1

AI for Lawyers

A Gentle Introduction

John A. McDermid, Yan Jia and Ibrahim Habli

I INTRODUCTION

This chapter introduces the basic concepts of artificial intelligence (AI) to assist lawyers in understanding in what way, if any, the private law framework needs to be updated to enable systems employing AI to be treated in an 'appropriate' manner. What 'appropriate' means is a matter for legal experts and ethicists, insofar as the law reflects ethical principles, so the chapter seeks to identify the technological challenges which *might* require a legal response, not to prejudge what such a response might be.

AI is a complex topic, and it is also moving very fast, with new methods and applications being developed all the time[1]. Consequently, this chapter focuses on principles that are likely to be stable over time, and this should help lawyers to appreciate the capabilities and limitations of AI. Further, it illustrates the insights with 'concrete' examples from current applications of AI. In particular, it discusses the state of the art in application of AI and machine learning (ML) and identifies a range of challenges relating to use of the technology where it can have an impact on human health and wellbeing.

The rest of the chapter is structured as follows. Section II introduces the key concepts of AI including ML and identifies some of the main types and uses of ML. Section III sets out a view of the current 'state of the art' in AI applications, the strengths and weaknesses of the technology and the challenges that this brings. This is supported by concrete examples. Section IV presents conclusions including arguing that a multidisciplinary approach is needed to evolve the legal framework relating to AI and ML.

[1] Z Somogyi, *The Application of Artificial Intelligence: Step-by-Step Guide from Beginner to Expert* (Springer 2021).

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

II ARTIFICIAL INTELLIGENCE: KEY CONCEPTS

The concept of AI is generally said to originate with Alan Turing[2] who proposed an 'imitation game' where a human held a conversation through a textual interface either with another human or a computer (machine).[3] If a human cannot distinguish the machine from another human, then the machine is said to have 'passed the test' – we now refer to this as the 'Turing Test',[4] although Turing didn't use that term himself. Technology has advanced to an enormous degree in the seventy years since Turing's original paper but the concept of a machine imitating a human remains valid and indicative of the aims of AI.[5]

A *Artificial Intelligence and Machine Learning*

First, we give a more direct definition of what is meant by AI and then introduce the concept of ML:

- AI involves developing computer systems to perform tasks normally regarded as requiring human intelligence, for example, deciding prison sentences,[6] or medical diagnosis.[7]

At the moment, there is no consensus on a standard definition of AI.[8] The European Commission's Communication on AI proposed the following definition of AI:

> Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g., voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g., advanced robots, autonomous cars, drones or Internet of Things applications).[9]

Some other definitions of AI tend to describe the technology in terms of its most widely used techniques, for example, ML, logic, and statistical approaches.[10]

[2] SB Cooper and J van Leeuwen (eds), *Alan Turing: His Work and Impact* (Elsevier 2013).

[3] A Turing, 'Computing Machinery and Intelligence' (1950) 59(236) *Mind* 433.

[4] J Moor (ed), *The Turing Test: The Elusive Standard of Artificial Intelligence*, vol 30 (Springer 2003).

[5] A Darwiche, 'Human-Level Intelligence or Animal-Like Abilities?' (2018) 61(1) Communications of the ACM 56.

[6] J Ryberg and JV Roberts (eds), *Sentencing and Artificial Intelligence* (Oxford University Press 2022).

[7] EJ Topol, 'High-Performance Medicine: The Convergence of Human and Artificial Intelligence' (2019) 25(1) *Nature Medicine* 44.

[8] R Calo, 'Artificial Intelligence Policy: A Primer and Roadmap' (2017) 51 *UCDL Rev* 399.

[9] <https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf>.

[10] T Madiega, 'Briefing, EU Legislation in Progress, Artificial Intelligence Act, PE 698.792' (*European Parliamentary Research Service*, January 2022) <www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf>.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

Early AI systems, often called expert systems,[11] were generally based on well-defined rules, and these rules were normally defined by humans reflecting knowledge of the domain in which the system was to be used, for example, clinical decision-support tools that utilise a knowledge repository and a predefined ruleset for the prescribing of medications for common conditions.[12] ML is a form of AI, developing computer systems that *learn* to perform a task from training data, guided by performance measures, for example, accuracy.[13] ML is intended to generalise beyond the training data so the resultant systems can work effectively in situations on which they were not trained, for example, learning to identify the presence or absence of diabetic retinopathy from thousands of historic retinal scans labelled with outcomes.[14] We will use the term AI to include ML, but not *vice versa*.

It is common to distinguish 'narrow AI' from 'general AI', often referred to as artificial general intelligence (AGI).[15] The key difference is that 'narrow AI' is focused on a specific task, for example, recognising road signs, whereas AGI is not – indeed we would expect AGI to have the breadth of capabilities of humans including the ability to hold conversations, drive a car, interpret legal judgments, and so on. Modern AI systems can be classed as 'narrow' and some view AGI as unattainable[16] (see also the discussion of the 'state of the art' later).

ML is used in most modern AI systems as a cost-effective way of solving problems that would be prohibitively expensive or impossible to develop using conventional programming – and the key to this is the ability of ML to generalise beyond training data.[17] For example, an ML-based system used for medical diagnosis should work for any patient in the system's intended scope of application. This is similar to the way that humans apply their knowledge – doctors can treat patients they have not seen before, we can drive on new roads, including those that weren't built when we learnt to drive. This can be seen as generalising Turing's imitation game to a wider range of capabilities than textual communication.

[11] J Liebowitz (ed), *The Handbook of Applied Expert Systems* (CRC Press 2019).

[12] J Fox, N Johns and A Rahmanzadeh, 'Disseminating Medical Knowledge: The Proforma Approach' (1998) 14(1–2) *Artificial Intelligence in Medicine* 157.

[13] Most definitions of ML centre on learning from experiences that are captured via a training dataset. For example, Mitchell defines ML as 'the scientific study of computer algorithms that improve automatically through experience'. T Mitchell, *Machine Learning* (McGraw Hill 1997).

[14] Y Liu, PHC Chen, J Krause and L Peng, 'How to Read Articles that Use Machine Learning: Users' Guides to the Medical Literature' (2019) 322(18) *Jama* 1806.

[15] G Marcus and E Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust* (Vintage 2019).

[16] Despite the impressive performance of many ML-based systems that have exceeded human ability, say for object recognition in images, no 'new theory of the mind' has emerged that could be seen as paving the way for AGI. A Darwiche, 'Human-Level Intelligence or Animal-Like Abilities?' (2018) 61(10) *Communications of the ACM* 56.

[17] I Goodfellow, Y Bengio and A Courville, *Deep Learning* (MIT Press 2016).

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

B *Types of Machine Learning*

There are many different ML methods, but they can generally be divided into three classes.[18] We provide some contextual information then give descriptions of these three classes before giving some examples of different ML methods.[19]

Data plays a key role in ML and learning algorithms are used to discover knowledge or patterns from data without explicit (human) programming.[20] The result of learning is referred to as the ML model. The dataset used for training may be labelled, saying what each datum means, for example, a cat or a dog in an image, or it may be unlabelled.[21] The data is normally complex, and we will refer to the elements of each datum as features. The dataset is often split into a training set and a test set, with the test set used to assess the performance, for example, accuracy, of the learnt ML model.[22]

1 Supervised Learning

Supervised learning uses a labelled dataset and this *a priori* knowledge is used to guide the learning process. Supervised learning tries to find the relationships between the feature set and the label set. The knowledge extracted from supervised learning is often utilised for classification or for regression problems. Where the labels are categorical, the learning problem is referred to as *classification*.[23] On the other hand, if the labels correspond to numerical values, the learning problem is defined as *regression* problem.[24]

Figure 1.1 gives a simple illustration of the use of ML for object identification and classification. The ML models have classified dynamic objects in the image and placed bounding boxes around them; in general, such algorithms will distinguish different classes of vehicle, for example, vehicles from people, as this helps in predicting their movement. Here, regression may be used for predicting the future position or trajectory of a vehicle based on its past positions.[25]

[18] S Shalev-Shwartz and S Ben-David, *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press 2014).

[19] The descriptions and illustrations in the rest of this section are mainly drawn from: Y Jia, 'Embracing Machine Learning in Safety Assurance in Healthcare' (PhD thesis, University of York 2021).

[20] JC Mitchell and K Apt, *Concepts in Programming Languages* (Cambridge University Press 2003).

[21] R Raina and others, 'Self-Taught Learning: Transfer Learning from Unlabeled Data' (*Proceedings of the 24th International Conference on Machine learning*, June 2007) 759–766.

[22] R Ashmore, R Calinescu and C Paterson, 'Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges' (2021) 54(5) *ACM Computing Surveys (CSUR)* 1.

[23] G Haixiang and others, 'Learning from Class-Imbalanced Data: Review of Methods and Applications' (2017) 73 *Expert Systems with Applications* 220.

[24] A Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (O'Reilly Media 2019).

[25] A Benterki, M Boukhnifer, V Judalet and M Choubeila, 'Prediction of Surrounding Vehicles Lane Change Intention Using Machine Learning' (*10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, September 2019) 839–843.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

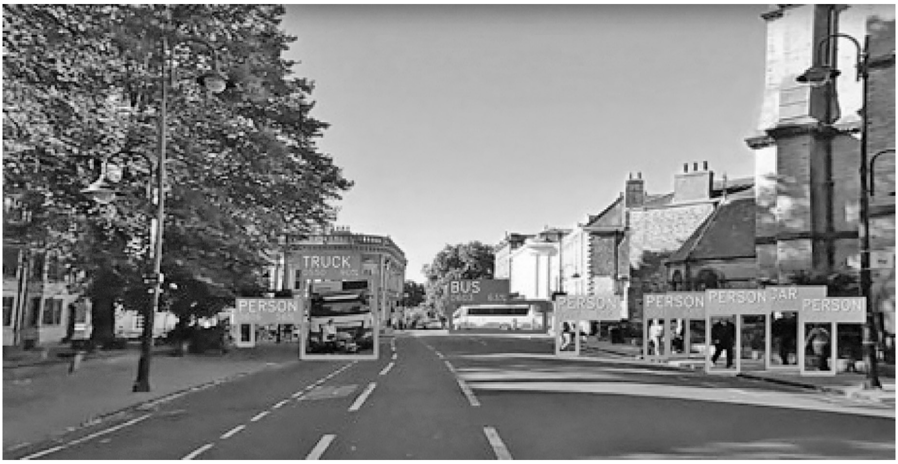


FIGURE 1.1 Object classification (courtesy of AAIP)
Note: Assuring Autonomy International Programme at the University of York, funded by the Lloyd's Register Foundation.

2 Unsupervised Learning

Unsupervised learning uses unlabelled data and can draw inferences from the dataset to identify hidden patterns.[26] Unsupervised learning is often used for clustering (grouping together related data) and finding associations among features. An active area of work is so-called 'self-supervised learning' (the self here is a computer, not a person) which learns good generic features from an enormous unlabelled dataset.[27] These features can then be used to solve a specific task with a smaller labelled dataset, that is, feeding into supervised learning.

The 'recommender' systems for online shopping systems produce outputs such as: 'people who bought this item also bought…'.[28] Practical recommender systems use a mixture of ML methods, and this may include unsupervised learning.[29] Thus, it is likely that most readers of this chapter will have used a system that employs unsupervised learning, without being aware of it.

[26] M Alloghani and others, 'A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science' in M. Berry, A. Mohamed, B. Yap (eds), *Supervised and Unsupervised Learning for Data Science* (Springer 2020) 3.

[27] D Hendrycks, M Mazeika, S Kadavath and D Song, 'Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty' (2019) *Advances in Neural Information Processing Systems* 32.

[28] I Portugal, P Alencar and D Cowan, 'The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review' (2018) 97 *Expert Systems with Applications* 205; M Beladev, L Rokach and B Shapira, 'Recommender Systems for Product Bundling' (2016) 111 *Knowledge-Based Systems* 193.

[29] See: Luciano Strika, 'K-Means Clustering: Unsupervised Learning for Recommender Systems' (*Towards Data Science*, 3 April 2019) <www.towardsdatascience.com/k-means-clustering-unsupervised-learning-for-recommender-systems-397d3790f90f> 18 August 2022.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

3 Reinforcement Learning

Reinforcement learning (RL) is a learning method that interacts with its environment by producing actions and discovering errors or receiving rewards.[30] Trial-and-error search and delayed reward are the most relevant characteristics of RL. In this class of learning, there are three primary components: the agent (the learner or decision-maker), the environment (everything the agent interacts with) and actions (what the agent can do).

The environment gives the agent a state (e.g., moving or stationary), the agent takes an action, then the environment gives back a reward as well as the next state. By analogy, this is like a children's game where one child is blindfolded (the agent), this child can move forwards, backwards, left and right (the actions) in a room (the environment) to find an object and is given hints (rewards), for example, warm, hot, cold, freezing, depending on how close they are to the object, by other children.

This loop continues until the environment gives back a terminal state and a final reward (perhaps some chocolate in the children's game), which ends the episode. The objective is for the agent to automatically determine the ideal behaviour in this environment to maximise its performance. Normally RL development is carried out in a simulated environment or on historical data before the agent is used in real-world applications, for example, optimising the treatment of sepsis.[31]

RL can be used in planning and prediction problems, for example, identifying safe paths for a robot to move around a factory, and recommending medication for a patient.[32] In constrained environments with concrete rules, for example, board games, RL has demonstrated outstanding performance. This is best illustrated by DeepMind's AlphaGo computer program that utilised RL, amongst other ML models, and defeated the world champion in the game of Go, which is much more complex than chess.[33]

C *Developing ML Models*

Following the identification and analysis of a problem in a specific context, ML models can be developed through three primary phases: data management, model learning and model testing.[34] Data management involves collecting or creating, for example, by simulation, data on which to train the models. The data needs to be representative of the situation of interest, for example, roads for autonomous vehicles (AVs),[35] patient treatments and outcomes in healthcare, and perhaps successful and unsuccessful cases

[30] RS Sutton and AG Barto, *Reinforcement Learning: An Introduction* (MIT Press 2018).

[31] M Komorowski and others, 'The Artificial Intelligence Clinician Learns Optimal Treatment Strategies for Sepsis in Intensive Care' (2018) 24(11) *Nature Medicine* 1716.

[32] I Kavakiotis and others, 'Machine Learning and Data Mining Methods in Diabetes Research' (2017) 15 *Computational and Structural Biotechnology Journal* 104.

[33] DeepMind, 'AlphaGo' <www.deepmind.com/research/highlighted-research/alphago>.

[34] Ashmore, Calinescu and Paterson (n 22).

[35] We use the AVs term to embrace all driver assistance system, for example, adaptive cruise control, that reduce the need for the driver to engage in the dynamic driving task whether or not they are 'fully autonomous'.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

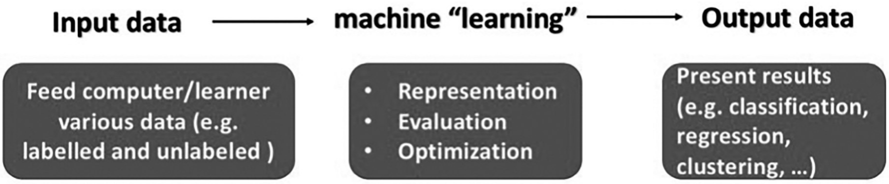

Input data
machine "learning"
Output data
Feed computer/learner various data (e.g. labelled and unlabeled)
• Representation
• Evaluation
• Optimization
Present results (e.g. classification, regression, clustering, ...)

FIGURE 1.2 A simple illustration of machine learning process
Jia, 'Embracing Machine Learning in Safety Assurance in Healthcare' (n 19).

for a legal assistant. As well as splitting a dataset into a training dataset and a test data set, a validation dataset can also be used for model parameter tuning during model learning. For model learning itself, it is necessary to consider how to represent the knowledge derived from the training data, that is, what type of ML method to use, how to evaluate the ML model performance and then how to optimise the learning process. This is illustrated in Figure 1.2.

In model testing, the performance of the ML models is assessed using various metrics on the test dataset. It is easiest to explain this concept by considering classification of objects for an AV. The ML model output is therefore the assessed class for the observed object. For simplicity, assume we are only interested in identifying dynamic objects, that is, those that can move, and distinguishing them from static objects. In this case, we can have:

- True positive – correct classification, for example, a person is identified as a dynamic object.
- True negative – correct classification, for example, a lamppost is not identified as dynamic.
- False positive – incorrect classification, for example, a statue[36] is classified as dynamic.
- False negative – incorrect classification, for example, a person is not classified as dynamic.

It is common to convert these cases into rates and measures,[37] for example, a true positive rate (TPR), which is the proportion of positives correctly identified, that is, true positives, out of all the positives. Similarly, other measures are defined, for example, accuracy, which is the proportion of correct outputs (true positives plus true negatives) out of all the ML model outputs.

Some ML methods, for example neural networks (NNs), produce a score or probability qualifying the output,[38] for example, dynamic object with 0.6 probability. If the threshold in this case was 0.5, then the output would be interpreted as saying that the object was

[36] Although, of course, statues might move when being installed or if being toppled in a revolution or other form of protest – but this simply serves to show the difficulty of the problems being addressed by ML.

[37] T Fawcett, 'An Introduction to ROC Analysis' (2006) 27(8) *Pattern Recognition Letters* 861.

[38] MA Nielsen, *Neural Networks and Deep Learning*, vol 25 (Determination Press 2015).

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

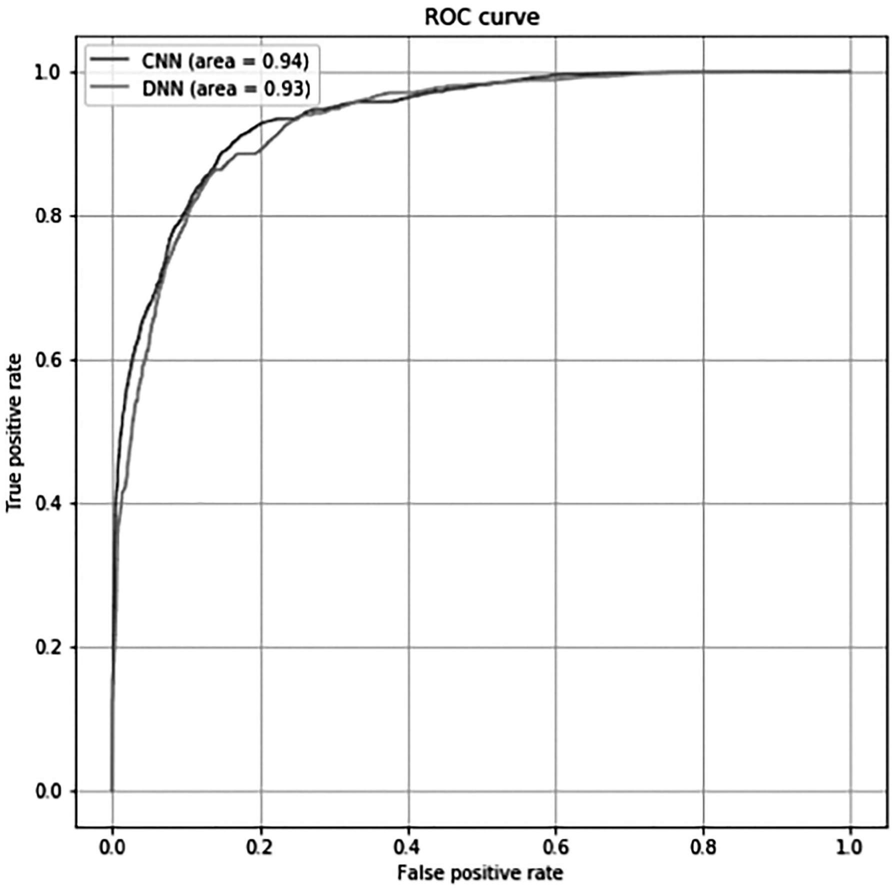


FIGURE 1.3 Illustration of ROC
JA McDermid, Y Jia, Z Porter and I Habli, 'Artificial Intelligence Explainability: The Technical and Ethical Dimensions' (2021) 379(2207) *Phil. Trans. R. Soc. A.*

dynamic. However, if the threshold was set at 1, then the use of the NN would never give a positive output (saying the object was dynamic), thus the TPR would be 0, and so would the false positive rate (FPR). Similarly, a threshold of 0 would mean that everything was treated as positive, so both TPR and FPR would be 1. Intermediate thresholds, for example 0.5, would give a different value for TPR and FPR. TPR and FPR are combined into a measure known as the receiver operating characteristic (ROC)[39] which plots TPR vs. FPR as the threshold varies with different values, see Figure 1.3 for an example. It is also common to use the area under the curve ROC (AUC-ROC) to report the model performance, and the closer the AUC is to 1, the better the performance is.[40]

[39] The origin of the concept was in the development of radars in the 1940s, hence the slightly unintuitive name.

[40] T Fawcett, 'An Introduction to ROC Analysis' (2006) 27(8) *Pattern Recognition Letters* 861.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

The AUC-ROC can be used to compare different ML models to choose the best one for a particular application. Figure 1.3 illustrates the use of a ROC curve for this purpose, comparing a convolutional neural network (CNN) with a fully connected deep neural network (DNN).[41] A random ML model (prediction) would produce a diagonal line on the ROC and the AUC-ROC would be 0.5. A perfect ML model would give an AUC-ROC of 1, and the 'curve' would follow the axes in the diagram. The example in Figure 1.3 shows that the two ML models have similar performance as measured by the AUC-ROC.

The intent of the evaluation criteria for ML models is to illuminate how well the model performs, contrasting desired behaviour with erroneous or undesirable behaviour.

In practice, development of ML models is highly iterative[42] and model developers frequently build and test new models, evaluating them to see if the performance has improved. Once ML models are put into operation they may still be updated, for example if new data is available.

D *Examples of ML Methods*

There are many ML methods as we mentioned earlier. The aim here is to illustrate the variety and their capabilities to inform the discussion on the use of ML methods later, and on strengths, limitations, the state of the art and challenges in Section III.

Some of the more widely used ML methods are:

- NNs – a network of artificial (computer models of) neurons, inspired by the human brain.[43] NNs are good at analysing complex data, for example images, and can be used supervised, for example, with labelled images, or unsupervised.[44] There are many variants, for example, CNN and fully connected DNN as illustrated in Figure 1.3.
- Random forest (RF) – a collection of decision trees which is normally more robust (less susceptible to error in a single input) than a single decision tree.[45] Usually, RF is developed using supervised learning, and they are well-suited to decision problems, for example, for clinical diagnosis.[46]

[41] NNs have an input layer (of neurons), and an output layer with hidden layers in between. Here, the fully connected DNN means all of the hidden layers are fully connected. CNN means that at least one of the hidden layers uses convolution instead of being fully connected.

[42] R Hawkins and others, 'Guidance on the Assurance of Machine Leaning in Autonomous Systems (AMLAS)' (2021) <arXiv:2102.01564>.

[43] MA Nielsen, *Neural Networks and Deep Learning*, vol 25 (Determination Press 2015).

[44] M Alloghani and others, "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science' in M. Berry, A. Mohamed, B. Yap (eds), *Supervised and Unsupervised Learning for Data Science* (Springer 2020) 3.

[45] M Belgiu and L Drăguţ, 'Random Forest in Remote Sensing: A Review of Applications and Future Directions' (2016) 114 *ISPRS Journal of Photogrammetry and Remote Sensing* 24.

[46] KR Gray and others, 'Random Forest-Based Similarity Measures for Multi-Modal Classification of Alzheimer's Disease' (2013) 65 *NeuroImage* 167.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

- Probabilistic graphical models (PGMs) – a graph of variables (features) of interest in the problem domain and probabilistic relationships between them. There are several types of PGM including Bayesian networks (BNs) and Markov networks.[47] They can be used both supervised and unsupervised.

Generally, the learnt models, most notably DNNs, are very complex and 'opaque' to humans, that is, not open to scrutiny. PGMs are more amenable to human inspection, and it is possible to integrate human domain knowledge into PGMs. The primary difference between DNNs and PGMs lies in the structure of the machine learnt model in that PGMs tend to reflect human reasoning more explicitly, including causation.[48] This aids the process of interrogating the model for understanding the basis of its output. This level of transparency is harder to achieve with DNNs and therefore the majority of the techniques that are used to explain the output of DNN models rely on indirect means,[49] for example, examples and counterfactual explanations.[50]

E *Uses of ML Models*

There are many uses of ML models. Some are embedded in engineered systems, for example, AVs, whereas others are IT systems, that is, operating on a computer, phone, or similar device.

AVs are an example of embedded ML. AVs often use ML for camera image analysis and understanding, for example classifying 'objects' into dynamic vs static, and identifying subclasses of dynamic objects – cars, bicycles, pedestrians, and so on. Typically, the systems employ a form of NN, for example, CNNs.[51] Many employ conventional computational methods of path planning (local navigation) but some use RL to determine safe and optimal paths.[52]

ML is increasingly being proposed for use in healthcare for both diagnosis and treatment;[53] most of such applications are IT systems. Some of the systems also involve image analysis, for example, identifying tumours in images, with performance exceeding that of clinicians in some cases.[54] There are online systems, for

[47] J Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann 1988); D Lowd and A Rooshenas, 'Learning Markov Networks with Arithmetic Circuits' (2013) 31 *Artificial Intelligence and Statistics* 406.

[48] J Pearl and D Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books 2018).

[49] J McDermid, Y Jia, Z Porter and I Habli, 'Artificial Intelligence Explainability: The Technical and Ethical Dimensions' (2021) 379(2207) *PhilTrans*.

[50] S Wachter, B Mittelstadt and C Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31 *Harv JL & Tech* 841.

[51] S Ren, K He, R Girshick and J Sun, 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks' (2015) 28 *Advances in Neural Information Processing Systems* 91.

[52] AE Sallab, M Abdou, E Perot and S Yogamani, 'Deep Reinforcement Learning Framework for Autonomous Driving' (2017) 19 *Electronic Imaging* 70.

[53] E Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (Hachette 2019).

[54] EJ Hwang and others, 'Development and Validation of a Deep Learning Based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs' (2019) 69(5) *Clinical Infectious Diseases* 739–747.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

example, from Babylon Health,[55] which employs an ML-based symptom checker. Applications which recommend treatments are also being explored, for example, delivery of vasopressors as part of sepsis treatment.[56]

Legal uses of AI are also IT systems. The applications include predicting the outcome of tax appeals[57] and helping with the production of legal letters which correctly phrase non-expert text in support of claims, and other legal actions.[58] Some of the benefit of such tools arises from currently available computational power to trawl large volumes of documents, and there are now commercially available tools that use ML (including supervised and unsupervised learning) to find appropriate legal documentation to support a case.[59]

III STATE OF THE ART AND CHALLENGES

AI, particularly ML, has enormous potential. As noted above, this arises out of its ability to generalise from the data used for training to new situations; this is perhaps the strongest justification for the use of the term 'intelligence'. However, some would argue that the potential hasn't been fully realised.[60] The aim in this section is to try to characterise the state of the art in the use of ML, noting that it differs across application domains, and to identify some of the challenges in achieving more widespread use of the technology. The focus here is on technical and ethical issues, rather than on legal challenges.

A *State of the Art*

ML is already pervasive in a range of online applications (IT systems). As indicated above, online platforms, which many use daily, such as Google search and online shopping, make massive use of ML.[61] Arguably, Google's search engine is one of the most impressive applications of ML providing extensive results to arbitrary textual queries in a very short space of time. This is all the more impressive as the learning is necessarily unsupervised. As well as good algorithms, this is made possible by access to massive computational power in data centres (sometimes referred to as 'cloud computing').[62]

[55] www.emed.com/uk 18 August 2022.

[56] Y Jia and others, 'Safety-Driven Design of Machine Learning for Sepsis Treatment' (2021) 117 *Journal of Biomedical Informatics* 103762.

[57] <www2.deloitte.com/nl/nl/pages/tax/articles/tax-i-outcome-predictions-dutch-tax-cases.html>.

[58] <www.donotpay.com>.

[59] <www.luminance.com>.

[60] Demis Hassabis, 'Royal Society Lecture on the History, Capabilities and Frontiers of AI' <www.royalsociety.org/science-events-and-lectures/2018/04/you-and-ai-history/>.

[61] See: <www.blog.hubspot.com/marketing/rankbrain-guide>.

[62] R Buyya, J Broberg and AM Goscinski (eds), *Cloud Computing: Principles and Paradigms* (Wiley & Sons 2010).

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

Such capabilities are becoming 'commoditised' and companies, for example, Amazon Web Services,[63] now provide access to data centres as a commercial offering. Further, the software to build ML applications is now widely available. For example, TensorFlow,[64] originally developed by Google is readily available; it can be used to build applications with a wide range of ML models including NNs, although it still requires extensive programming skills; there is also support for developing popular classes of system such as recommenders.

Further, there is a growing availability of skills to develop such systems with most computer science departments in universities teaching ML at undergraduate and postgraduate level. Thus, the ingredients are there for widespread development of AI and ML applications.

Most application domains where ML is being applied can be viewed as emergent or nascent. Whilst there are examples of systems, for example, in healthcare and legal practice, their adoption is not widespread. We will illuminate some of the reasons for this when we consider challenges.

There has been work on ML in embedded systems for some time, for example, in robotics, but the 'autonomous vehicle challenge' set up by the US Defense Advanced Research Projects Agency ('DARPA') about fifteen years ago can perhaps be seen as prompting a step-change in research in this area.[65] Although there is work using ML systems across transportation and in other sectors, for example, factory automation,[66] mining[67] and robotic surgery,[68] perhaps the greatest investment and development has been seen in AVs. Waymo (a spin off from Google) is now offering a 'ride hailing' service known as Waymo One;[69] whilst this service is only available in limited areas, for example, in Phoenix Arizona,[70] the service does operate without a human driver and the vehicles have now operated for about 20 million miles on the roads.[71] Waymo has also now forged partnerships with several automotive Original Equipment Manufacturers ('OEMs'), for example, Jaguar Land Rover.[72] However, whilst extremely impressive, the systems are not 'perfect',

[63] <www.aws.amazon.com/>.

[64] <www.tensorflow.org>.

[65] M Buehler, K Iagnemma and S Singh (eds), *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic*, vol 56 (Springer 2009).

[66] DH Kim and others, 'Smart Machining Process Using Machine Learning: A Review and Perspective on Machining Industry' (2018) 5(4) *International Journal of Precision Engineering and Manufacturing-Green Technology* 555.

[67] Z Hyder, K Siau and F Nah, 'Artificial Intelligence, Machine Learning, and Autonomous Technologies in Mining Industry' (2019) 30(2) *Journal of Database Management (JDM)* 67.

[68] M Bhandari, T Zeffiro and M Reddiboina, 'Artificial Intelligence and Robotic Surgery: Current Perspective and Future Directions' (2020) 30(1) *Current Opinion in Urology* 48.

[69] <www.waymo.com/waymo-one/>.

[70] At the time of writing, services were being extended to San Francisco to 'trusted testers', see for example: <www.arstechnica.com/gadgets/2021/08/waymo-expands-to-san-francisco-with-public-self-driving-test/>.

[71] <www.reuters.com/article/us-autonomous-waymo-idUSKBN1Z61RX>.

[72] <www.theverge.com/2018/3/27/17165992/waymo-jaguar-i-pace-self-driving-ny-auto-show-2018>.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

and there have been several examples of vehicles getting confused or 'stuck', for example, by traffic cones.[73]

Note that these systems are computationally expensive (particularly for image analysis)[74] and are only practicable because of the availability of super-computer levels of performance at affordable prices.[75] Further, computational power is doubling roughly every eighteen months[76] which should facilitate the broader adoption of ML.

B *Challenges*

There are many challenges in developing ML-based systems, so that they can be used with confidence that their behaviour will be sound, safe, legal, and so on, where their use can give rise to harm. The aim here is to identify some of the key technical challenges and to outline some of the possible approaches to addressing these challenges.

First, and most fundamentally, there is a transfer of decision-making or responsibility for recommending a course of action from a human to a computer and its ML components. From a legal perspective, this raises issues about agency and liability which are discussed elsewhere in this volume.

Second, humans have a semantic model, for example, know what a bicycle is and its likely behaviour; computers, even those incorporating ML, do not have these models.[77] Similarly, humans have contextual models, for example, know what a roundabout is and the effects on driver behaviour, and the ML does not.[78] These semantic and contextual models allow humans to generalise beyond their experience to reliably deal with new situations. However, for systems using ML the lack of such models can contribute to 'gaps' between what is required and what is achieved, which may be significant in engineering, ethical and legal terms.[79] The solution to this is to encode enough additional information in the systems to cope with the limitations in the ML components to enable effective operation – note that this is potentially feasible as we are considering 'narrow AI' not AGI[80] – but, as the example of the Waymo getting stuck encountering traffic cones shows, doing this remains a major challenge.

[73] <www.vice.com/en/article/y3dv55/waymo-self-driving-car-gets-stuck-by-cones-drives-away-from-assistance>.

[74] F Dufaux, 'Grand Challenges in Image Processing' (2021) 1 *Frontiers in Signal Processing* 3.

[75] <www.qblocks.medium.com/how-much-did-it-cost-to-build-the-fastest-supercomputer-in-the-world-8e9e30a56f60>.

[76] This claim is often made in reference to Gordon Moore's prediction that the number of components in an integrated circuit would double every two year (referred to as 'Moore's law'), <www.britannica.com/technology/Moores-law>.

[77] A Darwiche, 'Human-Level Intelligence or Animal-Like Abilities?' (2018) 61(10) *Communications of the ACM* 56.

[78] C Paterson and others, 'DeepCert: Verification of Contextually Relevant Robustness for Neural Network Image Classifiers' (*International Conference on Computer Safety, Reliability, and Security*, September 2021) 3–17.

[79] S Burton and others, 'Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective' (2020) 279 *Artif Intell* 103201.

[80] G Marcus and E Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust* (Vintage 2019).

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

Third, the way the ML systems work, generalising from training data, identifies correlations not causation.[81] A recent study[82] used ML to assess the relationship between body shape and income, and identified correlations which differ across genders, for example that obesity in females correlates with lower income. It would be a mistake, however, to infer that body shape *causes* low income – it may be that those on low income cannot afford a good diet and that might lead to obesity. Further, there may be other causally relevant factors that have not been considered in the ML model. This does not mean that the ML model is wrong; just that care needs to be taken when acting on the outputs of the ML model.

Fourth, the learnt ML models are 'opaque', that is not amenable to human scrutiny.[83] This means that it is hard to understand why the ML models produce their outputs. This can, in turn, give rise to doubts – why was that recommendation made, and was it biased? This has legal implications, for example, in terms of complying with the General Data Protection Regulations,[84] as well as ethical ones in terms of fairness. A partial solution is via so-called explainable AI methods, where simpler approaches are used to make the workings of the ML model human interpretable.[85] One of the most commonly used explainable AI methods is feature importance which illustrates the relative weight of each input feature for the ML model as a whole (global importance) or for a particular output (local importance).[86] This is illustrated in Figure 1.4, for a system concerned with weaning intensive care patients from mechanical ventilation. Here, the longer bars show greater influence of that input feature on the ML model output, with those bars close to zero length being of least importance.

This figure is for the two ML models shown in Figure 1.3. The two ML models have similar performance as shown in Figure 1.3, but the feature importance is quite different. Clinicians can judge the relevance and validity of these weightings to see which, if either, of the ML models is preferable. It is also notable that gender, ethnicity, and age are close to zero (low importance) for the CNN but age and gender in the fully connected DNN are relatively important, so this model might be thought to show bias. Care needs to be taken here. Age and gender might be clinically relevant, so a judgement about whether a system is biased or not needs to be

[81] JG Richens, CM Lee and S Johri, 'Improving the Accuracy of Medical Diagnosis with Causal Machine Learning' (2020) 11(1) *Nature Communications* 1.

[82] S Song and S Baek, 'Body Shape Matters: Evidence from Machine Learning on Body Shape-Income Relationship' (2021) 16(7) *PLoS One* e0254785 <https://doi.org/10.1371/journal.pone.0254785>.

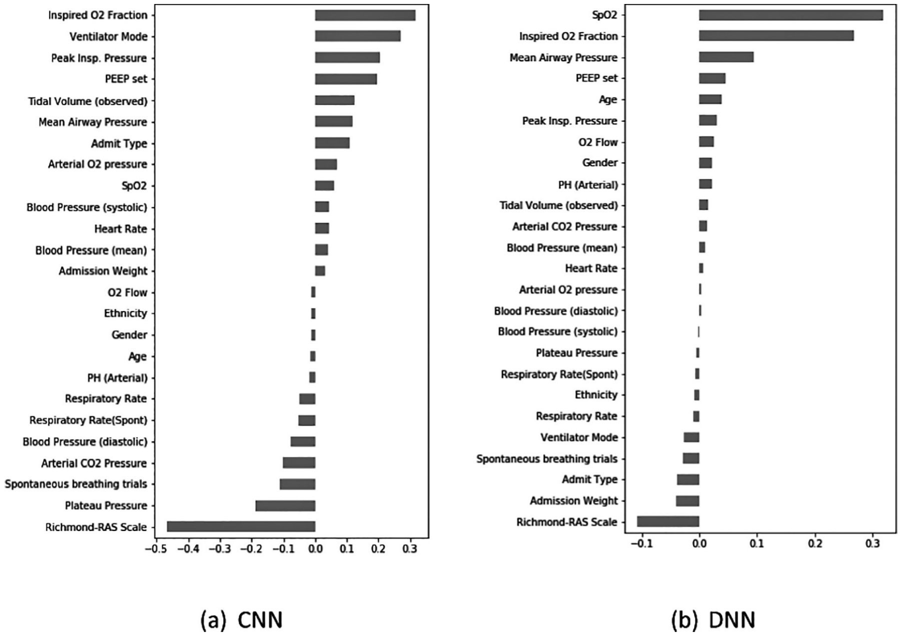
[83] C Molnar, 'Interpretable Machine Learning' (Lulu.com, 2021).

[84] C Kuner and others, 'Machine Learning with Personal Data: Is Data Protection Law Smart Enough to Meet the Challenge?' (2017) 7(1) *International Data Privacy Law* 1, 1–2.

[85] D Doran, S Schulz and TR Besold, 'What Does Explainable AI Really Mean? A New Conceptualization of Perspectives' (2017) <arXiv preprint arXiv:1710.00794>.

[86] LH Gilpin and others, 'Explaining Explanations: An Overview of Interpretability of Machine Learning' (*IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, October 2018) 80–89.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press



(a) CNN (b) DNN

FIGURE 1.4 Global feature importance for CNN and fully connected DNN McDermid, Jia, Porter and Habli, 'Artificial Intelligence Explainability: The Technical and Ethical Dimensions' (n 49).

considered carefully; in this case, ethical considerations need to be treated alongside clinical ones.

Fifth, there is an issue of trust and human control over the system employing ML. As noted above, some ML systems produce outputs with a probability; in all cases, there is uncertainty in the accuracy of the results.[87] Users should be (made) aware of this intrinsic uncertainty. However, even if they are aware, there can be automation bias where users tend to trust the system's outputs without questioning them.[88] Further, the user might have no practical way of cross-checking the output of the ML system – they might not have access to the 'raw' data and there may simply be insufficient time to assess the data and to intervene. Such issues might, in part, be addressed using techniques such as explainable AI methods but there remain legal and ethical issues, for example, the ethical conditions for carrying responsibility might not be met for those who carry legal responsibility for the effects of using the system[89].

[87] MA Nielsen, *Neural Networks and Deep Learning*, vol 25 (Determination Press 2015).

[88] R Parasuraman and V Riley, 'Humans and Automation: Use, Misuse, Disuse, Abuse' (1997) 39(2) *Human Factors* 230.

[89] Burton and others (n 79).

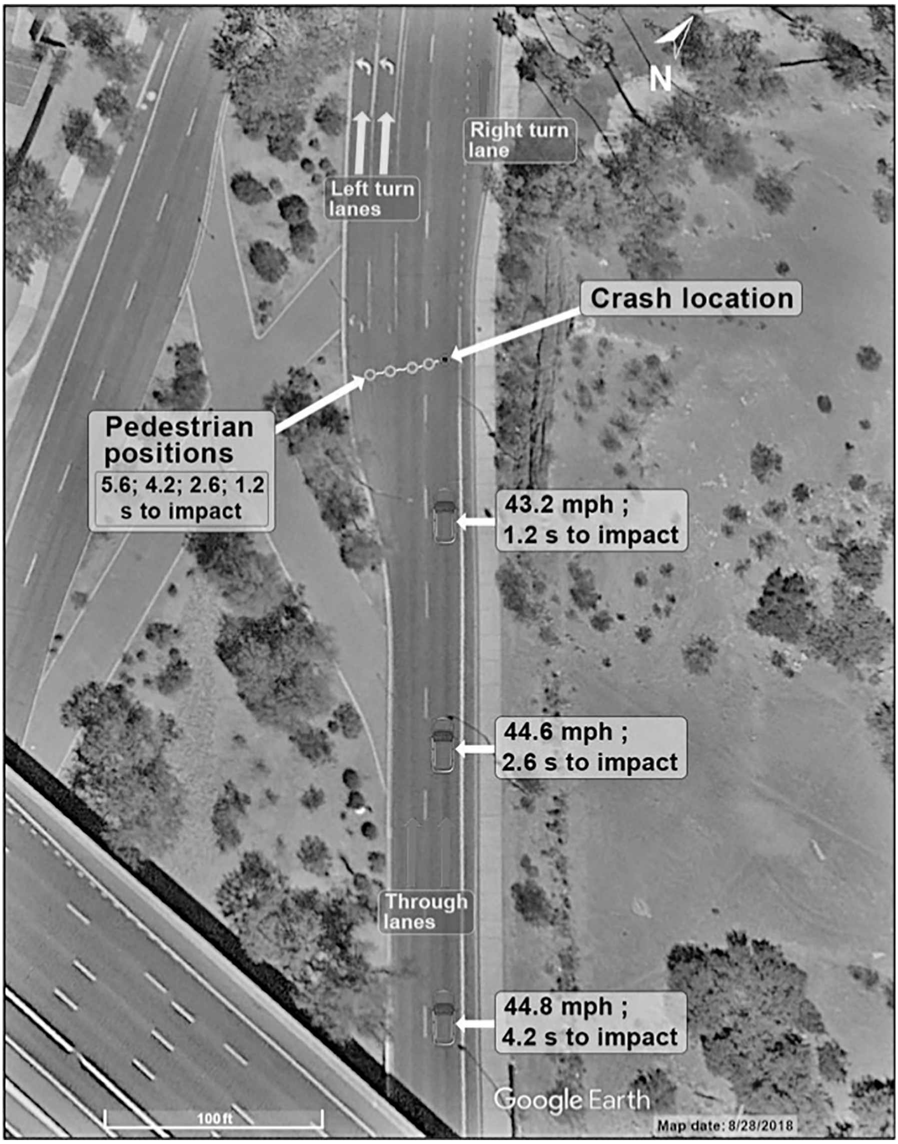


FIGURE 1.5 Partial timeline in Uber Tempe accident

Sixth, many embedded systems operate in situations where they can pose a threat to human health or safety, for example, in unmanned aircraft for reconnaissance.[90] However, this is perhaps most apparent with AVs although it can arise in other cases, for example, robotic surgery. Figure 1.5 presents a partial timeline for the accident

[90] JA McDermid, Y Jia and I Habli, 'Towards a Framework for Safety Assurance of Autonomous Systems' (CEUR Workshop Proceedings, August 2019) 1–7.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

caused by an Uber ATG vehicle in March 2018 in Tempe Arizona that led to the death of Elaine Herzberg.[91] This example enables us to illustrate the importance of some of the concepts introduced earlier.

Figure 1.5 shows the positions of Elaine Herzberg and her bicycle (labelled as pedestrian) and the Uber ATG vehicle (shown in green) at four times prior to the impact. The Highway Accident Report published by the National Transportation Safety Board stated that the Automated Driving System 'never accurately classified her as a pedestrian or predicted her path'.[92] Critically, the predicted motion depended on the classification so when she was on one of the left turn lanes and classified as a car, she was predicted to leave the main road. Her movement history was discarded each time the vehicle reclassified her so at no time was her trajectory predicted as crossing the road. An impending collision was predicted 1.2S before the actual accident took place but the system did not act automatically (due to a concern over false positives leading to unnecessary emergency braking) with the expectation that the safety driver would respond. The safety driver (Rafaela Vasquez) didn't initiate timely braking – reportedly she was not paying attention, perhaps due to lack of training or due to automation bias (the vehicle had already successfully navigated the 'circuit' on which she was driving once). However, it may have been the case that she had insufficient time to react – see the previous discussion about legal and ethical responsibility. Uber was found to have no (legal) (criminal) case to answer for the accident, but the safety driver is facing a trial for negligent homicide.[93] There is no currently accepted solution to assuring the safety of autonomous systems.[94] There is relevant work on the assurance of the ML components of autonomous systems[95] but this remains an active area of research.

Finally, ML models can be set up to continue learning in operation – sometimes referred to as online learning.[96] This is, of course, analogous to the way humans learn. Most current ML-based systems learn off-line with the ML models being updated periodically by the developers (perhaps via over-the-air updates in the case of AVs).[97] As systems move towards online learning this introduces new challenges including how to assure continued safety, and it raises further questions about human control and agency.

[91] National Transportation Safety Board, 'Collision between Vehicle Controlled by Developmental Automated Driving System and Pedestrian' (2019) *NTSB Tech Rep* <www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf>.

[92] Ibid.

[93] <www.bbc.co.uk/news/technology-54175359>.

[94] McDermid, Jia and Habli 'Towards a Framework for Safety Assurance of Autonomous Systems' (n 90).

[95] R Hawkins and others, 'Guidance on the Assurance of Machine Leaning in Autonomous Systems (AMLAS)' (2021) <arXiv:2102.01564>.

[96] GI Parisi, 'Continual Lifelong Learning with Neural Networks: A Review' (2019) 113 *Neural Networks* 54.

[97] J Bauwens, 'Over-the-Air Software Updates in the Internet of Things: An Overview of Key Principles' (2020) 58(2) *IEEE Communications Magazine* 35.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press

IV CONCLUSIONS

AI, especially ML, is already a key component of many systems affecting society – not least online search and other online services. The capability of current ML systems and the trends in the power of computer systems means that these uses are likely to expand over time from current applications which are predominantly IT systems to include embedded systems, for example, in AVs, implantable medical devices and manufacturing. Further, the range of application domains is likely to expand. These capabilities bring with them challenges in technical, ethical and legal terms.

Technically, the biggest challenge is to develop and assure systems employing ML models so that they can be used with confidence that they are safe and have other desirable properties, including being free from bias. This links to the broader issues of trust and the ability for humans to exercise informed control or consent when this is appropriate. There are many legal questions, including those around the notion of agency and liability. This is a complex and intellectually challenging area, but also one requiring urgent attention since systems employing ML models are already being used and there is potential for considerable growth in applications.

This chapter has tried to give an accessible (gentle) introduction to the concepts of AI and ML for lawyers. Some technical details have been presented, for example, explaining the concept of feature importance for ML models, to give an idea of the depth and subtlety of the issues raised by the use of AI and ML models. It is hoped that this makes clear the need to take a multi-disciplinary approach to studying and evolving the legal framework relating to AI and ML and gives an adequate basis to help lawyers engage in constructive discussions with technical specialists.

https://doi.org/10.1017/9781108980197.002 Published online by Cambridge University Press