How to cite:

R. Nock, “Algorithms, neural networks and other machine learning techniques”
in *Closer to the Machine: Technical, social and legal aspects of AI,* pp 79—102
Office of the Victorian Information Commissioner, 2019

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.

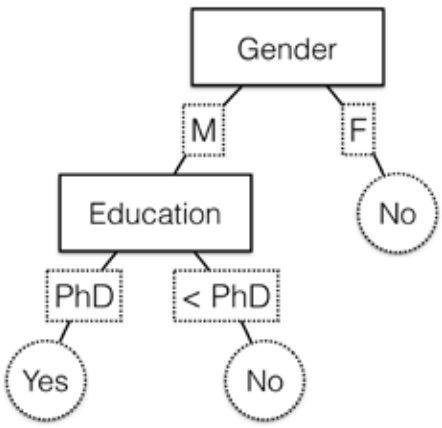

Figure 2

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.

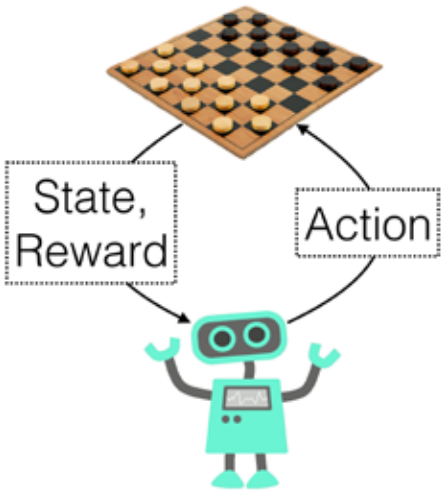

Figure 3

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]

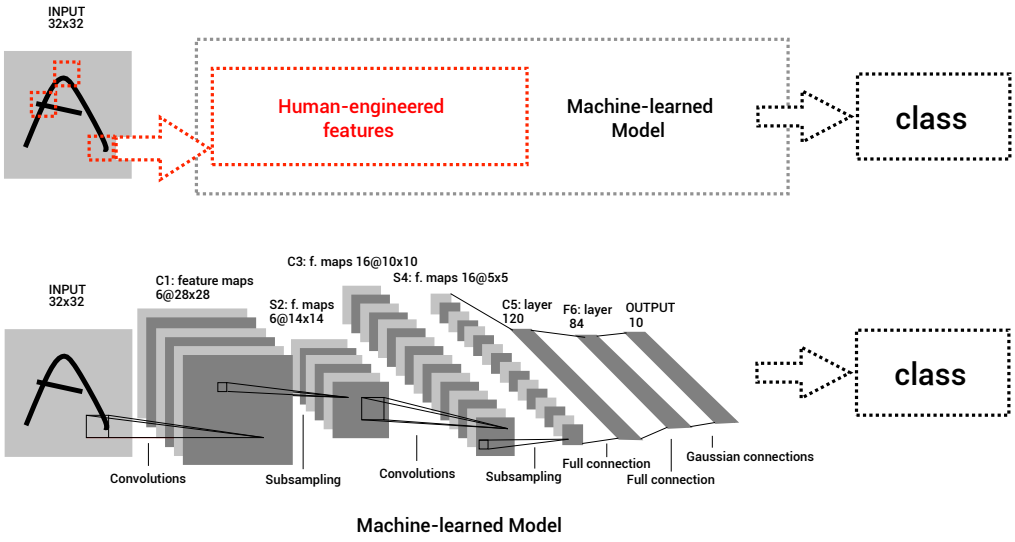

Figure 4

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.

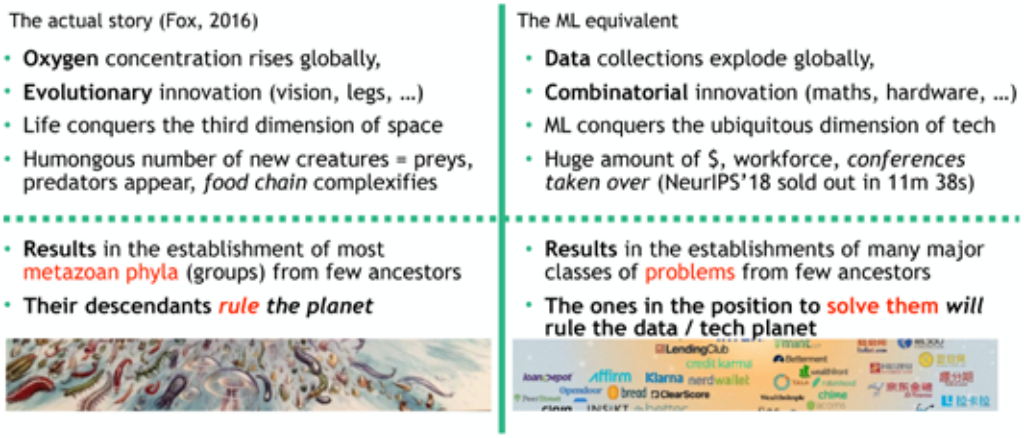

Figure 5

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*.

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*.