant technical levy on machine learning, to make sure that learning parallels the performances of the non-private case, for example, to make sure that the final model is still accurate enough.

Learning and inference everywhere (and an unexpected consequence)

Consider a follow-up example regarding privacy: what would happen if, for example, a person had their personal information on a device (a smartphone) and wanted to run a hiring model directly on the smartphone to check whether they would be a potential hire for a specific company (such a model could be provided by a third party, helping people to find a job). *On-device learning or inference* (which means we just run the model on our device, like in our hiring example) is getting a lot of attention, even in the research community, for the simple reason that even if it is just to locally run a model, one needs to pay attention not just to privacy but also to the constraints of the device, that are not necessarily capable of running models as big as the ones we now see in

[ix] They are sometimes called 'Gradient marketplaces'.

deep learning. Considerations on storage, communication and energy consumption are important on such devices, and such constraints are becoming a major challenge for the field, especially as people are now beginning to consider all possible devices in the Internet of Things. In fact, it was recently revealed that the global energy footprint of machine learning is spectacular, as training some of the most complex deep learning models (with hundreds of millions of parameters) bears a carbon footprint that far exceeds that of the whole life of a car.[196] Because of this, we can expect much more efficient machine learning algorithms, even outside the market of mobile devices or 'intelligent' Internet of Things appliances.

Machine learning in an adversarial world

Another problem that has become crucial given the rapidly growing interface that machine learning has with society and the public sphere at large is *adversarial tampering*. Consider the setting of our hiring company, learning a model using its own data to predict whether a candidate is to be contacted for an interview. Suppose that the algorithm used is accurate and fair, not biased. What could possibly go wrong? One possible answer: *data poisoning*. Knowing the algorithm that is going to be run to build a model, it would be possible to locally influence the predictions of the model it is going to learn, with a simple protocol: figure out the eventual slight changes to make in the database to ensure that the model learned overall looks the same (as it would be without doing anything) but radically changing its prediction on a few targeted candidates, with the objective to make sure they get (or do not get) interviewed.[x]

Worse than local bad results: distorting the fabric of reality

Data poisoning is a simple example of what could come out of the Pandora's box of possible misuses of machine learning, whether accidental or made on purpose. Another example, which has recently made it to the headlines, is a breakthrough utilising the potential of deep learning to *generate* complex data. In this case, the machine learns how to generate new (and realistic) images, sounds, text, and the like. Let us stick to the image case for simplicity. The way these techniques work is interesting in itself. Somehow, they work in *reverse* to the way deep learning was originally designed; instead of taking raw images and converting them to simple machine learned features useful for classification, by passing through learned layers of progressive abstraction, we start from such simple abstract features, typically randomly sampled, and then go the opposite way to create more and more realistic features through sets of layers, until the last layer where, suddenly, a fully realistic image appears. This technique is a *generative model*.

[x] This subject is also covered in detail in *Data security and AI*.

Modern generative models were born in 2014 and were recognised as a breakthrough for computer science as part of the ACM Turing Award 2019.[197] This recognition came even faster than the recognition of the earlier work of Geoffrey Hinton.[198] An original use of the technique came equally quickly: to show that the machine could become an artist.[xi] Unfortunately, also equally fast-paced was (mis)use of generative models, in a now infamous piece of technology that some people believe could threaten the core of democracy: 'deepfakes'.[199]

There is now clearly an arms race around deepfakes, to generate them and detect them, and if the technology is still too expensive for the layman to generate realistic content, it is a completely different story for more powerful actors like state actors.[200] It is beyond the scope of this chapter to explore this further, but it is worth mentioning that the technology was developed initially by somehow *implementing* this arms race in the machine. Indeed, in the original training framework, training involves two competing players – a *generator* (which is the system we want) and a *discriminator*, which is used against the generator. The generator is jointly trained with the discriminator, the latter trying to guess between the generated content and a set of ground truth – if we want a generator as good as Picasso, then the ground truth could contain the complete set of work from the famous painter. As the generator gets better and better, it becomes harder for the discriminator to tell the generated data and the ground truth apart. Ultimately, our generator becomes the perfect forger for new Picasso artwork! Or, if the ground truth contains the set of television interviews of a President, then the generator learns how to generate new interviews that never existed and, with a little bit of experience from the persons running the whole system, the generator can forge not just random interviews but new interviews *with a purpose* – precisely deepfakes.

Not everybody agrees on the potential impact of deepfakes – from classical propaganda to threats of 'infocalypse' and the distortion of reality – but it seems reasonable to believe that, in the same way as many disruptive technologies could be used for opposite (good/bad) purposes, the same may happen in the use of machine learning against the spread of deepfake messages, for instance, using machine learning to detect deepfakes. This will contribute to making trust a fundamental part of the deployment of machine learning.

Superhuman performances and where they are deployed

The deepfakes example shows how machine learning has become efficient in solving the problem at hand. It should be clear from this last part of the chapter that the field of machine learning is now growing *horizontally* as well, bringing more and more (distinct) problems to solve to the table of researchers and engineers.

[xi] For an example, see 'Edmond de Belamy, from La Famille de Belamy'.

The deepfake problem is not the only problem for which machines are reaching human or superhuman performances on complex tasks, but it is fortunately not always a source of concern. On the entertaining side, the successes of Checkers and Chess automation have been followed by renewed interest in reinforcement learning, and subsequent breakthroughs have occurred in which the machine learning part has been substantially improved – not just the hardware component as was essentially the case for IBM's Deep Blue. One such breakthrough, AlphaGo, which again uses deep neural networks, achieved the remarkable ability to be able to train a machine Go player without any other information than the game's rules to start with, training itself from the sole observation of games. It was able to reach superhuman performance in just a few days of self-training.[201] There is little doubt that these recent advances in reinforcement learning will have significant impact in other fields, in particular, robotics.

On the more sober side, we now know that just an excerpt of Facebook data can basically allow a machine to know us better than our own family.[202] Independently of the considerations of this chapter, this invites a different kind of question than the ones classically asked when a data breach happens, namely, *what could be achieved with this kind of data, what could we do with it, and what could be learned from it?*

Still, we need better machine learning

But the machine is – unfortunately – still not perfect in circumstances where we wish it were. For example, we know that deep learning models are sometimes brittle to classification:[203] slightly altering a road sign with a change that would make no difference for a human can produce dramatic changes in the output of a deep neural network for computer vision. Making machine learning more *robust* is a very important challenge for the field. The *application* of machine learning in such areas as autonomous cars will also be an important challenge for regulators.

After the Cambrian explosion of machine learning

It is appropriate at this point to come back to the Cambrian analogy, and now try to complete it, as shown in Figure 5. We now know better what happened during the Earth's Cambrian explosion, and it is easy to make a more complete analogy with machine learning, where oxygen becomes data and the technology gets to conquer a dimension of technology previously unavailable, because the proper infrastructure for data collection and storage, and the necessary computational power, was not available. There is, as shown in Figure 5, considerable heat and excitement in the field, as exemplified by the fact that one of its two major conferences (NeurIPS, 'Advances in Neural Information Processing Systems') was sold out faster than some rockstar concerts in 2018 – and, it turns out, for a large crowd of 8,000+ registrants.

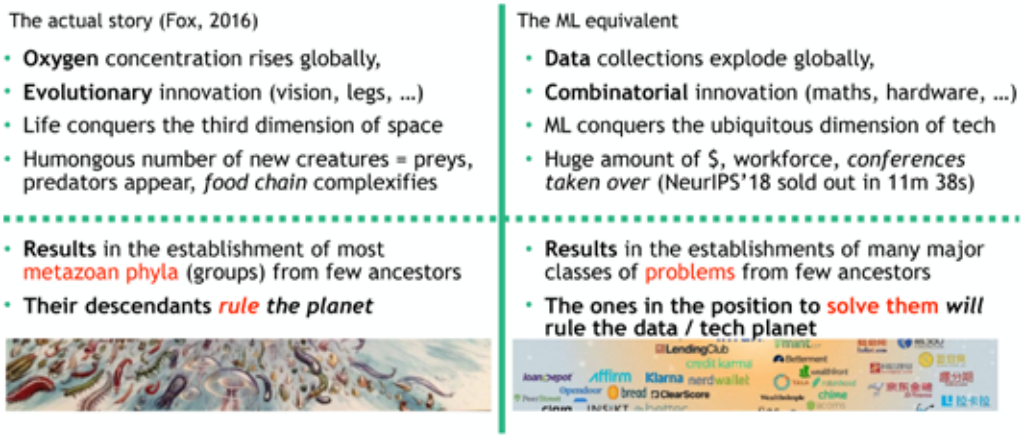


Figure 5

The parallel with the Cambrian explosion (left) for machine learning (right) is in fact quite striking if we make the effort to go until its end, risking a speculative answer on the future of machine learning (Fox, D. (2016). 'What sparked the Cambrian explosion?', Nature, Vol. 530, pp 268-270).

What is interesting is what comes next. If the current state of Cambrian paleontology is accurate, the Cambrian explosion saw the rise of *predators* – literally born in the food pantry of evolution. One should be careful of drawing a parallel with machine learning, but nonetheless there is a lot of opportunistic behaviour that is observable in the field, especially on its industrial side.

In particular, there is currently a rise in the interest of collecting data whose machine learning-based exploitation should prove far more valuable than Facebook-level data: medical data. It is arguably more valuable because one's preferences as stored in Facebook will inevitably change through years. On the contrary, the one who possesses the medical data of people – and in particular its lowest level description, as in genetic sequences – possesses them forever.[xii]

There's no doubt that machine learning technology will be here to lead science breakthroughs on such data. One can only hope that the lessons from the past successes, threats and failures will contribute to shaping good practices and safe usage for our ever-more-personal information to be used, because it suggests that the ones in position to solve the related problems will be in the position to rule our tech planet. We are probably, from this standpoint, witnessing the beginning of an age that is going to reshape our relation to technology, in part under the influence of machine learning.

[xii] And of course, one's genetic data also potentially discloses information about other people as well, forever.

The toolbox to make this work at proper scale

To finish on a positive note, from a technical standpoint, the field of machine learning embraced mathematics early as a strong backup field to safeguard its algorithms and theories. This obviously started with statistics but rapidly spread to a host of different mathematical horizons and theories. I believe mathematics will be instrumental in contributing to safely developing the field further. *This will be an absolute necessity*.

Biography

Richard Nock is Adjunct Professor at the Australian National University and researcher at Data61. He leads the Machine Learning Research Group of Data61. He obtained his PhD and an accreditation to lead research (HDR) in Computer Science from the University of Montpellier (France). His research interests include machine learning, privacy and information geometry at large.

DATA SECURITY AND AI

Benjamin Rubinstein

This chapter examines the security and privacy implications of deploying AI in real-world commercial systems. For more than a decade, academic research has demonstrated susceptibility of learning systems to attack, across a wide array of machine learning methods and numerous application domains. While much of the excitement in AI is justified by the capacity of machine learning to deliver actionable insights from data, these demonstrations suggest that policy and business leaders should be mindful of unintended consequences. This chapter will consider attacks that have been demonstrated against AI systems, including those 'outside the lab'. While *adversarial machine learning* is a young and active area of research, we will overview common patterns for defence that have emerged. Pointers to AI being used within attacks on digital systems will be provided, highlighting the dual-use nature of artificial intelligence technology. The chapter concludes with recommendations for policy and decision makers.

Introduction

Artificial Intelligence is enjoying a renaissance. Driven largely by progress in machine learning, available data, and massive computational power, AI is finding meaningful applications from medical diagnostics, to fintech, and autonomous vehicles. However, the very adaptability that is driving success in such applications, also exposes machine learning systems to manipulation and unintended consequences.

The latest incarnation of research on machine learning in adversarial environments is known as *adversarial machine learning*.[204] Challenges in adversarial learning include:

1. What attacks could be launched against machine learning systems in real deployments?
2. How susceptible are learning systems to such attacks?
3. How could AI systems be designed to be less vulnerable to attack?

In its current form, adversarial learning dates at least to Lowd and Meek in 2005.[205] A 2013 paper, *Intriguing properties of neural networks*, further raised awareness by

demonstrating sensitivity of deep neural networks – contradicting the narrative that large neural networks are incredibly accurate in most situations.[206] However, many of the concepts of adversarial learning pre-date much of modern machine learning and AI altogether.

Taxonomy of attacks on machine learning systems

A weakness of systems that rely on data for their operation – and AI certainly falls into this category – is that they are susceptible to attacks based on that data. A popular taxonomy of attacks on the security[207] and privacy[208] of machine learning identifies threat models of possible attackers. *Threat models* describe a hypothetical adversary – their capabilities, knowledge, and goals. The concept is used for risk assessment and prioritisation. These are outlined below.

Influence

- **Causative** attacks manipulate the learning process with control over training data.
- **Exploratory** attacks gain knowledge about how algorithms and models work or influence their predictions without affecting training data.

Security violation

- **Confidentiality** attacks obtain information from the machine learning system, compromising the privacy of the training or test data it uses.
- **Integrity** attacks induce false negatives, for instance to evade detection.
- **Availability** attacks cause denial of service – or result in an AI system being switched off due to legitimate behaviour being flagged as malicious – usually via false positives.

Specificity

- **Targeted** attacks focus on causing a different outcome for a specific instance. For example, getting a spam filter to block a specific email.
- **Indiscriminate** attacks encompass a wide class of instances. For example, getting a spam filter to block all emails.

Where the security violation reflects an attacker's goal, specificity adds nuance. As originally worded, integrity and availability violations implicitly focus on attacks

on classifiers.[xiii] However, the notion of integrity applies beyond classification: any unwanted manipulation of the output of machine learning corresponds to a breach of integrity. For example, shifting sales accounting from one quarter to another could manipulate forecasts of quarterly profits or commodity prices[209] – an integrity attack on autoregression.[xiv]

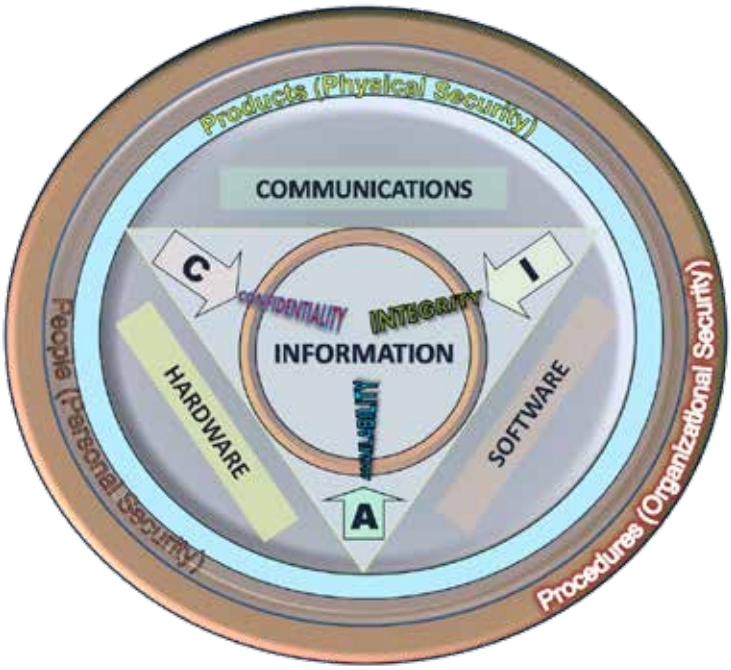


Figure 6

C.I.A. triad of information security. Image credit (CC-SA-3) John Manuel Kennedy Traverso.

Likewise, availability attacks cause a machine learning pipeline to become so dysfunctional on legitimate use cases, that the defender incurs significant cost owing to the use of AI; or data processing is switched off altogether, making subsequent attacks easier to launch. *Confidentiality*, *Integrity*, and *Availability* taken together form the C.I.A. triad of information security (see Figure 6), which has been extended and widely applied across cybersecurity. While the C.I.A. triad has been criticised as being too focused on *data*, it works well where data is a primary concern, as in AI.

Two considerations omitted by the taxonomy have emerged as important to adversarial machine learning threat models.[210] The first refines the focus of the influence axis to include limitations faced by the attacker, and the attacker cost model.

[xiii] Classification is a popular machine learning task in which each input instance must be placed into one of a discrete set of categories or classes. For instance, whether or not a given picture contains a kitten.

[xiv] In statistics, autoregression is a method of using past information to predict future information.

- **Data transformation form and cost:** exactly what transformations can an attacker realistically apply to data, be it when the model is being trained or being tested? Where there are a range of feasible transformations, each with different degrees of magnitude, how much cost is the attacker willing to go to in achieving their intended security violation?
- **Attacker knowledge:** how much information does the attacker possess about the learning process, the learned model, and the existing data sampling process? At its most basic, knowledge includes white box knowledge of a learned model under attack, as opposed to black box attacks, which assume no knowledge of the learned model or training process.

Confidentiality

Confidentiality attacks seek to reverse the learning or prediction process to breach privacy of input data. While privacy attacks in general have long been known in practice and in privacy literature, attacks specific to machine learning have emerged only in the past few years.

Model inversion attacks

For a model output of interest – such as a face classifier's detection of a police officer – if we can find an input data point that leads to the target output, then we can identify characteristics of individuals in the data set.[211] The process of finding the input data point is called *model inversion*.

For example, in an analysis of models trained to recommend treatments based on patient genetics and history, researchers were able to leverage public population statistics and model inversion to accurately predict patients' genetic markers based on their recommended treatments.[212]

It has been argued that model inversion in general only produces average or indicative instances that yield a target model output – not an actual input such as from a training set – and in the case of medical treatment inversion, assumes attacker knowledge of sensitive information, for example, a patient's Warfarin dose.[213] However, in the case of so-called *extreme classification* – multiclass classification over enormous numbers of classes – any element of a class may lead to privacy breach.[214] For example, in a facial recognition system, each class corresponds to an individual; and while model inversion is unlikely to reveal an actual image of a target person, it may synthesise a realistic 'average' image.

Membership inference attacks

Membership inference attacks aim to determine whether a given datum, such as an individual person, was in the training set of a learned model. Recently, such attacks have achieved remarkable success. One example reached 90% accuracy inferring membership against the commercial Google Prediction API, in a completely black box setting without knowledge or access to the training data, its characteristics or distribution, or the specific algorithms used by the learning system. The attack used a capability to probe the system with inputs, as designed. Similar results have been reported against Amazon Machine Learning, and include high accuracy inference of membership of particularly privacy-sensitive health data – over 70% accuracy on a Texas hospital discharge dataset.[215]

Such attacks are not only effective, but simple to implement: they involve training secondary 'shadow' models for which training dataset membership is known, then training a third 'membership inference' classifier against these models to make accurate predictions as to whether a candidate belongs to the (hidden) original training set.

Mitigations for membership inference attacks essentially involve limiting the target classifier's specificity or accuracy on training data: memorising training data is bad for leaking membership.[216] A brief introduction to differential privacy is included below; notably, differentially-private learning algorithms are guaranteed to be secure against membership inference,[217] and are robust against model inversion.[218]

Classical privacy attacks

While the above attacks on confidentiality target AI systems specifically, it is appropriate to view privacy attacks on data analysis more broadly. Often the boundary between 'AI' and 'algorithm' is unclear, and it is important not to ignore less sophisticated attacks by mislabeling a target system as complex and somehow robust. Classical attacks include, but are not limited to, those outlined below.

- **Linkage attacks** re-identify a dataset by joining it to another dataset. For example, the complete Netflix movie watching history of some individuals was revealed by taking an anonymous dataset containing de-identified Netflix movie watching histories and ratings, and combining it with similar information found on IMDB user profiles.[219] Where unique information is common between datasets (such as when an individual gives the same ratings to the same movies on both Netflix and IMDB), this is a simple database 'lookup'. Typically, a new data release may be incorrectly considered 'de-identified' by removing personal information, even though it carries identifying (and potentially sensitive) attributes such as health services or prescription medicines. A second pre-existing source might contain personal information and share some pattern

with the first dataset, such as unique combinations of child ages.[220] However, individuals need not be uniquely identified to suffer harm from data releases: for example, if they are identified in a group that is homogeneous in a sensitive attribute such as HIV.[221] The potential for linkage attacks should serve as a significant risk to releasing unit-record-level micro data.[222]

- **Frequency attacks** exploit available population-wide occurrence frequencies of individuals in a data release, to identify individuals by reversing hashing or encryption.[223] For instance, consider an encryption scheme used to hide names during privacy-preserving record linkage. Encrypted, the names are individually indistinguishable from random information. However, when sorted by frequency in the encrypted release, they will closely align with any available source of unencrypted names also sorted by frequency (e.g. from birth records).
- **Differencing attacks** exploit available information of a time related nature across data release series. For instance, if it is known that a target passenger is the only person who could board a bus at a certain remote location (near their house), one can take aggregated, 'de-identified' count data from public transport logs and find the difference between the number of people before and after the specified stop to determine whether the target had boarded at any given time.[224]
- **Reconstruction attacks** seek to reconstruct sensitive attributes of privately-held unit record-level data, from aggregate statistical releases. The United States Census Bureau recently performed a large-scale reconstruction to assess their existing disclosure control mechanisms of aggregating census data for external release.[225] This involved running a standard mathematical optimisation procedure that found 'most likely' attribute values from the raw data that could have produced 2010 statistics released by the Bureau. Their analysis showed highly accurate reconstruction, demonstrating insufficiency of existing protocols. As a result, the 2020 U.S. Census will be fully differentially private.

Differential privacy

Differential privacy has recently emerged as the leading data protection framework for releasing statistics or AI models, derived from sensitive data, to untrusted third parties.[226] Differential privacy leverages randomisation: a differentially-private mechanism takes a data set and, for example, adds random noise, then outputs the result. The average results can be very accurate, but details about an individual are obscured by randomness.[227] Differential privacy is not itself an algorithm, nor is it a property of a release, but rather a property of *release mechanisms*.

To provide differential privacy, a mechanism's outputs must exhibit some randomness. Where each record in a dataset represents information on an individual, arbitrary changes to a single record cause limited change to the probability of releasing any

particular output. Differential privacy has been adopted for the U.S. 2020 Census[228] and has been deployed within services by Google,[229] Apple,[230] Uber,[231] and a Transport for NSW data release.[232]

Differential privacy's success is owed in part to three beneficial factors:

1. **A strong security property:** the presence, absence, or attribute values of any individual input record are indistinguishable based on a differentially-private release. This security property is guaranteed even in the face of auxiliary information on a target record of interest or on other input records, for example via linkage attack, even when an attacker has access to large or even infinite computational resources, and even when the attacker is knowledgeable of the differentially-private mechanism's inner workings. This is in stark contrast to existing protection measures, such as *k-anonymity*, which do not offer any security property but instead offer only qualitative protections, tend to be vulnerable to linkage attack, and tend to focus on the released data, not the release process.
2. **Generic mechanisms as powerful building blocks:** while designing new differentially-private mechanisms from scratch can be challenging, a growing number of available building-block mechanisms are simple to implement and have well-understood privacy and utility guarantees. Examples include:

    - Laplace[xv] and Gaussian mechanisms for releasing numeric data;[233]
    - the exponential mechanism for private optimisation;[234]
    - objective perturbation applicable to many learning methods;[235]
    - the sparse vector technique for releasing realistic synthetic datasets;[236]
    - many more, as overviewed by Dwork and Roth.[237]

    Collectively these off-the-shelf mechanisms are known as *generic mechanisms*. Many of these generic mechanisms make non-private data analysis privacy preserving. To do so, they require calculation of how sensitive the target analysis is to changing any input record. The more *sensitive* an analysis is, the more randomisation the mechanism must employ to achieve a desired level of privacy. Using larger input dataset sizes or releasing fewer output values often reduces this sensitivity and increases privacy without utility loss. While sensitivity calculation can be involved, recently techniques[238] and open-source tools[239] have become available to automate this process.

[xv] The Laplace distribution is very similar to the more familiar Normal distribution. It is in the same exponential family, it has heavier tails than the Normal, and is implemented as standard in all major statistical software packages.

3. **Rules of composition:** to build up more complex data analyses such as machine learning systems, off-the-shelf generic mechanisms can be run in sequence, with the outputs of one mechanism fed into a subsequent generic mechanism. Such workflows may involve multiple mechanisms accessing the sensitive dataset in its entirety (*sequential composition*), or only on distinct partitions (*parallel composition*). Composition rules account for the way the privacy cost grows when differentially-private mechanisms are combined.[240] Combinations of off-the-shelf generic mechanisms have produced numerous differentially-private versions of common machine learning algorithms, including linear classifiers,[241] non-linear kernel methods,[242] deep neural networks,[243] database queries,[244] and much more.

Federated learning

Where differential privacy can defend against untrusted release recipients who try to reconstruct, re-identify or link against data releases, classical cryptographic technologies target a different but complementary threat model: storage of data or models on untrusted devices, transmission across untrusted channels, or computation on untrusted services. The latter could be computation on a cloud provider like Microsoft Azure, or potentially collaborative computation across multiple devices held by different people – the focus of *federated learning*.

As an example, Google researchers have developed a secure aggregation protocol,[245] whereby Google users collaborate on a global model with their own data. Users do not share their data, but rather use their data to update a local model, which is then communicated under encryption to a centralised server. Without decryption, the server cannot distinguish these updates from completely random data, however with *homomorphic encryption*,[xvi] the server can perform (blinded) computations on such updates. The secure aggregation protocol only permits the server to decrypt and include these updates in its global model if hundreds or thousands of users contribute similar updates.

While a recently proposed *local differential privacy* model protects against untrusted data curators,[246] cryptographic protocols have an advantage that input data can be decrypted exactly as encrypted. Sometimes the large computational resources that can be required by homomorphic protocols limit practical application of this technique.

[xvi] Homomorphic encryption is a way to protect information by encrypting it while still allowing some mathematical operations to be performed on it.

Integrity and availability

The issues related to integrity and availability of machine learning differ markedly to those related to confidentiality: far more attention has been paid to integrity and availability attacks than privacy attacks on learners. Unfortunately, much less is known about general effective defences.

The discussion in this chapter is largely agnostic to the type of security violation (integrity versus availability), as many of the same issues arise in both. However, the distinction tends to be most relevant to binary classification, where it makes sense to discuss false positives and false negatives in the first place;[xvii] the literature focuses on integrity attacks (induced false negatives or evasions) as a greater threat to learning systems.

Adversarial examples

After a model is created by training, an attacker may seek to modify input points to manipulate the model's predictions. For example, altering a malware sample until it is not detected by anti-virus software is an integrity attack attempting *evasion*.[247] Adversarial examples – exploratory, or test-time attacks – typically begin with a regular instance (such as a malware sample) and proceed to modify that instance to alter the model's prediction of it. Test-time attacks have been demonstrated in a wide variety of domains, including network security,[248] and email spam filters.[249]

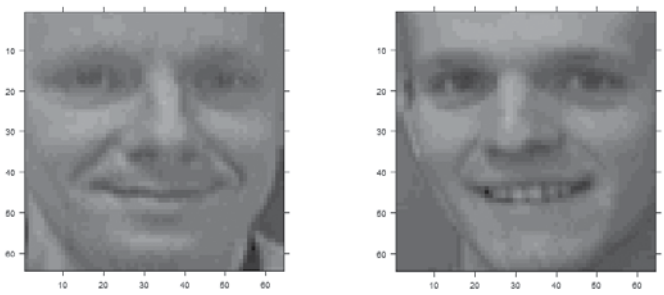
Recently, much activity in adversarial learning research has been initiated by the success of adversarial examples against deep neural networks, particularly in computer vision[250] – images can be modified by an imperceptible amount for a human but cause them to be egregiously misclassified (this is demonstrated in an example on page 114). This phenomenon has been reproduced by numerous researchers and is a strong rebuke to the popular narrative that deep learning necessarily generalises well in such domains.[xviii]

Attacks on vision systems represent a potential safety threat to self-driving vehicles, for example. Early attacks that manipulated trained classifiers at the time of making predictions relied on researcher knowledge of machine learning to carefully craft adversarial examples. Today, adversarial examples can be automated by framing the search for adversarial examples as an optimisation problem.[251] This is straightforward, with freely available tools for automatic differentiation.[252]

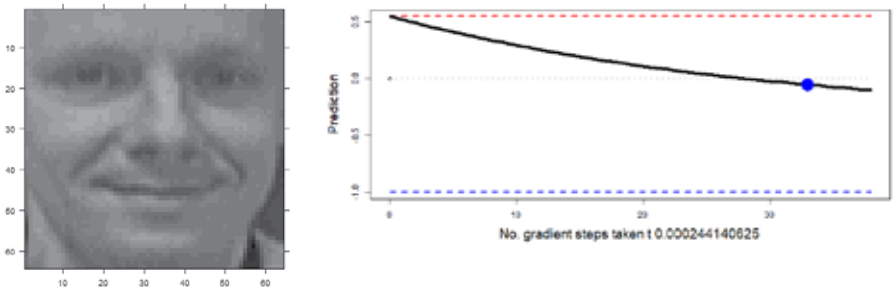
[xvii] In the context of binary classification, a false positive (negative) is an instance that is incorrectly predicted as a positive (respectively negative) by the classifier. In security, the malicious class is conventionally labelled positive.

[xviii] Generalisation, in the context of machine learning, refers to a model's ability to properly handle new, unseen data.

A simple experiment demonstrates the phenomenon of adversarial examples. Here we have trained a support vector machine using the popular OpenCV open-source computer vision library, to accurately classify face images (well positioned in similar lighting) as belonging to one of two people. Examples of the subjects follow:



In under a few seconds, we apply automatic differentiation to modify the left example by a small but strategically chosen number of pixels. The following imperceptibly perturbed image fools the classifier into misclassifying the image of the first subject as the second subject. The curve below displays the classifier's prediction score changing as more gradient steps are taken, until ultimately the classification flips.



Images sourced from the Yale Face Database.

Adversarial examples can also be made effective with little or no knowledge of the target classifier. Effective attacks have been demonstrated in *black box* settings – without knowledge of model, learning algorithm or training data. In this kind of attack, a surrogate model is first trained on (hopefully) similar training data as used by the target model – or potentially on training data labelled by the target model. Then adversarial examples are formed by a *white box* attack against this surrogate model – attacks that make use of the inner structure of model.

In their experiments attacking commercial service Clarifai.com, researchers from Shanghai Jiao Tong University and University of California Berkeley achieved 76% success rates.[253] Similar attacks on models hosted by Google and Amazon have yielded misclassification rates of over 96% and 88% respectively.[254] In some cases, even *universal perturbations* may be possible:[255] while most adversarial perturbations are used to cause a single image to be misclassified, universal perturbations can be used to cause a range of images to be misclassified.

Adversarial examples are not limited to laboratory settings. They have been demonstrated in the physical world with cell phone images of adversarial examples fooling learned models,[256] universal *adversarial patches* that can be printed as stickers,[xix 257] large wearable badges that can hide people from people detectors in video surveillance,[258] and in speech-to-text with 100% success rates reported against Mozilla's DeepSpeech system with perturbed audio waveforms 99.9% similar to original waveforms but recognised as arbitrary text.[259]

Poisoning attacks

Poisoning attacks involve adjusting the inputs with new data to deliver toxic results. While less attention has been paid to attacks that poison training data to influence the learned model, results demonstrate that learning models can be highly susceptible to such *causative attacks*.[xx] As an example of an early *availability attack*,[xxi] an attack on the open-source SpamBayes email spam filter presented a compelling threat model, in which just 1% of messages that were clearly malicious or spam included 'good words', such as those found in a dictionary or online newsgroup data set.[260] Assuming such emails were correctly labelled by the user as spam, and trained on, the spam filter subsequently targeted close to one in two legitimate messages as spam (a false positive), rendering the filter unusable.

Machine learning models for detecting anomalous traffic, such as denial-of-service attacks, on backbone networks, have also been successfully poisoned.[261] The threat model in such *integrity attacks*[xxii] leverages the online nature of learning-based security appliances: continuous training over time is necessary to allow the anomaly detector to adapt to changing characteristics of benign data. So long as poisoning is subtle enough not to be flagged as malicious, it is likely to continue to be used to train security systems. The idea can be taken to the extreme in a 'boiling frog attack',

[xix] The stickers can be placed on objects to cause a classifier to misclassify them, for example, an adversarial patch stuck to a stop sign could make an automated vehicle misclassify it as a speed limit sign.

[xx] Causative attacks manipulate the learning process with control over training data.

[xxi] Availability attacks cause denial of service or result in an AI system being switched off.

[xxii] Integrity attacks induce false negatives.

in which poisoning levels slowly increase over time as the detector becomes more and more influenced to permit malicious data through. Again, early work was carefully crafted by hand, by machine learning experts.

Poisoning attacks have also been demonstrated against popular malware detection tools.[262] Motivated by concern in the machine learning supply chain, researchers from New York University have demonstrated how to create deep neural network backdoors – called *trojans* – such that deployed models can retain state-of-the-art performance on normal data, but in the presence of a 'trigger', the model reliably misclassifies.[263] Their demonstration included an 'accurate' street sign detector that would be reliably fooled when a small sticker was placed in the corner of the sign. Such attacks have consequences for self-driving cars.

Defensive strategies

In contrast with the privacy attack setting, where differential privacy and cryptographic protocols together provide strong defensive technology, little is available in the way of strong, universal, defences against poisoning attacks and adversarial examples. Many speculative defences have been proposed, however almost all have quickly succumbed to attacks. For example, a group of Berkeley researchers famously broke seven adversarial defence papers accepted to the International Conference on Learning Representations conference between the time papers were posted online and the time the conference took place.[264] While the discipline of robust statistics[265] has established techniques that can provably resist outliers (at training time), the forms in which data is poisoned do not reflect threat models in adversarial machine learning.

There is a possibility that, in time, these results will be adapted. Currently *adversarial training* is the most effective and generally applicable strategy: the defender augments their training data by including adversarial examples produced by a hypothetical adversary.[266]

Finally, in a recent paper, researchers from Carnegie Mellon University's Bosch Center for AI proposed the concept of *certified adversarial robustness*, which connects integrity defences to differential privacy by repeatedly feeding test points modified with random noise into a susceptible classifier.[267] The resulting randomised classifier's predictions are accompanied by certifications of robustness: no perturbation within a small radius of a given input could change the classifier's predictions. Where certifications can be made, better understanding of the effectiveness of tampering is possible.

AI-enabled attacks on security and privacy

We have so far focused on security and privacy attacks on machine learning systems, however unintended consequences extend to the harmful use of AI systems more broadly.

Security attacks at scale

AI and machine learning offers unprecedented capabilities of automation. While pre-AI processes based on human decision making might individually be more accurate than current AI-based systems, the ability to make decisions *en masse* can change the value proposition for attackers. AI can speed up attacks or amplify the capabilities of small numbers of human actors. For example, consider *penetration testing* that uses AI planning. Penetration testing is the white hat counterpart to malicious (black hat) attacks; it aims to attack computer systems in order to find vulnerabilities so they can be fixed. Researchers have developed in-principle AI planning tools for automating the process of planning attacks on individual machines, accounting for uncertainty in results of actions and partial observability, then composing these attacks on whole networks.[268] These tools can strategically plan when to leverage known exploits against software on machines, and when to scan to reveal further configuration information.[xxiii] Unfortunately, such technologies could also benefit attackers and perhaps expand the scale and speed of attacks.

Privacy attacks driven by machine learning accuracy

Statistical machine learning underlies technologies with beneficial applications such as data linkage and facial recognition. Accuracy and scalability improve as the underlying AI progresses, and datasets and computing power availability grow. When leveraged indiscriminately, however, these machine learning applications can lead to significant invasion of privacy.

Privacy attacks are often driven by new machine learning applications. For example, *stylometric attacks*[xxiv] can re-identify online writing from among hundreds of thousands of potential authors:[269] train a massively-multiclass classifier with one class per author, on samples of known writing; an authoritarian government could then determine the authors of dissident online writing by applying this classifier to new writing samples. Such attacks leverage unique fingerprints of writing style over content, such as the use of punctuation, sentence structure, and word length. As such, they transfer from public writing to content written anonymously, even on distinct subjects.

[xxiii] Using *partially-observable Markov decision processes* (POMDPs).

[xxiv] Stylometry is the statistical analysis of the writing style of authors. Stylometric attacks attempt to identify authors based on their writing style.

GANs and 'deepfakes'

Generative adversarial networks (**GANs**) are a recent unsupervised learning approach. GANs comprise a pair of neural networks competing with one another: a *generative network* that fits a supplied unlabelled dataset and produces fake instances (like creating colour pictures from black and white ones) with the goal of fooling the discriminative network, which aims to correctly distinguish real from fake.[270] For example, a generator may be called upon to create an image of a cat, and the discriminator network will critique the generated image against an image set comprised of images of actual cats. The models interact until a threshold of acceptable accuracy is obtained – the discriminator decrees that the generated cat picture is sufficiently similar to an actual cat picture. GANs can be used to produce 'deepfakes' – synthesised videos of subjects speaking or performing an act that never actually happened, such as a fake video of a world leader delivering a speech, or a faked voice call from someone you know. Deepfakes, although in their infancy, can be viewed as attempts to attack public and journalistic integrity.

Recommendations

Having newfound knowledge of the many ways in which learning systems can be susceptible to attack, a corporate decision maker or government policy maker exploring deployment of AI systems might feel cautious. These non-exhaustive recommendations are designed to encourage thoughtful application of AI in practice.

1. **Identify relevant threat models:** in the face of unintended consequences of deploying AI systems in high-stakes decision making, or systems built upon privacy-sensitive data, it is imperative that possible threat models – hypothetical attackers with identified capabilities and goals – are not ignored. As a first step, practitioners should identify how an adversary might be incentivised to attack their system, what capabilities and goals they might possess, and which parties (such as providers of computing resources or recipients of shared data) can be fully or partially trusted.

2. **Evaluate systems against adversaries:** with possible threat models in hand, it is important to not only validate an AI system against its training data or test data held-out from the same data source, but to validate against new data sources, and against active adversaries. In the context of confidentiality, practitioners should attempt to re-identify, reconstruct or otherwise breach privacy of data given access to derived record-level data, statistics or model under proposal for release. In integrity or availability attacks, it is important to construct adversarial examples or poison training sets, to assess how a model that otherwise performs well, could be manipulated to do harm. In high-stakes or high-value settings, it

must not be assumed that attackers cannot exist, that attacks are too hard to be practical (without first attempting them), or that they cannot be effective.

3. **Assess risk relative to attack efficacy:** risk assessment of systems involving AI must be data driven, and should incorporate any findings from threat model identification and adversarial validation. For instance, when attempting to re-identify data, risk assessment should not classify privacy harms as minimal by assuming an attacker would only possess easily available public data – even if there are currently no identifiable parties that possess linkable data and an incentive to link. It is important that risk assessment be aligned with conservative threat models and validations. Moreover, any data security breach cost must not be externalised, but linked to social license, reputational and monetary risk.

4. **Identify unintended consequences:** even in the case where no attacks are identified as threatening to an AI system being deployed, it is critical to consider unintended negative consequences to individuals. In many cases, learning systems can be dual-use technologies with valuable legitimate applications, and uses that do harm.

5. **Adopt defensive technology:** adversarial machine learning is a fast-moving, active area of research. Practitioners with identified threat models of concern should employ available defensive technology wherever possible, such as adversarial training for integrity/availability attacks, and differential privacy in the case of data or model releases from privacy-sensitive data.

Biography

Ben is an Associate Professor in the School of Computing and Information Systems at the University of Melbourne. His research interests span machine learning, security and privacy, and databases, and is known for pioneering contributions to adversarial machine learning, differential privacy and statistical record linkage. He previously worked in the research divisions of Microsoft, Google, Intel and Yahoo! (all in the U.S.), followed by a short stint at IBM Research Australia. As a full-time Researcher at Microsoft Research, Silicon Valley, Ben shipped systems for record linkage in Bing and the Xbox360; his research has helped identify and plug side-channel attacks against the popular Firefox browser, and de-anonymise an unprecedented Australian 2016 Medicare data release. His work has been recognised through an Australian Research Council DECRA award, a Young Tall Poppy Science award, and membership of the Australian Academy of Sciences National Committee on Information and Communication Sciences. He holds a PhD in Computer Science from UC Berkeley.

REGULATING AI

Margaret Jackson

Artificial intelligence (**AI**) is already being used in many different sectors and industries globally. At this stage, the AI in use or being proposed is 'narrow' AI and not 'general' AI, which means that it has been designed for a specific purpose – say, to advise on sentencing levels or to select potential candidates for interview – rather than being designed to learn and do new things, like a human. This does not mean that narrow AI – generally non-conscious systems – may not be able to replicate human consciousness in recognising patterns.[271] Identifying patterns in large amounts of data is where AI excels.

While some of the development and deployment of AI systems is happening at a state or national level, there are concerns being expressed that AI development and ownership will be dominated by large global companies like Google, Facebook, Apple, Microsoft and Amazon.[272] Paul Nemitz cites four 'bases of digital power' to watch – lots of money, control of 'infrastructure of public discourse', collection of personal data and profiling, and the algorithms in a 'black box not open to public scrutiny'.[273] Each of these bases of power are possessed by the global companies who are investing considerably in AI development. What this means is, unless the international community is proactive in working together to create an acceptable and consistent framework of AI regulation, which can be adapted by individual nations, there is a risk that commercial interests will set the AI agenda and regulatory responses will be largely reactive.

This chapter explores how AI is being or could be regulated. It examines which existing regulations can apply to AI, which will need to be amended, and which areas might need new regulation to be introduced. Both national and international regulation will be discussed, but Australia is the main focus here. This chapter also examines the role of ethical codes and standards in handling AI challenges and discusses whether there is an appropriate regulatory and ethical framework for dealing with AI, one which will be able to handle future developments in AI technology.

What current regulation applies?

Laws are remarkably flexible and can often apply to new technology without the need for significant amendment. When computer technology was first introduced, only a few amendments to criminal legislation were required, to ensure that stealing information from a computer was theft of property, deceiving a machine like an ATM into giving you money was fraud, and changing data stored on a computer was forgery. The notion of computer trespass was criminalised through offences such as unauthorised access to a machine or destroying data without authority (with intent to commit a crime). Eventually, specific 'cybercrime' legislation needed to be introduced to deal with growing concerns about hacking and denial of service activities.[274]

Some laws introduced since the advent of computer technology are designed to be technology neutral, such as the *Privacy Act 1988* (Cth) (**the Privacy Act**) and the *Privacy and Data Protection Act 2014* (Vic), so that developments in new technology do not require new legislation. Many laws, such as the *Competition and Consumer Act 2010* (Cth), which contains the Australian Consumer Law (**ACL**), focus on the injury or loss suffered by the consumer due to actions of the seller, rather than on the type of technology sold by the seller which may have led to that injury. Manufacturers bear the responsibility for loss or damage. The users of computer systems such as banks, transport companies, airlines, hospitals, and so on, bear the risks for damage caused to their customers. They in turn may seek redress from suppliers of the product. This should not change if AI is involved in providing the service, or is part of the goods or products being sold, although there may be difficulty with integrated products in identifying which part of the supply chain – the designers, developers, or manufacturers of the different components – was the cause of the problem. The main approach to handling AI issues is to use current regulation as far as possible. A number of different areas using AI are discussed below.

Drones and driverless cars

Computer technology has been used for years in the automotive and aeronautics industries to provide assistance to drivers and pilots to improve safety. AI technology is enabling developers to replace humans either completely or partially in operating drones, and driverless and self-driving cars.

In the first instance, governments are dealing with both drones and driverless cars through amendments to existing legislation. With drones, in Australia, Part 101 of the *Civil Aviation Safety Regulations 1998*, which specifically regulates unmanned aircrafts, was amended in 2016 to introduce new rules about licensing, necessary permissions and notifications. In 2018, the Commonwealth government released a *Report on Regulatory Requirements that Impact on the Safe Use of Remotely Piloted Aircraft*

Systems, Unmanned Aerial Systems and Associated Systems, which reviewed the success of the new amendments and recommended new processes for dealing with risks.[275] In 2019, *Project US 18/09 Remotely Piloted Aircraft* (**RPA**), a scheme which is working on RPA registration and remotely piloted aircraft systems (**RPAS**) operator accreditation, commenced.[276] While the work around dealing with the growth of RPAS is being done primarily at Commonwealth level, it should be noted that it fits within the context of the international framework for civil aviation, the United Nations (**UN**) International Civil Aviation Organization.

A similar approach to using existing legislation is being taken with autonomous (driverless) cars and self-driving cars (in which a human driver is still in the car). Unlike autonomous drones, though, the regulatory approach to autonomous vehicles involves both Commonwealth and state governments. The National Transport Commission (**NTC**), which is funded by the Commonwealth, state and territory governments, has released guidelines for trialling autonomous vehicles, as well as discussion papers exploring issues around their use. These issues include options for amending existing legislation to cover autonomous vehicles, including trains, and approaches to providing appropriate motor accident injury insurance.[277]

Again, the discussion around how to handle the issues caused by autonomous vehicles is informed by similar investigations being undertaken globally, particularly in the United Kingdom (**UK**), the United States (**US**) and the European Union (**EU**). By 2017, six US states had introduced legislation dealing with autonomous vehicles and 19 others had similar bills under consideration.[278] Germany in particular has developed ethical guidelines for *Automated and connected driving*.[279] In the UK, the Centre for Connected and Driverless Cars has released a new Code of Practice on automated vehicle trialling.[280] There are also international groups such as the UN Global Forum for Road Traffic Safety, of which Australia is a member, who are examining rules for autonomous vehicles.[281]

Consumer protection

The ACL, nationwide legislation covering sale of goods to consumers, will apply to many aspects of AI systems involved in such sales. The ACL applies to goods like computer software that are provided to a consumer, either directly or embedded in a product provided to them. The *consumer guarantees* in the ACL include the requirement that goods are fit for purpose, and that they are of acceptable quality, which includes a requirement that the goods be reasonably 'safe'.[282] Customers who are injured or who suffer property damage as a result of unsafe goods can be compensated by manufacturers without having to prove that the manufacturer was negligent. However, at present, the ACL does offer the manufacturer some statutory defences, including that there was no defect in the goods when they were supplied,

that the state of scientific and technical knowledge at the time of supply did not enable the supplier or the manufacturer to discover the defect, or a cause independent of human control occurred after the goods left the manufacturer's control.[283] If manufacturers or designers claim that the pre-release testing of the AI system showed it worked as expected, but that the problem that led to injury or damage to a consumer developed after release in a way that was not expected, then the manufacturer might be able to avail themselves of the defences provided by the ACL. Whether these defences are appropriate for AI systems is discussed further below.

Many countries are reviewing whether existing product liability laws will apply to AI devices. The European Commission, for instance, produced a paper titled *Liability for emerging digital technologies* in April 2018, which examined whether the Product Liability Directive and the Machinery Directive adequately covered issues with AI.[284] Both of these directives provide for strict liability in the event of loss or damage. Generally, the paper decided that the Directives were adequate for the current state of development in technology but that further examination was needed, in particular around defectiveness, burden of proof, and management of risk.[285] Product liability legislation also covers 'goods', 'products' and 'manufacturers', and whether self-conscious AI in particular may require different terminology. This issue is discussed further below.

Government AI decisions

The discussion above has been focused on consumer protection issues, but most of the legal cases that have arisen to date involving AI have not arisen in the consumer context. They have primarily involved the use of AI by government departments. In Australia, there are already a number of instances where human decision makers have delegated decision making to computers. For example, the *Migration Act 1958* (Cth) states that the Minister may "arrange for the use, under the Minister's control, of computer programs for any purposes for which the Minister may, or must … make a decision; or exercise any power…".[286] There are another 22 sections in a range of acts that allow government departments to deem a decision by a computer system to be a decision made by a designated officer, for instance, the *Therapeutic Goods Act 1989* (Cth) and the *Social Security (Administration) Act 1999* (Cth).[287]

Decisions made by Australian Government Ministers, departments and agencies can be reviewed by the Administrative Appeals Tribunal if allowed for under the relevant Commonwealth legislation. Similar state and territory bodies, such as the Victorian Civil and Administrative Tribunal, fulfil similar roles. The Commonwealth Human Rights Commission and its state and territory counterparts can also investigate complaints alleging discriminatory actions by government departments in some cases. Complaints against public sector decisions can also be lodged with the relevant Ombudsman,

as occurred with the 2016 Centrelink project, 'Robodebt'. Robodebt was designed to check for overpayment of social services by matching Centrelink data with Australian Taxation Office data, which resulted in 20,000 people being falsely accused of fraud. The Commonwealth Ombudsman investigated numerous complaints and has now issued two reports containing recommendations for improvements to be implemented by the Departments of Human Services and Social Services, noting that the design of the system suffered from limitation purpose, and risk management and overall planning were held to be poor.[288]

Courts provide further avenues for appeal. Two appeals have been lodged with the Federal Court against penalty assessments raised by Robodebt.[289]

Anti-competitive behaviour

The ACL makes it an offence for a person to deceive or mislead a consumer or a business.[290] A similar section in the *Competition and Consumer Act 2010* (Cth) (**CCA**) applies to the conduct of companies in trade and commerce.[291] No intent is needed to be shown in a case of misleading behaviour. It is arguable that the actions of an AI could fall under these sections. AI can fabricate and manipulate data, either because of its programming or because of the data it has been fed or has collected.

AI may also be involved in anti-competitive behaviour, particularly for monitoring pricing of competitors and making price decisions. In 2015, the US Department of Justice prosecuted a seller in the Amazon marketplace for collusion with other sellers to fix the price of posters sold online, by sharing and jointly implementing dynamic pricing algorithms.[292] While no similar cases have occurred in Australia, the Australian Competition and Consumer Commission (**ACCC**) Chair, Rod Sims, in an address titled *The ACCC's approach to colluding robots*, stated that he considered the anti-competitive provisions of the CCA able to handle cases involving price algorithms and collusion.[293] Mr Sims referred in particular to two new provisions in the CCA. One of these provisions – the 'concerted practices' provision, provides that a corporation may not "…engage with one or more persons in a concerted practice that has the purpose, or has or is likely to have the effect, of substantially lessening competition".[294] The ACCC Guidelines on the new provision describe a concerted practice as one "where competitors substitute cooperation with each other for the uncertainties of competition".[295] The other provision to which Mr Sims referred prohibits a firm with a substantial degree of market power from engaging in conduct that has the purpose, effect or likely effect of substantially lessening competition in a market.[296] Mr Sims believes these two new sections would give the ACCC the appropriate powers to address collusion by AI systems.

Privacy protection and big data

The next area of existing law to be discussed is that of privacy protection and big data. Data is the key driver of AI systems and, in Australia, the Privacy Act and relevant state and territory privacy legislation apply to the collection, use and disclosure of personal information. Some also cover privacy breaches and information security. The fact that personal information is collected and used by AI does not affect the operation of the acts, designed to be technology neutral.

However, privacy legislation in Australia and overseas has struggled a little to cope with the advent of 'big data', that is, the enormous amounts of data collected by organisations and governments, much of it generated by individuals online. Many of the difficulties with big data have arisen because the organisations using it are not necessarily the same organisations that collected it, so that informed consent, an important requirement in privacy law, becomes difficult if not impossible to obtain, where personal information is involved. For example, where the personal information in big data sets has been collected via social media or other websites, CCTV and similar surveillance technology, or online cookies, any consent, let alone informed consent, is not possible depending on the processes used to collect the personal information.

Organisations and governments have often struggled to process big data due to its size and complexity. AI, however, has the capacity to analyse big data and is able to recognise patterns in data which could lead to identifying individuals from the data analysed. The UK Information Commissioner's Office and the Office of the Australian Information Commissioner have released guides to assist organisations and governments with big data analytics.[297] These guides focus on embedding 'privacy-by-design' into the early stage of AI development, rather than 'unpredictability by design' which can result if data is fed into algorithms without pre-defined queries.[298] The guides restate the importance of privacy impact assessments at the beginning of AI projects, for transparency about processing and for minimisation in data collection.

There have been some amendments to privacy laws internationally to strengthen them in light of the ongoing growth in the collection of vast amounts of personal data. The *EU General Data Protection Regulation 2016* (**GDPR**) is designed to be technology neutral as well, and so applies to new technology such as AI. It focuses on informed consent, more protections around the collection of sensitive data, breach notification and two new rights for individuals – the right to be forgotten and the right to data portability – although these two new rights became gradually weaker as the lengthy negotiations over the new GDPR took place. The GDPR can now impose heavy penalties on organisations and governments that breach its provisions. However, doubts have been raised about the applicability of the new GDPR to AI technology with its associated issues such as a lack of transparency about decision-making and difficulties in obtaining consent.[299]

The Commonwealth Government has announced its intention to strengthen the Privacy Act by increasing penalties for serious and repeated interference with an individual's privacy. It intends to introduce a right for individuals to ask technology and social media companies to cease using and disclosing their personal information, and a new code of conduct for social media and online platforms covering collection and use of personal information.[300] These amendments will address some of the issues relating to big data and social media, but do not appear to address specific AI-related issues, such as how to ensure that the human rights of individuals whose data is being collected and used by AI is protected if obtaining consent from the individuals becomes impracticable.

Proposals for new specific AI regulation, particularly related to general AI

So are AI systems different from other software, resulting in a need for specific legislation to deal with safety and security issues that might arise? Three stages have been proposed as possible areas that need addressing – specification, robustness, and assurance[301]

Stage one – specification – covers the design of the AI system. Was the design appropriate for the purpose it was intended to fulfil? While the initial stage of AI design, deciding the purpose of it, is not particularly different from designing any computer system, the architectural design of AI software often involves the selection of appropriate data, and tuning and training of a neural network, rather than coding in a programming language. How the algorithm then develops may be in ways not intended or expected from the original intent. This leads to the second stage.

Robustness, the second stage proposed, is also expected of the design and implementation of any computer system. Has the designer built into the system appropriate ways to deal with risk, to incorporate margins of acceptable unpredictability, and adequate failsafe mechanisms? The main challenge with AI systems, particularly those intended to be dynamic, is understanding how and why the AI acted as it did, and if its actions or decisions were anticipated.

The third stage – assurance – covers monitoring the performance of the system and enforcing the controls and safeguards built into the system, including interrupting the system and closing it down. Again, this would be expected in any computer system design. With AI, while the design objective might be appropriate, the data as accurate as possible and the neural network fully trained, the results may not be what was expected or intended, able to be explained, or even foreseen.

Explainability and foreseeability

The questions of 'explainability' and 'foreseeability' of actions by AI systems are two key issues to be considered. Machine learning AI systems, for instance, are a result of the provision of data, training and tuning in how to analyse the data, resulting in an algorithm which is then used, say, to predict prices. The algorithm may change as new or different data is received. From this starting point, the system develops its own conclusions regarding analysis of the data. Despite testing, it is not always possible to foresee if the AI algorithm will operate as expected. For example, in 2016, Microsoft developed an AI system called Tay to engage and chat with people. It appeared on Twitter, but after 16 hours online Tay was making racist and inflammatory statements as a result of online interaction with other tweeters.[302]

As noted earlier, risks relating to injury or damage are borne by the owners or implementers of an AI system. They in turn may seek redress from suppliers of the product. A statutory defence in the ACL is available to manufacturers, who may claim that the product, the AI, worked as designed when tested before release, and that it was beyond the control of the manufacturer to foresee all it might do when in operation. There has been a change from what is developed to how it evolves. At this stage, it is possible for the manufacturer to claim that they have shown 'reasonable care' in designing an AI system. But is 'reasonable care' or avoidance of foreseeable harm in designing an autonomous AI an adequate standard? Is a different test required and should a specific liability regime be established for AI? Is an AI system a 'product' at all?

The difficulty in understanding how algorithms operate has been described as the 'black box' problem.[303] The 'black box problem' is a response by some AI developers to questions of why they cannot explain how the AI operates and why it did what it did. This issue of explainability is also relevant when considering ways in which consent for the use of personal information can be obtained from individuals; if the developers and users of AI cannot understand how it works, it is almost impossible for an individual to consent to it, as it cannot be understood.[304]

Another argument against explainability is that algorithms are commercial in confidence and cannot be disclosed to others.[305] In *Cordover and Australian Electoral Commission*,[306] the Administrative Appeals Tribunal upheld the Australian Electoral Commission's (**AEC**) refusal to release code of a computer program that it used to read and count Senate ballot papers, as it claimed that the code was a 'trade secret', used for the AEC's fee-for-service function.[307]

The argument that AI actions might not be foreseeable, understandable or accessible has led to calls for transparency around algorithms. Both the US Congress and Senate introduced a Federal *Algorithmic Accountability Act* in April 2019. Applying only to

companies earning over $USD 50 million per year, the Act would make the Federal Trade Commission responsible for evaluating automated systems that had been classified as 'highly sensitive'. In addition, companies would be required to evaluate algorithms for a range of issues such as bias, discrimination, and or security and privacy risks. While the US Senate seems unlikely to approve the Bill for political reasons at this time, the issue will remain on the national agenda, particularly as some US states are trying to respond to citizen concerns.

New York City Council passed an algorithm transparency law in 2017 (Local Law 49) which mandated that a task force be established to study the use of algorithms by New York agencies and to develop recommendations about how "information on agency automated decision systems may be shared with the public" and how agencies can address any harm caused. Progress to date by the Task Force has been slow.[308] Washington State has also drafted algorithmic accountability bills, which if passed, would require algorithms to be made available by vendors of AI systems for government agency or third-party testing, auditing or research.[309]

Bias

As noted above, issues associated with a lack of transparency lead to concerns about bias and discrimination by AI systems. Bias can arise in two ways with AI. The first is when bias, conscious or unconscious, is incorporated into the design of the AI system and the algorithm that it will use. The second is when the data that is provided to the AI contains biases. Amazon provides an example of the importance of ensuring that the data provided to an AI is not biased. Amazon had developed an AI recruitment tool but found that it was biased against women and was more inclined to select males rather than female applicants. The AI system had been provided with the resumes of successful applicants over the previous ten-year period, but this data reflected the dominance of males working in information technology. The AI is no longer used.[310]

AI systems are being used widely in employment recruitment, in setting penalties in lower courts, in checking for fraud in government payments, for facial recognition in areas such as education and policing, and in legal advice work. The private sector offers more challenges than the public sector as there are limited avenues to complaints processes. In recruitment situations, appeals by non-employers against decisions made to interview or appoint applicants are generally not available. In some cases, anti-discrimination laws (such as the *Age Discrimination Act 2014* (Cth) and the *Sex Discrimination Act* (2004) (Cth) and state and territory equivalents) might be used.

The UK's Centre for Data Ethics and Innovation (**CDEI**), created by the UK Government to advise on AI, has announced that one of its first activities will be to undertake reviews into bias in a number of sectors, starting with the finance sector, and moving

onto local government, recruitment, and crime and justice over the next two years.[311] The use of AI for decision-making in all these areas has the potential to adversely affect individuals if bias is embedded in the algorithms being used.

One regulatory solution to concerns about bias with AI decision making that has been proposed is that there should be human involvement in decisions by AI which affect human rights. The EU GDPR provides a right to individuals "not to be subject to a decision based solely on processing",[312] so that some human contribution to the decision making is needed. However, this right, which was originally described as a 'right to explanation', was watered down after lengthy compromises, and its operation now is fairly restricted. For example, it does not apply if the individual gives explicit consent. It also does not apply if the decision is necessary for entering into a contract, say, for employment.[313] However, Australia's Privacy Act does not include any similar right.

Electronic personality

There have been suggestions made that a solution to the problem of foreseeability and lack of transparency is to recognise some form of 'electronic personality' for AI and robots. These suggestions are an attempt to address the related issues of liability, rather than an attempt to grant legal human status to AI systems. For instance, while Estonia is considering granting legal status to robots,[314] or 'kratts',[xxv] it has done so in the context of allocating liability for damages, having rejected introducing sector-based liability regulation and "opting for algorithmic liability instead".[315] A draft bill is being prepared for discussion.

Similarly, the European Parliament approved a *Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics*.[316] Article 59(f) of that Resolution states that the EU should consider:

> *creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently*

With this approach, it is envisaged that awarding some form of legal status to AI systems would enable a specific compensation scheme for loss and damage caused by the AI to be established for that AI, separating it from the company that developed or initiated it. There is no doubt that, as AI technology develops, there will be a need for new forms of accountability and liability for business and consumer related

[xxv] 'Kratts' are part of Estonia folklore and are servants built from hay or old household items.

damages and injuries resulting from AI activity, but whether it takes the form of legal personhood will require much more debate.

Strict liability

Another suggestion has been to introduce strict liability (no fault) laws.[317] Unlike the European Union, the product liability laws in Australia have limited strict liability. Often, a remedy for damages is sought under tort law. It will apply if a person suffers injury or loss as a result of the negligent actions of another. For negligence to be proven, the injured person has to show that the other party owed them a duty of care, that they failed in that duty, that the risk of harm was reasonably foreseeable, and that the injury was caused by the failure of the person. However, it is only in cases where the activities that led to injury of a person could be called 'ultra-hazardous' activities – such as those involving fire, hazardous materials and so on – that the injured person does not have to prove negligence.[318]

If strict liability is extended to cover all AI activities, then it may be that insurance cover will be needed, creating compensation pools funded by developers and implementers of AI. Examples of similar compensation pools are the Victorian Motor Car Traders Guarantee Fund, established by act, funded by motor car trader licence fees, and intended to compensate individuals who suffer loss due to actions of car sellers,[319] or the various worker's compensation insurance schemes in Australia which are funded by employers to cover employees work injuries.[320]

Intellectual property

Apart from issues around liability for loss and damage, and bias, there arise questions around intellectual property (IP). New IP might be created once an AI system is operational. For instance, the developer of the AI system may provide the technical expertise, while the user of the system may provide the data without which the AI cannot operate. New IP may then be created. In Australia, copyright exists in works resulting from human intellectual effort, not from computer generated works.[321] Some countries, like the UK and New Zealand, have amended their copyright law to grant ownership of computer generated works for copyright purposes to the person who made the arrangements for the work to be undertaken.[322]

Ethical guideline proposals

There is no shortage of ethical guidelines for AI being proposed, some by specific professional and industry groups,[323] others by individual countries,[324] the EU,[325] and the large global technology companies.[326] It is clear that there is national and international agreement that such guidelines will be vital for the protection of human rights, including protection against discrimination and bias.

In Australia, Data61, a division of the Commonwealth Scientific and Industrial Research Organisation (CSIRO), released a discussion paper in 2019 titled *Artificial Intelligence: Australia's Ethics Framework*, which summarises relevant legislation and ethical principles relating to AI, both Australian and overseas. It provides a number of case studies illustrating issues with possible AI bias in decision-making, automated data decisions in government settings, transparency issues and the need for human oversight, and predictive systems in health, policing and insurance. The paper proposes eight core principles for an ethical AI Framework.[327] These principles are:

1. The system must generate net benefits
2. The system will do no harm
3. There will be appropriate regulatory and legal compliance
4. Privacy protection will be ensured
5. The system will be fair
6. There will be transparency and explainability
7. There will be a process to contest decisions
8. The people and organisations responsible for the AI will be identifiable and accountable

The final section of the paper briefly discusses possible ways in which the ethical framework could be implemented, including through impact assessments, review processes, risk assessments, best practice guidelines, education and training standards, AI monitoring, and recourse mechanisms.[328]

The European Commission High-Level Expert Group on Artificial Intelligence (**AIHLEG**) has also released *Ethics Guidelines for Trustworthy AI*. The AIHLEG guidelines propose a human-centric approach, with the key question to be asked: how will this AI help humans? Humans will need to be able to trust the AI. To achieve 'trustworthy' AI, three components must be satisfied throughout an AI system's entire life cycle. First, the AI system must be lawful, complying with all applicable laws and regulations; second, it should be ethical, adhering to ethical principles and values; and, third, it must be robust from a technical and social perspective.[329] The guidelines are intended to be voluntary and to provide a broad and general horizontal framework, which should be supplemented by sectorial approaches, say, in areas such as medical health.

In the AIHLEG guidelines for trustworthy AI, the framework to be implemented envisages three stages – first, establishment of four ethical principles; second, implementation of seven key requirements; and third, the assessment of trustworthy AI through operationalising the key requirements. The four ethical principles are based on fundamental human rights, many of which are contained in existing legal requirements.

These principles are:[330]

1. Respect for human autonomy
2. Prevention of harm
3. Fairness
4. Explicability

The guidelines acknowledge that there may be tensions between these four principles. The example they provide of this tension is when AI systems are used for 'predicative policing', usually involving surveillance activities. While this might prevent harm, it may also impinge on individual privacy and liberty, and this tension will need to be considered by AI designers and operators.

The seven key requirements are:[332]

i. Human agency and oversight
ii. Technical robustness and safety
iii. Privacy and data governance
iv. Transparency
v. Diversity, non-discrimination and fairness
vi. Societal and environmental wellbeing
vii. Accountability

These requirements are all of equal importance and are interrelated. They apply equally to all stakeholders involved with AI, including developers, deployers and end-users.[333] Again, tensions may arise between the requirements, but how these tensions are dealt with must be rationally considered, and the solutions acknowledged and evaluated, with accountability clearly stated.[334]

The guidelines also address technical and non-technical methods to be used in realising trustworthy AI. Technical methods include ensuring ethics and the rule of law are incorporated into design from the beginning, mechanisms for fail-safe shutdowns, and appropriate testing and validation.[335] Non-technical methods include regulation, codes of conduct, standardisation, certification, accountability through governance systems, and the use of diversity and inclusive design teams.[336]

The final section of the AIHLEG guidelines discusses processes for assessing trustworthy AI. It contains a six-page assessment list which is to be piloted with stakeholders from the public and private sectors throughout 2019, with a revised version due in early 2020.[337]

Other ethical guidelines contain similar principles, a number of which have been based on the AIHLEG guidelines, such as *The European Group on Ethics in Science & New Technologies Statement on AI*.[338] The AI 4People's *Ethical Framework for a Good AI Society* project has five principles derived from a survey of 37 different sets of ethical principles, including those in the AIHLEG guidelines.[339] The International Conference of Data Protection and Privacy Commissioners also released the Declaration on Ethics and Protection in Artificial Intelligence in 2018.[340]

The OECD Council on Artificial Intelligence approved the *Recommendation on Artificial Intelligence* in 2019.[341] In Section 2 of the Recommendation, the Council recommends that member countries should implement the suggested national policies and engage in international co-operation using the following five principles contained in Section 1:

1. Inclusive growth, sustainable development and wellbeing
2. Human-centred values and fairness
3. Transparency and explainability
4. Robustness, security and safety
5. Accountability

All of the ethical guidelines developed or being developed are intended to be voluntary non-binding guidelines. Some are aimed at professional groups; others are documents intended for national or international impact. Human rights and data privacy are key rights in all of the ethical guidelines, as is the need for transparency (or explicability).

Criticisms of these approaches at industry and national levels centre around the weaknesses of self-regulation, while those of international level raise concerns about the effectiveness of 'soft law' approaches to international issues. However, there are a number of examples of internationally agreed guidelines which have achieved global impact, and which have resulted in legislative adoption by numerous nations. One such example is the 1980 *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, which was incorporated into the Privacy Act and influenced many other data protection laws around the world.

Another example of a set of international guidelines that eventually led to widespread adoption is that of research ethics. Starting with the *1947 Nuremberg Code for research on human subjects*, formulated by the American judges sitting on the Nuremberg Tribunal – and leading to the *1964 Declaration of Helsinki* developed by the World Medical Association to govern medical research – ethical principles governing the conduct of medical, biomedical and social science research involving human subjects apply to research activities in most countries. The international acceptance and adoption of research ethics involving human subjects has been a longer and more tortuous journey than that of data protection – and is still continuing – but it does

illustrate the need for initial international expression of what is acceptable.
In areas of important social and economic change that will have a global impact, such as what is occurring with AI systems and technology, ethical frameworks offer a way to introduce agreed principles that support the introduction and implementation of the technological change, whilst at the same time setting acceptable standards to lessen any negative impact on humans. The frameworks that have been proposed by the various international bodies and organisations are necessary to provide a readily understood context for individuals to accept that there are constraints within which AI development and implementation will occur. At this stage, there are numerous AI ethical frameworks being proposed. Fortunately, there is considerable overlap in the principles comprising each framework.

The next stage in development of an AI ethical framework is international agreement on one set of principles.

Introducing a layered approach for regulation

It is evident that there is not one solution to regulation of AI, as there is already a mix of state and national legislative approaches, plus a myriad of ethical principles at all levels. The Data61 discussion paper on an ethics framework for AI notes that an ethics framework alone is not enough to deal with the issues of AI; it should be "one part of a suite of governance mechanisms and policy tools which can include laws, regulations, standards and codes of conduct".[342]

The various approaches – or layers – that are necessary to regulate AI are discussed below.

Level one

The first level of regulation will be voluntary and should take the form of internationally agreed ethical guidelines. As mentioned above, there are already a lot of international, national and industry ethical frameworks being suggested and there is considerable overlap in the content of the various guidelines. The majority call for transparency, accountability, safety, assurance of human rights – particularly relating to personal information – and lack of discrimination. While it is important that there is consensus about what the key ethical principles to govern the design, development and use of AI are, this does not mean that guidelines must be exactly the same. Overall acceptance of the key principles should suffice at this early stage of AI use and development.

Level two

The second level of the approach to regulation will also result in non-binding developments. There must be internationally agreed technical standards for human-centric design of AI, which can then be adopted by nations. The Institute of Electrical and Electronics Engineers (**IEEE**) is developing AI related standards. Eleven areas have been identified, including one on data privacy and another on transparency of AI systems.[343] The International Organization of Standardization (**ISO**) has also commenced similar work, with three published ISO standards and 11 ISO standards under development.[344] Whilst its standards are voluntary, ISO standards are usually considered by courts to be best practice. Standards Australia is involved with some of these developments.

Level three

The third level is national or state legislation that ensures there is adequate consumer protection, product security and data protection legislation. These laws will need to address what can be delegated to an AI system and what cannot. Many countries have started to review existing legislation in these areas to either ensure that existing legislation will apply, or otherwise that new amendments might be needed in some areas. The EU has been a leader in such a process.

Level four

The fourth level will be for specific areas in which AI is to be used – such as automated vehicles, drones and smart services – to have appropriate legislation, whether through amendment of current legislation or through enactment of new legislation.

Level five

The fifth level is to provide some form of national governance oversight to ensure there is accountability of AI developers and deployers. The Australian Human Rights Commission has suggested the establishment of a Responsible Innovation Organisation (RIO).[345] The RIO would have investigatory powers similar to the ACCC, could develop standards, regulation and have powers similar to the Australian Information Commissioner around monitoring, compliance and penalties, operate a certification scheme for AI systems, and adjudicate complaints.[346] It would be common sense, though, to incorporate some human involvement in decision making in all cases involving AI, such as through the establishment of an AI Ombudsman.

The UK Government is establishing both an industry-led AI Council and advisory body called the Centre for Data Ethics & Innovation. The UK's Information Commissioner is to develop a new framework for auditing artificial intelligence tools.[347]

Level six

Finally, the sixth layer is the provision of appropriate insurance schemes to assist with overall industry risk management, through levies on AI manufacturers, developers and deployers of such systems. Insurance companies are involved at present in assessing risk in relation to driverless cars in particular, and some governments have already acted to ensure there will be no gaps. In the UK, for instance, the *Automated and Electric Vehicle Act 2018* has been enacted to address a gap in both insurance and public liability coverage. This act is not considered to be implemented immediately but has been passed in anticipation of the introduction of driverless cars in 2021.[348]

Conclusion

Regulation of AI has already commenced. Challenges associated with certain types of AI currently being tested and introduced, like driverless cars and drones, are being addressed through amendments to existing national and state legislation, but with input from global forums and working groups.

Further, many nations are reviewing their laws around product liability, and in particular, consumer protection, to determine whether any amendments might be needed to address AI-related issues. Governments and industry both have key roles to establish clear frameworks for developers, deployers and users, and to determine gaps in existing regulation.

There is growth in the uses of AI by governments, most appropriately included in relevant legislation. The private sector, too, has expanded its use of AI. In many cases, this growth has been built on the understanding that there are protections from misuse of personal information contained in privacy laws, however, the exceptions and the breadth of the rights contained in the extended uses of personal information appear to have weakened those protections. This needs to be addressed by regulators.

The lack of transparency about how AI systems operate, and the associated issues around foreseeability around AI decisions, are the biggest challenges for regulating AI, and these issues are unlikely to be dealt with without the introduction of new legislation. General AI involves bigger challenges than those posed by narrow AI, particularly as machine consciousness develops, if it does so. There will be increasing problems of foreseeability, from what is developed to how it evolves. It will be vital that human rights and the rule of law are protected.

The manufacturer might argue that at the time of placing the goods on the market or providing the system to the user, no vulnerability with the AI was known; or that they

do not control the algorithms they have developed. It is likely that the concept of strict liability for damage or loss may need to be extended. Insurance and compensation pools may be needed. Control and liability issues will need to be addressed as more advanced AI is developed. It will likely not be one single regulatory solution, but rather a layered approach to AI regulation that is needed at various levels. Ethical guidelines will be an important part of that framework, as will the involvement of human oversight in decisions made by AI.

Ethical AI principles and guidelines are being developed at international, national and industry levels. While there is considerable overlap in the principles being developed, there does need to be international support for universally agreed principles.

What is missing with the AI ethical principles, which occurred with the OECD data protection guidelines and research ethics frameworks, is a clear international leader. With data protection, the OECD, which now represents 34 members plus 16 adherents, took that role. With research ethics, both the United Nations and the World Medical Association provided leadership. At this stage, the EU AI ethical principles appear to be the most developed and have been incorporated into other proposed frameworks, but the approval of the *Recommendation on AI* by the OECD may see this change.

There are considerable challenges facing Australia and the rest of the global community in ensuring that the concepts embedded in the various ethical principles – such as transparency, explainability, impact assessments, risk assessment and review processes of AI, and avenues for recourse against decisions by AI – are operationalised effectively and appropriately. The possible future advent of AI that may achieve consciousness lends urgency to these endeavours.

AI developments offer potentially destructive and detrimental impacts on the lives of individuals globally. A workable and globally accepted AI regulation and ethical framework should be able to restrict the harm to individuals and society without stifling development that can benefit the same groups with advances in health and security, and ensuring economic security and stability. Regulation does not need to be a barrier to development, but it does need to be consistent.

Biography

Professor Jackson conducts research in the areas of computer law, Big Data, data protection and privacy, and artificial intelligence. She is the co-author (with Dr G Hughes) of Private Life in a Digital World, Thomson Reuters 2015; co-author (with Dr M Shelly) of Electronic Information and the Law, Thomson Reuters 2011, author of A Practical Guide to Protecting Confidential Business Information, LawBookCo 2003, and author of Hughes on Data Protection in Australia, LawBook Co 2001.

REFERENCES

UNDERSTANDING AI

1. Turing, A. (1950). 'Computing Machinery and Intelligence', *Mind*, Vol. LIX, No. 236, pp. 433–460.
2. Walsh, T. (2017). *It's Alive!: Artificial Intelligence from the Logic Piano to Killer Robots*, Black Inc.
3. Su, J., Vasconcellos Vargas, D. & Sakurai, K. (2017). 'One pixel attack for fooling deep neural networks'.
4. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M. & Zhao, S. (2019). 'Applications of machine learning in drug discovery and development', *Nature Reviews Drug Discovery*, Vol. 18, pp. 463–477.
5. Rao, A. S. & Verweij, G. (2017). 'Sizing the prize: What's the real value of AI for your business and how can you capitalise?'.

A MATTER OF PERSPECTIVE:

Discrimination, bias and inequality in AI

6. Solonec, T. (2000). 'Racial discrimination in the private rental market: Overcoming stereotypes and breaking the cycle of housing despair in Western Australia', *Indigenous Law Bulletin*, Vol. 5, No. 2, p. 4; Australian Human Rights Commission. (2002). Chapter 2, *Annual Report 2001-2002*; Australian Human Rights Commission. (2009). *DDA Conciliation: Goods, Services and Facilities*.
7. Blair, D. & Bernard, J. R. L. (eds.), *Macquarie Pocket Dictionary* (3rd ed): 'discriminate'.
8. *Street v Queensland Bar Association* (1989) 168 CLR 461, 570, Gaudron J.
9. Krywko, J. (2017). 'Siri can't talk to me: The challenge of teaching language to voice assistants', *Ars Technica*.
10. Rees, N., Rice, S. & Allen, D. (2018). *Australian Anti-Discrimination and Equal Opportunity Law* (3rd ed), The Federation Press, p. 52.
11. Toratani, M., Konno, M., Asai, A., Koseki, J., Kawamoto, K., Tamati, K., Li, Z., Sakai, D., Kudo, T., Satoh, T., Sato, K. Motooka, D., Okuzaki, D., Doki, Y., Mori, M., Ogawa, K. & Ishii, H. (2018). 'A convolutional neural network uses microscopic images to differentiate between mouse and human cell lines and their radioresistant clones', *Cancer Research*, Vol. 78, No. 23, p. 6703.
12. Buolamwini, J. & Gebru, T. (2018). 'Gender shades: Intersectional accuracy disparities in commercial gender classification', *Conference on Fairness, Accountability and Transparency*.
13. Rees, N., Rice, S. & Allen, D. (2018). *Australian Anti-Discrimination and Equal Opportunity Law* (3rd ed), The Federation Press, p. 52.
14. *Waterhouse v Bell* (1991) 25 NSWLR 99; *Daniels v Hunter Water Board* (1994) EOC 92-626.
15. Rees, N., Rice, S. & Allen, D. (2018). *Australian Anti-Discrimination and Equal Opportunity Law* (3rd ed), The Federation Press, p. 53.
16. Rees, N., Rice, S. & Allen, D. (2018). *Australian Anti-Discrimination and Equal Opportunity Law* (3rd ed), The Federation Press, p. 53.
17. O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Allen Lane, pp. 17–18.
18. Rees, N., Rice, S. & Allen, D. (2018). *Australian Anti-Discrimination and Equal Opportunity Law* (3rd ed), The Federation Press, p. 144; *Equal Opportunity Act 2010* (Vic), s 9(1).
19. Aronson, M. & Groves, M. (2013). *Judicial Review of Administrative Action* (5th ed), Thomson Reuters Australia, p. 610.
20. Hughes, J. M., Michell, P. A. & Ramson, W. S. (eds.) (1993). *The Australian Concise Oxford Dictionary* (2nd ed): 'bias'.
21. O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Allen Lane, pp. 20–21.
22. Office of Diversity and Outreach. 'State of science on unconscious bias', *University of California San Francisco*.
23. Office of the Victorian Information Commissioner. (2019). 'Submission to DIIS on Artificial Intelligence: Australia's Ethics Framework Discussion Paper'.
24. Friedler, S. A., Scheidegger, C. & Venkatasubramanian, S. (2016), 'On the (im)possibility of fairness', pp. 1-2.
25. Aronson, M. & Groves, M. (2013). *Judicial Review of Administrative Action* (5th ed), Thomson Reuters Australia, p. 610; Groves, M. (2017). 'The unfolding purpose of fairness', *Federal Law Review*, Vol. 45, No. 4, pp. 653-679.
26. Rees, N., Rice, S. & Allen, D. (2018). *Australian Anti-Discrimination and Equal Opportunity Law* (3rd ed), The Federation Press, pp. 12-17.
27. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J. & Schwartz, O. (2018). 'AI Now Report 2018', *AI Now Institute, New York University*; Chen, I., Johansson, F. D. & Sontag, D. (2018). 'Why is my classifier discriminatory?', *Advances in Neural Information Processing Systems*; Kim, M. P., Ghorbani, A. & Zou, J. (2018). 'Multiaccuracy: Black-box post-processing for fairness in classification'; Lahoti, P., Weikum, G. & Gummadi, K. P. (2018). 'iFair: Learning individually fair data representations for algorithmic decision making'.
28. Dale, S. (2015). 'Heuristics and biases: The science of decision-making', *Business Information Review*, Vol. 32, No. 2, p. 93; Bodenhausen, G. V. (1990). 'Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination', *Psychological Science*, Vol. 1, No. 5, p. 319.
29. For example, segregating prisoners with HIV/AIDS from other prisoners: *NC v Queensland Corrective Services Commission* [1997] QADT 22.
30. For example, refusing to provide an interpreter for a person with a hearing impairment who uses Auslan: *Woodforth v Queensland* [2017] QCA 100.
31. Australian Human Rights Commission. (2014). *Supporting working parents: Pregnancy and return to work national review*.

32. Danziger, S., Levav, J. & Avnaim-Pesso, L. (2011). 'Extraneous factors in judicial decisions', *Proceedings of the National Academy of Sciences*, Vol. 108, No. 17, p. 6889; however, the results have been disputed by Keren Weinshall-Margel and John Shapard, who suggest that the legal representation of prisoners may have more influence than the hunger of parole officials. See Weinshall-Margel, K. & Shapard, J, (2011). 'Overlooked factors in the analysis of parole decisions', *Proceedings of the National Academy of Sciences*, Vol. 108, No. 42, E833.

33. Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J. & Handelsman, J. (2012). 'Science faculty's subtle gender biases favor male students', *Proceedings of the National Academy of Sciences*, Vol. 109, No. 41, p. 16474.

34. Banerjee, R., Reitz, J. R. & Oreopoulos, P. (2018). 'Do large employers treat racial minorities more fairly? An analysis of Canadian field experiment data', *Canadian Public Policy*, Vol. 44, No. 1, p. 1; Booth, A. L., Leigh, A. & Varganova, E. (2012). 'Does ethnic discrimination vary across minority groups? Evidence from a field experiment', *Oxford Bulletin of Economics and Statistics*, Vol. 74, No. 4, p. 547; Chohan, U. W. (2016). 'Skin deep: Should Australia consider name-blind resumes?', *The Conversation*.

35. Lattice. (2017). 'How to reduce unconscious bias at work'; Uhlmann, E. L. & Cohen, G. L. (2007). '"I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination', *Organizational Behavior and Human Decision Processes*, Vol. 104, No. 2, p. 207.

36. Cowgill, B. (2018). 'Bias and productivity in humans and algorithms: Theory and evidence from resume screening', *Columbia Business School, Columbia University*.

37. Erel, I., Stern, L. H., Tan, C. & Weisbach, M. S. (2018). 'Selecting directors using machine learning', *National Bureau of Economic Research*.

38. Wharton Gates, S., Perry, V. G. & Zorn, P. M. (2002). 'Automated underwriting in mortgage lending: Good news for the underserved?', *Housing Policy Debate*, Vol. 13, No. 2, p. 369.

39. Khaitan, T. (2015). *A Theory of Discrimination Law*, Oxford University Press, pp. 130–132.

40. Henman, P. (2004). 'Targeted!: Population segmentation, electronic surveillance and governing the unemployed in Australia', *International Sociology*, Vol. 19, No. 2, p. 173.

41. Lane, S. (2017), 'Interview with Christian Porter, Minister for Social Services', *AM, Australian Broadcasting Corporation*.

42. Henman, P. (2004). 'Targeted!: Population segmentation, electronic surveillance and governing the unemployed in Australia', *International Sociology*, Vol. 19, No. 2, pp. 174-175.

43. Tay, L. (2012). 'Immigration Targets "problem Travellers" with Analytics', iTnews; Ajana, B. (2015). 'Augmented borders: Big Data and the ethics of immigration control', *Journal of Information, Communication & Ethics in Society*, Vol. 13, No. 1, p. 58.

44. Australian Human Rights Commission and World Economic Forum. (2019). *Artificial Intelligence: Governance and Leadership*.

45. Dastin, J. (2018). 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters*.

46. Senate Standing Committees on Community Affairs. (2019). 'Design, scope, cost-benefit analysis, contracts awarded and implementation associated with the better management of the social welfare system initiative', *Australian Parliament*, pp. 34–35.

47. Angwin, J. & Parris, T. (2016). 'Facebook lets advertisers exclude users by race', *ProPublica*.

48. Sonnad, N. (2018). 'US border agents hacked their "risk assessment" system to recommend detention 100% of the time', *Quartz*.

49. Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, pp. 39–40.

50. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). 'Machine bias', *ProPublica*; Dressel, J. & Farid, H. (2018). 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances*, Vol. 4, No. 1, p. 5580.

51. Palmiter Bajorek, J. (2019). 'Voice recognition still has significant race and gender biases', *Harvard Business Review*.

52. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J. & Schwartz, O. (2018). 'AI Now Report 2018', *AI Now Institute, New York University*, p. 25.

53. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J. & Schwartz, O. (2018). 'AI Now Report 2018', *AI Now Institute, New York University*, p. 25

54. Metz, R. (2016). 'Why Microsoft accidentally unleashed a neo-Nazi sexbot', *MIT Technology Review*.

55. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J. & Schwartz, O. (2018). 'AI Now Report 2018', *AI Now Institute, New York University*, p. 20; Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, St. Martin's Press, pp. 141–143.

56. Rice, S. (2013). 'Basic instinct: The heroic project of anti-discrimination law', Roma Mitchell Oration.

57. Rice, S. (2013). 'Basic instinct: The heroic project of anti-discrimination law', Roma Mitchell Oration.

58. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J. & Schwartz, O. (2018). 'AI Now Report 2018', *AI Now Institute, New York University*, p. 25.

59. *IW v City of Perth* (1997) 191 CLR 1, 59, 63.

60. Allen, D. (2009). 'Reducing the burden of proving discrimination in Australia', *Sydney Law Review*, Vol. 31, No. 4, p. 579.

61. O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Allen Lane, p. 8.

62. Nunes, I. & Jannach, D. (2017). 'A systematic review and taxonomy of explanations in decision support and recommender systems', *User Modeling and User-Adapted Interaction*, Vol. 27, No. 3-5, p. 393.

63. Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, pp. 80–83; Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J. & Schwartz, O. (2018). 'AI Now Report 2018', *AI Now Institute, New York University*, p. 11.

64. Pasquale, F. (2017). 'Toward a fourth law of robotics: Preserving attribution, responsibility, and explainability in an algorithmic society', *Ohio State Law Journal*, Vol. 78, p. 1243; Nunes, I. & Jannach, D. (2017). 'A systematic review and taxonomy of explanations in decision support and recommender systems', *User Modeling and User-Adapted Interaction*, Vol. 27, No. 3-5, p. 393.

65. Miller, K. (2017). 'Connecting the dots: A case study of the Robodebt communities', *Australian Institute of Administrative Law Forum*, No. 89, p. 50.

66. Kaminski, M. E. (2019). 'The right to explanation, explained', *Berkeley Technology Law Journal*, Vol. 34, No. 1, p. 189.

67. Rees, N., Rice, S. & Allen, D. (2018). *Australian Anti-Discrimination and Equal Opportunity Law* (3rd ed), The Federation Press, p. 767.

68. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J. & Schwartz, O. (2018). 'AI Now Report 2018', *AI Now Institute, New York University*; Chen, I., Johansson, F. D. & Sontag, D. (2018). 'Why is my classifier discriminatory?', *Advances in Neural Information Processing Systems*; Kim, M. P., Ghorbani, A. & Zou, J. (2018). 'Multiaccuracy: Black-box post-processing for fairness in classification'; Lahoti, P., Weikum, G. & Gummadi, K. P. (2018). 'iFair: Learning individually fair data representations for algorithmic decision making'.
69. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J. & Schwartz, O. (2018). 'AI Now Report 2018', *AI Now Institute, New York University*, p. 8.
70. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J. & Schwartz, O. (2018). 'AI Now Report 2018', *AI Now Institute, New York University*, p. 5.
71. Office of the Victorian Information Commissioner. (2019). 'Submission to DIIS on Artificial Intelligence: Australia's Ethics Framework Discussion Paper', p. 9.
72. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J. & Schwartz, O. (2018). 'AI Now Report 2018', *AI Now Institute, New York University*, p. 18-22.
73. Haack, P. & Sieweke, J. (2018). 'The legitimacy of inequality: Integrating the perspectives of system justification and social judgment', *Journal of Management Studies*, Vol. 55, No. 3, p. 486.
74. Australian Government Workplace Gender Equality Agency. (2019). 'Australia's gender pay gap statistics'.
75. *Re Lifestyle Communities Ltd (No 3)* (2009) 31 VAR 286; [2009] VCAT 1869, [137]-[141], [287]-[288].
76. Pound, A. & Evans, K. (2019). *Annotated Victorian Charter of Rights* (2nd ed), Thomson Reuters (Professional) Australia Limited, p. 116.
77. Belbin, R. (2018). 'When Google becomes the norm: The case for privacy and the right to be forgotten', *Dalhousie Journal of Legal Studies*, Vol. 26, p. 17.
78. Selinger, E. & Hartzog, W. (2014). 'Obscurity and Privacy', *Social Science Research Network*.
79. *Charter of Human Rights and Responsibilities Act 200*6, s 13.
80. *WBM v Chief Commissioner of Police* (2010) 27 VR 469; [2010] VSC 219, [51]-[57].
81. Hill, K. (2012). 'How Target figured out a teen girl was pregnant before her father did', *Forbes*.
82. Henman, P. (2004). 'Targeted!: Population segmentation, electronic surveillance and governing the unemployed in Australia', *International Sociology*, Vol. 19, No. 2, p. 179.
83. *Segerstedt-Wiberg v Sweden* (2007) 44 EHRR 2; [2006] ECHR 597, [105]-[107].
84. *Caripis v Victoria Police* [2012] VCAT 1472, [76]; *R (Countryside Alliance) v Attorney General* [2008] AC 719; [2007] UKHL 52, [17].

ALGORITHMIC TRANSPARENCY AND DECISION-MAKING ACCOUNTABILITY:
Thoughts for buying machine learning algorithms

85. Nissenbaum, H. (1996). 'Accountability in a computerized society', *Science and Engineering Ethics*, Vol. 2, pp. 25-42.
86. Australian Government. (2007). *Automated Assistance in Administrative Decision-Making: Better Practice Guide*; Australian Government. (2007). *Automated Assistance in Administrative Decision-Making: Better Practice Guide*: Summary of Checklist Points.
87. Burrell, J. (2016). 'How the machine thinks: Understanding opacity in machine learning algorithms', *Big Data & Society*, Vol. 3, No. 1, pp. 1-2.
88. Goodman, E. P. (2019). 'Smart algorithmic change requires a collaborative political process', *The Regulatory Review*.
89. Wickens, C., Clegg, B. A., Vieane, A. Z. & Sebok, A. (2015). 'Complacency and automation bias in the use of imperfect automation', *Human Factors: The Journal of Human Factors and Ergonomics Society*, Vol. 57, No. 5, p. 728; Skitka, L., Mosier, K. & Burdick, M. D. (2000). 'Accountability and Automation Bias', *International Journal of Human-Computer Studies*, Vol. 52, No. 4, p. 701.
90. Patel, F., Levinson-Waldman, R., DenUyl, S. & Koreh, R. (2019). 'Social media monitoring: How the Department of Homeland Security uses digital data in the name of national security', *Brennan Center for Justice*.
91. Harcourt, B. (2006). *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*, University of Chicago Press.
92. Perry, W. L., McInnis, B., Price, C. C., Smith, S. C. & Hollywood, J. S. (2013). 'Predictive policing: The role of crime forecasting in law enforcement operations' *Rand Corporation*; Uchida, C. (2013). 'Predictive policing' in Bruinsma, G. & Weisburd, D. (eds), *Encyclopedia of Criminology and Criminal Justice*, Springer, p. 3871; Stroud, M. (2014). 'The minority report: Chicago's new police computer predicts crimes, but is it racist?', *The Verge*; Nicholson, J. (2014). 'Detroit law enforcement's secret weapon: Big data analytics', *Venture Beat*.
93. Winston, A. (2018). 'Palantir has secretly been using New Orleans to test its predictive policing technology', *The Verge*.
94. Sentas, V. & Pandolfini, C. (2017). 'Policing young people in NSW: A study of the suspect targeting management plan', *A Report of the Youth Justice Coalition NSW*; McLean, A. (2018). 'Why Australia is quickly developing a technology-based human rights problem', *Tech Republic*; Seccombe, M. (2017).'"Predictive" policing in NSW', The Saturday Paper.
95. *NSW Legislative Assembly Questions and Answers No 162*. 2018.
96. *DEZ v Commissioner of Police, NSW Police Force* [2015] NSWCATAD 15.
97. Scassa, T. (2017). 'Law enforcement in the age of big data and surveillance intermediaries: Transparency challenges', *SCRIPTed*, Vol. 14, No. 2, p. 239.
98. Dressel, J. & Farid, H. (2018). 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances*, Vol. 4, No. 1.
99. O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Random House; Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, St Martin's Press.
100. Robinson, D. G. (2018). 'The challenges of prediction: Lessons from criminal justice', I/S: *A Journal of Law and Policy for the Information Society*, Vol. 14, No. 2, p. 151.
101. *State v Loomis*, 2015AP157-CR (WI, 2016).

102. Joh, E. E. (2017). 'The undue influence of surveillance technology companies on policing', *NYU Law Review*, Vol. 92, p. 101.

103. Administrative Review Council. (2004). *Automated Assistance in Administrative Decision Making: Report to the Attorney General*; Australian Government. (2007). *Automated Assistance in Administrative Decision Making: Better Practice Guide*.

104. [1943] 2 All ER 560.

105. *Re Smith & Australian Securities and Investments Commission* [2014] AAT 192; *B & L Whittaker Pty Ltd and ASIC and Anor* (2014) 106 IPR 361; *Boyce and Australian Securities and Investments Commission* [2015] ATT 768; *Stasiw v ASIC* [2015] AAT 328; *Re Swinburne v ASIC* [2014] AAT 602.

106. Wroe, D. (2018). 'Top official's "Golden Rule": In border protection, computer won't ever say no', *Sydney Morning Herald*.

107. Finkel, A. (2018). 'What kind of society do we want to be?', Keynote for Human Rights Commission Human Rights and Technology Conference.

108. *Council Directive 95/13/EC of 23 November 1993 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive)* [1995] OJ L281/31, art 15; *General Data Protection Regulation* [2016] OJ L 119/1, Art 22.

109. Mendoza, I. & Bygrave, L. A. (2017). 'The right not to be subject to automated decisions based on profiling', in Synodinou. T. E., Jougleux, P., Markou, C. & Prastitou, T. (eds), *EU Internet Law: Regulation and Enforcement*, Springer, p. 77.

110. Bayamlioglu, E. (2018). 'Transparency of automated decisions in the GDPR: An attempt for systemisation', Working Paper; UK Information Commissioner's Office. (2018). *Guide to the General Data Protection Regulation*.

111. *Bundesgerichtshof* [German Federal Court of Justice], VI ZR 156/13, 2014 reported in (2014 BGHZ) in the so-called SCHUFA case concerning the use of automated credit-scoring systems, concerning DPD Art 15.

112. Hildebrandt, M. (2019). 'Privacy as protection of the incomputable self: From agnostic to agonistic machine learning', *Theoretical Inquiries in Law*, Vol. 20, No. 1, p. 83.

113. Brauneis, R. & Goodman, E. (2018). 'Algorithmic accountability for the smart city', *Yale Journal of Law and Technology*, Vol. 20, p. 103.

114. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)* [2016] OJ L 119/1, Recital 63.

115. Selbst, A. D. & Barocas, S. (2018). 'The intuitive appeal of explainable machines', *Fordham Law Review*, Vol. 87, p. 1085.

116. Kroll, J. A. Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G. & Yu, H. (2017). 'Accountable algorithms', *University of Pennsylvania Law Review*, Vol. 3, p. 633.

117. Reisman, D., Schultz, J., Crawford, K. & Whittaker, M. (2018). 'Algorithmic impact assessments: A practical framework for public agency accountability', *AI Now Institute*.

118. Tutt, M. (2017). 'An FDA for algorithms', *Administrative Law Review*, Vol. 69, p. 83.

119. Citron, D. K. (2008). 'Technological due process', *Washington University Law Review*, Vol. 8, p. 1249.

120. Citron, D. K. & Pasquale, F. (2014). 'The scored society: Due process for automated predictions', *Washington Law Review*, Vol. 89, No. 1, p. 1.

121. New York City Council. (2018). *A Local Law in Relation to Automated Decision Systems Used by Agencies*, Pub L No 2018/049.

122. Powles, J. (2017). 'New York City's bold, flawed attempt to make algorithms accountable', *NewYorker*.

123. Goodman, B. W. (2016). 'A step towards accountable algorithms? Algorithmic discrimination and the European Union General Data Protection', Paper presented at the 29th Conference on Neural Information Processing Systems.

124. Borocas, S. & Selbst, A. (2016). 'Big data's disparate impact', *California Law Review*, Vol. 104, pp. 671-733.

125. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). 'Machine bias', *ProPublica*.

126. Narayanan, A. (2018). 'Translation tutorial: 21 fairness definitions and their politics', Tutorial delivered at Fairness, Accountability and Transparency Conference 2018, New York, citing Chouldechova, A. (2017). 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments', *Big Data*, Vol. 5, No. 2, p. 153.

127. Kleinberg, J., Mullainathan, S. & Raghavan, M. (2019). 'Inherent trade-offs in the fair determination of risk scores', Paper presented at the 8th Innovations in Theoretical Computer Science Conference.

128. Verma, S. & Rubin, J. (2018). 'Fairness definitions explained', *ACM/IEEE International Workshop on Software Fairness*, p. 1.

129. Kusner, M. J., Loftus, J. R., Russell, C. & Silva, R. (2017). 'Counterfactual fairness'.

130. Goodman, B. W. (2016). 'A step towards accountable algorithms? Algorithmic discrimination and the European Union General Data Protection', Paper presented at the 29th Conference on Neural Information Processing Systems, citing Dodge, Y. (2003). 'Interaction effect', *Oxford Dictionary of Statistical Terms*, Oxford.

131. Henrique-Gomez, L. (2019). 'Centrelink cancels 40,000 Robotdebts, new figures reveal', *The Guardian*.

132. Norris, C. & L'Hoiry, X. (2014). 'What do they know? Exercising subject access rights in democratic societies', Paper presented at the 6th Biannual Surveillance and Society Conference.

133. Wachter, S., Mittelstadt, B. & Floridi, L. (2017). 'Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation', *International Data Privacy Law*, Vol. 7, No. 2, p. 76.

134. Selbst, A. & Powles, J. (2017). 'Meaningful information and the right to explanation', *International Data Privacy Law*, Vol. 7, No. 4, p. 233.

135. Abdul , A., Vermeulen, J., Wang, D., Lim, B. Y. & Kankanhalli, M. (2018). 'Trends and trajectories for explainable, accountable, and intelligible systems: An HCI research agenda', Paper presented at the ACM Conference of Human Factors in Computing Systems.

136. Gunning, G. (2017). 'Explainable artificial intelligence (XAI)', DARPA/I20.

137. Samek, W., Wiegand, T. & Müller, K. R. (2017). 'Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models', *ITU Journal: ICT Discoveries*, Special Issue No 1, p. 1.

138. Besold, T. R. & Uckelman, S. L. (2018). 'The what, the why, and the how of artificial explanations in automated decision-making'.

139. Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). '"Why should I trust you?" Explaining the predictions of any classifier', Paper presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 1135.

140. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D. & Wood, A. (2017). 'Accountability of AI under the law: The role of explanation'.

141. Wachter, S., Mittelstadt, B. & Russell, C. (2018). 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR', *Harvard Journal of Law & Technology*, Vol. 31, p. 841.

142. Mittelstadt, B., Russell, C. & Wachter, S. (2019). 'Explaining explanations in AI', Paper presented at 2019 Fairness, Accountability and Transparency Conference.

143. Lipton, Z. C. (2016). 'The mythos of model interpretability', Paper presented at the 2016 Workshop on Human Interpretability in Machine Learning.

144. Mittelstadt, B., Russell, C. & Wachter, S. (2019). 'Explaining explanations in AI', Paper presented at 2019 Fairness, Accountability and Transparency Conference, Atlanta, GA.

145. Edwards, L. & Veale, M. (2017). 'Slave to the algorithm? Why a 'right to explanation' is probably not the remedy you are looking for', *Duke Law & Technology* Review, Vol. 16, p. 18.

146. Miller, T., Howe, P. & Sonenberg, L. (2017). 'Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences'; Miller T. (2017). 'Explanation in artificial intelligence: Insights from the social sciences'.

147. Yeung, K. & Weller, A. (2018). 'How is "transparency" understood by legal scholars and the machine learning community?', in Bayamlioglu, E. et al (eds), *Being Profiled: Cogitas Ergo Sum (10 Years of 'Profiling the European Citizen')*, Amsterdam University Press, p. 36.

148. Gonzalez Fuster, G. (2018). 'Transparency as translation in data protection', in Bayamlioglu, E. et al (eds), *Being Profiled: Cogitas Ergo Sum (10 Years of 'Profiling the European Citizen')*, Amsterdam University Press, p. 52.

149. Pasquale, F. (2018). 'Odd numbers', *Real Life Magazine*.

150. Katz, Y. (2017). 'Manufacturing an artificial intelligence revolution'.

AI IN THE PUBLIC INTEREST

151. Sentryo. (2017). 'The 4 industrial revolutions'.

152. Hashimoto, Y., Murase, H., Morimoto, T. & Torii, T. (2001). 'Intelligent systems for agriculture in Japan', *IEEE Control Systems Magazine*, Vol. 21, No. 5, pp. 71-85.

153. Chen, D. L. (2019). 'Machine Learning and the Rule of Law', in M. Livermore and D. Rockmore (eds.) *Computational Analysis of Law*, Santa Fe Institute Press (forthcoming).

154. Royal Astronomical Society. (2019). 'Deep-CEE: The AI deep learning tool helping astronomers explore deep space', *ScienceDaily*.

155. Ekins, S. (2016). 'The Next Era: Deep Learning in Pharmaceutical Research,' *Pharmaceutical Research*, Vol. 33, No. 11, pp 2594-603.

156. Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., Henke, N. & Trench, M. (2017). 'Artificial Intelligence: The Next Digital Frontier?', *McKinsey Global Institute*.

157. Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P. & Malhotra, S. (2018). 'Notes from the AI frontier: Applications and value of deep learning'. *McKinsey Global Institute*.

158. Austroads. (2016). 'Congestion and Reliability Review'.

159. Advanced Data Analytics in Transport team. (2019). 'How data science can help you beat traffic congestion', *Data 61, CSIRO*; Analytics Magazine. (2018). 'Big Data helps city of Dublin improve public bus transportation and reduce congestion'; Wen, T., Mihaita, A. S., Nguyen, H., Cai, C. & Chen, F. (2018). 'Integrated incident decision-support using traffic simulation and data-driven models', *Transportation Research Record*, Vol. 2672, No. 42, pp. 247-256.

160. Li, Z., Zhang, B., Wang, Y., Chen, F., Taib, R., Whiffin, V. & Wang, Y. (2014). 'Water pipe condition assessment: A hierarchical beta process approach for sparse incident data', *Machine Learning*, Vol. 95, No. 1, pp. 11–26.

161. Li, Z., Zhang, B., Wang, Y., Chen, F., Taib, R., Whiffin, V. & Wang, Y. (2014). 'Water pipe condition assessment: A hierarchical beta process approach for sparse incident data', Machine Learning, Vol. 95, No. 1, pp. 11–26; Zhou, J., Sun, J., Wang, Y. & Chen, F. (2017). 'Wrapping practical problems into a machine learning framework: Using water pipe failure prediction as a case study', *International Journal of Intelligent Systems Technologies and Application*s, Vol. 16, No. 3, pp. 191–207.

162. Whiffin, V., Crawley, C., Wang, Y., Li, Z. & Chen, F. (2013). 'Evaluation of machine learning for predicting critical main failure', *Water Asset Management International*, Vol. 9, No. 4, pp. 17–20.

163. Data61, CSIRO. 'Helping to maintain Sydney Harbour Bridge'.

164. Polizzi, G. & Liebman, A. (2019). 'AI in Australia's electricity sector', *Electrical Comms Data*.

165. Fraunhofer-Gesellschaft. (2019). 'Artificial intelligence automatically detects disturbances in power supply grids', *PhysOrg*.

166. Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A. & Conway, D. (2016). *Robust Multimodal Cognitive Load Measurement*, Springer International Publishing.

167. Marr, B. (2018). 'How is AI used in education - Real world examples of today and a peek into the future, *Forbes*.

168. Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., Henke, N. & Trench, M. (2017). 'Artificial intelligence: The next digital frontier?', *McKinsey Global Institute*.

169. Burt, C. (2019). 'Researchers develop AI method for movement identification and tracking without facial recognition', *Biometric Update*.

170. Meng, A. (2015). '"World's first" facial recognition ATM unveiled in China', *South China Morning Post.*

171. Schneider, K. (2018). 'Big change coming to the way we fly', *News.com.au*.

172. Datatilsynet, The Norwegian Data Protection Authority. (2018). 'Artificial Intelligence and privacy'.

173. Radebaugh, G. & Erlingsson, U. (2019). 'Introducing TensorFlow privacy: Learning with differential privacy for training data', *Medium*.

174. McMahan, H.B., Andrew, G., Erlingsson, U., Chien, S., Mironov, I., Papernot, N. & Kairouz, P. (2018). 'A general approach to adding differential privacy to iterative training procedures'.

175. Yang, Q., Liu, Y., Chen, T. & Tong, Y. (2019). 'Federated machine learning: Concept and applications', *ACM Transactions on Intelligent Systems and Technology*, Vol. 10, No. 2, pp. 1-19.

176. Zhou, J. & Chen, F. (eds.) (2018). *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, Springer International Publishing.

177. Lee, J. D. & See, K. A. (2004). 'Trust in automation: Designing for appropriate reliance'. *Human Factors*, Vol. 46, No. 1, pp. 50-80.

178. Carrasco, M., Mills, S., Whybrew, A. & Jura, A. (2019). 'The citizen's perspective on the use of AI in government', *Boston Consulting Group*.

179. Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J. & Hajkowicz, S. (2019). 'Artificial Intelligence: Australia's Ethics Framework, *Data61, CSIRO*.

ALGORITHMS, NEURAL NETWORKS AND OTHER MACHINE LEARNING TECHNIQUES

180. Huang, J. (2017). 'AI is eating software', *Nvidia*; Parloff, R. (2016). 'Why deep learning is suddenly changing your life', Fortune; Stevens, M. (2019). 'AI for research pragmatists: What it means and what you can use it for today', *Market Research Summit*, London.

181. Goodfellow, I. (2019). 'Adversarial Machine Learning', 7th *International Conference on Learning Representations*.

182. Turing, A. (1950). 'Computing Machinery and Intelligence', *Mind*, Vol. LIX, No. 236, pp. 433-460.

183. Valiant, L. (2010). ACM Turing Award.

184. Samuel, A. (1959). 'Some studies in Machine Learning using the Game of Checkers', *IBM Journal of Research and Development*, Vol. 3, No. 3, pp. 210-229.

185. Mullins, J. (2007). 'Checkers 'solved' after years of number crunching', *NewScientist*.

186. Samuel, A. (1959). 'Some studies in Machine Learning using the Game of Checkers', *IBM Journal of Research and Development*, Vol. 3, No. 3, pp. 210-229.

187. Rumelhart, D. E., Hinton, G. & Williams, R. J. (1986). 'Learning representations by back-propagating errors', *Nature*, Vol. 323, pp 533-536.

188. Hinton, G., Bengio, Y. & LeCun, Y. (2019). ACM Turing Award.

189. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278-2323.

190. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278-2323.

191. Vance, A. (2018). 'This man is the godfather the AI community wants to forget', *Bloomberg Businessweek*; HochreiterS. & Schmidhuber, J. (1997). 'Long short-term memory', *Neural Computation*, Vol 9, No. 8, pp. 1735-1780.

192. MIT Technology Review. (2014). 'The revolutionary technique that quietly changed machine vision forever'.

193. Dastin, J. (2018). 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters*.

194. Tashea, J. (2017). 'Courts are using AI to sentence criminals. That must stop now', *Wired*.

195. Worland, J. (2016). 'Microsoft takes Chatbot Offline after it starts Tweeting Racist Messages', *Time*.

196. Hao, K. (2019). 'Training a single AI model can emit as much carbon as five cars in their lifetimes', *MIT Technology Review*; Ausick, P. (2019). 'The dirty expensive secret of artificial intelligence and machine learning', *24/7 Wall St*; Strubell, E., Ganesh, A. & McCallum, A. (2019). 'Energy and policy considerations for deep learning in NLP', *57th Annual Meeting of the Association for Computational Linguistics*.

197. Hinton, G., Bengio, Y. & LeCun, Y. (2019). ACM Turing Award.

198. Rumelhart, D. E., Hinton, G. & Williams, R. J. (1986). 'Learning representations by back-propagating errors', *Nature*, Vol. 323, pp. 533-536.

199. Parkin, S. (2019). 'The rise of the deepfake and the threat to democracy', *The Guardian*.

200. Parkin, S. (2019). 'The rise of the deepfake and the threat to democracy', *The Guardian*.

201. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T. & Hassabis, D. (2017). 'Mastering the game of Go without human knowledge', *Nature*, Vol. 550, pp. 354-359.

202. Youyou, W., Kosinski, M. & Stillwell, D. (2015). 'Computer-based personality judgments are more accurate than those made by humans', *Proceedings of the National Academy of Sciences USA*, Vol. 112, No. 4, pp. 1036-1040.

203. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. & Song, D. (2018). 'Robust physical-world attacks on deep learning visual classification', *IEEE International Conference on Computer Vision and Pattern Recognition*.

DATA SECURITY AND AI

204. Biggio, B. & Roli, F. (2018). 'Wild patterns: Ten years after the rise of adversarial machine learning', *Pattern Recognition*, Vol. 84, pp. 317-331; Joseph, A. D., Nelson, B., Rubinstein, B. & Tygar, J. D. (2019). *Adversarial Machine Learning*, Cambridge University Press; Vorobeychik, Y. & Kantarcioglu, M. (2018). *Adversarial Machine Learning*, Morgan & Claypool.

205. Lowd, D. & Meek, C. (2005). 'Adversarial learning', *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 641-647.

206. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2013). *Intriguing properties of neural networks*.

207. Barreno, M., Nelson, B., Sears, R., Joseph, A. D. & Tygar, J. D. (2006). 'Can machine learning be secure?', *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, pp. 16-25.

208. Joseph, A. D., Nelson, B., Rubinstein, B. & Tygar, J. D. (2019). *Adversarial Machine Learning*, Cambridge University Press.

209. Alfeld, S., Zhu, X. & Barford, P. (2016). 'Data poisoning attacks against autoregressive models', *Thirtieth AAAI Conference on Artificial Intelligence*.

210. Huang, L., Joseph, A., Nelson, B., Rubinstein, B. & Tygar, J. (2011). 'Adversarial machine learning', *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, pp. 43-58.

211. Fredrikson, M., Jha, S. & Ristenpart, T. (2015). 'Model inversion attacks that exploit confidence information and basic countermeasures', *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322-1333.

212. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D. & Ristenpart, T. (2014). 'Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing', *23rd USENIX Security Symposium*, pp. 17-32.

213. McSherry, F. (2016). 'Statistical inference considered harmful'.

214. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. (2017). 'Membership inference attacks against machine learning models', *2017 IEEE Symposium on Security and Privacy*, pp. 3-18.

215. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. (2017). 'Membership inference attacks against machine learning models', *2017 IEEE Symposium on Security and Privacy*, pp. 3-18.

216. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. (2017). 'Membership inference attacks against machine learning models', *2017 IEEE Symposium on Security and Privacy*, pp. 3-18.

217. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. (2017). 'Membership inference attacks against machine learning models', *2017 IEEE Symposium on Security and Privacy*, pp. 3-18.

218. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D. & Ristenpart, T. (2014). 'Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing', *23rd USENIX Security Symposium*, pp. 17-32.

219. Narayanan, A. & Shmatikov, V. (2008). 'Robust de-anonymization of large sparse datasets', *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pp. 111-125.

220. Culnane, C., Rubinstein, B. & Teague, V. (2017). 'Health data in an open world'.

221. Culnane, C., Rubinstein, B. & Teague, V. (2017). 'Health data in an open world'.

222. Office of the Victorian Information Commissioner. (2018). 'Protecting unit-record level personal information: The limitations of de-identification and the implications for the *Privacy and Data Protection Act 2014'*.

223. Culnane, C., Rubinstein, B. & Teague, V. (2017). 'Vulnerabilities in the use of similarity tables in combination with pseudonymisation to preserve data privacy in the UK Office for National Statistics' privacy-preserving record linkage'.

224. Culnane, C., Rubinstein, B. & Teague, V. (2017). 'Privacy assessment of de-identified Opal data: A report for Transport for NSW'.

225. Garfinkel, S., Abowd, J. & Martindale, C. (2018). 'Understanding database reconstruction attacks on public data', *Queue*, Vol. 16, No. 5.

226. Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2006). 'Calibrating noise to sensitivity in private data analysis', *Theory of Cryptography Conference*, pp. 265-284.

227. Office of the Victorian Information Commissioner. (2018). 'Protecting unit-record level personal information: The limitations of de-identification and the implications for the *Privacy and Data Protection Act 2014'*, p. 16.

228. Garfinkel, S., Abowd, J. & Martindale, C. (2018). 'Understanding database reconstruction attacks on public data', *Queue*, Vol. 16, No. 5.

229. Erlingsson, U., Pihur, V. & Korolova, A. (2014). 'Rappor: Randomized aggregatable privacy-preserving ordinal response', *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054-1067.

230. Tang, J., Korolova, A., Bai, X., Wang, X. & Wang, X. (2017). 'Privacy loss in Apple's implementation of differential privacy on MacOS 10.12'.

231. Johnson, N., Near, J. & Song, D. (2018). 'Towards practical differential privacy for SQL queries', *Proceedings of the VLDB Endowment*, pp. 526-539.

232. Culnane, C., Rubinstein, B. & Teague, V. (2017). 'Privacy assessment of de-identified Opal data: A report for Transport for NSW'.

233. Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2006). 'Calibrating noise to sensitivity in private data analysis', *Theory of Cryptography Conference*, pp. 265-284.

234. McSherry, F. & Talwar, K. (2007). 'Mechanism design via differential privacy', *48th Annual IEEE Symposium on Foundations of Computer Science*, pp. 94-103.

235. Chaudhuri, K., Monteleoni, C. & Sarwate, A. (2011). 'Differentially private empirical risk minimization', *Journal of Machine Learning Research*, Vol. 12, pp. 1069-1109.

236. Lyu, M., Su, D. & Li, N. (2017). 'Understanding the sparse vector technique for differential privacy', *Proceedings of the VLDB Endowment*, pp. 637-648.

237. Dwork, C. & Roth, A. (2014). 'The algorithmic foundations of differential privacy', *Foundations and Trends in Theoretical Computer Science*, Vol. 9, No. 3-4, pp. 211-407.

238. Rubinstein, B. & Alda, F. (2017). 'Pain-free random differential privacy with sensitivity sampling', *Proceedings of the 34th International Conference on Machine Learning*, pp. 2950-2959.

239. Rubinstein, B. (2017). 'diffpriv open-source R package'.

240. Dwork, C. & Roth, A. (2014). 'The algorithmic foundations of differential privacy', *Foundations and Trends in Theoretical Computer Science*, Vol. 9, No. 3-4, pp. 211-407.

241. Chaudhuri, K., Monteleoni, C. & Sarwate, A. (2011). 'Differentially private empirical risk minimization', *Journal of Machine Learning Research*, Vol. 12, pp. 1069-1109.

242. Rubinstein, B., Bartlett, P., Huang, L. & Taft, N. (2012). 'Learning in a large function space: Privacy-preserving mechanisms for SVM learning', *Journal of Privacy and Confidentiality*, Vol. 4, No. 1, pp. 65-100.

243. Abadi, M., Chu, A., Goodfellow, I., McMahan, H., Mironov, I., Talwar, K. & Zhang, L. (2016). 'Deep learning with differential privacy', *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308-318.

244. Johnson, N., Near, J. & Song, D. (2018). 'Towards practical differential privacy for SQL queries', *Proceedings of the VLDB Endowment*, pp. 526-539.

245. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H., Patel, S., Ramage, D., Segal, A. & Seth, K. (2017). 'Practical secure aggregation for privacy-preserving machine learning', *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175-1191.

246. Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D. & Wang, T. (2018). 'Privacy at scale: Local differential privacy in practice', *Proceedings of the 2018 International Conference on Management of Data (SIGMOD)*, pp. 1655-1658.

247. Kolosnjaji, B., Demontis, A., Biggio, B., Maiorca, D., Giacinto, G., Eckert, C. & Roli, F. (2018). 'Adversarial malware binaries: Evading deep learning for malware detection in executables', *26th European Signal Processing Conference (EUSIPCO)*, pp. 533-537.

248. Tan, K., Kevin, K. & Maxion, R. (2002). 'Undermining an anomaly-based intrusion detection system using common exploits', *International Workshop on Recent Advances in Intrusion Detection*, pp. 54-73.

249. Wittel, G. & Wu, S. (2004). 'On attacking statistical spam filters', *Proceedings of the Conference on Email and Anti-Spam*.

250. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2013). 'Intriguing properties of neural networks'.

251. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G. & Roli, F. (2013). 'Evasion attacks against machine learning at test time', *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* pp. 387-402.

252. Baydin, A., Pearlmutter, B., Radul, A. & Siskind, J. (2018). 'Automatic differentiation in machine learning: A survey', *Journal of Machine Learning Research*, Vol. 18, pp. 1-43.

253. Liu, Y., Chen, X., Liu, C. & Song, D. (2017). 'Delving into transferable adversarial examples and black-box attacks', *International Conference on Learning Representations*.

254. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. & Swami, A. (2017). 'Practical black-box attacks against machine learning', *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*, pp. 506-519.

255. Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O. & Frossard, P. (2017). 'Universal adversarial perturbations', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1765-1773.

256. Kurakin, A., Goodfellow, I. & Bengio, S. (2018). 'Adversarial examples in the physical world', in Yampolskiy, R. V. (ed.), *Artificial Intelligence Safety and Security*, Taylor & Francis.

257. Brown, T., Mane, D., Roy, A., Abadi, M. & Gilmer, J. (2017). 'Adversarial patch'.

258. Thys, S., Van Ranst, W. & Goedeme, T. (2019). 'Fooling automated surveillance cameras: Adversarial patches to attack person detection', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

259. Carlini, N. & Wagner, D. (2018). 'Audio adversarial examples: Targeted attacks on speech-to-text', *IEEE Security and Privacy Workshops*, pp. 1-7.

260. Nelson, B., Barreno, M., Chi, F., Joseph, A., Rubinstein, B., Saini, U., Sutton, C., Tyar, J. D. & Xia, K. (2008). 'Exploiting machine learning to subvert your spam filter', *LEET '08 Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threat*, pp. 1-9.

261. Rubinstein, B., Nelson, B., Huang, L., Joseph, A., Lau, S.-h., Rao, S., Taft, N. &. Tygar, J. (2009). 'ANTIDOTE: Understanding and defending against poisoning of anomaly detectors', *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, pp. 1-14.

262. Chen, S., Xue, M., Fan, L., Hao, S., Xu, L., Zhu, H. & Li, B. (2018). 'Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach', *Computers & Security*, Vol. 73, pp. 326-344.

263. Gu, T., Dolan-Gavitt, B. & Garg, S. (2017). 'Badnets: Identifying vulnerabilities in the machine learning model supply chain'.

264. Athalye, A., Carlini, N. & Wagner, D. (2018). 'Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples', *Proceedings of the 35th International Conference on Machine Learning*, pp. 274-283.

265. Huber, P. J. (1981). *Robust Statistics*, John Wiley & Sons.

266. Goodfellow, I., Shlens, J. & Szegedy, C. (2014). 'Explaining and harnessing adversarial examples'.

267. Cohen, J., Rosenfeld, E. & Kolter, J. (2019). 'Certified adversarial robustness via randomized smoothing'.

268. Sarraute, C., Buffet, O. & Hoffmann, J. (2012). 'POMDPs make better hackers: Accounting for uncertainty in penetration testing', *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

269. Narayanan, A., Paskov, H., Gong, N., Bethencourt, J., Stefanov, E., Shin, E. & Song, D. (2012). 'On the feasibility of internet-scale author identification', *IEEE Symposium on Security and Privacy*, pp. 300-314.

270. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). 'Generative adversarial networks', *Proceedings of the International Conference on Neural Information Processing*, pp. 2672–2680.

REGULATING AI

271. Duval, Y. N. (2016). *Homo Deus: A brief History of Tomorrow*, Vintage Digital, p. 363.

272. Nemitz, P. (2018). 'Constitutional democracy and technology in the age of artificial intelligence', *Royal Society*, Vol. 367, No. 2133.

273. Nemitz, P. (2018). 'Constitutional democracy and technology in the age of artificial intelligence', *Royal Society*, Vol. 367, No. 2133, pp. 3-4.

274. *Criminal Code Act 1995* (Criminal Code), Part 10.7.

275. Senate Standing Committees on Rural and Regional Affairs and Transport. (2018). 'Regulatory requirements that impact on the safe use of Remotely Piloted Aircraft Systems, Unmanned Aerial Systems and associated systems', *Parliament of Australia*.

276. Civil Aviation Safety Authority. (2019). 'Remotely piloted aircraft (RPA) registration and RPAS operator accreditation scheme', Project US 18/09.

277. National Transport Commission. (2017). 'Automated vehicles in Australia'.

278. Stankovic, M., Gupta, R., Rossert, B. A., Myers, G. I. & Nicoli, M. (2017). 'Exploring legal, ethical and policy implications of artificial intelligence' (Draft), *Law, Justice and Development*.

279. Federal Ministry of Transport and Digital Infrastructure, Ethics Commission. (2017). 'Automated and connected driving'.

280. Centre for Connected and Driverless Cars. (2019). 'Code of Practice: Automated vehicle trialling'.

281. United Nations Economic Commission for Europe. (2019). 'Autonomous transport must be developed with a global eye'.

282. *Competition and Consumer Act 2010* (Cth), Schedule 2, *The Australian Consumer Law*, s 138.

283. *Competition and Consumer Act 2010* (Cth), Schedule 2, *The Australian Consumer Law*, s 142.

284. European Commission. (2018). 'European Commission Staff Working Document: Liability for emerging digital technologies'.

285. European Commission. (2018). 'European Commission Staff Working Document: Liability for emerging digital technologies, pp. 20-21.

286. *Migration Act 1985 (Cth), s 495A(1); Justice Perry, M. (2019). 'idecide: Digital pathways to decision', Federal Court of Australia.*

287. *Therapeutic Goods Act 1989 (Cth), s 7C(2); Social Security (Administration) Act 1999* (Cth), s 6A; Hogan-Doran, D. (2017). 'Computer says "no": Automation, algorithms and artificial intelligence in government decision-making', *The Judicial Review: Selected Conference Papers: Journal of the Judicial Commission of New South Wales*, Vol. 13, No. 3; Elvery, S. (2017). 'How algorithms make important government decisions', *The Age*.

288. Commonwealth Ombudsman. (2017). 'Centrelink's automated debt raising and recovery system: A report about the Department of Human Services' online compliance intervention system for debt raising and recovery'; Commonwealth Ombudsman. (2019). 'Centrelink's automated debt raising and recovery system: Implementation report'.

289. Henriques-Gomez, H. (2019). 'Centrelink robodebt scheme faces second legal challenge', *The Guardian*.

290. *Competition and Consumer Act 2010* (Cth), Schedule 2, *The Australian Consumer Law*, s 18.

291. *Competition and Consumer Act 2010* (Cth), s 131.

292. Department of Justice. (2015). 'Former e-commerce executive charged with price fixing in the antitrust division's first online marketplace prosecution', *Justice News of the US Department of Justice*.

293. Sims, R. (2017). 'The ACCC's approach to colluding robots', *Australian Competition and Consumer Commission*.

294. *Competition and Consumer Act 201*0 (Cth), s 45(1)(c).

295. Australian Competition and Consumer Commission. (2018). 'Guidelines on concerted practices', cl 1.3.

296. *Competition and Consumer Act 2010* (Cth), s 46.

297. Office of the Information Commissioner. (2017). 'Big data, artificial intelligence, machine learning and data protection'; Office of the Australian Information Commissioner. (2018). 'Guide to data analytics and the Australian Privacy Principles'.

298. Office of the Australian Information Commissioner. (2018). 'Guide to data analytics and the Australian Privacy Principles', p. 10.

299. Office of the Information Commissioner. (2017). 'Big data, artificial intelligence, machine learning and data protection'.

300. Gole, T., Burns, S., Caplan, M., Hii, A., McGregor, S., Sutton, L., Fai , M. & Yuen, A. (2019). 'Australia's privacy and consumer laws to be strengthened', *Lexology*.

301. Otega, P. A., Maini, V. & DeepMind Safety Team. (2018). 'Building safe artificial intelligence: specification, robustness, and assurance', *Medium*.

302. Hunt, E. (2016). 'Tay, Microsoft's AI chatbox, gets a crash course in racism from Twitter', *The Guardian*.

303. Term coined by Pasquale, F. (2015). *The Black Box society: the secret algorithms that control money and information*, Harvard University Press.

304. Office of the Victorian Information Commissioner. (2019). 'Submission in response to the *Artificial Intelligence: Australia's Ethic Framework* Discussion Paper', p. 3.

305. *Wisconsin v Loomis*, 881 NW 2d 749. (2016); Houston Federation of Teachers vs Houston Independent School District. (2017). Amended Summary Judgment Opinion, *US District Court of Southern District of Texas*; American Federation of Teachers. (2017). 'Federal suit settlement: End of value-added measures for teacher termination in Houston', Press Release.

306. [2015] AATA 956.

307. Miller, K. (2016). 'The application of administrative law principles to technology-assisted decision-making', *Australian Institute of Administrative Law Forum*, No. 86, pp. 28-29.

308. Lecher, C. (2019). 'New York's algorithm task force is fracturing', *The Verge*.

309. Pangburn, D. J. (2019). 'Washington could be the first state to rein in automated decision-making', *Fast Company*.

310. Dastin, J. (2018). 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters Business News*.

311. Centre for Data Ethics and Innovation. (2019). 'Centre for Data Ethics 2 Year Strategy', Independent report.

312. General Data Protection Regulation, Article 22.

313. General Data Protection Regulation, Article 22(2)(a) and (c).

314. Republic of Estonia. (2018). 'Estonia will have an artificial intelligence strategy'; Kaevats, M. (2018). 'AI and the Kratt momentum', *Invest in Estonia*; Tashea, J. (2017). 'Estonia considering new legal status for artificial intelligence', *ABA Journal*.

315. Republic of Estonia. (2018). 'Estonia will have an artificial intelligence strategy'.

316. European Parliament Legislative Observatory. (2015). 'Civil Laws for Robotics' (2015/2103(INL).

317. European Commission. (2018). 'European Commission Staff Working Document: Liability for emerging digital technologies';Barfield, W. (2018). 'Liability for autonomous and artificially intelligent robots', *De Gruyter*, Vol. 9, p. 198.

318. Burnie Port Authority v General Jones Pty Ltd (1994) HCA 13.

319. Consumer Affairs Victoria. (2019). 'Motor Car Traders Guarantee Fund'.

320. WorkSafe Victoria, 'How to register for WorkCover insurance'.

321. *Copyright Act 1968* (Cth), s 32(4).

322. *Copyright, Designs and Patents Act 1998* (UK), s 9(3); *Copyright Act 1994* (NZ), s 5(20(a)); Allens. (2019). 'AI Toolkit', p. 20.

323. Institute of Electrical and Electronics Engineers. (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*.

324. Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J. & Hajkowicz, S. (2019). 'Artificial Intelligence: Australia's Ethics Framework, *Data61, CSIRO*; UK House of Lords. (2018). 'AI in the UK: Ready, willing and able?'.
325. European Commission High-Level Expert Group on Artificial Intelligence (AIHLEG). (2019). 'Ethics guidelines for trustworthy AI'.
326. Shaw, G. (2019). 'The future computed: AI & manufacturing', Microsoft.
327. Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J. & Hajkowicz, S. (2019). 'Artificial Intelligence: Australia's Ethics Framework, *Data61, CSIRO*, p. 57.
328. Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J. & Hajkowicz, S. (2019). 'Artificial Intelligence: Australia's Ethics Framework, *Data61, CSIRO*, pp. 58-62.
329. European Commission High-Level Expert Group on Artificial Intelligence (AIHLEG). (2019). 'Ethics guidelines for trustworthy AI', p. 7.
330. European Commission High-Level Expert Group on Artificial Intelligence (AIHLEG). (2019). 'Ethics guidelines for trustworthy AI', p. 12.
331. European Commission High-Level Expert Group on Artificial Intelligence (AIHLEG). (2019). 'Ethics guidelines for trustworthy AI', p. 13.
332. European Commission High-Level Expert Group on Artificial Intelligence (AIHLEG). (2019). 'Ethics guidelines for trustworthy AI', p. 16.
333. European Commission High-Level Expert Group on Artificial Intelligence (AIHLEG). (2019). 'Ethics guidelines for trustworthy AI', p. 14.
334. European Commission High-Level Expert Group on Artificial Intelligence (AIHLEG). (2019). 'Ethics guidelines for trustworthy AI', p. 20.
335. European Commission High-Level Expert Group on Artificial Intelligence (AIHLEG). (2019). 'Ethics guidelines for trustworthy AI', p. 22.
336. European Commission High-Level Expert Group on Artificial Intelligence (AIHLEG). (2019). 'Ethics guidelines for trustworthy AI', pp. 22-23.
337. European Commission High-Level Expert Group on Artificial Intelligence (AIHLEG). (2019). 'Ethics guidelines for trustworthy AI', p. 24.
338. European Group on Ethics in Science and New Technologies. (2018). 'Statement on artificial intelligence, robotics and 'autonomous' systems', *European Commission*.
339. AI4People. (2019). 'An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations'.
340. International Conference of Data Protection and Privacy Commissioners. (2018). 'Declaration on ethics and data protection in artificial intelligence'.
341. OECD Council on Artificial Intelligence. (2019). 'Recommendation of the council on artificial intelligence'.
342. Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J. & Hajkowicz, S. (2019). 'Artificial Intelligence: Australia's Ethics Framework, *Data61, CSIRO*, p. 16.
343. Institute of Electrical and Electronics Engineers. (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*.
344. International Organization for Standardization. (2017). 'ISO/IEC JTC 1/SC 42: Artificial Intelligence'.
345. Australian Human Rights Commission and World Economic Forum. (2019). *Artificial Intelligence: Governance and Leadership*.
346. Australian Human Rights Commission and World Economic Forum. (2019). *Artificial Intelligence: Governance and Leadership, p. 16*.
347. Out-Law. (2019). 'AI audit framework on ICO agenda', *Pinsent Masons*.
348. Out-Law. (2018). 'Driverless cars insurance laws receives Royal Assent', *Pinsent Masons*.