GOVERNMENT USE OF ARTIFICIAL INTELLIGENCE IN NEW ZEALAND

The Law Foundation NEWZEALAND
SUPPORTING INDEPENDENT LEGAL THINKING.

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

ISBN 978-0-473-47442-3 (pbk)
978-0-473-47443-0 (pdf)

The Law Foundation NEW ZEALAND
SUPPORTING INDEPENDENT LEGAL THINKING.

GOVERNMENT USE OF ARTIFICIAL INTELLIGENCE IN NEW ZEALAND

Final Report on Phase 1 of the New Zealand Law Foundation's *Artificial Intelligence and Law in New Zealand* Project

COLIN GAVAGHAN

ALISTAIR KNOTT

JAMES MACLAURIN

JOHN ZERILLI

JOY LIDDICOAT

New Zealand Law Foundation | Wellington | 2019

CONTENTS

ACKNOWLEDGEMENTS

We extend warm thanks to Lynda Hagen and the New Zealand Law Foundation for their generous grant enabling our research to proceed.

We are grateful to Eddie Clark, Tim Dare, Katrine Evans, Toby Gee, Janine Hayward, Emily Keddell, Paul Roth, Katrina Sharples and Armon Tamatea for many constructive comments on earlier drafts of this report.

We would also like to express our gratitude to the following people, who participated in various workshops and informal colloquia throughout 2017–18: Nikolaos Aletras; Geoffrey Barnes; Janet Bastiman; Len Cook; Sam Corbett-Davies; Tom Douglas; Jamie Grace; Nigel Harvey; Maaike Helmus; William Isaac; James Mansell; Hannah Maslen; Julian Savulescu; Michael Veale; and Helena Webb.

Others who did not attend these events, but were generous with their time and resources, include: Briony Blackmore; Mady Delvaux; Marina Jirotka; Brent Mittelstadt; Kylie Reiri; and Sandra Wachter.

INTRODUCTION

This report was prepared as part of the New Zealand Law Foundation-funded project: *Artificial Intelligence and Law in New Zealand*. The overall focus of the report is on the regulatory issues surrounding uses of AI in New Zealand. But this is a big topic: there are many types of AI system, and many spheres within which AI systems are used (in New Zealand and beyond). In the design of our project, we wanted to identify some coherent sub-topics within this general area—and in particular, sub-topics where New Zealand could play an important role in a broader international discussion. We identified two such sub-topics; and accordingly, our project was divided into two Phases. The current document reports on Phase 1.

Phase 1 of the project focuses on regulatory issues surrounding the use of predictive AI models in New Zealand government departments. As discussed in Sections 1B and 1C, while there are many types of AI model, the concept of a "predictive model" picks out a reasonably well-defined class of models that share certain commonalities, and are fairly well characterizable as a regulatory target. We specifically focus on the use of predictive models in the public sector because we want to begin by discussing regulatory options in a sphere where the New Zealand Government can readily take action. New Zealand's Government can relatively easily effect changes in the way its own departments and public institutions operate. New Zealand has a small population, and its institutions are comparatively technology-literate, as evidenced by its membership of the D9 group of "digital governments". We believe New Zealand is well placed to serve as a model for how governments can "put their own house in order" in their uses of AI technology. New Zealand has often been able to provide a model to other countries in IT-related areas—for instance, in its early adoption of debit cards, its early use of online medical records, and more recently, in its integrated data infrastructure. We believe it can provide similar leadership in the area of government AI oversight.

Our emphasis in Phase 1 is certainly not intended to suggest that government uses of AI are in some way more important than commercial uses when it comes to regulatory discussions. If anything, we believe the opposite is the case: regulation of the AI technologies used by the multinational giants (Google, Facebook, etc.) is urgently needed. But we think there is a natural ordering to the topics of government AI and commercial AI. As just noted, government AI regulation is quite readily achievable, especially in small countries. Regulation of international tech giants is a much trickier issue. While some issues can be addressed separately within individual countries, many issues can only be dealt with in international agreements. Such agreements require lengthy international trade negotiations, so regulation of AI use by the big multinationals is going to happen over a longer timescale. At the same time—as also emphasised in Section 1B—the actual technologies used by multinational giants to build predictive models (deep networks, random forests, Bayesian models, regression techniques) are very similar to those used by governments. Experience in developing regulatory frameworks for these models in government is likely to be very helpful in tackling the harder problem of how to regulate their commercial use by the tech giants. Phase 1 of our project thus prepares us to participate in follow-up projects—perhaps international ones—focused on the regulation of commercial uses of AI by Google, Facebook and so on.

Phase 2 of the project will focus on the implications on employment of the increasingly widespread use of AI. Again, there will be a New Zealand focus. This research will primarily target commercial uses of AI. But we will not just be considering the technologies themselves: we are also interested in charting the social effects of these technologies, and considering whether regulation could help control some of these. While regulation of the technologies used by international companies requires a protracted international effort, it may be that we can find useful ways of controlling their effects locally (i.e. in New Zealand) within a shorter timeframe.

EXECUTIVE SUMMARY

Artificial intelligence or "AI" encompasses a wide variety of technologies. We focus on "predictive algorithms", an important class of algorithms that includes machine learning algorithms. The general concept of a "predictive algorithm" is useful for many regulation/oversight purposes, and covers a useful subset of the algorithms referred to as "AI" in recent public discourse.

The use of predictive algorithms within the New Zealand government sector is not a new phenomenon. Algorithms such as RoC*RoI in the criminal justice system have been in use for two decades. However, the increasing use of these tools, and their increasing power and complexity, presents a range of concerns and opportunities. The primary concerns around the use of predictive algorithms in the public sector relate to accuracy, human control, transparency, bias and privacy.

PRIMARY CONCERNS

Accuracy: There should be independent and public oversight of the accuracy of the predictive models being used in government. This is of central importance, but such information is not yet readily or systematically available.

Human control: Solutions such as requiring a "human in the loop" have an obvious appeal. But there is a risk that, if we do not approach them carefully, such guarantees could serve as regulatory placebos. In some situations, the addition of a human factor to an automated system may have a detrimental effect on that system's accuracy.

Nonetheless, a human in the loop can be useful: where automated systems are not reliable enough to be left to operate independently; where factors need to be considered that are not readily automatable; or in situations where a measure of discretion is for whatever reason desirable. If a general right to human involvement were deemed to be desirable, such provision should be accompanied by a "right to know" that automated decision-making is taking place (otherwise how would a person be able to demand human oversight in the first place?).

A legal obstacle to automated decisions may arise in public sector contexts, where statutory powers generally cannot be delegated or fettered without parliamentary approval. Statutory authorities that use algorithmic tools as decision aids must be wary of improper delegation to the tool, or otherwise fettering their discretion through automation complacency and bias.

Transparency and a right to reasons/explanations: New Zealand law already provides for a right to reasons for decisions by official agencies, primarily under section 23 of the Official Information Act. This is supported by judicial authority that such reasons must be understandable, both to a review body, to someone with vested interests in the decision and at least in some cases to the public at large. Where individuals affected have a right to an explanation, predictive tools used by government must support meaningful explanations. In cases where the workings of the system are inherently complex, this means augmenting the system with an "explanation system", geared to producing understandable explanations.

Independently of commonsense explainability, the algorithms used by government predictive models should be publicly inspectable. To ensure that agencies can comply with this requirement, policies should be adopted to ensure that algorithms are either developed "in house", or, when purchased from outside vendors, acquired on terms that allow for inspectability, so that neither their form nor conditions of sale preclude or obstruct details of the algorithm being made publicly available.

Bias, fairness and discrimination: "Fairness" in a predictive system can be defined in several ways. It may be impossible to satisfy all definitions simultaneously. Government agencies should consider the type(s) of fairness appropriate to the contexts in which they use specific algorithms.

Exclusion of protected characteristics from training data or input variables does not guarantee that outcomes are not discriminatory or unfair. For example, other variables can serve as close proxies for protected characteristics, and input data that appears innocuous can nonetheless be tainted by historic discrimination.

Privacy: In the realm of privacy and data protection law, we recommend that effect be given to more specific requirements to identify the purpose of collection of personal information (information privacy principle 3). New Zealand should also consider introducing a right to reasonable inferences along with better protections regarding re-identification, de-identification, data portability and the right to be forgotten (erasure).

OVERSIGHT AND REGULATION

There are various general approaches to regulation of AI. One involves the use of what is sometimes called "hard" law, in the form of legislation as interpreted and applied through court decisions; another involves self-regulatory models; a third involves a regulatory agency of some kind.

A range of legal protections—around accuracy, privacy, transparency and freedom from discrimination—already exist in New Zealand law, and all are likely to have important roles in the context of predictive algorithms. The possibility of strengthening or fine-tuning these rights so that they better respond to this technology is certainly worthy of consideration. In this regard, continued attention should be paid to international initiatives such as the European Union's General Data Protection Regulation for comparison.

While important, though, regulatory models that rely on affected individuals enforcing legal rights are unlikely to be adequate in addressing the concerns around increasing use of algorithms. Often, affected individuals will lack the knowledge or the means effectively to hold these tools and processes to account. They are also likely to lack the "wide-angle" perspective necessarily to evaluate their effect across populations.

In addition to individual rights models, then, some form of "top-down" scrutiny is likely to be required if the benefits of predictive algorithms are to be maximised, and their risks avoided or minimised. To that effect, we have proposed the creation of an independent regulatory agency.

There are several possible models for a new regulatory agency. These all have strengths and weaknesses. At present, there are very few international examples from which to learn, and those which exist are in their very early stages.

We have proposed a possible structure for how the new regulatory agency could work with government agencies. The new regulator could serve a range of functions, including:

- Producing best practice guidelines;
- Maintaining a register of algorithms used in government;
- Producing an annual public report on such uses;
- Conducting ongoing monitoring on the effects of these tools.

If a regulatory agency is to be given any sort of hard-edged powers, consideration will need to be given to its capacity to monitor and enforce compliance with these.

If the agency is to be charged with scrutinising algorithms, it must be borne in mind that these are versatile tools, capable of being repurposed for a variety of uses. Scrutiny should apply to new *uses*/potential *harms* and not only new *algorithms*.

Our study has been an initial exploration of the significant issues posed by some forms of artificial intelligence. These issues present risks and also unique opportunities for New Zealand to contribute new ways of thinking about and engaging with these new technologies. To minimize these risks and make the most of these opportunities we stress the need for consultation with a wide variety of stakeholders, especially those most likely to be affected by algorithmic decisions, including Māori and Pacific Islands people.

1. DEFINING THE TECHNOLOGY OF INTEREST

The broad aim of our study is to survey "uses of AI technology in the public sector", and discuss the ethical and legal implications of these. But how should we define the relevant technology? In this first chapter, we begin by considering this question. In Section 1A we argue that general definitions of AI, or of "algorithms", are not very helpful in identifying the technology of interest. In Section 1B, we propose a more precise and restricted technical definition of the systems we will focus on: namely, the class of "predictive systems". We argue that this definition identifies a coherent collection of systems in actual use within government departments, and elsewhere in the public sector, including systems at the forefront of the AI revolution, but also statistical tools that have long been in use. In Section 1C we note that our definition should also pick out a class of tools for which a well-defined set of ethical issues arise, so that laws invoking this definition can coherently target these issues. We argue that predictive systems raise a well-defined set of ethical issues of this kind: this alignment makes the class of predictive systems a useful one for regulators. However, it is not the only class that regulators need to reference. In Section 1D, we note some important cases where the relevant ethical issues apply to a wider group of systems and methods.

A. The challenge of defining AI technologies

Our aim is to discuss regulatory options for "AI systems" in use in the public domain. An important task from a legal perspective is to frame a *definition* of the intended target of regulation. Of course, it may be that existing laws or structures of more general application are found to be adequate. But if this is not the case, then some more precise definition of the technology in question will need to be offered.

In a context where a definition is considered necessary, consideration will have to be given to how tightly the definitional parameters are to be set. Courts, after all, need to know how to apply new rules, regulators need to know the boundaries of their remit, and stakeholders need to be able to predict how and to what the rules will apply. If a new legal rule, regulatory body or code of practice were to be created to respond to "AI algorithms" in government, for instance, it would be important to know exactly which algorithms are in scope.

We suggest that seeking to restrict the scope of any rules to those algorithms which use "artificial intelligence" techniques is unlikely to clarify matters. As recognised by the House of Lords Select Committee on Artificial Intelligence in its 2018 report *AI in the UK: Ready, Willing and Able?* (2018, [9]): "There is no widely accepted definition of artificial intelligence. Respondents and witnesses provided dozens of different definitions".

New Zealand's AI Forum (2018, p. 14) has also acknowledged the problem in its report *Artificial Intelligence: Shaping a Future New Zealand:*

> "The term "Artificial Intelligence" is notoriously hard to define, spanning a wide range of reference points from data science, machine learning and conversational interfaces right through to debate about whether AI will displace jobs and lead to science fiction scenarios."

Of course, many situations that have come to pass started off as "science fiction scenarios". The challenge of how to define these terms is a real one. The AI Forum report settles for a characterisation of AI as "advanced digital technologies that enable machines to reproduce or surpass abilities that would require intelligence if humans were to perform them" (2018 p. 26). This is a simple and useful definition for the purposes of public debate, but it is less clear that it would suffice to identify a clear regulatory target. We might wonder what sort of digital technologies count as sufficiently "advanced". Most people think of the perceptual capacities of a self-opening door and the arithmetic capacity of a pocket calculator are too simple to count as AI. Yet denying that these count as advanced digital technologies seems to owe more to the fact that we are used to them than to facts about the complexity of their workings or manufacture. After all, the great majority of people cannot explain how these machines actually work. In a word, they *aren't* simple, and yet the intuition that they should be excluded from AI regulation seems to be strong (we're willing to bet). So a definitional threshold that refers to the "advanced" nature of a technology is probably not going to help regulators sharpen their focus.

The challenge is not made easier by what some perceive as a disconnection between technical uses of terms from the manner in which they are more widely understood. For example, one influential article by Mittelstadt and colleagues (2016) makes the claim that "Any attempt to map an 'ethics of algorithms' must address this conflation between formal definitions and popular usage of 'algorithm'".

As the influential law and technologies writer Roger Brownsword has warned, the risk of "descriptive disconnection" is ever-present in any attempt to regulate a technology that is still in an emergent stage. In particular, the disconnect poses a risk of definitional over-specificity and deprives regulators of the sort of flexibility necessary to respond to a fast-moving technology. Brownsword (2008, p. 27) puts it this way:

> "the details of the regulatory regime will always reflect a tension between the need for flexibility (if regulation is to move with the technology) and the demand for predictability and consistency (if regulatees are to know where they stand)."

This tension creates something of a dilemma for those charged with making the rules, since

> "the more that regulators (in an attempt to let regulatees know where they stand) try to establish an initial set of standards that are clear, detailed and precise, the more likely it is that the regulation will lose connection with its technological target (leaving regulatees unclear as to their position)." (2008, P. 27)

While "descriptive disconnection" refers to the situation where the technology takes a different *form* from what was anticipated by regulators, what Brownsword calls "normative disconnection" refers to the scenario where the technology is put to an unanticipated use, particularly one that poses different risks or ethical concerns. In that situation, it may be that the rules as originally framed remain connected to the technology itself, but are a bad "moral fit" for the manner in which it is used.

This is not to say that any attempt at anticipatory regulation is doomed to fail. Neither is it to say that any definition must be framed as loosely as possible; overly vague definitions, after all, are likely to serve as poor action guides for those seeking to be guided by the rules.

One possible response to this dilemma is to foreswear an all-purpose definition, and adopt a definition suited to a particular risk or problem. In October 2018, Internal Affairs and Stats NZ released their *Algorithm Assessment Report*, which chose to concentrate on "operational algorithms". The report defined these as:

> "analytical processes [which] interpret or evaluate information (often using large or complex data sets) that result in, or materially inform, decisions that impact significantly on individuals or groups." (STATS NZ 2018, P. 4)

This approach combines a very broad definition of the technical processes under scrutiny, with a very broad definition of the impacts that qualify them for inclusion. The report still explicitly excludes certain kinds of algorithm, namely "Algorithms used for policy development and research" and "Business rules", which we will consider further in Section 1B. But given that our focus is on "AI" algorithms in government, we will attempt to formulate a slightly tighter definition, both of the technical processes that are the object of our study, and of the relevant kinds of impact that might need to be regulated and the form this regulation might take. We will propose a technical definition in Section 1B, and sketch the range of relevant impacts in Section 1C.

B. A technical focus: "Predictive models"

The tools we would like to focus on are "predictive analytics" models, or simply "predictive models". This technical definition will encompass many systems that are squarely at the centre of the current "AI revolution"—but it also extends to systems that have been in use in governments for decades (and sometimes longer). It's a definition that emphasises the *continuity* of computer modelling work within government, and indeed within industry too. This continuity is important in policy terms, because there are already regulations and conventions

surrounding the use of traditional models: to the extent that the new models resemble the traditional ones, some of these regulations and conventions may extend to the new models. Our presentation will also emphasise that the AI models used in government departments (the focus of the current report) are technically very similar to those used in modern commercial settings, for instance by Google, Amazon and Facebook. Regulation of the use of AI models in government may look quite different from regulation of their use in industry, due to the very different social functions of the institutions in these two spheres—but it's helpful to understand that the models that are the object of regulation in these areas are basically the same.

Policymakers sometimes think of AI models as recent arrivals in the fields of government and commerce. It's certainly true that AI models are newly *prominent* in these fields, but it's a mistake to think of them as new as such. There is a long tradition of statistical modelling in both government and commerce—and the statistical models used historically in these fields have much in common with the AI models that are now becoming prevalent.

In this section, we will give a simple introduction to modern AI models, targeted at policymakers who are new to the subject, and with a focus on predictive models. Our approach will be to start by introducing an earlier generation of statistical models. This approach is helpful because the earlier models are simpler. But it is also helpful in highlighting what the new AI models have in common with the earlier models, and how they differ from them.

Simply put, predictive models are models which make predictions about some unknown variable, based on one or more known variables. A "variable" can be any measurable aspect of the world. For instance, a predictive model might predict a person's weight based on their height. Such a model would be useful if for some reason we can readily get information about people's height, but not about their weight (and we are nonetheless interested in their weight). Note that a predictive model doesn't have to predict an occurrence in the future. The unknown variable might relate to the current moment, or even to times in the past. The key thing is that we need to guess it, because we can't measure it directly. More neutrally, we can define a predictive model as a tool that can make guesses about some *outcome variable*, based on a set of *input variables*.

To build a predictive model, the key ingredient is a *training set* of cases where we know the outcome variable as well as the input variables. In the above example, the training set would be measurements of the height and weight of a number of sample people. There is a (loose) relationship between people's height and weight. The training set provides information about this relationship. A predictive model uses this information to compute a general hypothesis about the relationship between height and weight, which it can use to make a guess about someone's weight given their height. A training set is in essence a database of facts about known cases: the larger this database, the more information is provided about the outcome variable. We will take it as part of the definition of a "predictive model" that it is derived from training data, through a training process of some kind.

A brief history of predictive models

Mathematical methods have been in use for centuries to guess an unknown variable by consulting a database of known facts. The first serious applications were in the insurance industry. The Lloyd's register, developed in 1688, which assessed the likely risks of shipping ventures, is a well-known early example (Watson 2010). Equitable Life, the earliest company to use "actuarial" methods to predict life expectancy, was founded in 1762 (Ogborn 1962). The earliest predictive models with relevance to government date from around this time. For instance, in the 1740s, the German statistician Johann Süssmilch used data from church records to devise a model that used the availability of land in a given region to predict marriage age and marriage rate (and through these, birth rates) (Kotz 2005). Governments have been maintaining databases of information about their citizens from time immemorial, largely for the purposes of assessing their tax obligations. As the science of predictive modelling developed, these databases could be reused for other government functions, particularly those relating to financial planning. The British Government employed its first official actuary in the 1830s: an employee who worked in naval pensions, and is credited with saving the government hundreds of millions of pounds in today's money (Clarke 2018). Already at that time, the predictive models used in government mirrored those used in commerce—a trend that continues to this day.

To begin with, developing predictive models involved calculations done by hand, using databases stored in written ledgers. Computers can help in two ways: they facilitate the storage of large amounts of data, and they can perform calculations automatically. Predictive models are now routinely implemented as computer programs that consult databases held in computer memory.

When computers were first introduced, their only users were governments and large corporations, due to their great expense. Both companies and governments quickly started to develop computer-based predictive models. For instance, the FICO corporation in the US, which specialises in predicting credit risk, produced its first computerized model of risk scores in 1958. The US government used computers to predict missile trajectories in the 1940s (Weik 1961), to predict weather in the 1950s (Platzman 1979), and to predict suitability of military personnel for missions in the 1960s (Roomsburg 1988). Governments across the world have been using predictive models for a wide variety of purposes ever since.

We will introduce modern predictive AI systems by first presenting some older predictive statistical models, and then showing how the AI models extend (and differ from) these.

Actuarial tables

The Equitable Life company made a novel contribution to insurance when it produced a table showing for each age, the probability that a person will die at that age (based on available mortality statistics), and computing an associated insurance premium for that age. This kind of "actuarial table" is a simple form of predictive model. Its innovation lay in systematically charting probabilities for each possible age. As actuarial science progressed, more complex tables were developed, taking into account factors other than age, so that more accurate predictions could be made, and more accurate premiums charged.

Geometric approaches to predictive modelling

A drawback with actuarial tables is that they treat ages as discrete categories, and then compute probabilities for each age separately, without regard for one another. But age varies continuously, and the probability of dying varies smoothly as a function of age. It is useful to be able to think about probabilities *geometrically*, as functions on graphs, so that probability models can be expressed in the terminology of Cartesian geometry.

For instance, we can define a mathematical function that maps a variable "age" represented on the x axis of a graph onto a probability of dying represented on the y axis, as shown in Figure 1. This idea was pioneered by the nineteenth century actuary Benjamin Gompertz, who found that the probability of dying can be modelled quite accurately by a simple function, examples of which are shown in Figure 1. The four curves in this figure show the function with different sets of parameter values.

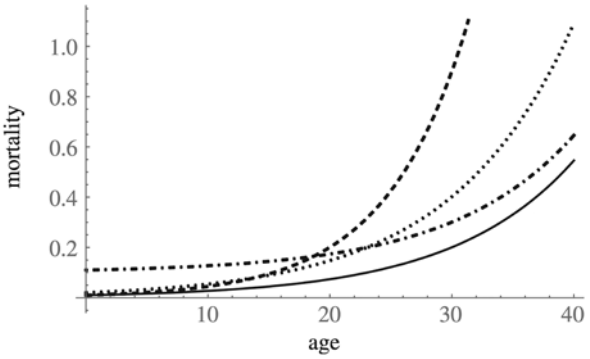


Figure 1. Examples of the Gompertz curve: a mathematical function mapping age onto probability of dying.

We still have to use actual training data to estimate these functions, of course. But now, estimation involves setting the parameters of a continuous mathematical function—"fitting a function" to some data—rather than estimating many separate probabilities. On this approach, age is modelled as a "random variable": we don't know its actual value, but we have a mathematical function that models the "distribution" of its *possible* values.

Regression models

An important class of predictive model that uses a geometric conception of probability is the regression model. The first such model was the linear regression model developed in the early nineteenth century (independently, by Gauss and Adrien-Marie Legendre). The key idea in linear regression is to model the relationship between variables with a mathematical function. For instance, recall from our earlier discussion that there is a "loose" relationship between height and weight. To quantify this relationship, we can gather a set of known height-weight pairs, to serve as a "training set" for our model. An example training set is shown in Figure 2, as a set of data points on a two-dimensional graph, where the *x* axis depicts height, and the *y* axis depicts weight.

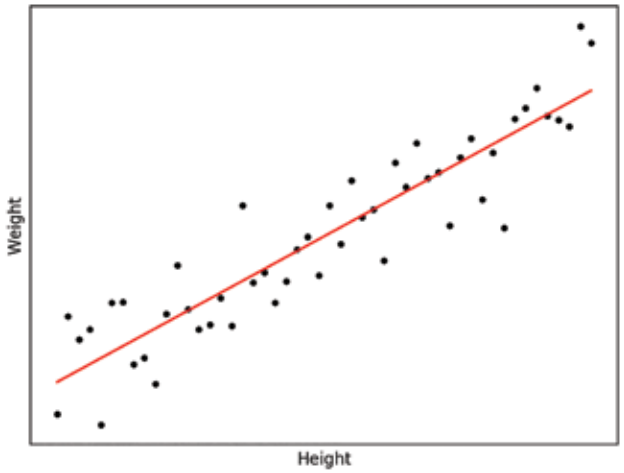


Figure 2. A function defining a relation between people's height and weight, learned by linear regression from the datapoints shown in black.

Crucially, having represented the training set as points in a graph, we can learn a mathematical function that maps *every possible height* onto a weight, as shown by the red line in Figure 2. (The training points include "noise": miscellaneous factors that are relevant to weight, but aren't included in the model.) The line gives an answer even for heights which aren't included in the training set, and thus can be used to estimate weights for people who aren't exactly like those in the training set. It does this by smoothly *interpolating* between the known points. Note that the line doesn't go through many of the training points. (It might not go through any at all.) In the presence of noise, we have to make our best guess. Linear regression is a mathematical way of making a best guess by finding (or "defining") the function that makes least error in its approximation of the training points. Note that a "function" is just a particular kind of "model". So a function learned by regression from a set of datapoints can be used as a "predictive model" of the kind that is our focus.

Modern regression models

Regression is a key technique in modern statistical modelling. The basic method outlined above has been expanded on in a multitude of different ways. For instance, linear regression models can involve many variables, not just two. If we have exactly three variables, datapoints can be visualised in a three-dimensional space, and regression can be understood as identifying a three-dimensional *plane* that best fits the points. (For more than three dimensions, we have "hyperplanes" that are hard to visualise, but the mathematical techniques are just the same.) Moreover, regression modelers are free to decide how *complex* the function that fits the training datapoints should be. In Figure 2, the function is constrained to be a straight line, but we can also allow the function to be a curve, with different amounts of squiggliness[1]—or, in three dimensions or more, planes with different amounts of "hilliness". Furthermore, regression techniques can also be used to model relationships between variables in circumstances where outcome variables can take a number of discrete values (unlike the above examples, where, for a given input variable like "height", only *one* outcome variable is calculated, such as "weight"). This is done in "logistic regression" models. Finally, there are many varieties of regression model specialised for particular tasks. An important variety for many government applications is "survival analysis", which is a method for estimating the likely amount of time that will elapse before some event of interest happens. The original applications of these methods were in drug trials, where the "event of interest" was the death of a patient under some form of therapy. But there are many applications in government planning where it is very useful to have a way of predicting how far in the future some event may be, for different people or groups. Again, the same regression methods are used both within government and industry.

We should also note that regression models don't *have* to be used for prediction. Scientists who use it are often just interested in stating relationships between variables in some domain. A scientist might, for instance, want to be able to state as an empirical finding that "there is a relationship between height and weight". Methods relating to regression can be used to quantify the strength of this relationship. (These methods were pioneered by Francis Galton and Karl Pearson in the early twentieth century in their concept of "correlation".) Often, scientists' main agenda in using regression models is to identify the strength of the correlations between variables in a dataset—again, purely for the purpose of describing the data succinctly, and therefore understanding the domain of study. Nonetheless, regression models can be used to make predictions about unknown variables in a domain, based on a sample of known cases. And the stronger the correlations are between variables, the better these predictions will be.

This raises an important point in relation to discussions about AI. The field of AI has had a dramatic impact on predictive modelling, an impact due only to a particular

1. Or more technically, "polynomials" of different degrees.

branch of AI—the field known as "machine learning" (our focus in this report). Current commentators often refer to "machine learning" as if it's a new phenomenon. But regression models are, fundamentally, machine learning techniques. They take a finite set of training instances to learn a general model, which can be applied to unseen cases. Even an actuarial table can be considered a simple machine learning technique, in that it can make predictions about cases beyond those used to construct it. What has changed over the years is that our machine learning models have become more complex, and as a result more powerful.

In what follows, we will introduce some newer machine learning techniques that are often associated with the field of AI: decision trees, neural networks, and Bayesian models. In fact, all of these techniques have a history in statistics as well as in AI: only neural networks squarely originated within AI. All share a focus on the *process of learning*, which distinguishes them from regression modelling, where the focus is on fitting mathematical models to data. They also share a focus on complex data, where it's not obvious which mathematical models are to be fitted.

Decision trees

Decision trees are a fairly simple machine learning method, and are often used to introduce machine learning models. A decision tree is a set of instructions for guessing the value of some outcome variable, by consulting the values of the input variables one by one. A toy example in the domain of criminal justice (a domain that features prominently in this report) is shown in Figure 3. This decision tree provides a way of guessing whether a prisoner up for bail will reoffend (the outcome variable), based on whether they have behaved well in prison, and whether they committed a violent offence (two toy input variables).

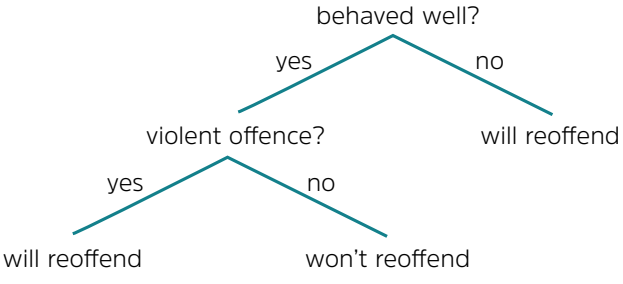


Figure 3. A simple decision tree for predicting a prisoner's reoffending.

The tree states that if the prisoner didn't behave well, they will reoffend, regardless of whether their original offence was violent. If they did behave well, they will reoffend if their original offence was violent, and they won't if it wasn't. (We use this crude example for illustration, but decision trees—admittedly with many more variables—are widely used to aid with practical decisions in public policy, as we will describe later.)

The key task in decision tree modelling is to devise an algorithm that *creates* a good decision tree from the training data it is given.[2] (Such an algorithm models a learning *process*: there's no comparable model of a learning process in regression modelling.) In the classic algorithm, we build the decision tree progressively, starting from the top: at each point in the tree, we find the input variable that supplies the most "information" about the outcome variable in the training set, and add a node consulting that variable at that point. Stepping back from this particular algorithm, it's also useful to see decision tree modelling as embodying a particular approach to statistics that starts with the data, and asks pragmatically how we can best make use of it.

An attractive feature about a decision tree is the procedure for reaching a decision can be readily understood by humans: at base, a decision tree is just a complex "if-then" statement. Understandability is an important attribute for machine learning systems making important decisions. However, modern decision tree models often use multiple decision trees, embodying a range of different decision procedures, and take some aggregate over the decisions reached. "Random forests" are the dominant model of this kind at present. For various reasons, these aggregate methods are more accurate. There is frequently a tradeoff between a machine learning system's explainability/complexity and its predictive performance. (This is true for regression models as well, incidentally.)

Decision trees provide a useful opportunity to introduce the problem of *overfitting* in machine learning (and statistical modelling more widely). Say we have a training set of prisoners for whom we record the values of many variables. We could readily construct a large decision tree that lets us decide whether prisoners will reoffend based on this training set. But reoffending is likely to be at best a "loose" function of these variables:

2. For example, you can imagine the sort of data that would lead an algorithm to create the decision tree in Figure 3: data on good behaviour, episodes of violence, and so on.

they are unlikely to correlate perfectly with reoffending. Put another way, the training set is likely to contain a considerable degree of *noise*, which is not helpful to the machine learning system, as it won't carry over to new prisoners. If we are not careful, our decision tree will simply be a detailed *description of the training set*, rather than a model of the complete population of prisoners from which the training set is drawn. There are various ways of preventing overfitting, which we won't describe here—but it is important to be aware that this problem can arise.

We can also use decision trees to introduce the concept of a "classifier", which is widely used in machine learning. A classifier is simply a predictive model whose outcome variable can take a number of discrete values. These discrete values represent different classes that the input items can be grouped into. Decision trees have to operate on variables with discrete values, so they can easily implement classifiers. Our decision tree for reoffending can be understood as a classifier that sorts prisoners into two classes: "will reoffend" and "won't reoffend". (To implement classifiers with regression techniques, we must use logistic regression models, which were specifically designed to handle discrete outcome variables.)

Neural networks

Neural networks (sometimes called "connectionist" networks) are machine learning techniques that are loosely inspired by the way brains perform computation. A brain is a collection of neurons, linked together by synapses. Each neuron is a tiny, very simple processor: the brain can learn complex representations, and produce complex behaviour because of the very large number of neurons in the brain, and the even larger number of synapses that connect them together. Learning in the brain happens through the adjustment of the "strength" of individual synapses. (The strength of a synapse determines how efficiently it communicates information between the neurons it connects.) We are still far from understanding how this learning process works, and how the brain represents information.

A neural network is a collection of neuron-like units that perform simple computations and can have different degrees of activation. These units are connected by synapse-like links, that have adjustable weights. (It's important to emphasize that in most practical networks, these "units" and "links" are only based in the loosest terms on real neurons and synapses.) While neural networks were originally developed by researchers interested in biological learning processes, they are now also widely used by people who have no interest in brain modelling, simply because of their power as learning devices.

There are many different types of neural network—but networks that learn a predictive model are predominantly "feedforward networks" of the kind illustrated (very roughly) in Figure 4. A feedforward network used to learn a predictive model has a set of input units which encode the input variables of training items (or test items); it has a set of output units which encode the outcome variable for these same items; and it has a set of intermediate "hidden units". Activity flows from the input units, through the hidden units, to the output units: through this process, the network implements a function from input variables to the outcome variable, just like a regression model or a decision tree model. Often there are many "layers" of hidden units, each connected to the layer before and the layer after. (The network in Figure 4 has one hidden layer.)

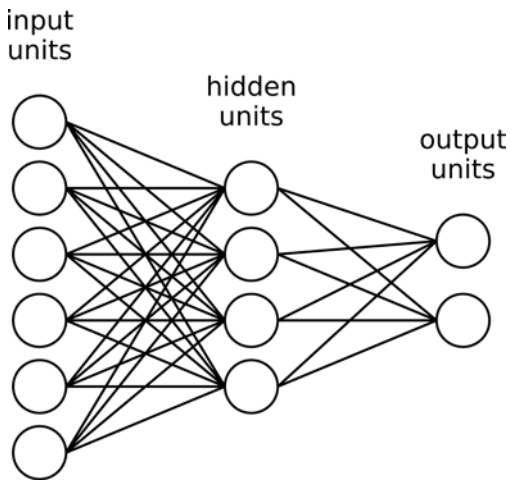


Figure 4. A simple feedforward network. (A template for a "deep network".)

To give a simple example: imagine the network in Figure 4 is a very simple image classifier that takes a tiny image comprising 6 pixels, and decides whether these pixels represent an image of type A or type B. The intensity of each pixel would be encoded in the activity of one of the input units. The activity of the output units encodes the type, in some designated scheme. (For instance, type A could be encoded by setting the activity of one unit to 1, and the other unit to 0, while type B could be encoded by setting the activity of the former unit to 0, and the latter unit to 1.)

There are many different learning algorithms for feedforward networks. But the basic principle for all of them is "supervised learning". In this learning algorithm, we begin by setting the weights of all the links in the network to random values. The network then implements a function from inputs to outputs in successive rounds of training. (In our case, it makes guesses—essentially random guesses—about the type of each training image.) To train the network, we present the training inputs to the network one at a time, in each case computing what its guesses are. Crucially, we compute the "error" of each guess by comparing the network's *actual* output values to the output values it *should* have produced, making small changes to the weights of all links in the network so as to *reduce* that error. This gradually improves the performance of the network.

All the smarts in a supervised learning algorithm relate to how to tweak its weights so as to reduce error. A big breakthrough in this area is the technique of "error backpropagation", which was invented (or at least made prominent) in 1986 by David Rumelhart and colleagues at the University of California, San Diego. This algorithm allowed the weights of neurons in a network's hidden layer(s) to be sensibly adjusted. The invention of backpropagation led to a wave of academic interest in neural networks, but not to immediate practical effect. The development of "deep networks" in the late 1990s and early 2000s was due partly to a number of separate extensions to the algorithm, and partly to the huge increases in computing power that occurred around that time. The most salient feature of deep networks is that they have many layers of hidden units (unlike the single layer in Figure 4). There are now many varieties of deep networks, deployed in many different areas of machine learning. Deep networks of one sort or another are often the best performing models. The field of machine learning has in fact undergone a paradigm shift: the majority of researchers in this area currently focus their attention on deep networks. There are several software packages that support the implementation, training and testing of deep networks (of which the most prominent at the moment is Google's TensorFlow). These packages have undoubtedly helped to consolidate the new paradigm, and their ongoing development has helped to progress it.

Deep networks still rely crucially on supervised learning and backpropagation. The main thing to understand about this training regime is that it *gradually* improves the network's performance by making small changes to its weights. The training process is often construed by imagining a large multidimensional space that represents all the possible weight combinations the network can have. Each point in this space represents a different model (or "function") the network can express. To describe how supervised learning selects the model which best represents the training set, we can envisage a space with one additional dimension representing the network's total error on the training set. Within this larger space, we can define an "error surface": a hilly landscape, where peaks correspond to high error, and troughs to low error. In backpropagation, the network starts at a random place in this landscape, and simply follows the slope of the landscape downwards, by many small steps, until it can't go any further down. It's quite possible that there are points in the landscape with lower error: backpropagation is only guaranteed to find a "local minimum". This rather heuristic method of finding the best model is very different from the regression algorithm described earlier, which *provably* finds the function with the very lowest error on the training set. What makes deep networks competitive with regression is that their hidden units allow them to learn their *own* internal representation of the training items. Regression, and other machine learning techniques, don't have this ability.

It should be noted that deep networks are also susceptible to overfitting. In fact they are more susceptible, because the models they learn can be very complex. (In graphical terms, they can describe almost arbitrarily "curvy" or "squiggly" functions.) But many of the techniques that address overfitting for other machine learning methods also work for neural networks—and there are other useful techniques that are specific to neural networks.

Nonetheless, a significant drawback with deep networks is that the models they learn are so complex that it is essentially impossible for humans to understand their workings. Humans have a reasonable chance of being able to understand a decision tree (or even a set of such trees), or to understand a regression model that succinctly states the relationships between variables. But they have no chance of understanding how a deep network computes its output from its inputs. If we want our machine learning tools to provide human-understandable explanations of their decisions, we need to supplement them with additional tools that generate

explanations. The development of "explanation tools" is a growth area of AI, and sufficiently important that we will discuss explanation systems separately in this section (see below). They will also feature in our discussion in Section 4B.

Bayesian models

Thomas Bayes formulated his famous theorem around the time the first actuarial tables were produced: "Bayes' theorem" was published posthumously in 1763. The theorem was applied in various scientific fields, notably through the work of Pierre Laplace—and in fact, was used implicitly by Alan Turing in some of his code-breaking work (Fienberg 2006). But its widespread use in predictive modelling is relatively recent: like neural networks, practical Bayesian models require a lot of computing power, and it was only in the 1980s that they became popular as a machine learning technique.

Bayesian models can be introduced in various ways. One approach is to focus on Bayesian models as models of a learning process whereby an initial hypothesis is refined, or updated, when new data arrives. Another approach is to focus on the Bayesian conception of probability as modelling the degree of belief that an observer has in some proposition, given the facts she has encountered, plus some initial predisposition. We will pursue a third approach, which sees Bayesian models as a practical way to address a problem often encountered when gathering training data for predictive models.

Say we are interested in building a model that predicts the presence or absence of some disease, given a large set of symptoms. It is often complex *combinations* of symptoms that diagnose diseases: the relevance of individual symptoms is hard to quantify. More generally, we are often interested in predicting a cause variable from its observable effects. (Remember, the "predicted" variable in a predictive model can be in the past.) Again, the relevance of individual effects in diagnosing the cause is hard to quantify.

We can illustrate the difficulty by considering the case of disease diagnosis a little further. If we have a training set of people with the disease we are interested in, and people without the disease, we can readily estimate the probability of each individual symptom given the presence or absence of the disease. But what we *want* to estimate is the probability of the disease, given some particular combination of symptoms. Using standard models for estimation (e.g. regression), this would require us to gather a training set in which there are people with every possible combination of symptoms. (In fact, to make accurate estimates, we would need *large numbers* of people with each combination of symptoms, so we are not misled by individual cases.) As we increase the number of symptoms in our model, this quickly becomes impossible.

Bayesian models address this problem. Bayes' theorem allows us to infer the probability of a cause given its effects—the probability we want, but which is hard to estimate—from the probabilities of individual effects given the cause (which are easier to estimate). For instance, the theorem lets us infer the probability of a disease given a particular combination of symptoms, from the probabilities of each symptom given the presence or absence of the disease. Bayes' theorem is of great practical use in many cases where we wish to predict (or diagnose) some hidden cause from a set of readily observable effects.

Bayesian models are often used in classification tasks. For instance, spam filters, which classify email documents into the categories "spam" and "non-spam" are often implemented as Bayesian models. In this context, the category of a document is the "cause" which is hidden to the user, and the effects of this cause are the observable words in the document. Using a training set of known spam and non-spam documents, we can readily estimate the probability of any given word appearing in a spam document, and in a non-spam document. We can then use these probabilities, together with Bayes' theorem, to infer the probability of a document being spam or non-spam, given the words it contains.

Explanation tools for complex predictive models

Modern predictive models operating in real-world domains tend to be complex things, regardless of which of the above machine learning methods they use. If we want to build a predictive system that can convey to a human user *why* a certain decision was reached, we have to add functionality that goes beyond what was needed to generate the decision in the first place. The development of "explanation tools" that add this functionality is a rapidly growing new area of AI.

The basic insight behind the new generation of explanation tools is that to understand how one predictive model works, *we can train another predictive model to reproduce its performance* (see Edwards &

Veale 2017 for a good review). While the original model can be very complex, and optimised to achieve the best predictive performance, the second model can be much simpler, and optimised to offer maximally useful explanations. These "model-of-a-model" explanation systems have the additional benefit that they provide an account of how a system arrived at a given decision without revealing any of its internal workings—essentially, by treating it as a "black box". Some of the most promising systems in this space are ones that build a local model of the factors most relevant to any given decision of the system being explained (see e.g. Ribeiro et al. 2016). How best to configure a model-based explanation tool is still an active research question—but the availability of such systems should certainly inform current discussions of transparency in AI tools. We will return to this topic in Section 4B.

Non-predictive algorithmic tools used in government

Our review has focused on "predictive models" and their use in government. Before we conclude, we will briefly note some other types of computer tool in widespread use in government departments, which will ultimately fall outside the scope of our recommendations.

"Optimisation systems" are sophisticated computer programs that are designed to find an optimal solution to some complex problem featuring many variables. Common applications are in timetabling, route planning or resource allocation. In a typical application, there is a large multidimensional space of possible solutions, and a way of computing the "goodness" of any given solution: the optimisation algorithm has to find the solution with the best "goodness". For instance, it might have to find the "best" bus route that passes through a certain set of stops. (Often, the algorithm searches for an optimal solution by exploring a hilly "goodness surface", similar to the "error surface" explored by neural network learning algorithms, travelling upwards to find peaks, rather than downwards to find troughs.) The goodness of a bus route might take into account measures such as how much petrol is used, how much time is taken, and the likely demand of people at stops. These define the algorithm's "goodness" function. Typically, the goodness function is a weighted sum of several such measures. (How the different measures are weighted is often an ethically-laden issue: for instance, stops in neighbourhoods that rely more on buses might be weighted more heavily than other stops.)

Optimisation algorithms are an important tool for government planners, but they are not systems that make decisions or judgements about individuals in the way that predictive models can. In a sense, they make a single "decision" about how to solve a particular problem. While this decision may determine government strategy in some particular area, and thus have importan t relevance for groups of people, such as populations (or subpopulations), it never affects people as individuals, the way predictive tools frequently do. The New Zealand government's recent *Algorithm Assessment Report* (Stats NZ 2018) distinguished helpfully between "operational" algorithms, which make decisions about individuals based on their personal circumstances, and other decisions that could be called "strategic", which are used to inform policy development or research and don't impact on individuals directly in this way. The report's focus was on operational algorithms—and ours will be too.

"Ranking systems" are also extensively used in government. A ranking algorithm takes a set of items as input, and ranks them according to some predefined criterion. For instance, an algorithm might rank patients waiting for surgery, using a criterion capturing some measure of urgency. The criterion of a ranking algorithm is like the goodness function of an optimisation algorithm: it is typically defined as a weighted sum of attributes. In a system ranking patients for surgery, it might take into account the seriousness of a patient's condition, the likelihood of the surgery being successful, the patient's age, and so on. As with the goodness function in optimisation, everything hangs on the attributes picked, and the weights assigned to them. The aim is often to use weights that deliver rankings similar to those of human experts.

Ranking algorithms certainly "make decisions" about individual people—often very important ones. They are certainly "operational" algorithms, in the *Algorithm Assessment Report's* sense. But they don't make predictions that can be independently verified; and they don't typically need to be trained. For this reason, we distinguish them from predictive systems.

An important class of computer algorithms to mention is that of "business rules". While optimisation systems can be very complex, business rules are at the other end of the spectrum: they are simple pieces of code, that automate routine procedures. A payroll program may implement a number of business rules. Such systems are in very common use: they have no autonomy of their

own, and they certainly don't involve AI techniques. The *Algorithm Assessment Report* excluded these from its remit, and we will too.

We should finally note that we include within our class of "predictive models" models that advocate some course of action, rather than a prediction *per se*. For instance, a model that advocates a range of medical interventions given patients' circumstances is advising on courses of action, rather than predicting facts or events. But such systems are still trained on corpora of actions actually taken in different circumstances, and their training can use exactly the same methods as predictive models. They can be thought of as predicting human decisions about how to act.

Protocols for testing predictive models

It's essential to test predictive algorithms against independent data, as we have already noted. But it's important to stress that if a predictive algorithm is deployed it should be *regularly* tested, and if necessary retrained. It's essential that the items used to train an algorithm are *representative* of those on which the algorithm is deployed. If they are not, performance will decrease, and biases of various kinds may be introduced. This principle leads to many precepts in statistics—for instance, a system trained on one population should not be used on a different population. "Populations" can vary over space, but also over time, as people's behaviours and characteristics change. For any continuously deployed predictive algorithm, there should be a protocol for regular re-evaluation. There should also be a protocol for regular gathering of new training data, so that the system's data do not fall out of date. Sometimes this new training data requires human judgements uncontaminated by the predictive algorithm currently in use—which is something to bear in mind when predictive algorithms are deployed.

Often, a predictive algorithm can make several different *types* of error, which have very different implications for its use in the field. Consider a "binary classifier" that is trained to recognise members of one particular class. During testing, this classifier labels each test individual either as "positive" (a member of the class in question), or "negative" (not a member). If we also know the actual class of the test individuals, we can chart how often it is right and wrong in its assignment of these labels, and express these results in a "confusion matrix". An example of a confusion matrix is shown in Table 1. The classifier in this case is a system trained to predict fraudsters: it makes a "positive" response for people it predicts will commit fraud, and a "negative" response for everyone else.

	Did commit fraud	Did not commit fraud
Predicted to commit fraud	True positives	False positives (type 1 errors)
Not predicted to commit fraud	False negatives (type 2 errors)	True negatives

Table 1. A confusion matrix for a fraud detection algorithm.

The confusion matrix shows how frequently the system is right or wrong in both kinds of prediction. A "false positive" is a case where the system wrongly predicts someone to commit fraud; a "false negative" is a case where it fails to detect an actual fraudster. If a system is not perfect, there will always be a *tradeoff* between false positives and false negatives (and between true positives and true negatives). For example, an algorithm that judges everyone to be a fraudster will have no false negatives, while one judging no-one to be a fraudster will have no false positives. Importantly, in different domains, we might want to err on one side or the other. For instance, if we are predicting suitability for a rehabilitation project, we might want to err on the side of false positives, while if we are predicting guilt in a criminal case, we might want to err on the side of false negatives. For many applications, we would like the evaluation criterion for a classifier to specify *what counts as acceptable performance* in relation to the confusion matrix: that is, what sorts of error we would prefer the system to make.

One other evaluation metric we should introduce here is the "receiver operating characteristic" (ROC) curve. It is often possible to *tune* a binary classifier, so it occupies different points on the tradeoff between false positives and false negatives. For instance, we could tune a fraud detection algorithm to err on the side of finding all fraudsters, or to err on the side of making no false accusations, and to all settings in between. If we want to compare two classifiers, we often want to do so in a way that considers performance for all possible tuning settings. The ROC curve for a binary classifier is a graph plotting its false positive rate against its true positive rate for a variety of settings. A classifier that always returns a negative result has a false positive rate of 0, and by the

same token, a true positive rate of 0 as well. A classifier that always returns a positive result has a false positive rate of 1, and a true positive rate of 1. For classifiers that can be tuned to these extremes, and all settings in between, we can plot a full ROC curve. In other cases, we might have to plot performance for just some settings, and estimate the rest of the curve. In either case, the metric we are interested in is the "area under the curve" (or AUC), which charts the performance of the classifier at all points on the false positives/negatives tradeoff.

Conclusion

In this section we have presented the main varieties of predictive models currently used by government departments around the world. Our aim has been to emphasize the continuity of modelling techniques: today's models are extensions of predictive models that have been in use since the start of the computer age, and in some cases, well before that. We have also emphasised that while these models are often referred to in current discussions as "AI" models, they are often equally well described as "statistical" models. (We will mostly use the term "AI model" in what follows, because of our project's explicit focus on AI.) The main novelty of modern AI predictive models is that they often perform better than traditional models, partly because of improvements in techniques, and partly due to the many new data sources that are coming online. For this reason, they are becoming more widely adopted in government departments (and elsewhere). A positive feedback loop is in evidence. Because the new models are successful, considerable effort is expended to improve the software implementing these models, making it more efficient and easier to use. Practitioners are also increasingly being taught about these models, and trained in the use of the associated software. This in turn leads to further proliferation in the use of the models, further improvements in the software, and so on.

In the case of modern AI technologies, it has to be said that commercial companies have taken the lead in development and adoption. But government departments are catching up, and are increasingly deploying these tools. The new interest shown by policymakers is partly due to the genuinely novel features of these tools, and partly due to the new prevalence of these tools in government operations (and elsewhere). Both of these factors should indeed occasion a renewed interest in the use of AI tools in government among regulators and policymakers.

C. Ethical and regulatory issues arising for predictive models

In defining a class of AI systems to serve as a regulatory target, we need to identify a class which is coherent from a technical perspective—but also one that is coherent as regards the set of ethical issues that arise, and the types of regulation that are necessary. Regulation will dictate a particular *approach* to systems of the identified type: certain aspects of their performance that must be scrutinised, or evaluated, or controlled, in order to address the relevant ethical issues. It is therefore important that the same ethical issues arise for all instances of these systems.

We believe that a reasonably well-defined set of ethical issues arise in relation to government use of predictive models. That is, we believe that predictive models are a well-defined *technical* class of models—which nets in modern AI models as well as older statistical models—but we also believe they are a well-defined *ethical* object of study. This convergence of technical and ethical definitions makes them a helpful category of algorithms for regulators.

The ethical issues at play will be discussed at length in Chapter 4, after the wider context is set in Chapters 2 and 3. But we will briefly summarise the issues here. First, if predictive models are being used to aid humans in making decisions, how can we ensure that the interaction between humans and machines is optimal? In particular, how can we ensure users don't become *passive* partners in this interaction? Second, if models are contributing to decisions about people, how can those affected obtain explanations about the decisions that were taken? Third, how can we ensure that the machine's decisions are not biased, for or against some particular group? Cutting across all of these questions is the wider question, *how well is the machine performing?* As discussed in the previous section, there are standard ways of *evaluating* predictive models on unseen test data, and looking for various kinds of error: false positives, false negatives, and so on. Our key point for the moment is that all these questions can be coherently asked of any predictive model, regardless of how it is implemented: whether it is a random forest, or a neural network, or a regression model, is of little importance. If the target of regulation is "predictive models", it should be possible to propose regulatory mechanisms that apply in sensible ways across these different technologies.

D. A wider view of government algorithms and mechanisms

We certainly don't want to suggest that the ethical issues that arise for predictive systems *only* arise for this type of system. Many of them apply more widely—and regulatory options should certainly reflect this. In this section, we will flag two important ways in which the ethical issues we focus on extend to a wider group of algorithms and mechanisms.

Firstly, many tools that subserve "policy development and research" (as flagged in the *Algorithm Assessment Report*) raise important ethical issues. Optimisation and ranking algorithms, in particular, can raise issues around bias. For instance, an algorithm optimizing bus transport routes might not take into account special circumstances for certain groups. But the issue of human-machine interaction in decision-making does not apply in the same way. Policy development and research tools are used by government experts, rather than by caseworkers in the field. As they are not used as part of a routine practice, there is little danger of users becoming passive partners. The fact that users are experts also somewhat obviates the requirement for explanations of decisions for these systems. Finally, there is no analogue for these systems of the performance evaluations that can be conducted for predictive models, to see how well they have learned their task. They are not trained to perform a task in the same way as predictive models, so different evaluation methods must be applied.

Secondly, there are some decision-making mechanisms in government domains which have a component of automation, *even though they are not implemented in computers*. To illustrate, we will briefly consider a very simple tool for structured decision-making. The Youth Offending Risk Screening Tool (YORST) was first deployed by the New Zealand Police in 2007. It appears in the *Algorithm Assessment Report* but is an "algorithm" only in a weak sense of the term. The YORST is a simple checkbox form (see Appendix 1). The answers for each question are summed at the bottom of the form providing a risk score. Crucially, this process does not (as far as we are aware) result from any automated training regime: as well as being implemented by hand, it was also devised by hand. The YORST is not a "predictive model", as we define the term, because it isn't trained on data using automated means. And it certainly seems unlikely that anybody would count it as a form of artificial intelligence. Yet some of the ethical issues it raises are similar to those associated with bona fide predictive models that provide much more sophisticated automated support for decisions in government.

The YORST is similar to tools used in many jurisdictions for the case management of at-risk young offenders. External review (Mossman 2010) has found it to be a reliable evidence-based tool for risk assessment in that the factors it measures are reliable indicators of reoffending in young people. In the context of case management, the tool seems to be well-designed and fit for purpose. Yet, when in the lead up to the 2017 election, the National Party proposed using the YORST as a tool for screening young offenders into "boot camps", public debate turned to a variety of aspects in which the YORST seemed problematic. The proposed policy was to apply to any "Young Serious Offender". This was a new category of young person who had committed a serious offence, scored highly on the YORST, and had offended after being in a youth justice or adult custodial facility. Public concerns quickly coalesced around the algorithmic component of this test.

These concerns were understandable. One issue was that while the YORST does not use ethnicity as a variable, many of the variables it employs, while predictively accurate, are also proxies for ethnicity. Māori youth are more likely to come from families that have been subject to Child Youth and Family notifications. They are also more likely to live in poor neighbourhoods. So while the YORST follows accepted practice by focusing on the causes of criminal behaviour rather than on ethnicity (Gottfredson & Snyder, 2005), the proposed policy seemed likely to result in boot camps disproportionately populated by Māori youth. And while the stated aims of the policy were to turn the lives of young people away from crime, in the short term it seemed likely to exacerbate rather than alleviate existing racial inequality in New Zealand (Harris 2017). A second issue arises from the sources of data that the YORST employs. Part B of the tool focuses on "Peer Group Factors" and part C on "Family Factors" including family violence and living situation (see Appendix 1). This led to concern that the proposed policy was effectively punishing young people, not for their own behaviour, but for the behaviour of those around them.

So an "AI" like YORST poses substantial risks, even though it's technically speaking "AI lite", and these risks are no less serious than those posed by more "heavy-

duty" AI tools utilising more advanced machine learning methods. In light of that, there is clearly value in crafting regulations so that the kind of harms or risks posed by a technology, rather than its architectural design or modelling principles, dictates whether and how it may be used. Conversely, regulation which *does* specifically define and target a type of artificial intelligence should generally only be appropriate where the issues raised by that technology are not found elsewhere.

For example, in Chapter 4, we survey the risks associated with using automated decision-making tools in government contexts and elsewhere. These include difficulties in retaining meaningful control over decision-making, algorithmic opacity, bias, privacy intrusions, and so on. Most of these issues are not unique to applications of any one system that is used by the New Zealand Government. Appropriate levels of transparency in public decision-making are essential no matter who or what is making the decisions, and no matter whether a tool uses random forests or Bayesian methods to arrive at its decisions. Data protection principles must be respected regardless of the technology that is used. Discrimination on prohibited grounds is no more allowed by a deep learning tool than by a simple decision tree. And so on. In all these cases, the key point is to identify the relevant risks the decision-making technologies pose, and to respond to these risks.

To summarise: while for some purposes it is useful to pick out a particular class of technologies (like the class of predictive models), the development of technology-specific regulations is likely going to be insufficient to tackle all the various issues discussed in this report. The issues of bias and fairness which arise from the use of a simple tool like the YORST are exactly the same as those which arise from the use of complex random forest and deep learning tools. A technology-specific ("AI-centric") approach to such issues would probably miss simple tools such as the YORST altogether. As well as attempting to define regulations that apply specifically to predictive models, we should keep in mind that many of the ethical issues that apply to this class of model may also apply much more widely within operation of government departments.

2. CURRENT AND PROJECTED USE

A. The New Zealand Government's algorithmic stocktake

Recent years have seen significant attention drawn to the use of algorithms by the New Zealand Government and Crown entities. Specific applications, including those by ACC and Immigration NZ, have attracted criticism from the media and academic commentators (e.g. Johnston 2017; Tan 2018).

Aside from these individual cases which have come to media attention, however, not much has been known about how—and how widely—algorithms are being used in New Zealand's public sector. In October 2018, Internal Affairs and Stats NZ took a first step in answering such questions. The *Algorithm Assessment Report* documented 32 algorithms being used for a variety of purposes across 14 agencies, including ACC, Department of Corrections, Department of Internal Affairs, Ministry of Social Development and New Zealand Police (Stats NZ 2018).

As we discussed in Chapter 1, the *Algorithm Assessment Report* focused primarily on "operational algorithms", those which "impact significantly on individuals or groups" (Stats NZ 2018, p. 7). Those used for policy development and research were excluded from the report, as were what it referred to as "business rules", which it defined as "simple algorithms created by people that use rules to constrain or define a business activity" (a definition similar to the one we gave in Section 1B).

The *Algorithm Assessment Report* presents a fairly upbeat perspective on government algorithm use. The Executive Summary begins with the claim that "All of the algorithms considered in this review are embedded in policies that deliver clear public benefit". Reassurances are given throughout about, for example, transparency and human review of decisions. The *Algorithm Assessment Report* does, however, note a degree of inconsistency across the agencies it examined. It makes a number of proposals for better practice, some of which we consider in the final section of this report.

The algorithms described in the *Algorithm Assessment Report* range from the very well established to the very recent. An example of the former is RoC*RoI (which stands for "Risk of re-Conviction x Risk of re-Imprisonment"). RoC*RoI is a predictive model developed primarily by New Zealand Department of Justice senior psychologists in the mid 1990s (Bakker et al. 1999; DoC 2009). RoC*RoI uses a formula involving a number of static variables (see Appendix 2), that is, factors that are not possible or reasonably practicable for the person being assessed to alter, such as age at first offending.

RoC*RoI scores have "been included in pre-sentence reports provided to judges at sentencing, and in reports to the Parole Board" and have played a part in allocating offenders to sentence management categories, which in turn determine eligibility for rehabilitation programmes and other services (DoC 2009, p. 14)

RoC*RoI is a relatively simple algorithm, using well-established statistical techniques—in particular, logistic regression models with a limited number of input variables. Nonetheless, the decisions that it informs could hardly be more important. Indeed, the application of predictive algorithms to matters of criminal justice are one of the most controversial aspects of their use.

According to the *Algorithm Assessment Report*, the Department of Corrections and Police "use algorithms less extensively [than other agencies], supporting their frontline staff to make decisions in certain specific circumstances as opposed to informing the majority of interactions" (Stats NZ 2018, p. 30). Other predictive tools have, however, been developed and used within the New Zealand corrections system, which are not listed in the *Algorithm Assessment Report*. These include the Automated Sexual Recidivism Scale (ASRS), a "validated actuarial measure of sex offender risk based on New Zealand data" (DoC 2009, p. 19). The ASRS "estimates the probability of sexual recidivism using electronically accessible static factors: enabling the identification of a subgroup with a significantly higher-than-average sexual recidivism rate" (Wilson 2013, p. 2). The recently revised version is known as the ASRS.

Even more recent assessment tools include the Dynamic Risk Assessment for Offender Re-entry (DRAOR), which is used by probation staff to measure dynamic (changeable) factors relevant to offender risk. Unlike the RoC*RoI and ASRS, the DRAOR is not automated, but rather, is scored manually. An interesting question for any regulatory or oversight system for "predictive algorithms" is whether it should apply only to automated tools, and not to manual tools like DRAOR. Possibilities of arbitrary distinctions and perverse incentives would need to be kept in mind, and we return to these in the final part of the report.

The *Algorithm Assessment Report* is not only concerned with algorithms in the criminal justice sector. In July 2018, the Accident Compensation Corporation announced the introduction of an automated claim system in which "ACC system uses statistical models and a rules engine to automate much of its current, manual, registration cover process" (ACC 2018b).

The new system is intended to automate and expedite the processing of the large majority of compensation claims—around 90% according to ACC—that it regards as straightforward. Claims that are more complex or sensitive will still be referred to staff members. The system relies upon two models, a Cover Decision Service and an Accident Description Service. The Cover Decision Service—the part of the system that determines whether a claim can be accepted without manual review—"uses two statistical models that work in tandem". The Probability of Accept model "predicts the likelihood that a claim would be approved, based on historical data", while the Case Complexity model—as the name would imply—predicts the complexity of a case. Both models use data such as injury diagnosis, claimant age and earner status (ACC 2018b, p. 7). Together, these models will produce a decision either to "auto-accept" the claim, or to refer it for manual review. The Accident Description Service classifies the text entered into the "free text" fields of an ACC claim form into one of a number of preset categories. These categories are not used in cover decisions: they are used for various descriptive purposes, such as injury prevention initiatives, and summary statistics, as well as for actuarial purposes.

The automated process uses logistic regression. ACC have explained that this was regarded as best meeting key requirements of transparency and flexibility. Regarding the latter, they explain that this will allow them to adjust inputs "in response to external or policy changes" (ACC 2018b, p. 9). ACC have repeatedly stressed that the automated system can only accept or refer claims: no claims will be declined on the basis of the automated process. They have also insisted that ethnicity and gender have been specifically excluded from the input variables. Nonetheless, the new system has attracted considerable criticism (Maude 2018).

B. The use of algorithms in the criminal justice system

The Government's *Algorithm Assessment Report* listed a fairly small number of algorithms being used in the New Zealand criminal justice system. We have already discussed three of these: YORST, RoC*RoI and ASRS-R. An international perspective, however, shows a much wider range of uses right across the criminal justice system: from policing to decisions about sentencing, parole and post-sentence detention.

The criminal justice system has probably proved to be the most controversial use of algorithms to date. A recent report from the UK human rights campaign group Liberty claimed that "predictive policing programmes entrench pre-existing inequalities while being disguised as cost-effective innovations in time of austerity—and their use puts our rights at risk" (Couchman 2019). It should be noted that some of the opposition to the use of such techniques is rooted in hostility to the idea of "predictive justice" more generally. Jamie Susskind (2018), for example, argues that there is "something philosophically problematic about restricting people's freedom on the basis of predictions about their future conduct".

In this section, we describe a range of these systems. None of them is currently in use in New Zealand, and we have no particular reason to believe that systems like these are being actively considered here. Nonetheless, the rapidity of their uptake in the USA, UK and other jurisdictions makes them relevant to New Zealand policy-makers.

There are four key junctures in the criminal justice system where predictive algorithms have already been deployed or at least trialled:

- Predictive policing (PredPol);
- Crime detection (VALCRI);
- Prosecution decisions (HART); and
- Post-conviction decisions (including sentencing, parole and post-sentence detention) (COMPAS).

We will discuss a number of these in turn.

PredPol

While we may think of the criminal justice system primarily in terms of detecting and responding to crime, algorithms are also being used to inform decisions "upstream" from actual offending. "Predictive policing" has been defined in an influential report as:

> "the application of analytical techniques—particularly quantitative techniques—to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions." (PERRY ET AL. 2013)

As the report makes clear, the use of statistical approaches to forecasting crime is a long-established practice. (Statistical methods have a long history of use right across government, as emphasised in Section 1B.) What is most novel in the new generation of crime forecasting tools is the size and variety of the datasets that are consulted. It is this novelty which is of primary interest for the purposes of our report.

Predictive policing can operate in a range of settings. Marion Oswald and colleagues (2018) list three different contexts in which algorithmic data can be used to assist or inform policing:

(i) predictive policing on a macro level incorporating strategic planning, prioritisation and forecasting;

(ii) operational intelligence linking and evaluation which may include, for instance, crime reduction activities; and

(iii) decision-making or risk-assessments relating to individuals.

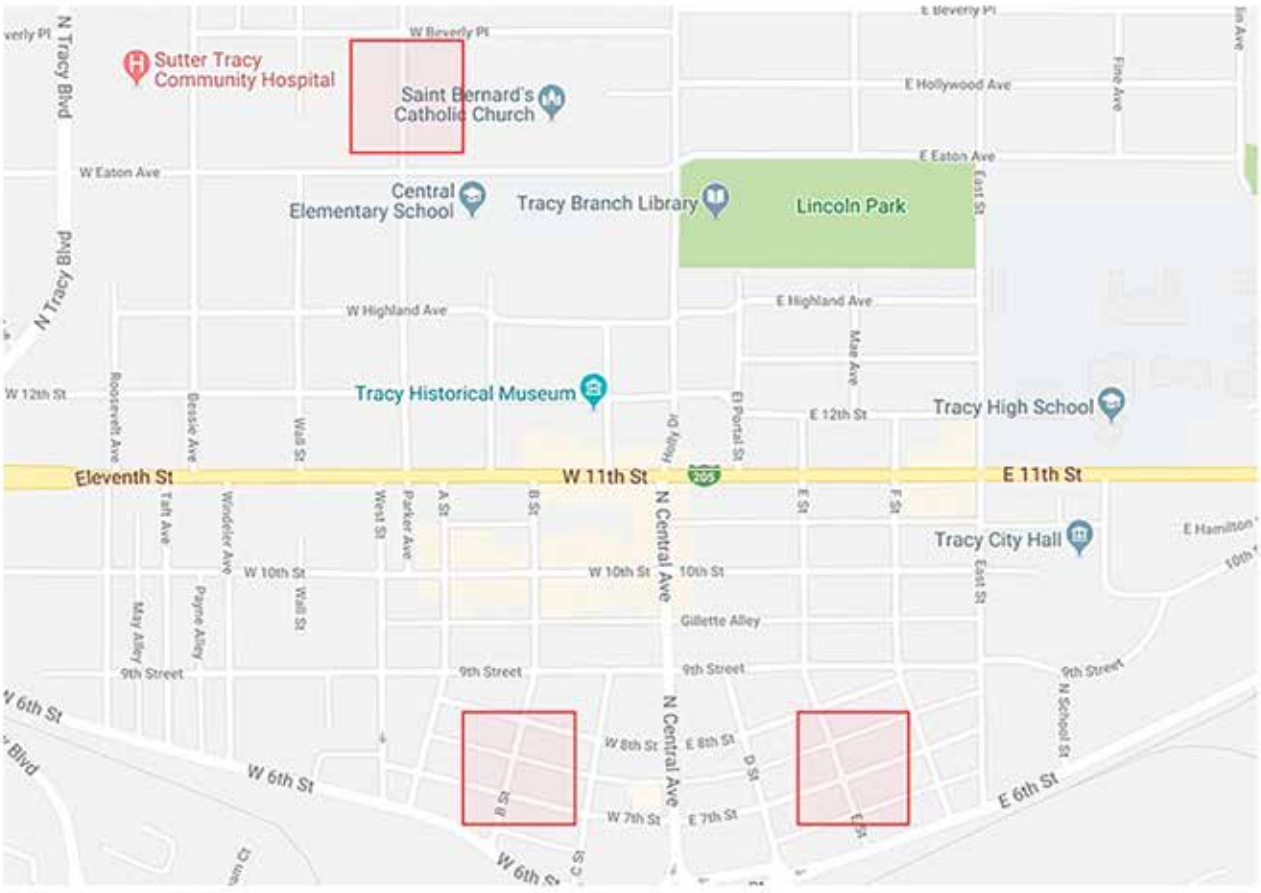
Perry et al.'s (2013) report offers its own taxonomy of potential applications:

> ***Methods for predicting crimes:*** these are approaches used to forecast places and times with an increased risk of crime.
>
> ***Methods for predicting offenders:*** these approaches identify individuals at risk of offending in the future.
>
> ***Methods for predicting perpetrators' identities:*** these techniques are used to create profiles that accurately match likely offenders with specific past crimes.
>
> ***Methods for predicting victims of crimes:*** Similar to those methods that focus on offenders, crime locations, and times of heightened risk, these approaches are used to identify groups or, in some cases, individuals who are likely to become victims of crime.

Probably the best known company name in predictive policing to date is PredPol. Incorporated in 2012, the company's origins go back to 2006, and a collaboration between the Los Angeles Police Department and a group of criminologists and mathematicians at the University of California, San Diego. Their objective was to try to use historical crime data to track crime "hotspots"—understanding and predicting their appearance and recurrence. Adopting an approach previously used in analysing earthquakes, the project's objective was "identifying the times and locations where specific crimes are most likely to occur, then patrolling those areas to prevent those crimes from occurring".

Predpol, it has been claimed, is now used by used by more than 60 police departments around the USA (Rieland 2019). Until November 2018, it was also used by Kent Constabulary in England. (Chowdhury 2018).

Using PredPol, high risk crime "hot-spots" are shown displayed as red boxes in a Google Maps window. Each box represents an area 150 square metres.



PredPol claims to use only three data points in its predictions: crime type, crime location, and crime date/time. Since "[n]o personally identifiable information is ever used", this, they claim, "eliminates the possibility for privacy or civil rights violations seen with other intelligence-led or predictive policing models".

This has not, however, spared Predpol—and other predictive policing technologies—from criticism. A joint statement by American Civil Liberties Union, NAACP and 14 other civil rights and related organisations identified a number of concerns about the use of such technologies—in particular, in relation to bias:

> "Predictive policing tools threaten to provide a misleading and undeserved imprimatur of impartiality for an institution that desperately needs fundamental change. Systems that are engineered to support the status quo have no place in American policing. The data driving predictive enforcement activities — such as the location and timing of previously reported crimes, or patterns of community- and officer-initiated 911 calls—is profoundly limited and biased." (ACLU ET AL. 2016)

The Brennan Center for Justice has expressed concerns about the self-fulfilling nature of the prophecies generated by such systems, claiming that:

> "most of these tools rely on historical policing data to generate their predictions; in the absence of meaningful oversight and transparency, the software may instead simply recreate and obscure the origins of racially biased policing."
> (BRENNAN CENTER 2018)

And in a similar vein, Kristian Lum and William Isaac (2016) have warned that:

> "Predictive policing software is designed to learn and reproduce patterns in data, but if biased data is used to train these predictive models, the models will reproduce and in some cases amplify those same biases. At best, this renders the predictive models ineffective. At worst, it results in discriminatory policing."

They too are concerned that Predpol and similar initiatives risk generating self-fulfilling prophecies:

> "Because these predictions are likely to over-represent areas that were already known to police, officers become increasingly likely to patrol these same areas and observe new criminal acts that confirm their prior beliefs regarding the distributions of criminal activity. The newly observed criminal acts that police document as a result of these targeted patrols then feed into the predictive policing algorithm on subsequent days, generating increasingly biased predictions. This creates a feedback loop where the model becomes increasingly confident that the locations most likely to experience further criminal activity are exactly the locations they had previously believed to be high in crime: selection bias meets confirmation bias."

Recent mathematical modelling by Ensign et al. (2018) appears to substantiate this risk:

> "Since such discovered incidents only occur in neighborhoods that police have been sent to by the predictive policing algorithm itself, there is the potential for this sampling bias to be compounded, causing a runaway feedback loop."

It is possible to imagine steps to mitigate against this kind of risk—for example, by only recording *reported* incidents of crime and excluding those *discovered* by police officers despatched in response to the prediction. Ensign et al. (2018, p. 11) suggest a possible solution "to counteract runaway feedback in predictive policing by appropriately filtering the inputs fed to the system". Such steps should be seriously considered by any law enforcement agency considering the deployment of these tools.

Another suggested risk is of displacement: a heavy police presence might reduce the rate of crime in an identified "hot spot" at the expense of "pushing" it elsewhere. Opportunistic offenders, it is thought, might target their efforts at less heavily policed areas, leaving the police engaged in a game of "whack-a-mole". Yet another concern relates to the potential erosion of relations between police and communities identified as crime "hot-spots". The Liberty report warns that

> "Focusing on abstract data, isolated from its human context, is at the expense of building proper community trust and understanding. This is particularly important in relation to over-policed communities."

The Brennan Center also raised concerns about the lack of transparency and accountability around the use of such systems. In June 2016, the Center filed a Freedom of Information Law request, seeking information about the use of predictive policing software by the New York Police department, "in the interest of better understanding and informing the public about the use of these systems". The request sought, among other things, communications between the NYPD and private developers of the software (predominantly in this case Palantir), details of inputs and outputs of the software, and records of its testing and utilisation. (The Center originally sought access to the "algorithm and code" used, but later narrowed its request.)

The NYPD refused to provide all of the information requested, claiming that it "would reveal non-routine techniques and procedures", and also violate nondisclosure agreements between NYPD and vendors bidding for contracts to supply predictive software products.

The Brennan Center filed suit. On 27 December 2017, the New York State Supreme Court found substantially in favour of the petitioner. It noted that the relevant statute places the burden of proof firmly with the respondent to demonstrate that the information sought fell within statutory disclosure exemptions. The NYPD had failed to discharge this burden, and therefore had to provide email correspondence with vendors and output data. The nature of the Center's changed request meant that the NYPD had not had sufficient time to respond to that part of the request, and the Court therefore rejected that part of the Center's case; that will need to be the subject of a new request. Finally, a decision about the trials of various vendors' products was to be conducted in camera, to allow the Court to determine whether they fell within permitted disclosure exceptions.

HART

Algorithmically-informed deployment decisions have proved controversial, but probably even more so is Oswald et al.'s (2018) third category: "decision-making or risk-assessments relating to individuals". This may arise in the context of predictive policing—granular assessments attempting to identify individual offenders. More typically, though, it will occur once an individual has come to the attention of law enforcement officials.

Some of these tools are used to inform decisions about prosecution. Durham Constabulary are currently involved in a trial of a Harm Assessment Risk Tool ("HART"). This is intended

> "to aid decision-making by custody officers when assessing the risk of future offending and to enable those arrestees forecast as moderate risk to be eligible for the Constabulary's Checkpoint programme."

namely, an "'out of court disposal' … aimed at reducing future offending" (Oswald et al. 2018). (In New Zealand terms, the Checkpoint programme is what we would refer to as "diversion".)

The aim of the HART tool is to identify those who present "an appropriate risk of offending" for inclusion on the Checkpoint programme. "Only those forecasted as Moderate Risk—who are expected to offend, but not in a seriously violent manner—are permitted into Checkpoint". At present, this assessment is conducted shortly after arrest by a custody officer.

The HART model uses a random forest technique (see Section 1B). It

> "is built using approximately 104,000 custody events over a five year period (2008-2012). It uses 34 different predictors to arrive at a forecast, most of which focus upon the prior offender's history of criminal behaviour. The random forest is constructed from 509 separate classification and regression decision trees (CART), which are then combined into the full forecasting model."

The use of the HART tool has been supported by ALGOCARE, a "decision-making guidance framework" (Oswald et al. 2018). Standing for Advisory-Lawful-Granularity-Ownership-Challengeable-Accuracy-Responsible-Explainable, Oswald et al. (2018) are fairly modest in their claims for the framework

> We appreciate that this framework does not provide any firm answers, nor do we claim to have covered every issue that may be relevant to the deployment of an algorithmic tool within policing or the wider public sector. In taking the first cautious steps into the use of algorithmic tools, Durham Constabulary is essentially engaging in experimental research, with the resultant requirement for ongoing testing and validation that such research entails.... Careful consideration of the factors set out in Algo-care should assist in reducing those uncertainties." (2018, P. 27)

This cannot be a once-and-for-all assessment however, as future impact is often uncertain, thus supporting our parallel proposal for new procedures to keep the proportionality of these technologies under review.

In April 2018, the Durham initiative came under scrutiny when civil liberties and privacy group Big Brother Watch revealed that Durham Constabulary "paid global data broker Experian for UK postcode stereotypes built on 850 million pieces of information". The group claimed that

> Experian's "Mosaic" links names to stereotypes: for example, people called "Stacey" are likely to fall under "Families with Needs" who receive "a range of benefits"; "Abdi" and "Asha" are "Crowded Kaleidoscope" described as "multi-cultural" families likely to live in "cramped" and "overcrowded flats"; whilst "Terrence" and "Denise" are "Low Income Workers" who have "few qualifications" and are "heavy TV viewers".

Silkie Carlo, Director of Big Brother Watch, argued that "for police to feed these crude and offensive profiles through artificial intelligence to make decisions on freedom and justice in the UK is truly dystopian". In a similar vein, the Liberty report warned that

> Running this data through individual risk assessment programs inevitably encourages a discriminatory and offensive association between factors such as family circumstances, income, class and the propensity to commit crime."

COMPAS

Of all the uses of algorithmic tools in criminal justice, perhaps the most scrutinised and debated has been COMPAS. First developed in 1998 by the company Northpointe (now Equivant), COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) was described by Northpointe as a "risk and needs assessment instrument", used by criminal justice agencies all across the USA "to inform decisions regarding the placement, supervision and case management of offenders" (Northpointe 2015).

As Northpointe explain:

> COMPAS has two primary risk models: General Recidivism Risk and Violent Recidivism Risk. COMPAS has scales that measure both dynamic risk (criminogenic factors) and static risk (historical factors)."

The exact models used by COMPAS are a trade secret. Essentially, Northpointe has only disclosed the types of machine learning technique that are used:

> the COMPAS risk and classification models use logistic regression, survival analysis, and bootstrap classification methods in a broad repertoire of prediction and classification procedures." (BRENNAN ET AL. 2009)

In recent years, COMPAS has been at the centre of two high profile controversies. In 2016, Eric Loomis brought a legal challenge against the use of COMPAS in the determination of his sentence (*Wisconsin v Loomis* 881 N.W.2d 749 (Wis. 2016)). Loomis argued that the use of the algorithm violated his due process rights. His case consisted of various strands, but of particular interest for our purposes are certain arguments directed at the use of the COMPAS algorithm. The first strand of his case was based on the absence of transparency around the algorithm:

> "Northpointe, Inc., the developer of COMPAS, considers COMPAS a proprietary instrument and a trade secret. Accordingly, it does not disclose how the risk scores are determined or how the factors are weighed. Loomis asserts that because COMPAS does not disclose this information, he has been denied information which the circuit court considered at sentencing."

Without access to this information, Loomis and his team could not verify or challenge the accuracy of the information used by the sentencing court.

Second, Loomis argued that

> "a circuit court's consideration of a COMPAS risk assessment amounts to sentencing based on group data, rather than an individualized sentence based on the charges and the unique character of the defendant."

Third, he argued that

> "because COMPAS risk scores take gender into account, a circuit court's consideration of a COMPAS risk assessment violates a defendant's due process right not to be sentenced on the basis of gender."

The Wisconsin Supreme Court rejected each of Loomis's arguments. With regard to the issue of transparency, it held that

> "Although Loomis cannot review and challenge how the COMPAS algorithm calculates risk, he can at least review and challenge the resulting risk scores set forth in the report attached to the [pre-sentence investigation]."

Regarding the issue of individualised sentencing, the Court was reassured by the fact that the COMPAS risk assessment was not the determinative factor considered at sentencing, and that human discretion would remain an important aspect: "we expect that circuit courts will exercise discretion when assessing a COMPAS risk score with respect to each individual defendant".

The claim of gender discrimination too was rejected. The Court held that "there is a factual basis underlying COMPAS's use of gender in calculating risk scores", to the extent that "any risk assessment tool which fails to differentiate between men and women will misclassify both genders". And "if the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose".

Loomis's challenge, then, was unsuccessful. The Court, however, expressed certain concerns about the use of COMPAS in sentencing decisions. It noted, for example, that

> "there is concern that risk assessment tools may disproportionately classify minority offenders as higher risk, often due to factors that may be outside their control, such as familial background and education."

This reflects a concern that has been articulated elsewhere in the literature around COMPAS. Sara Wachter-Boettcher (2017) has warned that

> "many of these questions [that are used in the COMPAS assessment] focus on whether people in your family or social circle have ever been arrested. According to Northpointe, these factors correlate to a person's risk level. But in the United States, black people are incarcerated at six times the rate of white people—often because of historical biases in policing, from racial profiling to the dramatically more severe penalties for possession of crack compared with possession of cocaine (the same drug) throughout the 1980s and 1990s. So if you're black—no matter how lawfully you act and how careful you are—you're simply a lot more likely to know people who've been arrested." (2017, PP. 126-127)

The Wisconsin Supreme Court also stressed that a COMPAS risk assessment informs only one of the aims of sentencing, and would be a "poor fit" for determining length or severity of sentence overall.

The decision in Loomis reflects the legal and constitutional position in Wisconsin and the USA. It does not provide much clarity about the likely legal status of a tool like COMPAS were it to be used in New Zealand. It is interesting to note, though, what the Wisconsin Supreme Court required to be told to any sentencing court seeking to use COMPAS in a sentencing decision:

(i) the proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are to be determined;

(ii) risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed;

(iii) some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism; and

(iv) risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations.

Earlier that year, the COMPAS algorithm was at the centre of another high profile controversy. The independent journalism organisation, ProPublica, conducted a study which looked at "more than 10,000 criminal defendants in Broward County, Florida, and compared their predicted recidivism rates with the rate that actually occurred over a two-year period" (Larson et al. 2016). The study

> "found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly tagged as low risk."

ProPublica's claims were widely reported, and Northpointe was quick to reply, publishing a report in which they "strongly reject the conclusion that the COMPAS risk scales are racially biased against blacks" (Northpointe 2016). In particular, Northpointe's report claims that

> "ProPublica made several statistical and technical errors such as misspecified regression models, wrongly defined classification terms and measures of discrimination, and the incorrect interpretation and use of model errors."

In particular, Northpointe claimed that ProPublica "focused on classification statistics that did not take into account the different base rates of recidivism for blacks and whites". Considered against a backdrop of those different base rates, the algorithm's predictions "are equally accurate for blacks and whites". (ProPublica issued a point by point response, published later that month, affirming their earlier conclusions.)

The details of the dispute are fairly technical. For some commentators, though, what lies at its heart are contested concepts of fairness and bias. Wachter-Boettcher has described the situation like this:

> "At Northpointe, fairness was defined as parity in accuracy: the company tuned its model to ensure that people of different races who were assigned the same score also had the same recidivism rates....At first glance, that makes intuitive sense. But parity in accuracy is only one measure of fairness." (2017, P. 127)

Sam Corbett-Davis and colleagues (e.g. 2016; 2017; 2018) have written quite extensively on the issue of fairness in predictive algorithms. In their (2016) *Washington Post* article about the dispute, they addressed the question of whether the COMPAS scores are "fair":

> "Northpointe contends they are indeed fair because scores mean essentially the same thing regardless of the defendant's race. For example, among defendants who scored a seven on the COMPAS scale, 60 percent of white defendants reoffended, which is nearly identical to the 61 percent of black defendants who reoffended. But ProPublica points out that among defendants who ultimately did not reoffend, blacks were more than twice as likely as whites to be classified as medium or high risk (42 percent vs. 22 percent). Even though these defendants did not go on to commit a crime, they are nonetheless subjected to harsher treatment by the courts. ProPublica argues that a fair algorithm cannot make these serious errors more frequently for one race group than for another."

As the authors note, the problem is that "it's actually impossible for a risk score to satisfy both fairness criteria at the same time" (except in the special case where the rates of recidivism happen to be identical for whites and blacks: certainly not the case in the actual data). In an important sense, they argue, a departure from Northpointe's definition of fairness—where a risk score should mean the same for all defendants regardless of race—would be highly problematic. All the same, they warn of the need to acknowledge that in cases like these, classification errors might disproportionately affect certain parts of the population, and where this is the case, "we have an obligation to explore alternative policies". (For a more detailed discussion of these issues, see Section 4C.)

Corbett-Davis et al. (2016) also allude to a problem that was also acknowledged by the Wisconsin Supreme Court in *Loomis*:

> "Northpointe has refused to disclose the details of its proprietary algorithm, making it impossible to fully assess the extent to which it may be unfair, however inadvertently. That's understandable: Northpointe needs to protect its bottom line. But it raises questions about relying on for-profit companies to develop risk assessment tools."

The challenge of commercial sensitivity in the context of algorithmic transparency is one that re-emerges frequently throughout our research, and it is one to which we return later in this report.

In the first part of this chapter (Section 2A), we have discussed the current state of play—as best as it can be ascertained—regarding government use of predictive algorithms in New Zealand. While most government departments are using these techniques in some capacity or another, the *Algorithm Assessment Report* found little in the way of consistency in terms of procedures around their use. But it also offered reassurance about their use, pointing out that, for example, they are not being used to make adverse judgments about affected individuals, and that they are augmenting rather than replacing human judgment.

In the second part of this chapter (Section 2B), we have discussed some of the more contentious use cases from other jurisdictions. We have elected to focus on examples within the criminal justice sector. While this is certainly not the only area of state decision-making that can have significant impacts on lives (decisions about immigration status and custody of children are also, of course, highly important), decisions that can result in the derivation of liberty must rank as among the most significant.

We have shown that the use of predictive algorithms in the criminal justice context remains highly controversial. To some extent, this reflects the more general suspicion of "actuarial justice"—of locking people up on the basis of what we think they will do in future, rather than what they have done in the past. Insofar as this is the basis for the concern, we note that it is not one that is particular to the use of predictive algorithms. But as the examples we have discussed show, the increasing use of predictive algorithms in this context is raising a number of concerns that merit serious attention.

The case studies we have selected for discussion here are by no means an exhaustive list of the predictive algorithms used in the criminal justice sector. It may be that, as the cases that have attracted most attention and controversy, they are not even especially representative. Nonetheless, we think there is merit in considering the sorts of problems, challenges and suggested solutions that have arisen elsewhere, with a view to informing the options for New Zealand.

3. THE PUBLIC DEBATE AND POLITICAL CONTEXT

There is great enthusiasm in New Zealand and elsewhere for the potential of artificial intelligence to enhance commerce, government, and everyday life:

> "Artificial intelligence technology is the next frontier. Its impact has been compared with the invention of electricity and according to the World Economic Forum, [it is] an important component of the Fourth Industrial Revolution."
> (AI FORUM 2018, P. 12)

Such descriptions give the impression that sectors of society that do not wholeheartedly embrace AI risk falling behind and missing out on great advantages on offer. But what exactly are these advantages and what do they mean for the current and future provision of public services in New Zealand?

A. Novel aspects of today's analytics

New Zealand's recent *Algorithm Assessment Report* opens with the observation that:

> "Algorithms have an essential role in supporting the services that government provides to people in New Zealand and in delivering new, innovative, and well-targeted policies to achieve government aims." (STATS NZ 2018, P. 4)

This should not be surprising. Contrary to popular perceptions of government departments, New Zealand, like other modern democracies, is often highly innovative in its delivery of public services (Rashbrooke 2018, p. 105).

Statistics has long been an aid to good government. Indeed, it was invented as a means of enhancing government, making it more accurate, efficient, and fair. The field was originally known in English as "political arithmetic" with the modern term "statistics" coming from the Latin "statista" meaning statesman and the German "statistik" often interpreted as the science of the state. But while the statistical ideas underpinning much AI are not new (as discussed in Chapter 1), the computerised and robotic systems on which they are implemented as well as the commercial environments in which they are developed and deployed are certainly new. There are in fact four novel features characteristic of modern predictive algorithms used by governments in New Zealand and internationally.

Complexity and opacity

Recent advances in machine learning have allowed for the development of predictive tools that are much more complex than the actuarial tables and checkbox forms common for much of the history of modern government. While this has greatly enhanced the accuracy of machine learning tools, it also limits the ability of government workers making operational decisions to *understand* the tools they are working with. While the idea of a decision tree is simple, it is not possible for users to envisage the operation of a random forest model (see Section 1B) composed of decision trees with millions of nodes such as the HART model (see Section 2B). Similarly a user might understand that a deep learning algorithm with a neural net architecture is able to learn patterns in the data which it reflects in the results it delivers. Nonetheless, unlike the traditional checkbox form, deep learning systems are essentially "black boxes" from the point of view of those that use them (but see Section 4B).

Commercial sensitivity

The success of artificial intelligence in many fields has resulted in the commercial development of predictive algorithms for government use. COMPAS is produced by the private American firm Equivant (formerly Northpointe). Although we have some broad-brush information about the machine learning techniques used in COMPAS (see Section 2B), commercial sensitivity means that its workings are effectively opaque not just to prisoners, but also to the corrections agencies that use it. (The opacity here is not due to complexity, but rather the inability to scrutinise the system's code, and find out what algorithms it is actually running—the problem we call "accessibility" in Section 4B). Of course the protection of intellectual property is a normal part of the commercial world. Nonetheless, this commercial opacity has been problematic for data subjects appealing its use (e.g. in *Loomis v Wisconsin*, discussed in Section 2B) and increasingly for the company itself. If it was in a better position to disclose

the workings of its algorithm, it might be better able to respond to recent studies claiming to demonstrate that the results produced by company's 137 question data collection tool can be bettered by an extremely simple algorithm based only on the offender's age, sex, and prior convictions (Dressel & Faird 2018) or that COMPAS can be outperformed by the untutored judgements of groups of non-experts provided with only very basic information (Angelino et al. 2017).

Availability

At the same time as the statistical instruments are becoming more complex and less transparent in their operation, they are also becoming more available and easier to use. Some of the very newest and most powerful machine learning systems are actually available free, as open source—for instance, Google's TensorFlow system—and every system developed by the "OpenAI" project (openai.com). On the data side, like most countries, New Zealand is capturing an increasing amount of it. In an era in which we are used to complex, powerful (and often proprietary) algorithms based on massive datasets accessible through our smart phones (e.g. Google's "pagerank"), it will seem remiss for government agencies not to employ similar methods to enhance the accuracy and fairness of operational decisions. This promises the efficient, timely and accurate delivery of public services in a way that was not previously possible.

The exciting opportunities offered by predictive algorithms are also fuelling increasing concern that advances in data science will make governments ever more efficient at collecting a growing volume of data about their citizenry (Susskind 2018) and that this increase in surveillance is likely to fall unevenly (Eubanks 2017), with groups that are already over-represented in various social statistics experiencing greater and greater levels of scrutiny.

Automation

Technology has long been deployed for the automation of routine tasks. Artificial intelligence promises to relieve workers of cognitive drudgery, just as industrial automation spared human labour from much physical drudgery. There is considerable scope across government for speeding up the resolution of routine requests without human intervention. Internationally there is increasing development of robotic process automation for the delivery of government services. These systems join together machine learning, basic business rules, computer vision, speech recognition and natural language processing to automate transactional tasks via a human-like interface. Such virtual assistants are increasingly being used in unsupervised triage roles in governmental contexts such as primary health care and immigration.

The "principles for the safe and effective use of data and analytics" developed by New Zealand's Privacy Commissioner and the Government Chief Data Steward, recommend that government use of predictive algorithms include "retaining human oversight" on the grounds that "analytical processes are a tool to inform human decision-making and should never entirely replace human oversight" (2017). However, in response to this principle, the *Algorithm Assessment Report* (Stats NZ 2018, p. 29) notes that "As technology continues to evolve, this will continue to be an area where government agencies must balance the importance of human oversight with possible efficiencies in service delivery". Indeed, the ACC's "Cover Decision" algorithm mentioned earlier is already configured to make decisions to award ACC cover in simple cases, without any human intervention. (See Section 4A for discussion of the efficacy and desirability of human supervision of complex algorithms.)

Thus a variety of computational, social and commercial factors make the new generation of predictive tools worthy of assessment as a new, powerful and potentially very beneficial mechanism for the delivery of government services. A further justification for the current focus on these technologies is that they have become an essential feature of New Zealand's "social investment" approach to the delivery of public services.

B. Social investment

A recent working group report for the New Zealand Treasury describes social investment as an investment in the welfare of New Zealanders predicated upon the successful use of information and technology:

> "Social investment is about improving the lives of New Zealanders by applying rigorous and evidence-based investment practices to social services. It means using information and technology to identify those people for whom additional early investment will improve long term outcomes, better understanding their needs and what works for them, and then adjusting services accordingly. What is learnt through this process informs the next set of investment decisions." (SCOTT ET AL. 2017)

Proposed in 2011, it is a whole-of-government approach to tackling persistent social problems affecting New Zealand's most vulnerable citizens. In the introduction to Social Investment: A New Zealand Policy Experiment, Jonathan Boston and Derek Gill note that, while there is disagreement about its effects, the intention of those developing the new approach was a major paradigm shift in the operation of New Zealand's provision of social services including:

> "greater reliance on integrated data, information sharing, risk profiling, actuarial analyses, outcomes-based contracting and joined-up services together with new institutional arrangements, a stronger focus on prevention rather than cure, and a drive for enhanced accountability." (2017, P. 11)

At the heart of social investment then is the prediction of outcomes for individuals as well as the segmentation of the population (Destremau & Wilson 2017) into groups of individuals with specific needs able to be met by newly designed policy interventions. Long range prediction of the effects of policy interventions and operational decision-making is of course epistemically challenging, but these are exactly the sort of tasks at which predictive algorithms are most competitive with humans: they can take into account large numbers of variables, and identify complex relationships between these (see e.g. De Baets & Harvey 2018 for a review). As a result it is possible to chart an increase in the use of these tools as a result of the implementation of the social investment approach (Hughes 2017, p. 162).

New Zealand is not the only country to employ social investment as an instrument for policy-making but its approach is unique. Where European programmes have focused on investment in education and in wealth redistribution, New Zealand has instead developed a framework based on (1) client segmentation, (2) intervention innovation and (3) governance to drive institutional change, along with pioneering the use of forward liability as a measurement tool (Destremau & Wilson 2017, p. 63). Social investment in New Zealand has also undergone significant changes. In the initial operation of the approach, the Crown's long term fiscal liability came to be a very important metric at the heart of New Zealand's welfare system (Boston & Gill 2017, pp. 18). In 2015-2016, the initial model was revised. The target population was broadened beyond working-age beneficiaries to include other at risk groups such as at risk children. There was some softening of the focus on long term fiscal liability. At the same time, use of actuarial analysis was increasingly to be replaced with use of predictive algorithms. After New Zealand's 2017 election there was considerable speculation that the incoming Labour government would scrap social investment altogether (Coughlan 2018a). This has not happened, although there *has* been a significant change in focus. The acting Chief Executive of the Social Investment Agency, Dorothy Adams, describes the change as a move from a social investment approach to an investing for social wellbeing approach.

> "The agency is broadening its focus from highly targeted interventions to broader measures of improving general wellbeing. There is less of a focus on fiscal measurement and much less of an emphasis on big data." (COUGHLAN 2018B)

Does this then spell a wholesale move away from the use of predictive analytics in New Zealand government? The agency has signalled its intention to employ more staff trained in qualitative analysis to work alongside its current "quant-heavy" workforce. This move away from quantitative analysis does not so far appear to be reflected in the operation and forward planning of the government agencies surveyed in the *Algorithm Assessment Report*. One explanation for this discrepancy is that the signalled changes at the Social Investment Agency reflect a leavening of its use of predictive analytics in strategic policy-making decisions. The *Algorithm Assessment Report* is specifically directed towards operational, rather than strategic decision-making, a domain in which the use of predictive algorithms offers a plethora of benefits (discussed below).

This report is not an evaluation of New Zealand's social investment approach (past or present), but it is an evaluation of predictive algorithms as an important component of the philosophical and technical underpinnings of policy-making as well as the day-to-day provision of government services to New Zealanders. In the context of the current mixed model for policy-making signalled by the Social Investment Agency, this report aims to provide useful advice about the scope for which, and context in which, government use of predictive algorithms can provide the greatest benefit for New Zealand.

C. Benefits claimed for predictive tools: A preliminary discussion

A PricewaterhouseCoopers report (PwC 2017) recently estimated that artificial intelligence could contribute up to US$15.7 trillion to the global economy in 2030 of which US$6.6 trillion would likely come from increased productivity. So what exactly are the benefits that make this new group of technologies so exciting for economists and how might these translate into better public policy and better provision of public services? New Zealand's Privacy Commissioner and the Government Chief Data Steward require that collection and use of public data must deliver "clear public benefit". The *Algorithm Assessment Report* sets out seven ways in which current government use of predictive analytics meets this requirement. (Box 1.)

Current public benefit from the use of predictive algorithms

- improved efficiency, which reduces cost for the taxpayer (for example, operational algorithms used by Inland Revenue to administer the tax system)
- streamlining processes to reduce the burden on members of the public (for example, the algorithm that enables streamlined passport renewal used by the Department of Internal Affairs)
- proactively targeting specific support to an individual based on data (for example algorithms used by ACC to improve client outcomes)
- supporting decisions which may be taken under complex or challenging circumstances (for example, the victim history scorecard the Police use to understand the cumulative harm a victim is subjected to)
- protecting New Zealand from risks and threats while enabling growing volumes of travel and trade (for example Immigration New Zealand and Customs algorithms that screen and assess passengers and goods at the border)
- providing empirical assessment to support a decision that identifies individuals who would benefit most from a new intervention or policy (for example, the NEET algorithm used by the Ministry of Social Development which uses a statistical predictive modelling tool to help identify those school leavers who may be at greater risk of long-term unemployment)
- providing assessment or forecasting to ensure policies are targeted properly and resourced adequately (for example the Social Housing Test Case developed by the Social Investment Agency).

Box 1. The Algorithm Assessment Report's list of ways in which predictive algorithms can deliver clear public benefit (Stats NZ 2018, pp. 26-27).

At a more abstract level, predictive algorithms in public decision-making have the potential to deliver five types of benefit: accuracy, objectivity, fairness, efficiency, and transparency.

Accuracy

Predictive algorithms have some clear advantages over human decision-makers in the accuracy of their judgements. They can take account of more input variables, and many more training examples, in a more systematic way (again, see De Baets & Harvey 2018 for a review). They can also disregard variables that aren't relevant, in ways that humans find hard to do. (This is well demonstrated in a study of stop-and-search decisions by New York police officers by Goel et al. (2016). Goel et al. showed that a statistical model of stop-and-search could find almost as many concealed weapons as human police officers, while searching far fewer people—and incidentally, being less biased against blacks and Hispanics.) Tools such as RoC*RoI implement a systematic analysis of decades of data. Although human employees can be made aware of relevant research, they will inevitably be less accurate at calculating probabilities where multiple factors affect a decision such as the likelihood that some individual will commit a crime. Learning algorithms can also *update* the functions they learn more readily than human reasoners when new data becomes available or new constraints on decision-making are imposed. But beyond these advantages, assessing accuracy remains a significant challenge.

In many contexts the efficacy of a decision-making procedure depends not just on now many errors it makes but also on what *type* of errors are most common. As discussed in Section 1B, to evaluate the performance of a binary classifier, data scientists often use a confusion matrix, which details the proportions of true positives, true negatives, false positives and false negatives in its decisions. As we noted earlier, if a system's performance isn't perfect, there will always be tradeoffs between false positives and false negatives (and true positives and true negatives). In different domains, different tradeoffs might be appropriate. Employing a predictive algorithm successfully often requires the developer to define an appropriate performance criterion in the confusion matrix, and then tune the algorithm to optimise for this criterion. But this will require policymakers to translate objectives characterised in terms of costs and benefits, and rights and duties, into recommendations that can guide software developers—a feat often easier said than done.

Assessing the accuracy of both human and machine reasoning is further complicated by the fact that correlations that hold within a dataset taken as a whole may be much less reliable when the dataset is partitioned. For example, although the accuracy of the COMPAS algorithm was found to be comparable to that of human experts making judgements about criminal recidivism, it was discovered to be very inaccurate when making predictions limited to individuals in the same dataset with a history of violent offending (Larsen et al. 2016).

Finally, actual levels of accuracy are often not very high, either for human or algorithmic decision-makers: an AUC of 0.7 for algorithmic systems is deemed acceptable in many contexts (see Section 1B for discussion of the AUC evaluation metric). So it is important that those interpreting the outputs of predictive algorithms understand the approximate level of accuracy they can expect from the tools they use. Similarly, those designing interventions should have in mind the *likelihood* of error as well as the types of error typical of the predictive algorithms used in relevant operational decision-making.

Incidentally, the common assumption that policy interventions offering a benefit need not meet the same ethical standards as punitive measures is only sustainable if overall levels of accuracy are sufficiently high. The *Algorithm Assessment Report* appears to make such an assumption in stating that:

> "Almost all participating agencies use operational algorithms to inform human decision-making, rather than to automate significant decisions. Where decisions are automated, these usually relate to automatic approvals or opportunities for people. None of the participating agencies described a circumstance where a significant decision about an individual that was negative, or impacted entitlement, freedom or access to a service, was made automatically and without human oversight." (STATS NZ 2018, P. 29)

Even in cases where unsupervised decision-making only delivers benefits to data subjects, sufficiently low accuracy in the system can mean that significant numbers of people in need of a benefit will fail to receive it. While this could be addressed statistically by making the relevant algorithms err on the side of generosity, the expense of this strategy makes it infeasible for low levels of overall accuracy.

Objectivity

One reason for preferring the use of predictive analytics in operational decision-making is that such algorithms are scientifically validated tools and as such they are objective in a way that unaided human decision-makers tend not to be. Predictive algorithms employ large datasets of measures of relevant variables. Their outputs are calculated by well-understood statistical techniques. Hence we can validate the use of particular algorithms within specific contexts. So described, they appear to be maximally objective methods for making high-stakes decisions affecting New Zealanders. But there are several reasons for caution about this optimistic assessment.

The objectivity of science is a complex ideal, difficult to characterise in the abstract and difficult to achieve in many practical settings (Reiss & Sprenger 2017). Philosophical characterisations of objectivity tend to see objective decisions as those that are not influenced by the values, perspectives, and interests of the individuals and social groups that make them. This idea that science can somehow jettison the perspectives of individual scientists (Nagel 1986) seems particularly implausible in the case of predictive analytics. Good data science is replete with judgements that are both intuitive and evaluative. For example, to develop an algorithm to aid in the amelioration of homelessness, a data scientists will have to decide on the best way to characterise homelessness in order to measure it. These sort of "what will count as X?" questions will often be evaluative, resting on moral and political judgements—in this case, why it is that homelessness is wrong, or that people have a right to adequate housing. Perhaps, most importantly, the data sets on which all predictive algorithms are based, effectively aggregate the past decisions of large numbers of individuals. The arrest and conviction records at the back of an algorithm such as PredPol are not, and cannot be, evaluated according to the objectivity of the decisions made by those involved.

None of this is intended to flag any inherent failing in data science. Rather it is to note that the objectivity of science *on the ground* is often difficult to evaluate and always a matter of degree. Hence it is not desirable for those working with algorithmic tools to be given the simplistic impression that the "objectivity" of predictive algorithms gives them an inherent advantage over human decision-making. Nor, given the current state of the technology, would we want high-stakes decisions about the lives of citizens to be wholly probabilistic, with no room for human discretion to protect data subjects from the operation of inexorable law. Of course, the decision-making of government employees should be evenhanded and free from various types of bias (see Sections 4C and 4F), but it is neither possible, nor perhaps wholly desirable, that it be completely objective. After all, as Aristotle pointed out long ago, injustice consists in the *equal* treatment of unequals as much as the unequal treatment of equals.

Fairness

The fact that algorithms can make decisions without the need for human intervention can be a benefit when impartiality is seen as important. This is particularly relevant in decisions about the distribution of scarce resources. New Zealand's Clinical Prioritisation Access Criteria (CPAC) tool is a ranking system designed to prioritise elective surgery in a crowded health system. It ranks the treatment of individual patients based on clinically developed criteria. The *Algorithm Assessment Report* describes CPAC as providing "a more equitable and consistent way of national prioritisation" (p. 22). "Equitable" here refers to the fairness of using an algorithm that knows nothing about each patient except those facts that are directly relevant to the likely benefit they will receive from the elective procedure in question.

Use of algorithms in such contexts demonstrates commitment to the Aristotelian characterisation of justice as treating like cases alike. But it's hard to avoid value judgements of some kind when building a ranking system. As discussed in Section 1B, a typical ranking system computes a weighted sum over the various factors to be considered. If the weights are directly specified, they build in the designers' intuitions about the relative importance of the different factors. If the weights are learned, so they approximate a "gold standard" ranking produced by expert humans, they

are essentially building in the experts' intuitions about relative importance, in a similar way. This definitional issue about the nature of fairness is compounded by recent studies suggesting that public conceptions of fairness are surprisingly complex and differ significantly from person to person (Webb et al. 2018). (For a more detailed discussion of these issues, see Section 4C.)

Efficiency

The very high volume of operational decisions required by many arms of government make efficiency and cost extremely important. So, it is not surprising that many of the "public benefits" of algorithmic decision-making listed in the *Algorithm Assessment Report* are essentially efficiency gains. It should be noted that efficiency is not just about cost. Many of the algorithms used by the New Zealand government are specifically designed to allow employees to spend more time on difficult cases and less time on simple ones. They have the added advantage of enhancing the working conditions of employees by relieving them from dull "mechanical" decision-making. (Although human employees should probably continue to perform some of the "mechanical" decisions, to provide new training data for the algorithms that run alongside them.) Efficiencies from the use of predictive algorithms might be expected to both speed up and enhance the accuracy of important interactions between citizens and government. This of course depends on the extent to which efficiency gains are used to enhance service rather than to cut cost by decreasing the number of employees delivering a particular service.

Securing efficiency gains are particularly important in a country like New Zealand that has small government and low tax compared to other OECD nations. This is important both because New Zealand has relatively little to spend on the delivery of public services and because our low level of public spending has arguably exacerbated a number of intractable social issues such as our extraordinarily high incarceration rate, particularly for Māori.

Transparency

The YORST (see Section 1D) is a publicly available algorithm. Publications available on the Corrections New Zealand website show the data collection form (see Appendix 1) which makes plain all the variables on which the algorithm operates as well as the very simple calculation on which it is based. However, as noted above, many algorithms are simply too complex for fine detail about their operation to be useful to citizens concerned about their use. Making details of algorithms such as RoC*RoI public can be beneficial if algorithms are well-designed and difficult to "game". They only way that offenders can enhance their RoC*RoI score is to commit fewer offences. It would be difficult to secure this transparency benefit for more accurate but more complex deep learning algorithms given the general challenge of making them explainable (but see Sections 1B and 4B for some promising strategies).

To summarise: New Zealand has a history of carefully designing predictive algorithms that secure significant public benefit. Nonetheless, the benefits provided by these algorithms are complex and often difficult to evaluate. None of the goals listed above should be pursued in isolation (it is relatively simple to make an algorithm more accurate at the expense of making it too crude to be useful) and the various benefits that predictive algorithms offer sometimes compete with one another (see Section 4C for discussion of the tradeoff between accuracy and fairness).

Impact on Māori

A discussion of the benefits of predictive algorithms in New Zealand must include benefits as viewed from diverse perspectives. For example, we are aware that there appears to be very little articulation of the benefits of use of predictive algorithms by government from Māori perspectives, including how these benefits would be defined and assessed. The *Algorithm Assessment Report* found little, if any, evidence of consultation about algorithmic use with affected Māori. This is a significant gap which must be addressed and which we consider could usefully be done now, before algorithmic use becomes even more prevalent.

Likewise, the use of predictive algorithms in immigration processing may be significant for people in particular territories in the Pacific region with which New Zealand has special legal relationships (such as the Cook Islands, Niue and Tokelau) and to those and other Pasifika peoples in New Zealand. However, we are not aware of any consultations to gain their views on the benefits of predictive algorithms in areas which may affect them.

4. CONCERNS ARISING FROM THE USE OF PREDICTIVE ANALYTICS IN GOVERNMENT

In this report we focus predominantly on criminal justice algorithms to illustrate the challenges and opportunities of predictive analytics. Our reasons for doing so are simple: criminal justice algorithms showcase many of the most pressing concerns which AI tools pose. The recent Liberty report (see Section 2B) on the use of algorithmic tools by various UK police forces complains of criminal justice algorithms "entrenching pre-existing discrimination"; it mentions offensive profiling techniques, limited transparency, breaches of privacy, and automation complacency (Couchman 2019). Liberty worries that these otherwise innovative and creative tools embed "discriminatory approaches in the system while adding a 'neutral' technological veneer that affords false legitimacy". The same can be said of many predictive algorithms in use today. In this chapter, we provide a survey of these concerns.

Before doing so, we recall, as noted in the Introduction, that the government has a special relationship with, and related set of obligations to, Māori, the indigenous people of New Zealand. The *Algorithm Assessment Report* acknowledged that the government's commitment to a Treaty of Waitangi based partnership with Māori should be reflected in its practice, by embedding a te ao Māori perspective into the development and use of algorithms, including "reflecting the taonga status of data that relates to Māori". This chapter is largely a literature review and summary of discussions with overseas experts: it is not, and does not purport to be, any sort of summary of views and experiences of Māori or any other part of the New Zealand public. We do, however, emphasize the necessity of more research in this area.

The first four sections (Sections 4A-D) are important because they relate in obvious ways to the public sector's use of predictive analytics. The last two sections (Sections E-F) are less relevant to the public sector, but they are too important not to mention at all.

A. Control, improper delegation and fettering discretion

The danger of human operators devolving responsibility to machines and failing to detect cases where they fail has been recognised for many years. There is no reason to believe machine learning tools will be any different in this respect. The problem is that, as automation becomes smarter and cheaper, its operators have to assume an increasingly supervisory role (Meister 1999; Strauch 2018). In aviation, for example, the role of the pilot appears to have become easier, but a closer look reveals that the pilot's role has been transformed rather than simplified, with the pilot now performing a crucial monitoring function (Baxter et al. 2012; cf. Stanton 2015). Likewise in financial trading, "[t]he human trader's role is now largely one of setting strategies and monitoring their execution" (Baxter et al. 2012, p. 68). Scholars in the field known as "human factors"—a branch of industrial psychology—have known for many years that the shift from operator to supervisor affects the operator in profound ways, not all of them good as far as human safety is concerned (Bainbridge 1983).

Automation has a significant impact on situation awareness (Stanton 2016). This is perhaps most clearly illustrated in respect of autonomous vehicles. Many semiautonomous vehicles issue "takeover requests" to the human driver when they require the human driver to resume control. But drivers in "autopilot" mode respond much more slowly to these requests than to stimuli occurring when they have full control (Stanton 2015; Cunningham & Regan 2018; Banks et al. 2018a; Banks et al. 2018b). Instantaneously transitioning from low to high workload poses great difficulties for most people (Walker et al. 2015). Another conundrum is that as the quality of automation improves, and the human operator's role becomes progressively less demanding, the operator "starts to assume that the system is infallible, and so will no longer actively monitor what is happening, meaning they have become complacent… [T]he operator assumes that the system is reliable and therefore failure detection deteriorates" (Pazouki et al. 2018, p. 299).

Related to automation complacency is automation bias, occurring when human operators "trust the automated system so much that they ignore other sources of information, including their own senses" (Pazouki et al. 2018, p. 299). Both complacency and bias "describe a conscious or unconscious response of the human operator induced by overtrust in the proper function of an automated system" (Parasuraman & Manzey 2010, p. 406). Decades of research confirm that these problems are both pernicious and potentially intractable (Banks et al. 2018b; Cunningham & Regan 2018; Greenlee et al. 2018). Somewhat alarmingly, they seem to afflict experts as much as novices, and are largely resistant to training (Parasuraman & Manzey 2010). Their effects may also be

observed beyond the limits of human-machine systems. For instance, it is well known that police officers, judges and jurors frequently overestimate the importance of forensic evidence—the so-called "CSI effect" (Marks et al. 2017; see also Damaška 1997).

Machine learning algorithms that require human supervision run into many of the same problems that human factors researchers have pointed out for decades (Cummings 2004). As a recent French report into artificial intelligence notes, "it is far easier for a judge to follow the recommendations of an algorithm which presents a prisoner as a danger to society than to look at the details of the prisoner's record himself and ultimately decide to free him. It is easier for a police officer to follow a patrol route dictated by an algorithm than to object to it" (Villani et al. 2018, p. 124). And as the AI Now Institute remarks in a recent report of its own: "[w]hen [a] risk assessment [system] produces a high-risk score, that score changes the sentencing outcome and can remove probation from the menu of sentencing options the judge is willing to consider" (AI Now 2018, p. 13). The Institute's report also offers a sobering glimpse into just how long such systems can go without being properly vetted. A system in Washington D.C. first deployed in 2004 was in use for 14 years before it was successfully challenged in court proceedings, the authors of the report attributing this to the "long-held assumption that the system had been rigorously validated" (AI Now, p. 14).

In her book, *Automating Inequality*, Virginia Eubanks (2017) notes the complacency that high tech decision tools can induce in the social services sector. Pennsylvania's Allegheny County introduced child welfare protection software as part of its child abuse prevention strategy. The technology is supposed to assist caseworkers deciding whether to follow up calls placed with the County's child welfare hotline. In fact, however, Eubanks relates how caseworkers would be tempted to adjust their estimates of risk to align with the model's. The Allegheny tool which features prominently in Eubanks' book in fact has links to New Zealand: it was developed by Rhema Vaithianathan and colleagues at AUT (Vaithianathan et al. 2013), and was originally intended for use in New Zealand. Anne Tolley, the Minister of Social Development at the time, refused to allow the tool to be trialled in New Zealand; her statement that children should not be treated as "lab rats" became something of a trope in local discussions about government AI ethics.

In light of concerns about the true extent of human control over algorithmic tools, an increasing number of human factors experts have concluded that, except in certain special circumstances, algorithmic decision tools should not be used in high-stakes or safety-critical decisions unless the systems are significantly "better than human" (see e.g.: Banks et al. 2018b; Cunningham & Regan 2018; Walker et al. 2015; Cebon 2015; but see earlier discussion in Chapter 3 on the difficulties of defining "accuracy" in a policy context). Even though no technology is really 100% reliable, on this view it does not matter, because the dangers posed by human complacency diminish to nothing when a system exceeds human performance by a substantial margin. How many systems achieve this standard is another question. Currently, autonomous vehicles do not approach this level of capability (Banks et al. 2018a; Banks et al. 2018b), but many subcomponents within standard (nonautonomous) vehicles clearly do, such as automatic transmission, automatic light control and first generation cruise control (Walker et al. 2015).

In more typical decision support settings, arguably medical diagnostic and legal case prediction software is approaching this better-than-human standard. There are at present AI systems which can distinguish between lung cancers and give prognoses more accurately than human pathologists armed with the same information, and systems which can spot Alzheimer's with 80% accuracy up to a decade before the first appearance of symptoms, a feat vastly outperforming the ablest human pathologist attempting anything similar (Bridge 2017). In the legal sphere, advances in natural language processing and machine learning have facilitated the development of case prediction software that can predict, with an average 79% accuracy, the outcomes of cases before the European Court of Human Rights when fed the facts of the cases alone (Aletras 2016). Most impressively, a similar system had better luck in predicting the rulings of the U.S. Supreme Court than a group of 83 legal experts, of whom almost half had previously served as the justices' law clerks (60% vs. 75% accuracy) (Brynjolfsson & McAfee 2017). If the disparity between the performance of such systems and that of well-trained and experienced human professionals widens any further, presumably it will not much matter if humans perfunctorily adhere to whatever these systems decide or advise in a particular situation.

Building on these ideas, some researchers advocate that in the short to medium term an optimal approach to the problem involves a *complementary* (and potentially dynamic) coupling between highly proficient (better-than-human) algorithmic tools and human agents working alongside one another, where the tools themselves are so reliable that there is effectively no need for oversight. This has been called the "DCAF" approach (dynamic/complementary allocation of function) (Zerilli et al. 2019). Complementarity is its key feature. The basic idea is that some systems clearly need to replace the human agent and be left to operate autonomously. Human-machine decision systems that contain automated subcomponents work best when the human operator is allowed to concentrate their energies on the chunks of the task better suited to human rather than autonomous execution—a setup which only avoids the problems of complacency and bias if the automated subroutines are handled by systems approaching near-perfect (better-than-human) dependability. Complementarity means humans and machines have clearly defined and clearly separated roles, where the human is effectively barred from interfering with the machine's outputs and in many cases even knowing what the outputs are. At the same time, DCAF emphasises that the allocation of functions should be flexible enough to support *dynamic* interaction, with hand-over and hand-back for shared competencies (as occurs when a driver disengages cruise control and thereby resumes control of acceleration).

In New Zealand, the Accident Compensation Corporation has largely relied on manual control for processing claims. In the past this has involved ACC staff members sorting through and assessing individual claims one by one. Even with improvements to case handling procedures over the years, such as technology allowing electronic submission, *all* claims have required some degree of manual processing (ACC 2018a). As we discussed in Section 2A, the ACC has introduced an improved claim registration and cover assessment process. It aims to make the claims approval process quicker and more efficient, removing the need for manual control in standard cases altogether. The ACC hopes that by harnessing the power of big data—12 million claims submitted between 2010 and 2016—it can both reduce the wait time for approvals as well as more efficiently distribute the more complex claims to ACC teams for final determination. An analysis of publicly available information about the workings of this system indicate that it exemplifies many virtues of the DCAF approach (Zerilli et al. 2019).

Whether systems that do *not* reach better-than-human levels of reliability can be used in high-stakes settings is difficult to answer straightforwardly. The problems of automation complacency and bias do not arise from the use of patently *suboptimal automation*, only from *generally dependable* automation. Therefore, depending on the exact nature of the human-computer interaction at issue, a *less*-than-reliable system *might* safely replace a human agent. Suboptimal tools may prove useful in circumstances where the tools have access to information to which the human does not, or otherwise "decide" things in ways that humans generally cannot. Such systems very literally augment human capacities: human and machine in effect *share* control. The clearest examples of this form of technology are the predictive risk systems used in law enforcement and policing, such as PredPol (for hot-spot policing) and COMPAS (predicting the likelihood of offender recidivism) (see Chapter 2). These systems answer questions of the form: How should we distribute police officers over a locality having such and such geographical characteristics? What is the likelihood that this prisoner will reoffend if released on parole? And so on. Often they use logistic regression or more advanced actuarial techniques to mine patterns from very large databases (see Sections 1B and 2B). This is not a feat unaided humans can hope to match. There are also some phenomena within human decision-making that algorithms can help to counteract—for example, decision fatigue and decision inertia, of which some of the classic studies actually involve judges' parole decisions (e.g. Danziger et al. 2011).

Nevertheless, we would urge that great caution be exercised before any form of suboptimal automation is adopted in high-stakes/safety-critical settings. Many of these systems (like COMPAS) are after all tools which have attained notoriety for their problematic biases and inherent technical limitations (e.g. Blomberg et al. 2010; Larson et al. 2016; Dressel & Farid 2018). And as some of these systems gradually begin to overcome their limitations, our worry is that the control problem will gradually re-emerge, taking human operators unawares and decision subjects along with them. It will be all too easy for a judge with decision fatigue, for example, to simply rely on what a predictive risk instrument "objectively" recommends.

Improper delegation and fettering discretion

So far we have not considered purely *legal* reasons for managing the control problem. But particularly in the public sector, where statutory office holders and public authorities may be conferred wide discretions by their enabling legislation, it is vital to ensure that the risks of automation complacency and bias are judiciously managed. Administrative and public law principles in the UK, Australia and New Zealand jealously guard the repository of statutory discretion. A power conferred upon a minister of the Crown, for example, not only *must* be exercised, it must be exercised *by the minister*. Public law principles effectively prohibit the delegation of statutory powers to third parties without express or implied authorisation in the decision-maker's enabling legislation. Likewise, these principles inhibit the authorised decision-maker from "fettering" their discretion, for instance, by blindly following company "policy" or other organisational protocols. Crucially, improper delegation is not limited to situations where a third party is relied upon to make a decision—such as would be the case were the Department of Immigration to outsource refugee-status determinations to a private company in order to better regulate its work load. Improper delegation may also occur when the relevant decision-maker merely *appears* to have turned their mind to the decision but in fact has simply rubber-stamped the advice of others, without coming to an independent view of the matter for themselves.

It is here that administrative law will have something to say about the use of algorithmic tools. As Marion Oswald explains:

> "A public body whose staff come to rely *unthinkingly* upon an algorithmic result in the exercise of discretionary power could be illegally "fettering its discretion" to an internal "home-grown" algorithm, or be regarded as delegating decision-making illegally to an externally developed or externally run algorithm, or having pre-determined its decision by surrendering its judgment." (OSWALD 2018, P. 14)

If the dangers of the control problem are as pronounced as our research suggests, there will inevitably come a point when complacency amounts to that very *unthinking-rubber-stamping* attitude which falls foul of public law prohibitions.

A tool which does not induce this complacency will not be in breach of these rules: in such a case the tool will be used as it was intended, that is, as an *aid* to decision-making. In the common law, statutory decision-makers were always free to inform themselves, consult widely, and so forth, with a view to producing high quality decisions. The use of algorithmic decision support tools that are used strictly *as* support tools are therefore unobjectionable from the point of view of administrative law. The law only looks askance at the *abdication* of decision-making responsibility. Unfortunately, the control problem strongly suggests that abdication is an ever-present danger (indeed *reality*) once automated decision tools reach a certain level of reliability.

The basic point here is that if Parliament confers power upon a person, office or statutory authority to decide some matter within a particular jurisdiction, a delegation of this power to a third party will be legal only if the legislation itself contemplates such delegation, either expressly or by necessary implication.[3] Therefore, to ward off future legal challenges against the use of algorithmic tools that are at risk of inducing complacency, it may be necessary to obtain express statutory authorisation for the "delegation". Of course this authorisation should only be sought where it is safe to do so—that is, where the dangers of complacency no longer exist because the tools are better-than-human in the domain of interest. But in those public sector decision-making contexts in which *discretion* is truly required—because not every contingency can be anticipated in the ordinary course of day-to-day administration—it is at best unclear and at worst doubtful whether algorithms will indeed be able to offer superior decision-making capabilities to those of humans. Once again, caution is advised.

3. In New Zealand, the State Sector Act 1988, Crown Entities Act 2004 and Local Government Act 2002 contain general delegation provisions. These Acts obviate specific statutory authority to delegate (i.e. entity-specific provisions in the relevant entity's enabling legislation), although a delegation under them still needs to be in writing. In the Australian context, see e.g. section 495A(1) of the Migration Act 1958 (Cth).

B. Transparency and the right to explanations

Transparency encompasses a number of distinct concerns about algorithms, perhaps befitting the politically and philosophically complex notion that it is (Meijer 2014; Dubnick 2014). At the broadest level transparency refers to accountability or answerability, indicating a general responsiveness to requests for information or a willingness to offer justification for actions taken or contemplated. This is a political, civic sense of the term, and embodies a dynamic or ongoing state of affairs. We expect our elected representatives to act in the public interest, and transparency stands for their commitment to do so. When government is open, answerable and accountable to its citizens, it is less tempted to become insular, self-serving or corrupt. Transparency in this broad sense is therefore prophylactic—a safeguard against the abuse of power. While all democracies notionally value this sense of transparency (Roberts 2006; Forssbæck & Oxelheim 2014; Lombrozo 2011; Heald 2006; Prat 2006), it is abstract and aspirational. From here, the notion branches out in at least three directions, each of which takes the concept into much more specific and less abstract terrain.

In one direction, transparency may be associated with moral and legal responsibility. This captures such familiar notions as blameworthiness and liability for harm. Here the sense of transparency is static, i.e. "once-for-all" or "point-in-time" (e.g. "Mary is liable for her negligence, and must compensate Amanda to the tune of $6000"). Unlike the broader notion, this sense of transparency is not intended to be dynamic, and nor is it necessarily public-interested (as the civic sense clearly is). Rather than prospectively *preventing* wrongdoing, it is *corrective* (and retrospective).

In a second direction, transparency retains its dynamic quality but relates more explicitly to the inspectability (or auditability) of institutions, practices and instruments. Here transparency is about mechanisms: How does this or that tool actually *work*? How do its component parts fit together to produce outcomes like those it is designed to produce? Algorithms can be "inspected" in two ways. First, we can enquire of their provenance: How were they developed, by whom, and for what purpose/s? This extends to procurement practices: How were they acquired, who commissioned them, and on what terms? This might be called *process* transparency. Second, we can ask of any algorithm: How does it work, what data has it been trained on, and by what logic does it proceed? This might be called *technical* transparency, and centres on the notion of *explainability*. Before any *particular* decision is reached using an algorithm, we may seek *general* explanations (*"ex ante"*): for instance, in a machine learning case, we can ask whether we're dealing with a decision tree, a regression algorithm, or some mixture. Information about the kind of algorithm we're dealing with can tell us quite a lot about its general principles of operation, and whether Algorithm A is better than Algorithm B. In the wake of any *particular* algorithmic decision, however, the questions posed can be more specific: Why did the algorithm decide *this* matter in *this* particular way? This is to seek a specific, individualised explanation for a decision (*"ex post"*). In both cases, though, it is important to realise that explainability immediately raises the question of *intelligibility*: Can anyone actually *comprehend* the explanation? Intelligibility is a distinctive sense of technical transparency-cum-explainability.

In a third direction, transparency denotes accessibility. Meaningful explanations of an algorithm may be possible, but they may not be *available*. Intellectual property rights might prevent the disclosure of proprietary code, or preclude access to training data, so that even if it were possible to understand how an algorithm operated, a full reckoning may not be possible for economic, legal or political reasons. Algorithms that are otherwise technically transparent may therefore be "opaque" for nontechnical reasons. Figure 5 depicts these various nested and interacting notions of transparency diagrammatically.

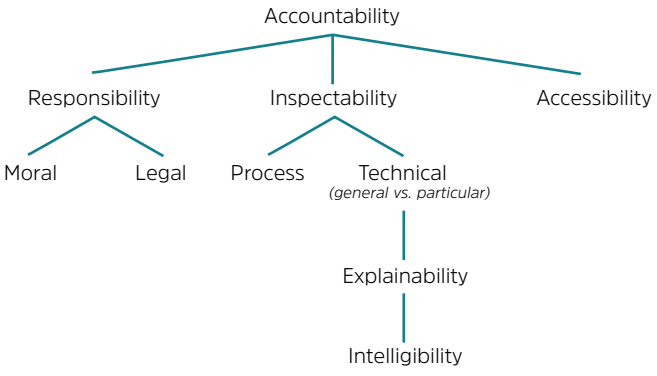


Figure 5. The various senses of transparency.

In the context of algorithms and machine learning, concerns have been raised about "transparency" in every one of these senses. The sense which most exercises policy specialists and a growing number of computer scientists, however, is technical transparency—specifically explainability and intelligibility (Miller 2017; Pasquale 2014; Edwards & Veale 2017). Here the worry is that because deep learning systems arrive at their decisions unaided, i.e. in a manner that is not specified in advance, it is not possible to interpret the system's internal processes except only approximately and imperfectly—and even this much is doubtful (Mittelstadt et al. 2016; Wachter et al. 2017a). As discussed in Section 1B, the neural networks that implement deep learning algorithms mimic the brain's own style of computation and learning: they take the form of large arrays of simple neuron-like units, densely interconnected by a very large number of plastic synapse-like links. During training, a deep learning system adjusts the weights of these links so as to improve its performance. If trained on a decision task, it essentially derives its own method of decision-making, much as we would expect of an intelligent system.

But there is the rub. In neural networks, these processes run independently of human control, so that transparency inevitably becomes an issue: it is simply not known in advance what computations will be used to handle unforeseen information.[4] Importantly, neither the operator nor the developer will be any the wiser in this respect. *Ex ante* predictions and *ex post* assessments of the system's operations alike will be difficult to formulate precisely. This is the crux of the complaint about the lack of transparency in today's algorithms. If we cannot ascertain exactly why a machine decides the way it does, upon what bases can its decisions be reviewed? Judges, administrators and departments of state can all supply reasons for their determinations. What sorts of "reasons" can we expect from an intelligent machine? Deep learning involves multiple hidden layers of processing that are fiendishly intricate and virtually impossible to unsnarl (Burrell 2016). Even certain simple algorithms which instantiate in the order of hundreds of rules "are very hard to inspect visually, especially when their predictions are combined probabilistically in complex ways" (Van Otterlo 2013).

On the other hand, some have worried that automated decision-making is being held to an unrealistically high standard (Zerilli et al. 2018). It is instructive to compare the kinds of explanation envisaged for predictive systems with those routinely provided by human agents. The latter do not yield the entrails of a decision, or "illuminat[e] the cognitive processes leading to…[a] conclusion", as the IEEE's Global Initiative on the Ethics of Autonomous and Intelligent Systems considers apt for algorithmic explanations (see p. 71 of version 2). It is true that human agents are able to furnish reasons for their decisions, but this is not the same as illuminating the cognitive processes leading to a conclusion. The cognitive processes underlying human choices, especially in areas in which a crucial element of intuition, personal impression and unarticulated hunches are driving much of the deliberation, are in fact far from transparent. Arenas of decision-making requiring, for example, assessment of the likelihood of recidivism, or the ability to repay a loan, more often than not involve significant reliance on factors beneath the level of conscious belief. As one researcher explains: "A large part of human decision making is based on the first few seconds and how much [the decision-makers] like the applicant. A well-dressed, well-groomed young individual has more chance than an unshaven, dishevelled bloke of obtaining a loan from a human credit checker" (Dutta 2017).

A large part of human-level opacity stems from the fact that human agents are also frequently *mistaken* about their real (internal) motivations and processing logic, a fact that is often obscured by the ability of human decision-makers to invent post hoc rationalisations. The upshot is that while full technical transparency may be unattainable, this may not be as much of a problem as it is sometimes made out to be: human agents, including judges and officials, are never expected to furnish low-level explanations for their decisions—descending to physical, biochemical or psychological explanations for their motivations and prejudices—so machines should not be either. Machines should offer explanations that are comparably situated vis-à-vis human reason-giving practices. This is at once a much less demanding standard of explainability than full technical transparency and one that computer scientists are in a much better position to meet.

4. For machine learning algorithms, it's useful to distinguish between the "learning algorithm" (e.g. a deep network) and the "learned algorithm" (which for a deep network is an impenetrable set of computations). We can readily specify the (second order) learning algorithm in great detail, but the (first order) learned algorithm is impenetrable.

Perhaps the most useful thing a decision subject wants to know is how different factors were weighed in coming to a final decision. It is common for human decision-makers to disclose these allocations, even if the inner processing logic leading to them remains obscure. Weights are classic exemplars of what has been called "intentional stance" logic—the sort of explanatory logic that human agents are disposed to offer for their decisions, cast in terms of beliefs and desires ("Mary *wants* to save money, so she is staying in tonight, because she *believes* that the restaurant will be set her back at least $100")—and one way for algorithmic decision tools to be held accountable in a manner consistent with human decision-makers is by having them divulge their weights (Montavon et al. 2017). As Edwards and Veale (2018) remark, "Extracting estimates of the weightings within a complex algorithm is increasingly possible, particularly if only the area 'local' to the query is being considered". This is because local terrain, "unlike the complex innards of the entire network, might display recognisable patterns" (Edwards & Veale 2018). It is therefore heartening to see the development of various model-agnostic explanation systems that provide pedagogical guidance, or "models-of-a-model" (Edwards & Veale 2017). (See Section 1B for discussion of explanation systems.)

An important question that all jurisdictions have had to consider is the availability and extent of a right to explanations for decisions reached partly or fully through automated means. What could make an otherwise straightforward issue a contentious one is that any such right necessarily accompanies a corollary "duty to give reasons", which some common law jurisdictions (such as the UK and Australia) do not recognise even for public officials exercising statutory functions (Aronson & Dyer 2013). The European Union's General Data Protection Regulation (GDPR), which entered into force in May of 2018, arguably does contain a general right to explanations (cf. Wachter et al. 2017b), at least for decisions reached *solely* through automated means, and this has inspired the hope that other jurisdictions lacking a clear right will follow suit.

Section 23 of New Zealand's Official Information Act 1982 has long provided citizens affected by government decisions with a "right of access…to reasons". We will discuss this provision further in Section 5A.

C. Algorithmic bias

Bias is a predisposition towards or against a particular thing, person or group, such as an ethnic group, social class, political party, religion or other demographic (such as an age group). While it implies being one-sided or closed-minded, many biases are simple heuristics and need not be pernicious (e.g. a bias against simplistic, ideologically-driven solutions to complex problems). When the bias in question is pernicious, unfounded, unreasonable or resistant to rational influence, it is *prejudice*. When prejudice is acted upon in such a way as to exclude, disadvantage or marginalise its subject, it is known as *discrimination*, which can be either *direct* (called "disparate treatment" in the USA), or *indirect* (called "disparate impact" in the USA).

Different disciplines have different takes on "bias". Statisticians have their own uses of the word, a fairly common one of which denotes "selection bias". This bias arises from over- or under-sampling members of particular groups. Thus a face recognition system trained on a data set that over-represents a particular racial group is going to have trouble recognising the faces of members from any of the under-represented groups. As this example shows, however, selection bias frequently intersects with more common (non-statistical) notions of discrimination and injustice.

A further complexity is that there are diverse *legal* meanings of terms such as "bias" and "discrimination". These concepts and definitions span administrative, human rights and other laws. For example, concepts of bias in administrative law (including judicial review) may relate to either, or both, the factors taken into account in decision-making (such as bias resulting in the taking into account of irrelevant factors or a failure to take into account relevant ones) and bias in the mind of the decision-maker (such as a known conflict of interest which the decision-maker has not declared or a prejudice against a particular party or pre-determined view on a particular topic which is the subject of the legal dispute). Depending on the type of bias and the particular form of decision-making, remedies may vary. For example, in administrative law, bias may result in a finding that a decision was unreasonable as a matter of law. Where this is due to a procedural flaw (for example, failure to take into account relevant information) the court might order the decision be made again so as to remedy the procedural flaw (for example, adducing

additional evidence or precluding the presentation of evidence that was previously taken into account). Alternatively, where the decision-maker is found to have acted with bias, the decision may be overturned and a fresh hearing ordered by the same or another decision-maker. In either case, if the bias is not found to have tainted the substantive decision, the original decision may be affirmed.

In New Zealand human rights law, concepts of equality, inequity, bias and discrimination are different again from those familiar in administrative law. Even within human rights law, these concepts and definitions also vary depending on whether the actors are government or private sector (for example the tests for discrimination under Parts 1A and Part 2 of the Human Rights Act 1993).

In addition to the complexity of existing concepts and definitions, both legal theory and general jurisprudence continue to evolve. The result is that concepts of bias and equality may change and improve over time. For example, in the early 1990s critical race feminist theorists developed concepts of intersectionality to identify and explain how race, gender, sexual orientation, disability and other forms of identity can intersect in the face of disadvantageous treatment to result in experiences of multiple forms of discrimination, rather than discrimination on one prohibited ground alone (Crenshaw 1993).

We outline these issues briefly in order to note that while detailed analysis of concepts and definitions of bias in the context of administrative and human rights law have not been the focus of this project, it is clear that much more work is needed to understand how these concepts and definitions sit alongside the other issues of bias that we have identified. This is especially important if the remedies or measures to address bias in each sphere are to be assessed against each other.

Some scholars distinguish further between two types of bias: intrinsic and extrinsic (Zerilli et al. 2018). Intrinsic bias resides in a system by virtue of its design, structure and rules of operation, or as a consequence of inputs effecting a permanent change in these features. Racial bias is a good example of intrinsic bias in human beings, because the connection with emotion is relatively clear (the emotion being fear) and emotion is a constitutive feature of personality (Pohl 2008; Angie et al. 2011). Furthermore, racist conditioning may affect long-term the way a person processes information and makes decisions. Extrinsic bias, on the other hand, derives from a system's inputs when they do *not* produce a permanent or long-term change in the system's internal structure and rules of operation. In these cases, false information may affect a system's outputs, but so long as the information is corrected, the outputs will be unbiased.

Overall, while it is true that an algorithm can be intrinsically biased (see below), extrinsic bias is probably the more immediate problem for AI (Friedman & Nissenbaum 1996; Johnson 2006). The so-called "dirty data" problem is an apt illustration. Errors and biases latent in data training sets tend to be reproduced in the outputs of machine learning tools (Barocas & Selbst 2015; Diakopoulos 2015). Moreover, because humans label much of the training data by hand, biases often creep in and taint algorithms through data labelling. The problem is compounded by copyright and intellectual property laws, which presently limit the access users have to better quality training data (Levendowski 2017).[5] Still, extrinsic bias in principle is less difficult to overcome than intrinsic bias. Most of the problems we have mentioned arise from the use of unrepresentative data sets. For instance, face recognition systems trained predominantly on Caucasian faces might reject the passport application photos of Asian persons, whose eyes appear closed (Griffiths 2016). Speech recognition systems, too, are notorious for being less accurate when processing female voices than male ones (Tatman 2016). Both situations arise from a failure to include representative members from all social groups in training data (an example of what we described earlier as selection bias). The obvious solution is to diversify the training sets (Klingele 2015; Crawford & Calo 2016). While there are political and legal barriers in the way of this, it is not as intractable a problem as the one posed by intrinsic human bias (Bezrukova et al. 2016; Plous 2003; Allport 1954).

Not all dirty data suffers from being unrepresentative in this way, however. For instance, a machine learning tool that disproportionately classifies African Americans as posing a greater risk of recidivism has probably learnt from a data set that reflects racial prejudices inherent in previous discriminatory patterns of policing (Larson et al. 2016; Lum & Isaac 2016; Crawford & Calo 2016). This would not count as intrinsic bias because the data do not affect the system's internal structure and rules of operation. But nor can such bias be said to originate from unrepresentative data, which can in theory be corrected by including more diverse ethnic groups in the training set.

5. Other factors impeding access include privacy law and income disparities.

There is another side to the story too. It seems that fairer algorithms are not possible that satisfy any more than one definition of fairness at a time, because "many notions of fairness are in conflict" (Corbett-Davies et al. 2017, p. 799; see also Chouldechova 2017; Hardt et al. 2016; Kleinberg et al. 2017). Three broad types of fairness definitions have been discussed.

(i) *Anti-classification* "stipulates that risk assessment algorithms not consider protected characteristics—like race, gender, or their proxies—when deriving estimates" (Corbett-Davies & Goel 2018, p. 2).

(ii) *Classification parity* "requires that certain common measures of predictive performance be equal across groups defined by the protected attributes. Under this definition, a risk assessment algorithm that predicts loan default might, for example, be required to produce similar false negative rates for white and black applicants" (Corbett-Davies & Goel 2018, p. 2). The latter would be an example of "error rate balance" (Chouldechova 2017).

(iii) *Calibration* "requires that outcomes are independent of protected attributes after controlling for estimated risk. For example, among loan applicants estimated to have a 10% chance of default, calibration requires that whites and blacks default at similar rates" (Corbett-Davies & Goel 2018, p. 2).

As Corbett-Davies & Goel assess the situation:

> "all three of these popular definitions of algorithmic fairness—anti-classification, classification parity, and calibration—suffer from deep statistical limitations. In particular, they are poor measures for detecting discriminatory algorithms and, even more importantly, designing algorithms to satisfy these definitions can, perversely, negatively impact the well-being of minority and majority communities alike.
>
> In contrast to the principle of anti-classification, it is often necessary for equitable risk assessment algorithms to explicitly consider protected characteristics. In the criminal justice system, for example, women are typically less likely to commit a future violent crime than men with similar criminal histories. As a result, gender-neutral risk scores can systematically overestimate a woman's recidivism risk, and can in turn encourage unnecessarily harsh judicial decisions.
>
> Enforcing classification parity can likewise lead to discriminatory decision-making. When the true underlying distribution of risk varies across groups, differences in group-level error rates are an expected consequence of algorithms that accurately capture each individual's risk....[C]alibration, while generally desirable, provides only a weak guarantee of equity. In particular, it is often straightforward to satisfy calibration while strategically mis-classifying individuals in order to discriminate. Indeed, the illegal practice of redlining in banking is closely related to such a discriminatory strategy. For example, to unfairly limit loans to minority applicants, a bank could base risk estimates only on coarse information, like one's neighbourhood, and ignore individual-level factors, like income and credit history. The resulting risk scores would be calibrated—assuming majority and minority applicants default at similar rates within neighbourhood—and could be used to deny loans to creditworthy minorities who live in relatively high-risk neighbourhoods." (2018, P. 2)

We mentioned in passing that algorithms can be intrinsically biased too (like humans). This is because algorithmic development is never an entirely objective, value-free endeavour: it will be influenced by a host of social and institutional norms, practices and attitudes that could well build bias into design. Such social and

institutional factors include—but are not limited to—the predominantly white, technically-educated and male composition of the field of AI (Crawford 2016). While these factors undoubtedly play a role in biasing algorithms, their effect is probably less palpable than extrinsically determined bias—such as dirty data. In any case, the two forms of bias are interconnected and originate from the same set of social and institutional conditions.

In the New Zealand context, this may be particularly important in relation to algorithms used in delivery of government services to Māori; for example, those which are trained on data that does not take account of Māori data concepts. Walter (2016) raises similar concerns in relation to how "Australia's racial terrain permeates statistics on Indigenous Australians" so that "In a seemingly unbroken circle, dominant social norms, values and racial understandings determine statistical construction and interpretations, which then shape perceptions of data needs and purpose, which then determine statistical construction and interpretation, and so on".

D. Informational privacy

The challenge here has been well stated by Lilian Edwards and Michael Veale (2017): "Machine learning and big data analytics in general are fundamentally based around the idea of repurposing data, which is in principle contrary to the data protection principle that data should be collected for named and specific purposes". Purpose limitation strikes at the heart of big data, and both the commercial and political imperatives justifying its use. While consent to data sharing has been the traditional way of cutting this Gordian knot, consent has become increasingly illusory and virtually meaningless in an age that has seen privacy clauses become longer, more complex, and effectively mandatory if a data subject is to have any hope of accessing basic services (e.g. medical and tax information, online banking, lifestyle apps, etc.).

A related set of problems arises from the curation of anonymised data sets, in which names and other identifying information have been removed from personal records. These data sets are typically used by commercial entities to construct "profiles" of types of people so that they can be more effectively targeted with advertising. While the obvious concern here is that such data, being anonymous, do not obviously fall within the reach of existing privacy provisions, the main worry concerns the potential for re-identification (Ohm 2010), which has become greater in the era of the Internet of Things and smart devices (Edwards and Veale 2017). It is also unclear whether the *controllers* of this data can be identified—obviously a crucial prerequisite to the enforcement of privacy rights for groups as much as for individuals; and whether collectives can, or should, have privacy rights in addition to individuals in these circumstances (Hildebrandt 2015; Mittelstadt 2017).

A major concern relates to the kind of consent which must be obtained (if any) for the use of inferred data—i.e. data inferred from *other* data whose collection has been consented to—and the rights of access data subjects have to such inferred data. In the era of machine learning, where inferences are the veritable stock-in-trade of advertising and social media business models, inferences potentially represent an enormous gap in the reach of existing data protection provisions. When a machine learning tool detects that a submitted tax return is fraudulent, for example, it is not doing so through directly ascertainable information, but on the basis of inferences drawn from directly ascertainable information (e.g. larger-than-usual losses reported in consecutive tax years). Indeed, in response to concerns about the intrusive nature of inferential analytics and perceived gaps in data protection laws in Europe, a new data right, the right to reasonable inferences, has been proposed (Wachter & Mittelstadt 2019).

Note that there are therefore at least two consent-related data protection issues thrown up by big data. One is the question of whether consent must be obtained from a data subject before an entity can *engage* in inferential analytics (i.e. whether purpose limitation prevents use of a data subject's information for the drawing of inferences). The other is the question of whether consent must be obtained for the *use* of inferred data (i.e. does inferred data count as data in the same way as primary data?).

Other information privacy issues relate to the application of well settled data protection standards to personal information used for machine learning and big data analytics. These issues include: obligations for informing an individual about when, how and what personal information is being collected and the intended recipients of the information (including third parties); how individuals can exercise their right to access information (to whom should an individual direct a request for access to data held about them?); who controls the information for the purposes of liability

for correction; exactly *where* obligations for ensuring accuracy should fall in the data chain (particularly when information is re-purposed or passed to a third party); what obligations for data deletion should be imposed (when must an agency delete information?); and the proper periods for data retention.

E. Liability and personhood

To the extent they can, settled principles of private law will continue to guide courts assigning liability for harms caused by technology. The traditional doctrine in tort, for example, would direct that if there is a problem with a mechanical component in an autonomous vehicle (say), and the problem arises from negligent assembly, the proper person to sue for any resulting damage is the manufacturer. However, if the problem resides in the vehicle's software, tracing liability may not be as simple as this. Machine learning algorithms learn to do things (including make mistakes) for themselves. This raises the prospect of a danger inherent in the software for which no human can be identified as directly responsible. It is true that product liability regimes generally hold the manufacturer responsible for defects even when the manufacturer could not have known of them given the state of knowledge at the time.[6] From this springs the thought that a software developer should always be liable in these situations—perhaps the software developer did not bake the glitch into the vehicle's code, but they should be liable regardless. One difficulty with this solution is that the analogy to product liability is fraught, the principles of which apply clearly enough to *products*, but not to *services*. It is not clear whether software is a product or a service.[7]

One way forward may be to recognise algorithms as juridical persons, which would presumably entail the creation of an insurance scheme enabling damages to be recovered from them (AlgoAware 2018, p. 33). A proposal along these lines is currently being debated in Estonia. New Zealand, of course, has a no-fault insurance scheme in the ACC, which should (or could be made to) provide cover for personal injuries inflicted by autonomous systems. For all other harms caused by autonomous systems, a separate or parallel insurance scheme could supplement the ACC's coverage, hand-in-hand with *de jure* (or merely *de facto*?) personality in algorithms.

In the late nineteenth and early twentieth centuries, corporate personality was an answer to a specific problem of regulation: how to encourage investment in the newly created joint stock companies. The answer the courts hit upon was to restrict the liability of investors to what each had agreed to pay as the price of company membership. Juridically the result was obtained by a fiction (i.e. that a company of people is itself a type of person). Today the question is whether artificial personhood is an acceptable solution to the *opposite* problem, namely, where there is no recognisable entity to whom liability unambiguously attaches (because "the machine did it"). But unlike the case of companies a hundred years ago, the recognition of technological personhood will probably have to come from parliament rather than the courts. This is because the issue will likely demand reflection on a more comprehensive set of factors and reference to a broader range of stakeholders than any single court can reasonably be expected to contend with. Vital contributions are required from software engineers, human factors psychologists, policy analysts and other social scientists—the sorts of contributions better canvassed through a consultative parliamentary committee process than through an amicus curiae or Brandeis brief.

This report limits itself to three observations regarding the feasibility of recognising legal personhood in algorithms. First, though strict liability for software developers is an apparently uncomplicated option—one which does not depend on the attribution of blame to a culpable entity, and does not require the recognition of artificial personhood in algorithms—it will almost certainly have repercussions for the tech industry as a whole, which can be expected to pass on the increased risk of liability, in the form of higher insurance premiums, to the aspiring owner of an autonomous vehicle (for example). The strict liability option should therefore only be chosen with the full acknowledgment and backing of the public. It is true that the UK, Australian and New Zealand legal systems have traditionally enjoined strict liability for the actions of substances (e.g. chemicals) or chattels (e.g. animals) that cause harm without the knowledge of their controller, a situation that might be thought directly parallel to the "mindless" but independent actions of machine learning algorithms which "learn" to classify and execute

6. Although under UK law, lack of technical and scientific knowledge at the relevant time is now a defence, the so-called "state of scientific knowledge" or "development risks" defence. See Consumer Protection Act 1987 (UK).
7. See e.g. *Computer Associates UK Ltd v The Software Incubator Ltd* [2018] EWCA Civ 518 at [67].

procedures without being explicitly programmed to do so by their developers. The chilling effect of strict liability on innovation may be reason enough, however, to reconsider its aptness in the circumstances.

Second, New Zealand has recently recognised natural entities as legal persons (Te Urewera and the Whanganui river). In the case of rivers, of course, the motivation is protection of the natural environment rather than restriction or expansion of financial liability. But there is precedent for an extension of the category of artificial persons within domestic law. Of course one of the requirements of personhood is identity. A company obtains its identity from registration in a companies register. Algorithmic personality presumably calls for a similar form of registration so that subsequent versions of an algorithm can be recognised as either the "same" or "new".

Third, contrary to the usual worry, if the problem posed by the new algorithms is not that they generate an accountability gap, but rather what has been called a "moral crumple zone"—in which humans risk being liable for harm-generating algorithmic decisions even when they have no effective control over them (Elish 2016)—then the problem to which artificial personhood may be a solution will more nearly resemble the nineteenth century problem of investor liability. There the problem was how to limit the liability of a visible agent (the owner of stock); here the problem would be how to shield a visible agent (the human element in a human-machine system) from the unfair burden of responsibility falling on their shoulders simply because there is no one else around to blame. Analogy with the strongest case for personhood in common law jurisprudence makes the case for artificial personhood in algorithms somewhat easier to maintain.

F. Human autonomy

How we are to understand and potentially regulate the influence of algorithms on the way we perceive the world is among the most important—perhaps *the* most important—of the questions that advanced artificial intelligence poses today. For instance, "nudging" algorithms filter information so that it appeals to users based on an in-depth understanding of their preferences. These preferences are reliably surmised through retail history, Facebook likes, Twitter posts, YouTube views, and the like. More worryingly, in the democratic sphere this technology potentially facilitates active manipulation through targeted political advertising. So-called "dark" ads can be sent to the very people most likely to be susceptible to them without the benefit—or even the possibility—of open refutation and contest which the marketplace of ideas depends on for its functioning. The extent of "perception-control" that new digital technologies make possible is really a first in history, and likely to concentrate unprecedented power in the hands of a few big tech giants and law-and-order-obsessed state authorities (Susskind 2018).

To date, very few jurisdictions have come up with strategies that convincingly curb this potential. In Australia, the Australian Competition and Consumer Commission's Preliminary Digital Platforms Inquiry (December 2018) has recommended that a regulatory body be given authority to monitor, investigate and publish reports on the operation of the algorithms used by large market players (those generating more than $A100 million annually from digital advertising). However, the Commission's recommendations stop short of indicating what powers any such regulator would have in the event it uncovered anticompetitive or discriminatory algorithms.

5. REGULATORY/GOVERNANCE STRATEGIES

In this chapter, we consider some of the actual and potential regulatory responses to the use of predictive algorithms by government. Some of these already exist in New Zealand, others have been implemented in other jurisdictions. A third cohort is at the stage only of being proposed or considered.

"Regulation" is a much discussed and somewhat contested term in legal literature. Some definitions extend the concept far beyond legal rules. A very influential definition proposed by Julia Black defines regulation as

> "regulation is the sustained and focused attempt to alter the behaviour of others according to defined standards or purposes with the intention of producing a broadly identified outcome or outcomes, which may involve mechanisms of standard-setting, information-gathering and behaviour-modification." (BLACK 2002, P. 26)

On this definition, regulation is not confined to legal prohibitions and orders, and neither is it limited to rules issues by government or parliament. As Brownsword and Goodwin say: "We should not assume that 'regulation' is co-extensive with 'law'; and we should not assume that 'regulators' are only those who are authorised to issue legal directives" (Brownsword & Goodwin 2012, p. 25) This expansive definition could even include "unintentional influence such as market forces" (Bennett Moses 2013, p. 4).

Other definitions are more restrictive, seeking to confine the concept to legal restrictions. According to Susy Frankel and John Yeabsley (2011):

> "Regulation takes many forms. Regulation includes legislation, legal rules, codes of practice (both formal and informal), and a combination of these. As such, it includes government regulation, regional and local government regulation and self-regulation."

We take no position on the wider notion of regulation, but for the purposes of this report, our focus will largely be on actions that could be taken by the New Zealand government or parliament. This is not to deny that factors such as market forces or cultural norms can be highly influential in dictating behaviour. But the question of how they might be used or controlled is beyond the scope of this particular project.

Our concern, then, is with the application of legislation, legal rules and codes of practice to the use of predictive algorithms in government. At present, New Zealand has no legal rules specifically directed at algorithms, or at artificial intelligence more generally, though this is not of course to say that there are no rules applicable to them.

Whether an algorithm- or AI-specific regulatory response is merited, or whether existing laws and norms could (perhaps with some modification) adequately regulate these technologies, is an important question. Before rushing to scratch-build a regulatory response to a new technology, it is prudent to consider what about it we think requires regulation, and to consider whether existing rules are sufficient to deal with that. As the US National Science and Technology Council (2016) proposed:

> "If a risk falls within the bounds of an existing regulatory regime…the policy discussion should start by considering whether the existing regulations already adequately address the risk, or whether they need to be adapted to the addition of AI."

And, as Brownsword and Goodwin noted, technologies rarely emerge into a complete legal vacuum:

> "Although there might be no part of the regulatory array that is specifically dedicated to the emerging technology, and although there might be gaps in the array, it will rarely be true to say that an emerging technology finds itself in a regulatory void." (BROWNSWORD & GOODWIN 2012, P. 64)

That said, we should not too readily discount the possibility that a new technology genuinely does pose concerns that are sufficiently new to merit a targeted response. As Lyria Bennett Moses says: "There is nothing illogical in arguing for technology-specific legislation, but it only makes sense to do so if the regulatory rationale is closely tied to the technology itself" (Bennett Moses 2013, p. 15). For a targeted regulatory response to be justified, then, there must be something about the regulatory target that sets it apart from the sorts of actions or decisions governed by existing rules. That may be because the technology, by its very nature, raises new ethical or political concerns, or it may be because the nature of the technology means that existing rules do not protect adequately against the risks that it poses.

Part of our task in this section is to consider the array of legal rules currently in place in New Zealand that might be relevant to decisions made or informed by predictive algorithms. Are these rules up to the task of addressing the concerns we identified in Chapter 4? Can they do so without unduly negating the advantages discussed in Chapter 3? Have they kept pace with the changing nature of the technology as outlined in Chapter 1?

It is also important, though, to lift our heads and survey the international scene. In recent years, a range of initiatives have been adopted or proposed in other jurisdictions, in attempts to address the sorts of risks and concerns we have identified. Some of these might merit serious consideration in a New Zealand context, and we examine some of the most influential or promising of those here.

Although this section will focus primarily on legal rules, the potential role of non-legal mechanisms should not be entirely ignored. In the latter part, therefore, we turn our brief attention to some other initiatives, including self-regulatory models.

Ultimately, our purpose here is not to recommend an ideal system of legal rules or regulatory control of predictive algorithms. The options we discuss all have much that could be said for and against them, and our main purpose is to present a balanced view of each. We do, however, begin with a few recommendations of a fairly general nature about regulatory initiatives.

Desiderata for regulation of AI

In considering the regulation of government use of predictive algorithms, we propose nine desiderata. Some of these are general principles akin to Lon Fuller's well-known desiderata for valid law, particularly scope of application, evenhandedness and certainty (Fuller 1964). Others are specifically directed toward the regulation of new technologies such as predictive algorithms. It is accepted that particular desiderata will be more important in particular contexts and that no type of regulation is maximally efficient at meeting all nine desiderata.

1. **Fairness:**
    - regulations should apply evenhandedly, i.e. to all those employing a technology with a particular attendant risk;
    - regulations should not decrease public wellbeing or stifle innovation by placing undue burdens on any particular sector of society.
2. **Parsimony:**
    - existing laws that cover an identified problem should be maximally utilised before new laws are proposed;
    - overlapping laws and regimes should be avoided where possible;
    - regulations should have broad scope where possible.
3. **Proportionality:**
    - regulations should be reasonably necessary to secure their stated purposes (inasmuch as means should be reasonably adapted to their ends);
    - regulations should not impose a higher burden than is necessary to achieve a stated outcome.
4. **Timeliness:**
    - regulators should give due consideration to the maturity of new technologies when assessing the appropriate point at which to regulate.

5. **Flexibility and Reviewability:**
    - regulations should be flexible and adaptable in light of technological change and should be subject to appropriately regular review.
6. **Certainty:**
    - regulations should enhance certainty for citizens who are data subjects, government employees, and other stakeholders.
7. **Oversight and appeal:**
    - regulations should be consistent with existing rights to appeal important decisions made about data subjects by public officials;
    - where appeal is impracticable or inappropriate, decision-making procedures should be subject to independent oversight.
8. **Social license:**
    - regulations should enhance public trust and confidence;
    - regulations should enhance political legitimacy.
9. **Externalities:**
    - regulations should prevent the externalisation of costs associated with new technologies such as decreasing the privacy or security of the public.

A. "Hard law" and individual rights

We begin with a range of what are sometimes called "hard law" responses: legislation (primary or delegated) and the decisions of courts charged with interpreting and applying them. In particular, we will consider various parts of the regulatory environment that are oriented towards the concerns we identified in Chapter 4: control, transparency, bias, privacy and the like.

As we will show, New Zealand law already has a number of rights and remedies that are relevant in this context. We will then consider some of the international initiatives that have been proposed or implemented. This will include the European Union's much-discussed GDPR, as well as proposals that have arisen in the USA.

Control

The Government's *Algorithm Assessment Report* placed considerable importance on the retention of human agency alongside algorithmic decisions:

> "where algorithms are material to decisions which affect people's lives in significant ways, it is reasonable to expect that a real person has exercised human judgement during the process and over the final decision." (STATS NZ 2018, P. 31)

This has certainly been claimed to be true of the RoC*RoI algorithm. Although it has been said of RoC*RoI that "It requires no clinical judgement or manual calculation" (Wilson 2013) provision for manual over-ride appears to have been made. In a 2009 report, the Department of Corrections acknowledged that, as

> "an offender's officially recorded criminal history does not always reflect the true extent of actual offending (and thus future risk), "professional over- ride" was also available to staff if other information existed to suggest high risks posed by an offender." (DOC 2009, P. 14)

An Official Information request for specific information about the provision to override a RoC*RoI score was submitted to Corrections in November 2018, but at the time of writing, no response has been published (see https://fyi.org.nz/request/8982-roc-roi-override?unfold=1).

Other jurisdictions seem to have followed the same assumption. The Wisconsin Supreme Court in *Loomis* (see Section 2A), in deciding that the use of the COMPAS algorithm in sentencing did not violate due process rights, noted their expectation "that circuit courts will exercise discretion when assessing a COMPAS risk score with respect to each individual defendant".

Perhaps the best known example can be found in the European Union's GDPR. Article 22 states:

> "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."

This right is not absolute; paragraph 2 sets out limited exceptions. Paragraph 3, however, requires that where those exceptions apply:

> "the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."

New Zealand has no right directly analogous to Article 22. However, it may be that other legal protections could indirectly contribute to the same result. In a 2010 case before the Privacy Commissioner,[8] a complaint was made about a fully automated transfer process between a debt collector and a credit reporter. The process was deemed to breach Principle 8 of the Privacy Act 1993, which requires an agency holding personal information to take reasonable steps to ensure that "the information is accurate, up to date, complete, relevant, and not misleading".

The Commissioner held that, to be compliant with the accuracy aspect of Principle 8, a manual notation had to be added to the record. In effect, a human had to be kept "in the loop". The complainant, however, was unable to show that breach of Principle 8 caused her any harm. This is a significant limitation of the Privacy Act regime. It is also unclear how many agencies are aware of this decision or requirement.

Although we have noted that there are no algorithm-specific laws in New Zealand, this may to some extent change with the passage of the Courts Matters Bill into law. This applies to what appears to be the first New Zealand statute-based automated electronic decision-making system. Of interest for present purposes, it also contains a right of human review. New sections 86DA to 86DD, provide for the Chief Executive to authorise an automated electronic decision-making system for setting fine payment arrangements, including a greater time for payment or payment by instalments. Where such a system is approved, the Chief Excutive must, pursuant to section 86A(4), also approve procedures for operating the system which must also include procedures for:

(a) setting the criteria for variations;

(b) identifying the information that will be sought from the individual;

(c) notifying the individual or their representative of the right to seek variation of the arrangement; and

(d) notifying of the right to review by a person of any automated decision.

New section 86DC provides that the Chief Executive may only approve an automated system if satisfied that "each system has the capacity to do any actions required with reasonable reliability" and "there is a process available under which a person affected by an action performed by an electronic system *can have that action reviewed by a person* authorised by the chief executive to review those actions, without undue delay" (emphasis added).

A possibility for New Zealand law, then, would be to "firm up" the right to "a human in the loop". How might this right operate? It could, for example, grant to an affected person a right to demand a human review of an automated decision. Or it could it place a prior restriction on all agencies to refrain from delegating such decisions to automated processes. The former seems to have the limitation of requiring the affected party to be aware that they have been subject to an automated decision, and as such, might require to be bolstered by a requirement on the agency to inform them of this.

8. PCC276 *Case Note 205558* [2010] NZPrivCmr 1.

A further question arises with regard to remedies for breach of this entitlement; it will not always be an easy thing for the affected person to demonstrate that they have been *adversely* affected by the absence of a human decision-maker in a particular decision.

More generally, and obviously, whether any such right is required depends on the desirability of requiring a "human in the loop". There are certainly contexts where the sorts of information provided by the algorithm are only part of what should inform the judgment, and where over-reliance upon it would therefore be problematic. As we saw in Section 2A, the Wisconsin Supreme Court acknowledged that the risk prediction provided by the COMPAS algorithm is only one factor informing a sentencing decision. In relation to the HART tool in Durham, Oswald et al. (2018) have noted that

> "The model simply does not have all of the information available to it, and can therefore only *support* human decision-makers, rather than replace them....With both their own local knowledge and their access to other data systems, custody officers will frequently be aware of other information that overrides the model's predictions, and they must apply their own judgement in deciding upon the disposition of each offender's case."

This has also been recognized by the New Zealand Court of Appeal. *Belcher v Chief Executive of the Department of Corrections*[9] concerned an appeal against the imposition of an Extended Supervision Order. The original granting of the Order had been supported by the results of "instruments measuring the likelihood of Mr Belcher reoffending including RoC*RoI, Static-AS and SONAR" (at [19]). In considering the relationship between the appellant's personal circumstances and the actuarial results, the Court made the following observation:

> "Obviously factors which have arisen post-release must be allowed for in an ESO assessment. For instance if the appellant had been rendered a tetraplegic as a result of a post-release accident, this would have presumably eliminated the likelihood of him reoffending and would undoubtedly have negated any adverse inferences which might otherwise have been drawn for actuarial assessments." (AT [90])

This is redolent of what some experts have referred to as the "broken leg problem". Derived from the work of psychologist Paul Meehl, this referred to a thought experiment whereby an actuarial prediction of the chances of someone attending the cinema on a given night is undermined by knowledge that they have broken their leg that day. In cases of that sort, the value of discretionary human input seems obvious.

Other instances are more contentious. Cathy O'Neill (2016) has argued for human involvement on the basis of the desirability of discretion:

> "you cannot appeal to a WMD. That's part of their fearsome power. They do not listen. Nor do they bend. They're deaf not only to charm, threats, and cajoling but also to logic—even when there is good reason to question the data that feeds their conclusions."

Yet at the same time, it could be argued that discretion is the mechanism that allows for prejudice, favouritism and nepotism. Might it sometimes be the case that less rather than more human discretion might be the better approach to bias?

9. [2007] 1 NZLR 507.

Interestingly, the evidence on this does not provide strong support for removing discretion. In a report into bias in the criminal justice system, Bronwyn Morrison found that the evidence for this was far from conclusive:

> "Responses directed towards reducing discretion implicitly assume that the key problem is too much discretion and that rule tightening will reduce discretion and therefore lead to less disproportionality. A number of research studies, however, have questioned this assumption and demonstrated that, in isolation from cultural, individual and broader organisational and/or social change, this type of response is unlikely to be successful in addressing disproportionate criminal justice outcomes." (MORRISON 2009, P. 113)

Specifically, Morrison pointed to "research from the United Kingdom [that] suggests that efforts to curb police discretion have not been particularly successful in reducing ethnic disproportionality" with stop and search (Morrison 2009, p. 111). Worse, she reported evidence that reducing judicial discretion in South Australia, through the introduction of mandatory sentencing, impacted most harshly on Aboriginal youth (Morrison 2009, p. 112).

The role of human discretion in perpetuating, mitigating or even exacerbating the disproportionate treatment of minority populations in the criminal justice system is a complex question, and one that obviously requires considerable empirical research. This is a particular concern in New Zealand, given the widely recognised over-representation of Māori in the criminal justice system (Waitangi Tribunal 2017). Clearly, though, it cannot be an area governed by unchecked assumptions.

We should also remain open to the possibility that, at least for certain kinds of decisions, algorithms are "better" than humans (see Section 4A)—and the advantages they offer could be curtailed by the insistence on human interference. Although Meehl recognized the "broken leg" problem, his research became famous chiefly for showing that, in a range of situations, algorithms made better predictions than even trained, expert humans. The Nobel Prize-winning psychologist and economist Daniel Kahnemann has also investigated the relative performances of algorithms and human experts. He has suggested a number of reasons for this, including that humans over-complicate sometimes simple decisions, and that we "are incorrigibly inconsistent" (Kahnemann 2011, p. 224).

The superiority of algorithms at certain kinds of calculations has been recognized by New Zealand's Court of Appeal. *R v Peta*[10] is another case that concerned an appeal against the imposition of an Extended Supervision Order. One of the reasons given by the Court for allowing the appeal related to an error attributable to human error in a calculation that should have been conducted electronically, leading the Court to direct that "In future, only electronically scored ASRS test results should be presented in evidence." (*Peta* at [63]).

Of course what is meant by "better" here and how it may be determined in different contexts is a live research question. In those cases where it is considered that retaining some human input into decisions is valuable, though, we should be wary of false reassurance. Again, it may be the case that, in certain contexts, automated decision-making should not be used at all—particularly where there is a reasonable concern that its impact on a decision will exceed its reliability. This is in part a political and ethical question, but one that should be informed by research from the fields of human-machine interaction, human factors and psychology. In Chapter 4, we considered some of the possible challenges around the questions of control and discretion in the context of algorithmic decisions.

Whatever decisions are made about when and whether humans should be kept "in the loop", it is important that the rules put in place around this are likely to promote the desired outcomes. A rule that is satisfied by a nominal human presence, effectively rubber-stamping an algorithmic decision, may serve only to offer false reassurance, and as such, be worse than no rule at all. On the other hand, guidelines recommending, for example, that decision-makers consult their own judgment first before consulting an algorithm, using the algorithm merely as a check on their intuitions, could assist in offsetting some of the effects of automation

10. [2007] NZCA 28.

complacency and bias. The Wisconsin Supreme Court in *Loomis* required that sentencing judges be given a list of warnings about COMPAS if they intend to rely on its predictions to inform their decisions. More empirical (human factors) research is required to see whether such approaches really do work.

Algorithmic bias

As we have noted, New Zealand presently has no law directed specifically at algorithmic decision-making, nor any law implemented with algorithms specifically in mind. The risk of bias and discrimination, however, is not unique to this context, and is likely to engage existing and more general provisions in New Zealand law.

Section 19 of the New Zealand Bill of Rights Act 1990 ("BoRA") provides that "Everyone has the right to freedom from discrimination on the grounds of discrimination in the Human Rights Act 1993". Section 20I-L of the Human Rights Act 1993 ("HRA") specifically apply this right to "the legislative, executive, or judicial branch of the Government of New Zealand" or "a person or body in the performance of any public function, power, or duty conferred or imposed on that person or body by or pursuant to law".

It is clear, then, that anyone subject to algorithmic decision-making by government agencies or courts has a legal protection from discrimination. Section 21(1) of the HRA lists those prohibited grounds, including sex, race, age and employment status—all factors that have occasioned controversy in relation to algorithmic decisions.

Some of the algorithms currently used in New Zealand make use of some of these protected categories. The *Algorithm Assessment Report* says that ethnicity is not a variable in the RoC*RoI algorithm. In fact, ethnicity was originally a variable (Bakker at al. 1997). In 2003, however, in response to concerns about the "negative connotations" of using that variable, Department of Corrections re-examined the role that ethnicity played in RoC*RoI's predictive accuracy. The Department "found that, because of the high correlation of ethnicity with other variables, the predictive accuracy of RoC*RoI could be maintained by recalibrating other variables and reducing the effect of the ethnicity variable to zero". Since then, the ethnicity variable has been set to zero (Waitangi Tribunal 2005).

It does not follow, though, that the use of the RoC*RoI algorithm poses no concerns about prohibited grounds of discrimination. For one thing, gender and age are also protected categories, and these still feature as variables in the RoC*RoI calculation. For another, the possibility must seriously be considered that other variables act as effective proxies for ethnicity. And concerns about inputs are only part of the story regarding discrimination and algorithms. We return to these considerations shortly.

Determining whether there has been a breach of a right involves several steps. As the Court of Appeal ruled in *Ministry of Health v Atkinson* [2012] NZCA 184, the first step is "to ask whether there is differential treatment or effects as between persons or groups in analogous or comparable situations on the basis of a prohibited ground of discrimination" (at [55]). As the Court went on to explain, though, not all differential treatment will be discriminatory (at [75]), and the differential treatment will satisfy this part of the test "if, when viewed in context, it imposes a material disadvantage on the person or group differentiated against" (at [109]).

Discriminatory treatment, then, must treat individuals or groups both differently and disadvantageously, on the grounds of a protected characteristic. But even this will not suffice to establish a rights violation. The right to be free from discrimination is not an absolute one, which means it may be permissibly restricted in some circumstances. According to section 5 of the BoRA, rights "may be subject only to such reasonable limits prescribed by law as can be demonstrably justified in a free and democratic society".

What would it take to satisfy the requirements of section 5? Tipping J in *R v Hansen* [2007] NZSC 7 set out a number of criteria that would have to be satisfied:

(a) does the limiting measure serve a purpose sufficiently important to justify curtailment of the right or freedom?

(b) (i) is the limiting measure rationally connected with its purpose?

(ii) does the limiting measure impair the right or freedom no more than is reasonably necessary for sufficient achievement of its purpose?

(iii) is the limit in due proportion to the importance of the objective? (at [104])

If these criteria are satisfied, then the use of variables such as ethnicity, sex and age may be justified. It is important to realise, though, that the burden will lie with the party seeking to justify a limit on the right.

As we saw in Section 2A, the use of gender as an input variable formed part of Eric Loomis's challenge against the use of the COMPAS algorithm in his sentencing. The Wisconsin Supreme Court rejected that part of his claim, holding that "if the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose".

Predicting the likely outcome of a discrimination challenge in circumstances like those in *Loomis* is not easy, in view of what has been noted to be "the relative dearth of decisions regarding s 19" (Butler & Butler 2015, p. 857). It seems likely, though, that if such a challenge were raised in New Zealand, the importance of including ethnicity, gender or age to the predictive accuracy of the algorithm would likely be a key consideration in determining whether the Hansen criteria were satisfied. If the discriminatory variable could be removed without sacrificing accuracy—as apparently was the case with RoC*RoI and with the ACC algorithm—then its inclusion will not be justified.

Even where the variable has *some* significance, it must be proportionate to the objective. When the state's objective is public protection from dangerous criminals, this may seem easily satisfied; other objectives will be harder to justify as being of sufficient importance. But even in the most compelling cases, the "reasonably necessary" requirement means that the objective could not reasonably have been fulfilled without the limiting measure.

Excluding prohibited grounds as input variables might seem like an obvious way of avoiding any suspicion of discrimination. But leaving aside the possible tradeoff in terms of predictive accuracy in some cases, this only avoids one form of discrimination. The HRA also extends to indirect discrimination, that is, conduct that has the *effect* of treating a person or group differently on the basis of a prohibited ground (section 65).

As we explained in Chapter 2, disparate effects on prisoners of different races was at the heart of ProPublica's criticism of the COMPAS algorithm. As this case shows, avoiding discriminatory inputs will not always guarantee against discriminatory outputs. As Edwards and Veale (2017) have warned, the danger exists that the

> "excluded variables are likely related to some of the variables that are included, e.g. transaction data, occupation data, or postcode. Put simply, if the sensitive variable might be predictively useful, and we suspect the remaining variables might contain signals that allow us to predict the variable we omitted, then unwanted discrimination can sneak back in."

In a New Zealand context, Emily Keddell has warned that

> "while we are offered the reassurance that ethnicity is not used as a variable in the ROC ROI algorithms, literally every other variable such as age at first offence, frequency of conviction, number of convictions will over-identify Māori as being high risk." (KEDDELL 2018)

Support for this view can also be found in Bronwyn Morrison's report for the Ministry of Justice:

> "Research has consistently shown that legal factors such as offence seriousness, evidentiary strength, offending history, the direct context of decision making, victim charging preferences, as well as extra-legal factors such as socioeconomic status account for most (but not all) of the variation between different ethnic groups." (MORRISON 2009, P. 12)

As Morrison explains, opinions vary as to how to account for these relationships. If we are to be confident that the algorithm is operating in a non-discriminatory manner, though, then we would also need to confident that the variables on which its decisions are based are not themselves the product of past discrimination. This is the "dirty data" problem that we referred to in Chapter 4. If, for example, Māori over-representation in prison

populations is at least partly attributable to historic discrimination in policing or sentencing practices, then there is a danger that that discrimination will "creep" back into decisions through reliance on variables that appear innocuous but which are heavily tainted by those attitudes and practices. Hence it is important to realise that simply excluding certain variables from the range of input factors is unlikely to ensure anything about the algorithm's outputs.

Guarding against problematic discrimination "sneaking in" will therefore involve ongoing monitoring of the algorithm's outputs and impacts. As a UN (2018) Special Rapporteur recently explained:

> "This involves, at a minimum, addressing sampling errors (where datasets are non-representative of society), scrubbing datasets to remove discriminatory data and putting in place measures to compensate for data that "contain the imprint of historical and structural patterns of discrimination" and from which AI systems are likely to develop discriminatory proxies."

As we discuss further later in this chapter, this may involve the capacity to take a "wide angle" perspective that may not be available to individuals. A serious commitment to monitoring human rights compliance may not be the sort of function that can be delegated to affected citizens.

Transparency and the right to explanations

As we have seen, a great deal of the concern about algorithmic decision-making relates to the potentially opaque nature of those decisions. After all, if we cannot understand how a decision has been arrived at, then we cannot check it for accuracy or bias, and importantly, we cannot challenge it if we think it is wrong or unfair.

In response to this concern, much has been made of what some claim is a "right to an explanation" in the GDPR. Whether such a right can in fact be discerned or inferred from the GDPR is the subject of considerable and ongoing academic debate (contrast Wachter et al. 2017b with Selbst & Powles 2017). Even if such a right does exist, a range of concerns have been raised about its usefulness in addressing the transparency problem (Edwards & Veale 2017).

As we mentioned in Section 4B, section 23 New Zealand's Official Information Act 1982 ("OIA") provides a right to reasons. If requested, these reasons are to be supplied in a written statement that includes any findings on material issues of fact and (subject to a few exceptions) a reference to the information on which the findings were based. Grounds for withholding such information are quite limited. But does this address concerns about the intelligibility of any explanation given about algorithmic outputs? As we have seen, in the context of algorithmic decisions, providing an explanation that is intelligible to affected persons and the general public could prove especially problematic.

In the context of the OIA, some reassurance has been provided by the High Court decision in *Vixen*, where it held that:

> "Where the legislature has specified that reasons must be given I should think those reasons must be sufficient to enable any body with a power of review to understand the process of thought whereby a conclusion was reached. Equally the reasons must allow those with vested interests, like those of the appellant, to so understand the basis for decisions as to be better informed in predicting that which is or is not within the law. Further, in this case the public has a general interest in knowing and comprehending the standards that the Board sees as important." (RE VIXEN DIGITAL LIMITED [2003] NZAR 418 AT [43])

Still, the OIA leaves citizens affected by private sector decision-making without a comparable right, and this is true whether the decisions are made by humans or algorithms. At least in the private sphere, New Zealand's laws lag considerably behind the EU's.

An alternative legal basis for access to information about a decision about personal information can be found in the Privacy Act 1993. The Act relates to information about oneself, held by private and public agencies. It applies to "personal information", which means information about an identifiable individual (section 2). Therefore, the Act applies if an individual is identifiable at any stage of the construction or use of the algorithm, including the results of an application of the algorithm using personal information about an identifiable individual. The application of an algorithm may not be enough on its own, because the algorithm does not appear to be personal information within the meaning of the Privacy Act. However, the processing of information about an identifiable individual by algorithm comes under the remit of the Act.

The Act's Information Privacy Principles ("IPPs") set out the requirements for the collection, storage, use and disclosure of personal information. Section 6 sets out 11 IPPs which cover the collection, holding and use of personal information (and a 12th IPP concerning unique identifiers which doesn't appear in an equivalent Act in any other jurisdiction).

Principle 6 states that:

(1) Where an agency holds personal information in such a way that it can readily be retrieved, the individual concerned shall be entitled—

(a) to obtain from the agency confirmation of whether or not the agency holds such personal information; and

(b) to have access to that information.

As with the OIA, the question of intelligibility of that information is likely to be highly relevant. Section 42 of the Act provides further details about how information must be provided. The recent decision in *Naidu v Australasian College of Surgeons* [2018] NZHRRT 234 provides some guidance on how this is likely to be applied.

Dr Naidu asked the College for access to personal information held about him in relation to an application for admission to a specialist medical training course. The request was not responded to within the statutory time period and, when finally complied with, the information included a score sheet with codes allocated to summaries of a referee's views about Dr Naidu's application. These scores were not in a form he considered meaningful (for example, what the score was out of or whether it was weighted).

The tribunal noted that paragraphs (c) and (d) of the Privacy Act's section 42(1) require information to be made available in a "form which can be comprehended". Considering the proper application of this section to the score results, the tribunal concluded that the College must provide "the key" which unlocks the information. The tribunal ordered the summary coding information be made available to Dr Naidu in a "meaningful" way, namely, "in a manner that is transparent, intelligible and easily accessible".

To satisfy the requirements of current New Zealand law, then, it appears that it will not suffice to provide information about the decision in a manner that is incomprehensible to all but experts. Instead, means must be found to render these explanations sufficiently clear for the individual themselves, the public, and any authority with the power to review that decision. The challenge of how best to achieve this is one that agencies will need to take seriously; certainly, ideas like "explainable AI" will merit consideration.

Equally important, though, will be ensuring that intellectual property rights and policies of suppliers do not hinder the agency's ability to give proper explanations. As far as we are aware, this has not yet presented the sort of problem in New Zealand that manifested in the Loomis case. The algorithms used by government agencies here are either manufactured in-house or (as in the recent ACC example) bought from local firms who are willing to make the details of their algorithms available to the public. As algorithms become more commonly employed, though, this is a consideration that will need to be kept in mind.

Informational privacy

Privacy concerns are a near-ubiquitous feature of the "information age" as noted by the New Zealand Human Rights Commission (2018). In the specific context of AI algorithms, the UN Special Rapporteur on the right to freedom of opinion and expression has written that:

> "AI-driven decision-making systems depend on the collection and exploitation of data, ranging from ambient, non-personal data to personally identifiable information, with the vast majority of data used to feed AI systems being somewhere in the middle

> — data that are inferred or extracted from personal data, or personal data that have been anonymized (often imperfectly).”
> (UN 2018, [34])

The Privacy Act 1993 is the starting point for a regulatory response to the issues around data collection and re-purposing (see Section 4D). The Act regulates the collection, access, use, correction, storage and deletion of personal information, and provides that, in general, personal information may only be used for the purpose for which it is collected (Principle 10(1)). This reinforces the obligations of agencies to be very clear, at the point of information collection from an individual, about the purposes for which the personal information is to be used, and to ensure that consent for these purposes has been given (Principles 1-4).

Unlike other jurisdictions, the New Zealand Privacy Act does not provide a means for “purpose setting”, not set out a standard for the specificity with which an agency must articulate the purpose or purposes for which information will be collected and used. However, the Office of the Privacy Commissioner generally construes collection and purpose provisions narrowly and has, for example, ruled unlawful a coercive collection that was not tied to a clear purpose (see, for example, *Inquiry into Ministry of Social Development Collection of Individual Client Level Data from NGOs (2017)*.

Unfortunately, given the nature of the technology at issue, the current regulatory regime is likely to prove unsatisfactory. While some concerns—such as those around inferential analytics—do not appear to have been considered in New Zealand at all, the application of existing laws to some of the other concerns we have mentioned is far from adequate.

For example, in relation to concerns about the re-purposing of information without consent, the Act provides that an agency holding information may use the information for a different purpose from that for which it was collected where an agency “believes on reasonable grounds” that the information has been “used in a form in which the individual concerned is not identified”, or where the information “is used for statistical or research purposes and will not be published in a form that could reasonably be expected to identify the individual concerned” (Principle 10(1)(f)). Principle 11 also permits disclosure of personal information to third parties in limited circumstances. Principle 11(h) is on the same terms as 10(1)(f), creating authority for sharing anonymised data sets. Principle 11(b) permits disclosure if the “source of the information is publicly available information” and “in the circumstances of the case, it would not be unfair or unreasonable to disclose the information”.

There are a number of problems here in relation to machine learning and the training of artificial intelligence tools. For one thing, it is unclear what constitutes “reasonable grounds to believe” that the information will be used in a form which will not identify the individual. In some cases the technology to enable re-identification may not exist at the time the information is used, but may emerge later. Risks of re-identification have increased with greater sharing of anonymised data sets and improvements in data analytics, with commentators debating whether it is now possible to guarantee anonymity by de-identification (see e.g. Cavoukian & Castro (2014) and Narayanan & Felten (2014)). While some concerns may be overstated, the law is not settled.

A further legal difficulty is how the authority to re-purpose information sits with the controller agency’s obligations not to keep information “for longer than is required for the purposes for which the information may lawfully be used” (Principle 9). Since anonymised information may lawfully be used for any purpose if the individual is not identifiable, the point at which the obligation to delete personal information arises is unclear. Authority to use personal information for “statistical and research purposes” holds so long as publication of the information is not “in a form that could reasonably be expected to identify the individual concerned”. The potential for discoveries of new uses for information combined with a “goldrush” approach to the potential of artificial intelligence creates strong commercial and other incentives for agencies to hold on to information as long as possible “just in case” they are able to use it for some other, potentially commercial or public benefit purpose in the future. Concerns with the inadequacy of current law in this area have prompted calls for a strengthened “right to be forgotten”, which would require deletion of personal information, including personal information used to train AI systems.

Even where it is possible to identify a breach of an information privacy principle, this is not enough, in most cases, for a finding of interference with privacy under the Privacy Act. The legal test requires both breach of an information privacy principle and harm that meets

the relevant legal threshold. In the context of big data analytics and machine learning, questions arise as to whether the harm from data re-purposing (or as a result of breach of any other privacy principle) can meet the legal tests for harm in various regulatory standards and, if it does not, whether other ethical issues may arise. More research is needed in this area, including on how tests of harm here compare and contrast to tests of harm in other areas of the law, such as those concerning discrimination, transparency, and consumer protection.

Algorithmic data protection impact assessments, closely modelled on the existing privacy impact assessment tools, have been developed by a variety of researchers and institutions as a practical means of providing both transparency and a means of accountability where personal information is used (Wachter 2018). However, there are no universal standards for privacy impact assessments. In New Zealand, a large number of government agencies publish privacy or algorithmic impact assessments. In the light of concerns about inadequate application of information privacy laws to data analytics and increased use of data analytics by government agencies, the New Zealand Government Chief Data Steward and the Privacy Commissioner recently issued a set of six "Principles for safe and effective use of data and analytics". The principles encourage agencies to: show clear public benefit; ensure data is fit for purpose; focus on people; maintain transparency; understand the limitations of analytics; and retain human oversight. The recent algorithm stocktake used these principles to assess algorithmic use, and found that some agencies have published information about algorithms they use. But the stocktake also concluded that many agencies do not comply with these principles, recommending that agencies consider making more detailed information available about the algorithms they use, including publication of operating code (which could enable technical peer review), and consulting with those likely to be affected by algorithmic use.

Regulators in New Zealand, Australia and the United Kingdom have warned of the perils of the current regulatory approach to information privacy concerns. In response, in 2016 the Australian Government proposed amending the Australian Privacy Act 1988 to create criminal offences and civil penalties as deterrents against attempts to re-identify personal information that has been published in anonymised form in government data sets. However, the Australian Bill is not yet law.

The United Kingdom has recently introduced two new offences. The first is an offence to re-identify information which has been anonymized, the second to knowingly or recklessly using information that has been re-identified unlawfully. The offences are subject to specific defences, including consent of the individual or the data controller, acted with reasonable belief of consent, for certain special purposes or where certain prescribed conditions for testing effectiveness, or notification to the Information Commissioner, are met (sections 171 and 172 of the Data Protection Act 2018).

The New Zealand Law Commission identified the emerging issue of re-identification as one that would need to be addressed in reform of the New Zealand Privacy Act. No progress was made and in 2016 the Privacy Commissioner reported to the Minister of Justice that new controls on re-identification of personal information were needed including protection against risks that personal information might be unexpectedly identified from information that had allegedly been anonymised or de-identified for the purposes of sharing in data sets. The Commissioner has also recommended three new measures: controls on those who receive de-identified information to better protect the privacy of individuals; more stringent requirements on agencies to take adequate steps to de-identify information before using or publishing it; and introduction of a new privacy principle conferring a right to erasure of personal information (the so-called "right to be forgotten").

A Privacy Bill introduced to Parliament in 2018 neither amends existing provisions nor introduces new provisions addressing these concerns. A number of submissions to the Bill have called for reforms, including new provisions prohibiting re-identification and prohibition on use of de-identified personal information of the kind we have mentioned. In March 2019 the Bill was reported back to the House, with no substantive amendments with respect to privacy issues we have highlighted such as transparency, the right to an explanation, right to erasure and re-identification protections.

The Ministry of Justice report on the Bill to the Select Committee summarised submissions about automated decision-making and the GDPR indicating that the majority of submissions favoured new and stronger privacy protections, but concluded:

> "The new rights in the GDPR raise interesting issues, including whether having a human involved as a check at the end of an automated process is an effective protection—some people think that humans tend to defer to the algorithm in that situation. "Explainable AI" is also a developing field and some people think that putting the burden on the affected individual to raise concerns about an algorithm is not realistic, and instead favour systems of algorithmic auditing. These issues require further consideration. We do not recommend any change to the Bill."

Many of the issues were deferred, slated instead for "future work on privacy reform" and deferred pending further policy analysis and consultation.

The Bill did, however, introduce some tightened requirements for transparency in collection of personal information, including from children and young people. Clause 18 introduces a new requirement for the Privacy Commissioner, in carrying out any of his or her functions, "to take account of cultural perspectives on privacy". This new requirement may also strengthen protection for taking account of te ao Māori perspectives on privacy issues with algorithms and other automated decision-making technologies.

Taking rights seriously in the AI era

Rights around accuracy, privacy, transparency and freedom from discrimination already exist in New Zealand law, and all are likely to have important roles in the context of predictive algorithms. The possibility of strengthening or fine-tuning these rights so that they better respond to this technology is certainly worthy of consideration, and it will be worthwhile to keep a watchful eye on international initiatives such as the GDPR for comparison.

As the UN Special Rapporteur on freedom of expression and opinion reported in late 2018, there are a number of steps that governments can take when procuring or deploying algorithms to ensure they act consistently with human rights instruments. These include undertaking human rights impact assessments, an approach we consider later in this chapter. The Rapporteur also makes the important observation that, to ensure ongoing compliance, these systems "should be subject to regular audits by external, independent experts" (UN 2018, [62]).

Such "top-down" scrutiny is likely to be necessary if human rights obligations are to be taken seriously. Leaving it to affected individuals to enforce their rights of privacy or against discrimination is unlikely to be an adequate method of ensuring compliance, not least because, as Virginia Eubanks has pointed out, many of those affected by algorithmic decisions "don't know that they are being targeted or don't have the energy or expertise to push back when they are" (Eubanks 2017, p. 6). On a similar note, Edwards and Veale (2017) have cautioned that "Individuals are mostly too time-poor, resource-poor, and lacking in the necessary expertise to meaningfully make use of these individual rights".

Edwards and Veale also warn against approaches that "rely too much on individual rights for what are too often group harms". These are concerns worth taking seriously. Those at the "sharp end" of algorithmic decisions will often not be well placed to interrogate the basis for those decisions, less still to challenge them. And where they are able to do so, their concern will typically (and understandably) be directed at how those decisions affected themselves and their families. Determine whether algorithms are impacting differentially on different groups will typically require a "wide angle" perspective that it may be unrealistic to expect from affected individuals.

We are also aware that concerns about indigenous rights and data sovereignty in relation to new forms of data collection and use have emerged, as Kukutai and Taylor note:

> "While indigenous peoples have long claimed sovereign status over their lands and territories, debates about "data sovereignty" have been dominated by national governments and multinational corporations focused on issues of legal jurisdiction. Missing from those conversations have been the inherent and inalienable rights and interests of indigenous peoples relating to the collection, ownership and application of data about their people, lifeways and territories."

Kesserwan (2018) notes that in the Canadian context discussion of AI could be assisted by indigenous concepts of what is human, including "what has spirit". Such concepts, she suggests, offer another way of conceptualising artificially intelligent personhood and might assist discussion of issues such as whether humans should have obligations of inter-generational stewardship in respect of AI.

We mention these as examples of how and why more research on the diversity of rights issues raised by algorithms, beyond individual rights, may be important and useful. With that in mind, we now turn to the next class of potential regulatory responses to the concerns around predictive algorithms.

B. Regulatory agencies

Legislation is only one of the tools available to address the concerns around algorithms. Regulatory bodies or agencies might also have an important role to play. Suzy Frankel and John Yeabsley have written of "a strong New Zealand tradition in establishing independent agencies at arm's length from the government" (Frankel & Yeabsley 2011, p. 6). It is to this part of the regulatory terrain that we now turn.

The 2017 Government Expectations for Good Regulatory Practice describe a "regulatory agency" as

> "any agency (other than courts, tribunals and other independent appeal bodies) that has any of the following responsibilities for the whole or part of a regulatory system: monitoring; evaluation; performance reporting; policy advice; policy and operational design; legislative design; implementation; administration; information provision; standard-setting; licensing and approvals; or compliance and enforcement." (NZ GOVERNMENT 2017)

Regulatory agencies come in a wide variety of forms, and have a wide array of remits and responsibilities. As with legislation, they could be specific to a particular technology or family of technologies (ACART, for example, exists to formulate policy for assisted reproductive technologies). Or they could be fashioned with a particular policy objective or value in mind (the Office of the Privacy Commissioner and the Human Rights Commission focus respectively on privacy and human rights issues, across a wide array of contexts).

Regulatory agencies can also be constructed with a wide array of powers. Some have compulsory inspectorate functions, others are able to hand out penalties and sanctions. Some can construct rules, while others exist to enforce or monitor compliance with rules constructed by others. Often, they will operate at a "softer" level, for example, in issuing best practice guidelines or codes of practice, or simply giving advice when requested. (ECART sometimes gives non-binding advice to fertility clinics in response to queries concerning ethical issues.) Which of these models would be best suited for the context of predictive algorithms is obviously going to be an important consideration.

In this section, we will examine some of these. We begin, though, by examining the suggested advantages of such an approach.

Advantages

In a 2016 article considering regulatory options for AI systems, Matthew Scherer set out some of the possible advantages of regulatory agencies over legislatures or courts when it comes to responding to emerging technologies.

- Agencies can be tailor-made for the regulation of a specific industry or for the resolution of a particular social problem. Policymakers in agencies can be experts with a background in the relevant field rather than the generalists that fill the ranks of courts and legislatures.
- They need not be bound by rules that restrict courts from conducting independent factual investigations and from making policy decisions based on broad social considerations rather than the facts of the specific case in front of them.

The idea of a regulatory agency with the requisite expertise in this area is obviously attractive. ACART may serve as a useful model here. Section 34(4) of the Human Assisted Reproductive Technology Act 2004 requires that the Committee must include:

(a) one or more members with expertise in assisted reproductive procedures;

(b) one or more members with expertise in human reproductive research;

(c) one or more members with expertise in ethics;

(d) one or more Māori members with expertise in Māori customary values and practice and the ability to articulate issues from a Māori perspective;

(e) one or more members with the ability to articulate issues from a consumer perspective;

(f) one or more members with expertise in relevant areas of the law; and

(g) one or more members with the ability to articulate the interests of children.

What an analogous range of expertise look like for predictive algorithms is a topic that could merit further consultation, but on the face of it, expertise in computer science, data analytics, law and ethics seem like obvious inclusions. But some degree of input from those most likely to be adversely affected by algorithmic decisions would be vital. Given the extent of disproportionate Māori representation in some of the areas we have considered, a member who is well placed to speak of those experiences would be indispensable. A requirement for the agency to take account of cultural perspectives when carrying out its work would also strengthen likely engagement with the diversity of communities that may be affected by use of predictive algorithms.

Scherer's second point is that regulatory bodies can be proactive rather than merely reactive. They could be empowered to take steps without the necessity of someone already having been harmed or disadvantaged. Furthermore, their decisions could be more broad ranging than the facts of the case before them, a limitation that significantly restricts the effect of court decisions. The point about courts is echoed by Frankel and Yeabsley (2011), who point out that regulating through court decisions is unlikely to be well suited to a small country like New Zealand, which will typically not have many test cases.

Of course, it is true that legislatures are also free to make rules in an "upstream" regulatory phase, and they are not confined to any particular set of facts. It is almost a cliché in emerging technology regulation, though, that the process of legislative reform is too slow to respond to fast-changing technologies. While major questions of policy should properly be dealt with at that level, a range of more detailed decisions could better be handled by more agile regulatory mechanisms.

As Scherer acknowledged, regulatory bodies or agencies might also have certain disadvantages, not least—relative to legislators—the lack of accountability to the public. Nonetheless, "In the context of AI", he concluded, "this makes agencies well-positioned to determine the substantive content of regulatory policies" (Scherer 2016).

Regulatory approaches in other jurisdictions

A review of regulatory approaches to AI in other jurisdictions reveals a very diverse approach. By December 2018, 26 countries (including the European Commission) had developed some form of national AI strategy or undertaken a national assessment of AI implications including France, Germany, the United Kingdom, Japan, China, Russia, Kenya, India, South Korea, Sweden, the United States of America and Singapore. The number of such strategies is rapidly increasing, but no country has yet regulated AI in general. In this section we give a brief overview of some of the approaches, particularly in relation to creating new regulatory bodies.

In 2017, only seven countries had national AI strategies, whereas by the end of 2018, this number had jumped to 26, including some multi-lateral strategies. Most strategies focus on adoption and promotion, with ethical and social implications the subject of research and consultation rather than specific policies or regulation. Most were developed by, or propose, expert bodies of various kinds, such as steering groups, research institutions, public/private think tanks, task forces, consortiums or multi-stakeholder advisory groups. The functions of these various bodies varies widely: from research, to development, economic impacts, skills and training, increasing scientific talent, and social implications.

A significant number of strategies propose the establishment of some kind of centre for AI, for example, Finland has established the Finnish Centre for AI which is a partnership between Aalto and Helsinki universities

with the goals of increasing AI research, skills and industry collaboration. Most of these kinds of centres are not focused on regulatory issues. Some countries have established mechanisms to carry out research or to examine particular issues. India, for example, developed a national strategy focused on scientific with Centres of Research Excellence in AI and creating AI applications of social importance. Germany, in addition to a comprehensive strategy for AI development, has established a commission to inquire into "Artificial Intelligence: Social Responsibility and Economic Potential" similar to a previous commission that had examined the ethics of autonomous vehicles.

Another trend is the establishment of ethics councils, or similar groups. For example, in June 2018 Singapore announced a number of new AI initiatives including a new Advisory Council on the Ethical Use of AI and Data. The function of the Council is to assist development of standards and governance frameworks for data use and ethics for the use of AI.

Some strategies are more comprehensive than others, and China has the most comprehensive national strategy, the *New Generation Artificial Intelligence Development Plan* (2017). The plan is unique for the breadth of its vision to create new theoretical models for AI and to develop intelligent AI infrastructure on which a variety of AI related services can be deployed. The plan proposes to develop these over three phases: develop AI to be alongside competitors by 2020; be world leaders in some fields by 2025; and by the primary global centre for AI innovation by 2030.

In France, the national initiative draws on the report *For Meaningful Artificial Intelligence: Towards a French and European Strategy* (Villani 2018) and was announced by President Macron in December 2018. The national strategy aims to address four challenges: building the AI workforce including research capability, creating open data policies to encourage adoption of AI technologies, creating a regulatory and financial framework to support development of "AI champions" (for example in the area of driverless cars) and, development of ethics to prevent discrimination or other arbitrary treatment. While President Macron's announcement of the national strategy (and associated 1.5 billion Euros funding package) mentioned regulation a number of times, no new regulatory bodies were announced and it remains to be seen whether any will emerge. However, the strategy includes the creation of an international panel of experts in artificial intelligence based on the model of the Intergovernmental Panel on Climate Change. The panel's aim will be to organise independent global expertise, with an initial focus on issues of transparency and fair use.

In late 2018 Australia, the United Kingdom and Canada announced new policy and regulatory proposals. Given the legal and other similarities between these two countries and New Zealand, we have taken a closer look at these developments for the purposes of considering the options for regulatory approaches in New Zealand.

The United Kingdom

The United Kingdom has a national AI strategy in the form of the *AI Sector Deal*, part of a larger initiative in UK industrial and economic policy. The UK House of Lords Select Committee also conducted a significant review of AI and its report, *AI in the UK: Ready, Willing and Able?* reviewing the economic, social and ethnical aspects of AI technologies and contained a range of recommendations. In addition, early work was done to map the roles of different existing UK regulatory bodies in order to determine whether to regulate and, if so, how might be best. For example, the role and powers of the Information Commissioner's Office and the CCTV Commissioner were considered along with other options. Ultimately, the government determined that a new body was needed, a decision which appeared to have broad, if not universal, support.

The UK Government Office for Artificial Intelligence ("OAI") was established in April 2018. While still developing its work programme, the OAI has a strong initial focus on promotion and adoption of AI across government. The mission of the OAI is to drive adoption of AI and its use to develop new services and uptake of related technologies. The OAI had already identified some barriers to adoption, particularly for small and medium-sized enterprises (SMEs) such as lack of access to data to use to develop AI services and SMEs being unaware of business models and how to calculate returns on investment. Tool kits and other resources to assist SMEs are planned. The Office is also monitoring whether there were any legislative barriers to adoption and working with others on innovative responses. For example, in response to concerns about intellectual property laws restricting access to data, the Open Data Initiative was establishing Data Trusts, where data could be shared.

The OAI work programme is still developing and workstreams include a proposed AI review, data, skills and adoption. The government AI review will comprise an audit of AI use (rather than a stocktake which has been done in New Zealand). The aims of the audit are to provide baseline information about existing of AI use, enable assessment of opportunities for AI adoption and to show that it is safe to use AI. Other areas of work for the OAI include a proposed procurement policy and a survey of the government's 3,000 data scientists to elicit potential areas for testing and deployment of AI related services.

The UK Government has recently established a new Interim Centre for Data Ethics and Innovation ("CDEI"). The Centre is still nascent, with a Ministerial consultation about its role starting in November 2018. The CDEI was set up with the support of the British Prime Minister, partly in response to public concerns about AI.

The CDEI is currently hosted in the Department of Culture, Media and Sport and aims to be a statutory body in the next 2-3 years if it is clear there are insufficient regulatory measures and a statutory body is needed, for example, for education or other purposes. CDEI is reviewing in more detail the current gaps in regulation and possible levers for regulation and will assess evidence and may propose regulation where it thinks this is needed. Eventually, CDEI will be an independent advisory body, rather than a regulator although it may advise about the need for regulation in a particular area or alternatively about regulations that are perceived as barriers to innovation. CDEI's main functions are to:

(i) analyse and anticipate—horizon scanning, looking for opportunities and risks;

(ii) deep dive into particular issues—initially at the issue of bias and micro-targeting and the harms it might have in particular domains (e.g. advertising promoting gambling to individuals at risk of addiction); and

(iii) bring people together—helping to develop networks involving government, commerce and diverse other groups.

The role of CDEI does not include developing ethical standards. Despite the likely lack of regulatory powers, CDEI do have some powers, for example to request information from agencies and give advice. In addition, insofar as regulation might be considered, our research meetings in the United Kingdom, revealed that there was more of an appetite for AI regulation in the UK than previously and some considered regulation sector-by-sector would be preferable, as some sectors were more advanced in deployment of AI than others.

Australia

The Australian Human Rights Commission ("AHRC") is currently investigating human rights and technology. In July 2018, the AHRC released a discussion paper asking, among other things, whether "Australia needs a better system of governance to harness the benefits of innovation using AI and other new technologies while effectively addressing threats to our human rights". The AHRC has also established an advisory group to assist its work in this area.

In early 2019 the AHRC and World Economic Forum released a white paper *Artificial Governance and Leadership* seeking feedback on which its proposal for governance of AI by a new body, a "Responsible Innovation Office". The AHRC made the proposal in response to consistent feedback that a new regulatory body was needed in relation to new technologies, developing a "straw" proposal for comment which, it says, is "unlike traditional oversight or compliance bodies".

Specifically in relation to AI, the AHRC suggests a range of functions the new body could carry out. For example, the proposed body would need to "establish a normative framework for the development and deployment of AI", have an inclusive governance structure (including participation by those most affected by new technologies), focus on AI in government and the private sector, and have a wide remit, for example to examine big data sets and related issues (such as ensuring collection and ownership of data sets was democratic). The paper recommends the body have both coercive and non-coercive powers, the ability to develop standards, a certification scheme for human rights compliant AI development, and power to investigate complaints. In addition, the proposed body would evaluate data sets, promote open data standards, build a repository of best practice for stakeholder consultation and engagement, and write and publish ethical codes of practice drawing on those developed elsewhere.

The white paper is seeking comment on:

- the nature and scope of the challenge for human rights protection posed by the rise of AI;
- whether Australia needs a new or existing organisation to lead in the promotion of responsible innovation in AI; and if so,
- what might be the aims, functions and roles of such an organisation.

As we noted in Section 4F, the Australian Competition and Consumer Commission ("ACCC") is also looking at consumer concerns about the impact of online search engines, social media, and digital platforms on competition in the advertising and media markets. The ACCC has identified a range of functions that a new regulatory body could perform, such as monitoring, investigating and reporting on discriminatory and anti-competitive conduct, and providing assurances to consumers, government and businesses on the performance and impact of algorithms and policies.

The issues being examined by the AHRC and ACCC reach much further than the remit of our research, which is focused more narrowly on artificial intelligence. In addition, some of the functions proposed by the ACCC (such as complaints about harmful online content) are already being performed in New Zealand by NetSafe, the Approved Agency for the purposes of the Harmful Digital Communications Act. The ACCC's final report is due to be released in mid-2019.

Canada

During 2018 the Canadian government developed a directive to federal government agencies on automated decision systems with applies to automated decision systems developed or procured after 1 April 2020. The directive is intended to be a guide for government agencies and its expected results are that:

a) Decisions made by federal government departments are data driven, responsible and comply with procedural fairness and due process requirements;

b) Impacts of algorithms on administrative decisions are assessed and negative outcomes, when encountered, are reduced;

c) Data and information on the use of automated decision systems are made available to the public, where appropriate

The directive applies to systems which provide external services, and to an "system, tool, or statistical models used to recommend or make an administrative decision about a client". The directive applies only to systems in production, not those in test environments, nor to any national security systems. The directive imposes five main requirements on those developing automated decision systems, namely:

a) Completion and publication of an algorithmic impact assessment (including any update that affects the initial assessment)

b) Transparency (including prominent, plain English notices that an automated decision system is in use, providing meaningful explanations of decisions, access to software components for review and audit and release of source code);

c) Quality assurance (including testing and monitoring, of outcomes and data quality, peer review, employee training, contingency systems, legal compliance and retention of human oversight);

d) Access to recourse, or remedy, for a client to challenge a decision;

e) Reporting by publishing information about the effectiveness and efficiency of the system.

The directive sets out definitions to assess the impact levels, together with detailed requirements for how each impact level should be addressed for each of the five areas. Decision levels range from those with an impact that is reversible and brief, to reversible and short-term, difficult to reverse and ongoing and (at the highest level) impacts with are irreversible and perpetual. The government has also established a website where more detail is provided about how to implement the directive, including the different expertise that should be included during design/ build and deployment/operational phases.

As a guide to assist government agencies the directive is a useful step and has been welcomed as a standard that the private sector could emulate. The directive appears to be a regulatory tool implemented within existing Canadian public sector laws and policies, including accountability systems, rather than a statutory tool accompanied by a new model of regulatory oversight.

A pharmaceutical model

An existing instance that has recently attracted some interest as a possible model for AI and algorithms is the pharmaceutical industry. Although specifics vary between jurisdictions, most countries have some sort of regulatory agency in place.

The suitability of this model has been advocated by several writers. Coravos et al. (2019) pose the following question:

> "For decades, pharma and biotech companies have tested drugs through meticulously fine-tuned clinical trials. Why not take some of those best practices and use them to create algorithms that are safer, more effective, and even more ethical?"

Coravos and colleagues are sceptical of the notion of a single AI regulator that could operate across all disciplines and use cases. Instead, they suggest that "oversight can and should be tailored to each field of application" (Coravos et al. 2019). The field of healthcare, they claim, would be one field "already well positioned to regulate the algorithms within its field". Writing in a US context, they offer the Federal Drug Agency as a potential model, while "[o]ther industries with regulatory bodies, such as education and finance, could also be responsible for articulating best practices through guidance or even formal regulation" (Coravos et al. 2019).

How would such an agency function? In an article that also made the case for an FDA-based model, Andrew Tutt considered a range of possible functions. These occupy a range of places on a scale from "light-touch" to "hard-edged". The agency could:

- act as a standards-setting body that coordinates and develops classifications, design standards, and best practices;
- classify algorithms into types based on their predictability, explainability, and general intelligence, but only subject the most opaque, complex, and dangerous types to regulatory scrutiny—thereby leaving untouched the vast majority of algorithms with relatively deterministic and predictable outputs (Tutt 2017, p. 107);
- establish guidance for design, testing, and performance to ensure that algorithms are developed with adequate margins of safety—that guidance, in turn, could be based on knowledge of an algorithm's expected use, types of critical versus acceptable errors it might make, and the suggested predicted legal standard to apply to accidents involving that algorithm (Tutt 2017, p. 108);
- promulgate guidance for developing algorithms that meet satisfactory standards of predictability and explainability (Tutt 2017, p. 109);
- require that technical details be disclosed, potentially pre-empting state-level trade secret protections in the name of public safety; and
- require that certain algorithms slated for use in certain applications receive approval from the agency before deployment.

The last of these suggestions would, on Tutt's analysis, be restricted for the most "opaque, complex and dangerous" uses. It could "provide an opportunity for the agency to require that companies substantiate the safety performance of their algorithms". Tutt also suggests that pre-market approval could be subject to usage restrictions. "Off-label use of an algorithm, or marketing an unapproved algorithm, could then be subject to legal sanctions".

Tutt's suggestion of a *use-based* approval system has much to recommend it. Regulation pitched at the level of particular algorithms seems likely to overlook the fact that these are in many cases highly flexible tools. An algorithm approved for an innocuous use could be repurposed for a much more sensitive or dangerous one.

The regulation of pharmaceutical substances in New Zealand falls under the Medicines Act 1981, and the New Zealand Medicines and Medical Devices Safety Authority (Medsafe). They are subject to pre-market approval, which sees them assessed for safety, quality and efficacy. Before they can be advertised or supplied, consent must be granted from the Minister of Health. In addition, post-market mechanisms exist to enable medicines to be removed from use if they are found to be unsafe or ineffective. Manufacturers and importers are also legally required "to report any substantial adverse events arising from the use (in New Zealand or overseas) of a medicine" (Medsafe website).

A regulatory scheme that is similar in several ways exists for the management of hazardous substances. These are regulated under the Hazardous Substances and New Organisms Act 1996 ("HSNOA"), various related regulations, and the Environmental Protection Agency.

The HSNOA creates a Hazard Classification System (section 74 (a)), which classifies substances according to various hazardous properties: for example, explosiveness, flammability and toxicity. If a substance falls below certain minimum levels for these properties, it falls outside the regulatory scheme. Those substances that reach the minimum hazard levels require regulatory approval before they can be imported into or manufactured in New Zealand. Hazardous substances can also be subject to a range of different "performance requirements", depending on the extent of hazards they pose. These can relate to a range of matters, including minimum degrees of hazard, packaging and disposal. As with the medicines scheme, mechanisms exist for post-market regulation.

"Trusted third party"

A more "light-touch" role for a specialist agency has been proposed by researchers at the highly influential Oxford Internet Institute. Recognising that a challenge to transparency arises from "sensitivity of trade secrets and intellectual property rights", they suggest that a "solution would allow for examination of automated decision-making systems, including the rationale and circumstances of specific decisions, by a trusted third party". This role could be discharged by expanding the powers of an existing supervisory authority, or by creating a new (European) regulator specifically for this purpose (e.g. see Wachter et al. 2017b, pp. 43-44).

Regulatory phase

Proposals for new regulators have not met with universal approval in the wider area of AI. This scepticism has taken a number of forms. For some, the issue is one of what Brownsword calls "regulatory phase". Simply put, the time for such regulation has not yet arrived.

Chris Reed is one commentator who has expressed this sort of regulatory scepticism. In his view:

> "any regulatory body needs a defined field of operation, and a set of overriding principles on the basis of which it will devise and apply regulation. Those principles will be based on mitigating the risks to society which the regulated activity creates. Until the risks of AI are known, at least to some degree, this is not achievable." (REED 2018, P. 2)

It is not entirely clear why a detailed list of risks posed by a new technology would need be known before a regulator can act effectively. In as much as this argument has merit, though, it seems to apply to AI in the wider sense. The distinct subset with which we are concerned—predictive algorithms, particularly as used by government—does plausibly pose a range of known risks, as we have outlined in Chapter 4. That this list may not be exhaustive should not detract from their viability as a regulatory target.

Relations with other regulatory agencies

Another reason for scepticism about the creation of a new regulator is that a variety of regulators already exist, whose remit is likely to overlap with the new body. This sort of consideration led the House of Lords Select Committee on AI to conclude that

> "existing sector-specific regulators are best placed to consider the impact on their sectors of any subsequent regulation which may be needed (House of Lords Select Committee on Artificial Intelligence." (2018, [386])

In a New Zealand context, this would be likely to involve at least the Office of the Privacy Commissioner and the Human Rights Commission. It may well also overlap with the role of Stats NZ.

In our prelude to discussing legislative provisions, we identified a number of desiderata for regulation of AI regulation. One of these, parsimony, includes the following aims:

- existing laws that cover an identified problem should be maximally utilised before new laws are proposed;
- overlapping laws and regimes should be avoided where possible.

There is reason to believe that these are also valid considerations with regard to regulatory agencies. Indeed, in its Expectations for Good Regulatory Practice, the New Zealand Government has explicitly noted that regulatory systems should be "well-aligned with existing requirements in related or supporting regulatory systems through minimising unintended gaps or overlaps and inconsistent or duplicative requirements" (NZ Government 2017, p. 2). To this effect, all regulatory agencies should

develop working relationships with other regulatory agencies within the same or related regulatory systems to share intelligence and co-ordinate activities to help manage regulatory gaps or overlaps, minimise the regulatory burden on regulated parties, and maximise the effective use of scarce regulator resources (NZ Government 2017, p. 5).

Could the function of supervising and regulating predictive algorithms in government be discharged by these existing agencies? We consider that there would be a number of obstacles to this approach.

(i) Diffuseness: a government agency intending to use a new algorithm or to put an algorithm to a new use would be required to liaise with all of these separate bodies.

(ii) Capacity: as the Lords Select Committee noted, asking existing regulators to assume the burden of monitoring algorithms could place a substantial additional burden on them (2018, [386]).

(iii) Expertise: it is not clear that any of the existing regulators currently possess expertise sufficient to allow them to scrutinise an algorithm in the manner that would be needed to evaluate or confirm its accuracy, or check it for bias. This is no criticism of these agencies; this is, after all, a fair way from their current function.

For these reasons, we believe that a strong case can be made for the creation of a specialist regulator to address predictive algorithms in government. It should be kept in mind, though, that such a regulator need not be "scratch-built". It may be possible to entrust these new regulatory functions to an existing body or agency. For example, the Harmful Digital Communications Act 2015 provided for the establishment of an "approved agency" to act as a first stop for complaints. Rather than create this agency de novo, the role was given to Netsafe, an "independent, non-profit online safety organisation". While we make no specific recommendation in this respect, we note that vesting this role with an existing agency (presumably with a commensurate increase in resources) is an option to be kept in mind. In any event, we strongly suggest that the potential for collaboration between that new regulator and the existing bodies we have identified should be explored and pursued.

Further thoughts on regulatory agencies

Regulatory agencies have often formed part of the overall regulatory landscape for new and emerging technologies. They are widely considered to offer a range of advantages, including expertise, and the capacity to respond relatively quickly to unexpected developments. On the other hand, they lack the accountability of an elected legislature. Mechanisms exist to mitigate this potential accountability gap, including restricting the range of decisions such an agency can make, and a requirement to consult prior to making policy decisions.

Agencies come in a wide variety of forms. At the lightest touch end of the spectrum, their role will be confined to issuing best practice guidelines and codes of practice, and to responding to requests for expert advice. The importance of such functions should not be underestimated and the presence of such guidance can play an important role in promoting safe and ethical practice.

One possible role for such an agency in New Zealand would be in providing a pre-implementation "safety check" for government use of predictive algorithms. For example, technical experts could validate their accuracy and transparency, while legal and ethical members would consider potential human rights or privacy breaches. This could be carried out prior to using a new product, or to employing an existing product for a new purpose. Although deciding whether the product

purpose was sufficiently "new" to merit a separate check could itself be a matter of contention, this is not unique to this context (for a discussion of the regulatory challenge around classifying nanomaterials as "new" material, see Gavaghan and Moore 2011).

An important consideration with the "safety check" model is the requirement for regular follow-ups. As with many other new technologies, ostensibly innocuous uses of predictive algorithms could have unexpectedly adverse consequences, that may not become apparent for months or years. A regular "warrant of fitness" may mitigate against this concern.

Finally, a relatively light-touch regulator could act as a trusted third party, in the manner suggested by Wachter and colleagues at the OII. We note, though, that if government agencies continue to avoid using algorithms where intellectual property might conflict with transparency, this is not a function that would be required.

Regulatory bodies with harder edged remits might be able to *demand* that algorithms are submitted for "safety checks". We have considered two models that already exist in New Zealand, in the context of, respectively, medicines and hazardous substances. Both make provision for a process of consents and conditions prior to importing or marketing. Some readers may question the extent to which these models are particularly relevant to the context of algorithms. The dangers from medicines and hazardous substances may seem to be of a different order to those posed by predictive algorithms, such that any analogous system of pre-market approval could seem excessive. On the other hand, while the harms caused by use of biased or inaccurate algorithms may not be as immediately obvious as those caused by unsafe medicines or flammable or toxic substances, they could be very substantial. When decisions relate to detention or release of prisoners, for example, or removing children from their families, the stakes are high.

Perhaps there is something to be said for a HSNOA-type model, whereby the first level of evaluation is to determine whether a substance (or in our context, an algorithm) presents new risks that reach a minimum threshold. The extent of regulatory scrutiny thereafter will depend on the outcome of that first evaluation. This would tally with Tutt's suggestion that "only subject the most opaque, complex, and dangerous types" of algorithms should be subject to regulatory scrutiny (Tutt 2017, p. 107). How that first level of "pre-regulatory" evaluation could be conducted is something to which we return in the next section.

Another question arises as to whether this should apply only to new products and purposes, or whether those predictive algorithms that are already in use should also be within scope. Given the fairly small number of algorithms identified in the government's *Algorithm Assessment Report*, we note that extending this power to those already in use may not be especially onerous for the new agency at this stage. It may also be that most or all of the algorithms currently in use would satisfy the first stage of pre-regulatory evaluation.

If it is decided to create or empower a regulatory agency with anything but the softest of edges, consideration will need to be given to its capacity to ensure its decisions are followed. As has been pointed out in relation to another emerging technology:

> "it should be borne in mind that even the most comprehensive regulatory framework will be an ineffective safeguard of public health if no effective mechanism exists to monitor and enforce compliance with it. This is what we identified as a third level regulatory gap." (GAVAGHAN & MOORE 2011)

This may involve provision for post-deployment monitoring.

C. Self-regulatory models

The creation and deployment of self-checking frameworks within agencies may have a valuable role to play, though we believe that this would be in conjunction with rather than as an alternative to independent regulatory oversight. We have already considered one overseas example earlier in the report: the ALGO-CARE framework used by Durham Constabulary alongside the HART tool (see Section 2B). But New Zealand researchers have created their own framework for use within government.

The "PHRaE"

The Privacy, Human Rights and Ethics Framework ("PHRaE") was developed by the Ministry of Social Development ("MSD") in conjunction with Professor Tim Dare from Auckland University. It is a questionnaire designed to aid project teams within the MSD developing and deploying new algorithmic tools intended to support operational decision-making. The PHRaE is interactive, requesting information from developers depending on details of the deployment and design of the tool under development. The MSD describes it as:

> "...a structured way of asking the right questions to make sure that we take into account privacy, human rights, and ethics from the very beginning of designing new services that are using personal information. This enables PHRaE risks to be designed out rather than risk being a barrier to implementation (or having to be accepted)."

The areas covered in the framework are:

- The intention behind any new use of data or the development of a new algorithm;
- The likely benefits and harms and to whom they would accrue;
- Whether the new use of personal information is necessary;
- Legal restrictions on the use of information already held by the MSD;
- Whether personal information would be used for the purpose for which it was collected;
- How and from whom new information would be collected;
- Design to ensure information is kept safe;
- Ensuring information used to make decisions is accurate;
- Whether people would be able to access their information;
- Whether the initiative would discriminate against some people;
- How this new use of information would be communicated to data subjects; and
- Whether personal information would be shared with others and if so why.

The interactive PHRaE tool is extremely detailed, requiring considerable research and analysis from developers as they progress through the design process. It also contains excellent resources explaining complex principles and technical and philosophical ideas required to assess the privacy, human rights, and ethical impact of new uses of data.

Evaluation

The *Algorithm Assessment Report* describes the PHRaE as an example of "good practice" in meeting the requirement that those deploying data and analytics be aware of the limitations of such tools set out in the Principles for Safe and Effective use of Data and Analytics (Stats NZ 2018, p. 29). We agree and note that it will help developers of predictive algorithms for government use meet all the Principles:

- Delivering clear public benefit;
- Maintaining transparency;
- Understanding limitations of various forms of data use;
- Retaining human oversight;
- Ensuring data is fit for purpose; and
- Focusing on people.

In assessing the PHRaE and other similar self-regulation frameworks, there are two fundamental questions to be answered. The first is: "should government departments use them". Related to this is: "How effective are they in addressing desiderata for regulation of government use of predictive algorithms?" The second question is about scope: "how many of the desiderata are addressed by this type of self-regulatory mechanism?"

One of the great advantages of the PHRaE is that it addresses a very broad array of issues very early on in the life of a predictive algorithm. This allows some problems to be "designed out" of the software and others to be addressed by the development of business rules that maximise benefit and minimise harm. Use of the interactive PHRaE tool provides a detailed record

of decisions made and problems addressed during the development phase. Its use is also likely to produce productive discussion between data scientists and policy analysts on exactly the sorts of issues on which there is likely to be misunderstanding between them. Use of the tool is likely to be a significant task for software developers but it is not so onerous as to stifle innovation. Compulsory use of the PHRaE within MSD will effectively provide a set of standards tailored to the work that that MSD does and to the particular contexts in which its employees make operational and strategic decisions. Moreover, while the PHRaE is designed for use by the MSD, tools like this could be used in a wide variety of government contexts.

So our short answer to the question of whether Ministries like MSD should use tools like the PHRaE, is "yes". That said, in-house self-regulatory tools like the PHRaE inevitably lack some of the advantages of other forms of regulation. As such, it could be that these work best in conjunction with, rather than as an alternative to, regulatory oversight. We also note the PHRaE is new and results not yet externally reviewed, so care should be taken not to over-emphasize it as a tool which can or does enable agencies to deal with all of the issues which arise. In addition, its detailed application may need specialist expert support (along with necessary resources). The efficacy of the PHRaE might be a useful area for further research, including to assess if changes or improvements are needed.

Addressing issues early in the design process has both pros and cons. Some issues that arise from the use of new predictive algorithms will be difficult to identify before the new algorithm is put to use. While we can predict a general risk for an algorithm to cause a feedback loop exacerbating existing inequality, the actual nature and extent of such harms can only be assessed by audit of the tool once it has been put to use and close study of its effects within the populations to which it is applied. These sorts of issues could be addressed by requiring regular review of algorithms in use including empirical study of risks originally "estimated" during the design process.

As noted in our Introduction, particular attention needs to be paid to government's obligations to Māori. This includes Māori views on the evaluation of the use of algorithms, particularly where this has been in areas of service delivery that disproportionately affect them. A general question for researchers in response to these issues is to consider whether the increasing use of algorithms exacerbates, reduces or disguises social inequalities, particularly for Māori. This will often not be a straightforward matter to evaluate but measures can be taken to ensure that such risks are at least identified and monitored, and mitigated or avoided as far as possible. The Algorithm Assessment Report found little, if any, evidence of consultation about algorithmic use with Māori affected by algorithmic use. This is a significant gap which, as we have said, must be addressed and which we consider could usefully be done now, before algorithmic use becomes more widespread.

Regular reviews and other assessments should also take into account the government's commitment to reflecting a Treaty of Waitangi based partnership with Māori in its practice. More work needs to be done to assess whether, and if so how, a te ao Māori perspective can be embedded into the development, use and evaluation of algorithms. This includes, how "the taonga status of data that relates to Māori" is reflected. Given the Ministry of Social Development's role in leading development of social investment policies (which disproportionately affect Māori) there is an opportunity for the Ministry to improve and strengthen its in-house work by engaging with Māori, including Māori data scientists and other experts, when using the PHRaE.

Normative disconnection (see discussion in Section 1A) occurs where a new technology is put to an unanticipated use, particularly one that poses different risks or ethical concerns. This is what would have happened had the YORST (see Section 1D) been employed for assigning young offenders into boot camps. The ethical implications of repurposing existing tools might be addressed to some extent by requiring that new uses of existing tools trigger a new assessment using the PHRaE (or other similar tools). It is unclear whether such a restriction would be politically viable.

In-house tools are of course not as visible as black letter law. So their operation is less transparent to the public in general and data subjects in particular. While they de facto provide well designed and flexible standards for the development of predictive algorithms, those standards are not visible to the community in a way that would provide certainty about government use of data and analytics. As such they only provide limited support for appeals of the use of particular algorithmic tools developed by government.

Finally, many of the issues addressed in this report set out in Chapter 4 are complex and their possible future effects on New Zealanders are poorly understood. New Zealand needs to decide how it will address issues such as algorithmic bias, hyper-surveillance, control, and transparency. As such we are not yet in a position to design these problems out of the tools we develop.

How might self-checking frameworks like PHRaE work with external and independent regulatory oversight? Further work will be required to consider precisely how this could be operationalised, but one possibility would look like this.

1. A government agency seeking to design or purchase a new algorithmic tool, or to use an existing tool for a new purpose, would first pass their proposal through a framework like PHRaE. This could either be a framework that is common across all government agencies; a framework that is tweaked for particular uses; or a framework designed for use within a particular agency. The new regulator could work with government agencies to help devise appropriate frameworks for this purpose.
2. That process would culminate in the production of a report on the proposed new algorithm/use, that would address issues such as accuracy, privacy, human rights compliance and transparency.
3. If the agency's internal review is not satisfied with the outcome of this internal assessment, then the proposal will be revised and resubmitted.
4. If/once the agency is satisfied, a report would be passed to the new regulator. This will address concerns with privacy, human rights compliance and transparency. In addition, it should supply an accuracy rating for the algorithm/proposed use.
5. If the regulator takes the view that the new tool/use posed no new risks, or if it takes the view that provision was being made for those risks to be managed adequately, no further pre-procurement/deployment regulatory steps will be required.
6. If the regulator is not so satisfied, then further information could be demanded, or conditions imposed on the use of the algorithmic tool. In presumably rare cases, permission to construct, purchase or use the new tool could be denied altogether.
7. The regulator will maintain a register of uses of predictive algorithms within government agencies. Those agencies will be required to conduct ongoing assessments of the use of those algorithms, and submit reports to the regulator at regular intervals—either every year or three years as required by the regulator.
8. The regulator will produce an annual public report on use of predictive algorithms within government. This report will make public the uses of predictive algorithms in its register, including input and output variables for each algorithm, with exceptions made in cases where this knowledge would enable "gaming" of the algorithm.

CONCLUSIONS AND RECOMMENDATIONS

STARTING ASSUMPTIONS

It is important to consider both the opportunities offered and the concerns presented by the use of predictive algorithms in government. It is also important, though, not to compare them with notionally perfect human decisionmakers. Human beings are subject to cognitive biases, logical fallacies, and a wide array of prejudices and errors. Equally, we must keep in mind that algorithms are not being introduced into a system that is in any sense perfect. In New Zealand, in common with the rest of the world, ethnicity, gender and social class are highly predictive of health and longevity, employment and economic security, and relationships with the legal system, including likelihood of imprisonment. The nature of these various relationships is complex and contested, but they have not been introduced de novo by the use of algorithmic decision-making.

Neither is the use of predictive algorithms within the New Zealand government sector entirely a new phenomenon. As we have shown, algorithms such as RoC*RoI have been in use for decades. However, the increasing use of these tools, and their increasing power and complexity, presents a range of concerns and opportunities. These include the potentially vast amount of information that can be factored into decisions; the opacity of the decision-making process; and the veneer of scientific objectivity that can cover the results.

We have sought to approach this topic from as neutral a starting position as we can: neither welcoming of nor hostile to the use of algorithmic tools, but keen to explore how their advantages can be maximised and their risks either avoided altogether or at least mitigated.

SCOPE

"Algorithms" come in many forms and variable degrees of complexity and transparency. We have suggested that law/regulation/ethical analysis should sometimes be targeted at the level of use/potential harm, rather than on the precise form of technology, as the latter approach risks regulatory disconnection. Nonetheless, the general concept of a "predictive algorithm" is useful for many regulation/oversight purposes, and covers a useful subset of the algorithms referred to as "AI" in recent public discourse.

ACCURACY

There should be independent and public oversight of the accuracy of the predictive models being used in government. This is of central importance, but such information is not yet readily or systematically available.

Acknowledging that algorithms of various types are used in many different circumstances generating diverse ethical concerns, government agencies should not employ single standards (such as AUC thresholds) for different algorithms used in different circumstances. Context-aware processes like the PHRaE are instructive.

Government should not assume that algorithms used to deliver positive interventions can be subject to lower levels of scrutiny than those delivering punitive or onerous interventions. At sufficiently low levels of accuracy, significant numbers of people in need of a benefit may fail to receive it. While tuning the algorithm to err on the side of false positives may be one way to address this problem, it will not always be financially feasible.

CONTROL AND HUMAN INPUT

Solutions such as requiring a "human in the loop" have an obvious appeal. The Government's *Algorithm Assessment Report* makes much of the role of human oversight with existing algorithms, and many commentators have spoken favourably of the right not to be subject to automated decisions within the GDPR. Our research, however, has led us to be more cautious about this approach.

In particular, there is a risk that, if we do not approach them carefully, such guarantees could serve as little more than regulatory placebos. Given the well-researched tendency of humans within human-machine systems to overestimate the value of a machine's outputs, a human in the loop may add nothing more than token reassurance.

In other situations, the addition of a human factor to an automated system may have a detrimental effect on that system's accuracy.

Nonetheless, there are situations where human involvement in automated decision-making is certainly valuable. This may be where the automated systems are not reliable enough to be left to operate independently, where factors need to be considered that are not readily automatable, or in situations where a measure of discretion is for whatever reason desirable.

Context-specific requirements for human oversight, such as is provided for in the Court Matters Bill, may be a more proportionate and useful response.

If a general right to human involvement were deemed to be desirable, such provision should be accompanied by a "right to know" that automated decision-making is taking place, akin to Articles 13-15 of the GDPR. Without such information, a right to demand human oversight would be meaningless to most affected parties.

Where human involvement or oversight is considered desirable, a number of measures can be employed that may reduce the risks of "automation bias". Guidelines recommending that decision-makers exercise their own judgment *before* consulting an algorithm could assist in offsetting some of the effects of automation complacency and bias. In these cases the algorithm would serve merely as a check on a decision-maker's intuitions. Guidance may also have a role to play. The Wisconsin Supreme Court in *Loomis* required that sentencing judges be given a list of warnings about COMPAS if they intend to have its predictions inform their decisions. While these measures are worth exploring, more empirical (human factors) research is required to see whether such approaches really do work.

A legal obstacle to automated decisions may arise in the public sector context, where statutory powers generally cannot be delegated or fettered without parliamentary approval. Statutory authorities that use algorithmic tools as decision aids must be wary of improper delegation to the tool, or otherwise fettering their discretion through automation complacency and bias. In those (rare) cases where systems are reliable enough to be used because they reach a better-than-human threshold, public sector bodies can get approval through New Zealand's general statutory delegation provisions (i.e. State Sector Act 1988, Crown Entities Act 2004 and Local Government Act 2002).

TRANSPARENCY AND A RIGHT TO REASONS/EXPLANATIONS

New Zealand law already provides for a right to reasons for decisions by official agencies, primarily under section 23 of the Official Information Act. This is supported by judicial authority that such reasons must be understandable, both to a review body, to someone with vested interests in the decision and at least in some cases to the public at large.

To ensure that agencies can comply with this requirement, policies should be adopted to ensure that algorithms are either developed "in house", or, when purchased from outside vendors, acquired on terms that allow for transparency, so that neither their form nor conditions of sale preclude or obstruct details of the algorithm being made publicly available.

To this end, government agencies' procurement policies should give preference to companies which are open (i.e. publish information) about their algorithms, rather than those who hide behind proprietary code.

Review of new algorithms or new uses of existing algorithms should pay particular attention to provision of explanations. It should be kept in mind that the form and level of detail of an explanation will be context-specific, and agencies should be prepared to explain decisions at both lay and expert levels.

BIAS, FAIRNESS AND DISCRIMINATION

"Fairness" can take several forms. It may be impossible to satisfy all definitions simultaneously. Government agencies should consider the type(s) of fairness appropriate to the contexts in which they use specific algorithms.

Exclusion of protected characteristics from training data or input variables does not guarantee that outcomes are not discriminatory or unfair. For example, other variables can serve as close proxies for protected characteristics, and input data that appears innocuous can nonetheless be tainted by historic discrimination.

Awareness of error rates is important, but it must be remembered that not all errors are equal; some impact disproportionately on certain parts of the population. Quite often, those will be parts with little political or economic power. In New Zealand, this is likely to include Māori people, as well as those in a range of vulnerable situations.

A regulatory response must include processes for addressing the risks of algorithmic bias due to historical injustices or to inaccuracies in existing datasets. Further research needs to be done to explore possible strategies for ameliorating such risks.

INFORMATIONAL PRIVACY

In the realm of privacy and data protection law, we recommend that effect be given to more specific requirements to identify the purpose of collection of personal information (information privacy principle 3).

Doubts persist about the status of inferred information. New Zealand is not unique in this respect; even in the post-GDPR European Union, a "right to reasonable inferences" has recently been proposed. Further consideration should be given to the adequacy of New Zealand law in this respect. Parliament should consider both a right to reasonable inferences (and a cognate right of *access* to these inferences), and whether inferred data should be afforded the same protections as primary data.

New Zealand should also consider introducing better protections regarding re-identification, de-identification, data portability and the right to be forgotten (erasure).

THE LIMITATIONS OF INDIVIDUAL RIGHTS MODELS

Individual rights are of course vital for any democracy, but we should be wary of relying exclusively on individual rights models that depend on affected parties holding predictive algorithms to account. Often, they will lack the resources to do so. Furthermore, individual rights models might offer limited efficacy in monitoring group harms.

OVERSIGHT

Government agencies should adopt or develop in-house processes to evaluate proposals to develop or procure new predictive algorithms. These should also apply when it is proposed to apply existing algorithms to a new purpose. These processes should evaluate a range of considerations, including accuracy, transparency, privacy and human rights impacts.

The PHRaE process being developed at MSD could serve as an instructive example in this regard, though we are aware that further research will be required on its use and efficacy. It is also likely that different agencies will require purpose built frameworks that are responsive to the particular concerns that arise in their contexts.

Internal processes should be sufficiently thorough to alleviate concerns, but this can be burdensome for staff having to navigate these processes. Provision should be made in workload models for this, and if necessary, training should be provided in the use of such tools.

Government should consider the establishment of a regulatory/oversight agency. This would work with individual government agencies who intend either to introduce a new predictive algorithm, or to use an existing predictive algorithm for a new purpose.

We have considered several possible models for the new regulatory agency. These all have strengths and weaknesses, but at this time we offer no detailed proposal as to the form it should take. At present, there are very few international examples from which to learn, and those which exist (such as the UK's CDIE) are in very early stages. We would welcome the opportunity to discuss this further with government and other regulatory agencies, and to contribute to the next stage of discussion about this.

We have proposed a possible structure for how the new regulatory agency could work with government agencies. This would involve

- a report from the internal review process described above to be provided to the new regulator;
- where new regulator decides that no new or non-trivial risks are being posed, or that they are being adequately managed, no further action will be required;
- where the new regulator is not so satisfied, it will be able to require answers to questions or additional steps to be taken;
- in presumably rare instances, the regulator will be able to deny permission for the new algorithm to be designed or acquired, or used in the manner proposed.

The new regulator could serve range of other functions, including

- Producing best practice guidelines;
- Maintaining a register of algorithms used in government;
- Producing an annual public report on such uses;
- Conducting ongoing monitoring on the effects of these tools.

We consider the last point to be important. The nature of these tools is such that a snap-shot assessment will be insufficient to ensure that concerns about, for example, control and bias are being adequately addressed.

Our preference is for a relatively "hard-edged" regulatory agency, with the authority to demand information and answers, and to deny permission for certain proposals. However, even a light-touch regulatory agency could serve an important function. The recent *Algorithm Assessment Report* acknowledged use of algorithms across NZ government to be somewhat piecemeal.

If a regulatory agency is to be given any sort of hard-edged powers, consideration will need to be given to its capacity to monitor and enforce compliance with these.

If the agency is to be charged with scrutinising algorithms, it must be borne in mind that these are versatile tools, capable of being repurposed for a variety of uses. Scrutiny should apply to new uses/potential *harms* and not only new *algorithms*.

CONSULTATION

We stress the need for consultation with a wide range of stakeholders across New Zealand society, especially with populations likely to be affected by algorithmic decisions, and with those likely to be under-represented in construction and training. This is likely to include those in lower socio-economic classes, and Māori and Pacific Island populations. Quite simply, they are likely to have insights, concerns and perspectives that will not be available to even the most well-intentioned of outside observers.

APPENDIX 1: THE YOUTH OFFENDING RISK SCREENING TOOL

New Zealand POLICE
Nga Pirihimana O Aotearoa

YOUTH OFFENDING RISK SCREENING TOOL

NAME							NIA Person ID No:		
(Child/YP):	Surname			First name(s)			File no:		
DOB		Age		Gender	Male	Female	Date RST Completed	by (QID)	
ETHNICITY	European	Pacific	Asian	Other					
	Maori	Iwi		Hapu					
Incident / Offence Code		Incident / Offence Description							

Part (A) Offending Factors

1 Time since last came to Police notice for their **offending**?

No previous	Over 2 yrs	1 to 2 yrs	Less than 1 yr	1 to 6 mths	Under 1 mth
0	1	2	3	4	5

D

2 Time since last came to Police notice for **incidents** (e.g. 1J, 2M, 1T) relating to them and/or serious behaviour incident at school?

No previous	Over 2 yrs	1 to 2 yrs	Less than 1 yr	1 to 6 mths	Under 1 mth
0	1	2	3	4	5

D

3 Highest level of previous intervention? (final outcome)

No previous	Noting	Warning	Alt. Action	FGC	Youth Court
0	1	2	3	4	5

4 At what age was offending first reported to Police (if first offence use current age)?

No offences	15+	14	13	10 to 12	Under 10
0	1	2	3	4	5

5 Rate the seriousness of the current primary offence using the youth offence rating tool (see A4 list).

Minimum	Minimum / Medium	Medium	Medium / Maximum	Maximum
1	2	3	4	5

D

6 Is the nature (MO) of current or previous offending of a concerning nature?

Very Low	Low	Medium	High	Extreme
1	2	3	4	5

D

Comments re Question 6:

Part (B) Peer Group Factors

7 Influential peers known to Police?

None	Very few known	Some known	Many known	All known repeat offenders	Unknown
0	1	3	4	5	0

D

Part (C) Education / Employment Factors (contact the school, but not the employer)

8 Current school / education / course or employment status

Full time well engaged	Full time some issues	Mostly attends	Irregular attendance	Stood down / suspended	Not attending (school / job)	Unknown
0	1	2	3	4	5	0

D

Part (D) Care & Protection History

9 Has a notification been made to CYF for this family or child / young person?

No	Notification concerning another sibling	Notification concerning this child / young person	Some form of intervention provided by Child, Youth & Family	Currently / previously in the custody of CYF (101 status)
0	2	3	4	5

Part (E) Alcohol and/or Drug Use

10 Is their use of alcohol or drugs causing concern? (consider the long term effects of the type of drugs used).

No concern	Slight	Moderate	Serious	Very Serious	Unknown
0	1	2	4	5	0

D

Part (F) Family Factors

11 If there are FAMILY VIOLENCE records in NIA for this family / address, what is the highest FV score?

Zero Records	Records, but no score	Score from 1 - 8	Score from 9 - 16	Score 17 or over
0	2	3	4	5

12 Where do they live? (socio economic area decile rating of local state primary school)

8 - 10	4 - 7	2 - 3	1	Transient / Motor Camp
0	2	3	4	5

D

13 Are there concerns in the living situation? e.g. parent / caregiver support and supervision of child / young person, parental mental health problems, drug and alcohol use, suspected child abuse and / or unrecorded family violence

None	Very minor concerns	Some concerns	Major concerns	Some major concerns	Young Person Transient	Unknown
0	1	2	3	4	5	0

D

Detail Concerns:	

14 Family members have offending history?

None	Parent(s) with minor history	Parent/s with major history (imprisonment)	Parent(s) have offended within past 12 months	Sibling(s) have offended within last 12 months	Unknown
0	2	3	4	5	0

Any General Comments:	

Information Sources

Spoken To	Child / young person	Parent / caregiver	School / course provider / MOE	Child Youth & Family	Other agency
This time					
Previously					
Not At All					

Scoring Instructions — Risk Screening

Questions				Answers			YORST Score	
No. of Questions			Max	Sum of the Scores (Above)	=		x 100 =	%
Not Answered:		Answered:	x 5	Max. Total for Answered Questions	=			

Dynamic Risk Factors

Static Factor Results		Dynamic YORST Score	
Sum of Dynamic Factors		x 100 =	%
Maximum Possible Total for Dynamic Factors	45		

Youth Aid Response — Your Station

Warning	AA	FGC	Youth Court	Police Youth Development	Other	

APPENDIX 2: ROC*ROI INPUT VARIABLES

Regression Variable (Risk of Reconviction)	Description/weighting
MaleFirstOffenderFree13	Log of time not in prison since age 13 for male first time offender
MaleReoffenderFree13	Log of time not in prison since age 13 for male re-offender
FemaleFirstOffenderFree13	Log of time not in prison since age 13 for female first time offender
FemaleReoffenderFree13	Log of time not in prison since age 13 for female re-offender
MaleEpisodesFree	Log of time between the two most recent sentence periods ("episodes") of male offenders
FemaleEpisodesFree	Log of time between the two most recent sentence periods ("episodes") of female offenders
Reoffender	Is this not a first offence – value 1 for yes, 0 for no
MaleReoffender	For males, is this not a first offence, for females, is this a first offence
MaleReoffenderSerious	History of offending based on seriousness, for males
FemaleReoffenderSerious	History of offending based on seriousness, for females
FemaleFirstOffenderSerious	Seriousness of offence for first time female offender
MaleFirstOffenderSerious	Seriousness of offence for first time male offender
MaleEpisodeCount	Total number of sentence periods ("episodes"), for male offender
FemaleEpisodeCount	Total number of sentence periods ("episodes"), for female offender
FemaleOffendingRate	Number of sentence periods / time not in prison, for females
MaleOffendingRate	Number of sentence periods / time not in prison, for males
PreAge13Offence	Any offence committed prior to age 13 – value 1 for yes, 0 for no
Male	Is offender male – value 1 for yes, 0 for no
FemaleReoffenderDrive	Is current most serious offence a traffic offence, and is it not a first offence for female offender – value 1 for yes, 0 for no
MaleReoffenderDrive	Is current most serious offence a traffic offence, and is it not a first offence for male offender – value 1 for yes, 0 for no
FemaleFirstOffenderDrive	Is current most serious offence a traffic offence, and is this a first offence for female offender – value 1 for yes, 0 for no
MaleFirstOffenderDrive	Is current most serious offence a traffic offence, and is this a first offence for male offender – value 1 for yes, 0 for no

Source: Blackmore 2019.

REFERENCES

Cases

Belcher v Chief Executive of the Department of Corrections [2007] 1 NZLR 507

Computer Associates UK Ltd v The Software Incubator Ltd [2018] EWCA Civ 518

Ministry of Health v Atkinson [2012] NZCA 184

Naidu v Australasian College of Surgeons [2018] NZHRRT 234

PCC276 Case Note 205558 [2010] NZPrivCmr 1

R v Hansen [2007] NZSC 7

R v Peta [2007] NZCA 28.

Re Vixen Digital Limited [2003] NZAR 418

Wisconsin v Loomis 881 N.W.2d 749 (Wis. 2016)

Statutes

NZ

Bill of Rights Act 1990

Court Matters Act 2019

Crown Entities Act 2004

Data Protection Act 2018

Hazardous Substances and New Organisms Act 1996

Human Assisted Reproductive Technology Act 2004

Human Rights Act 1993

Local Government Act 2002

Medicines Act 1981

Official Information Act 1982

Privacy Act 1993

State Sector Act 1988

Australia

Migration Act 1958 (Cth)

Privacy Act 1988 (Cth)

UK

Consumer Protection Act 1987

EU

General Data Protection Regulation 2016

Other

Stats NZ (2018) *Algorithm assessment report*. Available at: https://data.govt.nz/use-data/analyse-data/government-algorithm-transparency

Accident Compensation Corporation (2018a) *Improving the claim registration and approval process. Version 1.0.* 4 July 2018.

–––(2018b) *Statistical models to improve ACC claims approval and registration process. Version 1.0.* 21 August 2018.

AI Forum (2018) *Artificial intelligence: Shaping a future New Zealand.* Available at: https://aiforum.org.nz

AI Now (2018) *Litigating algorithms: Challenging government use of algorithmic decision systems*. Available at: https://ainowinstitute.org/litigatingalgorithms.pdf

Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D. & Lampos, V. (2016) Predicting judicial decisions of the European Court of Human Rights: A Natural Language Processing perspective. *PeerJ Computer Science* 2(93): 1-19.

AlgoAware (2018) *State of the art report: Algorithmic decision-making. Version 1.0*. December 2018.

Allport, G.W. (1954) *The nature of prejudice.* Cambridge, MA: Addison-Wesley.

American Civil Liberties Union et al. (2016) Predictive policing today: A shared statement of civil rights concerns. August 31. Available at: https://www.aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M. & Rudin, C. (2017) Learning certifiably optimal rule lists. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2017: 35-44.

Angie, A. D., Connelly, S., Waples, E. P., & Kligyte, V. (2011) The influence of discrete emotions on judgement and decision-making: A meta-analytic review. *Cognition and Emotion* 25(8): 1393-1422.

Aronson & Dyer (2013) *Judicial review of administrative action.* 5th edition. Sydney: Lawbook Co.

Bainbridge, L. (1983) Ironies of automation. *Automatica* 19(6): 775-779.

Bakker, L., O'Malley, J. & Riley, D. (1999) *Risk of reconviction: Statistical models predicting four types of reoffending* (Wellington: Department of Corrections).

Banks, V.A., Erikssona, A., O'Donoghue, J. & Stanton, N.A. (2018a) Is partially automated driving a bad idea? Observations from an on-road study. *Applied Ergonomics* 68: 138-145.

Banks, V.A., Plant, K.L. & Stanton, N.A. (2018b) Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. *Safety Science* 108: 278-285.

Barocas, S. & Selbst, A.D. (2015) Big data's disparate impact. *California Law Review* 104: 671-732.

Baxter, G., Rooksby, J., Wang, Y. & Khajeh-Hosseini, A. (2012) The ironies of automation…still going strong at 30? Proc. ECCE Conf. Edinburgh, Aug., pp. 65-71.

Bennett Moses, L. (2013) How to think about law, regulation and technology: Problems with "technology" as a regulatory target. *Law, Innovation and Technology* 5(1): 1-21.

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016) A meta-analytical integration of over 40 years of research on diversity training evaluation. Available at: http://scholarship.sha.cornell.edu/articles/974

Black, J. (2002) Critical reflections on regulation. *Australian Journal of Legal Philosophy* 27: 1-35.

Blackmore, B. (2019) *Developing transparency requirements for the operation of criminal justice algorithms in New Zealand.* Unpublished MA dissertation, University of Otago.

Blomberg, T., Bales, W., Mann, K., Meldrum, R. & Nedelec, J. (2010) Validation of the COMPAS risk assessment classification instrument. Center for Criminology and Public Policy Research College of Criminology and Criminal Justice Florida State University. Available at: https://arxiv.org/pdf/1311.2901.pdf

Boston, J. & Gill, D. (2017) *Social Investment: A New Zealand policy experiment.* Wellington: Bridget Williams Books.

Brennan Center for Justice (2018) Public deserves to know how NYPD uses predictive policing software. January 26. Available at: https://www.brennancenter.org/blog/court-rejects-nypd-attempts-shield-predictive-policing-disclosure

Bridge, M. (2017) AI can identify Alzheimer's disease a decade before symptoms appear. *The Times*, Sep. 20.

Brownsword, R. and Goodwin, M. (2012) *Law and the technologies of the twenty-first century.* Cambridge: Cambridge University Press.

Brynjolfsson, E. & McAfee, A. (2017) *Machine platform crowd: Harnessing our digital future.* New York: Norton.

Burrell, J. (2016) How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data and Society* 3(1): 1-12.

Butler, A. & Butler, P. (2015) *The New Zealand Bill of Rights Act: A commentary.* 2nd edition. Wellington: LexisNexis.

Coravos, A., Chen, I., Gordhandas, A. & Stern, A.D. (2019) We should treat algorithms like prescription drugs. *Quartz* February 15.

Cavoukian, A. & Castro, D. (2014) Big data and innovation, setting the record straight: Deidentification does work. Available at: http://www2.itif.org/2014-big-data-deidentification.pdf

Cebon, D. (2015) Responses to autonomous vehicles. *Ingenia* 62: 10.

Chowdhury, H. (2018) Kent Police stop using crime predicting software. *The Telegraph* November 27.

Chopra, S. & White, L.F. (2011) *A legal theory for autonomous artificial agents.* Ann Arbor: University of Michigan Press.

Chouldechova, A. (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Bog Data* 5(2): 153-163.

Clarke, M. (2018) Why actuaries are essential to government. UK Civil Service blog post Available at: https://civilservice.blog.gov.uk/2018/12/04/why-actuaries-are-essential-to-government

Corbett-Davies, S. & Goel, S. (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. Available at: https://arxiv.org/pdf/1808.00023.pdf

Corbett-Davies, S., Pierson, E., Feller, A. & Goel, S. (2016) A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post* October 17, 2016.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. (2017) Algorithmic decision making and the cost of fairness. Available at: https://arxiv.org/pdf/1701.08230.pdf

Couchman, H. (2018) *Policing by machine: Predictive policing and the threat to our rights.* London: Liberty.

–––Coughlan, T. (2018a) What becomes of Social Investment? Newsroom February 5. Available at: https://www.newsroom.co.nz/2018/02/04/80935/what-becomes-of-social-investment

(2018b) What now for English's data-crunching agency? *Newsroom* June 22. Available at: https://www.newsroom.co.nz/2018/06/21/127937/wellbeing-focus-for-english-s-data-crunching-agency

Crawford, K. (2016) Artificial intelligence's white guy problem. *New York Times* June 25, 2016.

Crawford, K. & Calo, R. (2016) There is a blind spot in AI research. *Nature* 538: 311-313.

Crenshaw, K. (1991) Mapping the margins: Intersectionality, identity politics and violence against women of colour. *Stanford Law Review* 43: 1241-1299.

Cummings, M.L. (2004) Automation bias in intelligent time critical decision support systems. *AIAA 1st Intelligent Systems Technical Conf.* (https://doi.org/10.2514/6.2004-6313).

Cunningham, M. & Regan, M. (2018) Automated vehicles may encourage a new breed of distracted drivers. *The Conversation*, Sep. 24.

Damaška, M.R. (1997) *Evidence law adrift.* New Haven: Yale University Press.

Danziger, S., Levav, J. & Avnaim-Pesso, L. (2011) Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108(17): 6889–6892.

De Baets, S. & Harvey, N. (2018) Forecasting from time series subject to sporadic perturbations: Effectiveness of different types of forecasting support. *International Journal of Forecasting* 34: 163-180.

Destremau, K. & Wilson, P. (2017) Defining Social Investment: Kiwi-style. In: S*ocial Investment: A New Zealand policy experiment*, eds. J. Boston, & D. Gill, pp. 35-73. Wellington: Bridget Williams Books.

Diakopoulos, N. (2015) Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3(3): 398–415.

DoC (2009) *What Works Now? A review and update of research evidence relevant to offender rehabilitation practices within the Department of Corrections.* Available at: https://www.corrections.govt.nz/__data/assets/pdf_file/0011/779006/What_Works_Now_Final_December_2009.pdf

Dressel, J. & Farid, H. (2018) The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4: 1-5.

Dubnick, M.J. (2014) Accountability as a cultural keyword. In: *The Oxford handbook of public accountability,* eds. M. Bovens, R.E. Goodin & T. Schillemans, pp. 23-38. New York: Oxford University Press.

Dutta, S. (2017) Do computers make better bank managers than humans? *The Conversation* October 17, 2017.

Edwards, L. & Veale, M. (2017) Slave to the algorithm? Why a "right to an explanation" is probably not the remedy you are looking for. *Duke Law and Technology Review* 16(1): 18-84.

–––(2018) Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"? *IEEE Security & Privacy* 16(3): 46-54.

Ensign, D. Friedler, S.A., Neville, S. Scheidegger, C. & Venkatasubramanian, S. (2018) Runaway feedback loops in predictive policing. *Proceedings of Machine Learning Research* 81: 1-12.

Eubanks, V. (2017) *Automating inequality: How high-tech tools profile, police, and punish the poor.* New York: St Martin's Press.

Fienberg, S. (2006) When Did Bayesian Inference Become "Bayesian"? *Bayesian Analysis* 1: 1-40.

Frankel, S. & Yeabsley, J. (2011) Introduction. In S. Frankel, ed. *Learning from the past, adapting for the future: Regulatory reform in New Zealand.* LexisNexis.

Forssbæck, J. & Oxelheim, L. (2014) The multifaceted concept of transparency. In: T*he Oxford handbook of economic and institutional transparency,* eds. J. Forssbæck & L. Oxelheim, pp. 3-31. New York: Oxford University Press.

Friedman, B. & Nissenbaum, H. (1996) Bias in computer systems. *ACM Transactions on Information Systems* 14(3): 330–347.

Fuller, L. (1964) The morality of law. New Haven, CT: Yale University Press.

Gavaghan, C. & Moore, J. (2011) *A review of the adequacy of New Zealand's regulatory systems to manage the possible impacts of manufactured nanomaterials.*

Glazebrook, S. (2010) Risky business: Predicting recidivism. *Psychiatry, Psychology and Law* 17(1): 88-120.

Goel, S., Rao, J., & Shroff, R. (2016) Personalized risk assessments in the criminal justice system. *American Economic Review: Papers & Proceedings* 106(5): 119–123.

Gottredson, D. M. & Snyder, H. N. (2005) The mathematics of risk classification: Changing data into valid instruments of juvenile courts. Washington, D.C.: Department of Justice, Office of Juvenile Justice and Delinquency Prevention.

Griffiths, J. (2016) New Zealand passport robot thinks this Asian man's eyes are closed. *CNN.com* December 9, 2016.

Hardt, M., Price, E. & Srebro, N. (2016) Equality of opportunity in supervised learning. *30th Conference on Neural Information Processing Systems* (NIPS 2016). Available at: https://arxiv.org/pdf/1610.02413v1.pdf

Harris, M. (2017) Boot camps won't sell politics to young people. *Newsroom*. Available at: https://www.newsroom.co.nz/2017/08/16/42890/boot-camps-wont-sell-politics-to-young-people

Heald, D. (2006) Transparency as an instrumental value. In: *Transparency: The key to better governance?* eds. C. Hood, & D. Heald, pp. 59-73. Oxford: Oxford University Press.

Hildebrandt, M. (2015) *Smart technologies and the end(s) of law: Novel entanglements of law and technology.* Cheltenham: Edward Elgar.

House of Lords Select Committee on Artificial Intelligence (2018) *AI in the UK: Ready, willing and able?*

Hughes, T. (2017) Prediction and Social Investment. In: *Social Investment: A New Zealand policy experiment,* eds. J. Boston, & D. Gill, pp. 161-181. Wellington: Bridget Williams Books.

Human Rights Commission (New Zealand) (2018) *Privacy, data, technology: Human rights challenges in the digital age.*

Johnson, J.A. (2006) Technology and pragmatism: From value neutrality to value criticality. *SSRN Scholarly Paper, Rochester, NY: Social Science Research Network.* Available at: http://papers.ssrn.com/abstract=2154654

Johnston, K. (2017) Privacy and profiling fears over secret ACC software. *NZ Herald* September 15.

Kahnemann, D. (2011) *Thinking Fast and Slow.* London: Penguin.

Keddell, E. (2018) How fair is an algorithm? A comment on the Algorithm Assessment Report. Available at: http://www.reimaginingsocialwork.nz/2018/12/how-fair-is-an-algorithm-a-comment-on-the-algorithm-assessment-report/

Kesserwan, K. (2018) How can indigenous knowledge shape our view of AI? *Policy Options Politiques* February 16, 2018. Available at: http://policyoptions.irpp.org/magazines/february-2018/how-can-indigenous-knowledge-shape-our-view-of-ai/

Kleinberg, J., Mullainathan, S. & Raghavan, M. (2017) Inherent trade-offs in the fair determination of risk scores. *8th Conference on Innovations in Theoretical Computer Science* (ITCS 2017). Available at: https://arxiv.org/pdf/1609.05807.pdf

Klingele, C. (2016) The promises and perils of evidence-based corrections. *Notre Dame Law Review* 91(2): 537-584.

Kotz, S. (2005) Reflections on Early History of Official Statistics and a Modest Proposal for Global Coordination. *Journal of Official Statistics* 21(2): 139-144.

Kukutai, T. & Taylor, J. (2016) Data sovereignty and indigenous people: Current practice and future needs. In: *Indigenous data sovereignty: Towards an agenda*, eds. T. Kukutai & J. Taylor, pp. 1-2. Canberra: ANU Press.

Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016) How we analyzed the COMPAS recidivism algorithm. *ProPublica.org* May 23, 2016.

Levendowski, A. (2017) How copyright law can fix artificial intelligence's implicit bias problem. *Washington Law Review* (forthcoming). Available at: https://ssrn.com/abstract=3024938

Lombrozo, T. (2011) The instrumental value of explanations. *Philosophy Compass* 6: 539.

Lum, K. & Isaac, W. (2016) To predict and serve? Bias in police-recorded data. *Significance* October 2016: 14-19.

Magee, H. (2013) The criminal character: A critique of contemporary risk assessment and preventive detention of criminal offenders in New Zealand. *Auckland University Law Review* 19: 76-98.

Marks, A., Bowling, B. & Keenan, C. (2017) Automated justice? Technology, crime, and social control. In: *The Oxford handbook of law, regulation, and technology,* eds. R Brownsword, E. Scotford & K. Yeung, pp. 705-730. New York: Oxford University Press.

Maude, S. (2018) OK computer? ACC claim process relies on bad data, breaches rights – lawyer. *New Zealand Doctor* July 18.

Mcquillan, D. (2015) Algorithmic states of exception. *European Journal of Cultural Studies* 18(4-5): 564-576.

Meijer, A. (2014) Transparency. In: *The Oxford handbook of public accountability,* eds. M. Bovens, R.E. Goodin & T. Schillemans, pp. 507-524. New York: Oxford University Press.

Meister, D. (1999) *The history of human factors and ergonomics.* Mahwah, NJ: Erlbaum.

Miller, T. (2017) Explanation in artificial intelligence: Insights from the social sciences. Available at: https://arxiv.org/pdf/1706.07269.pdf

Mittelstadt, B.D. (2017) From individual to group privacy in big data analytics. *Philosophy and Technology* (doi:10.1007/s13347-017-0253-7).

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016) The ethics of algorithms: Mapping the debate. *Big Data and Society* 16: 1-21.

Montavon, G., Bach, S., Binder, A., Samek, W. & Müller K.-R. (2017) Explaining nonlinear classification decisions with Deep Taylor decomposition. *Pattern Recognition* 65: 211.

Morrison, B. (2009) *Identifying and responding to bias in the criminal justice system: A review of international and New Zealand research.* Wellington: Ministry of Justice.

Mossman, E. (2010) Research to validate the New Zealand Police Youth Offending Risk Screening Tool (YORST)—Phase I. New Zealand Policy.

Nagel, T. (1986) *The view from nowhere.* New York: Oxford University Press.

Naryanan, A. & Felten, E.W. (2014) No silver bullet: De-identification still doesn't work. Available at: http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf

National Science and Technology Council (2016) *Preparing for the future of artificial intelligence.* Available at: https://publicintelligence.net/white-house-preparing-artificial-intelligence

Northpointe (2015) Practitioner's guide to COMPAS March 19. Available at: http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

–––(2016) COMPAS risk scales: Demonstrating accuracy, equity and predictive parity. July 8. Available at: http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

NZ Government (2017) Government Expectations for Good Regulatory Practice. Available at: https://treasury.govt.nz/sites/default/files/2015-09/good-reg-practice.pdf

Ogborn, M. (1962) *Equitable assurances: The story of life assurance in the experience of the Equitable Life Assurance Society, 1762-1962.* Routledge.

Ohm, P. (2010) Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review 57: 1701-1777*.

O'Neill, C. (2016) *Weapons of math destruction.* New York: Crown Publishing Group.

Oswald, M. (2018) Algorithm-assisted decision-making in the public sector: Framing the issues using administrative law rules governing discretionary power. *Philosophical Transactions of the Royal Society A* 376: 1-20.

Oswald, M., Grace, J., Urwin, S. & Barnes, G.C. (2018) Algorithmic risk assessment policing models: lessons from the Durham HART model. *Information & Communications Technology Law* 27(2): 223-250.

Parasuraman, R. & Manzey, D.H. (2010) Complacency and bias in human use of automation: An attentional integration. *Human Factors* 52(3): 381-410.

Pasquale, F. (2014) *The black box society: The secret algorithms that control money and information.* Cambridge, MA: Harvard University Press.

Pazouki, K., Forbes, N., Norman, R.A. & Woodward, M.D. (2018) Investigation on the impact of human-automation interaction in maritime operations. *Ocean Engineering* 153: 297-304.

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.

Platzman, G. (1979) *The ENIAC computation of 1950: Gateway to numerical weather prediction.* University of Chicago Press.

Plous, S. (2003) The psychology of prejudice, stereotyping, and discrimination. In: *Understanding prejudice and discrimination*, ed. S. Plous, pp. 3-48. New York: McGraw-Hill.

Pohl, J. (2008) Cognitive elements of human decision making. In: *Intelligent decision making: An AI-based approach*, eds. G. Phillips-Wren, N. Ichalkaranje & L.C. Jain, pp. 41-76. Berlin: Springer.

Powers, D.M.W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* 2(1): 37-63.

Prat, A. (2006) The more closely we are watched, the better we behave? In: *Transparency: The key to better governance?* eds. C. Hood, & D. Heald, pp. 91-103. Oxford: Oxford University Press.

PwC (2017) Sizing the prize: What's the real value of AI for your business and how can you capitalise? Available at: https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html

Rahwan, I. (2018) Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology* 20(1): 5-14.

Rashbrooke, M. (2018) *Government for the public good.* Wellington: Bridget Williams Books.

Reed, C. (2018) How should we regulate artificial intelligence? Philosophical Transactions of the Royal Society A 376: 20170360.

Reiss, J. & Sprenger, J. (2017) Scientific objectivity. *Stanford Encyclopedia of Philosophy*.

Ribeiro, M., Singh, S. & Guestrin, C. (2016) "Why Should I Trust You?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining* 1135-1144.

Rieland, R. (2019) Artificial intelligence is now used to predict crime. But is it biased? *Smithsonian.com* March 5.

Roberts, A. (2006) *Blacked out: Government secrecy in the information age*. New York: Cambridge University Press.

Roomsburg, J. (1988) Biographical data as predictors of success in military aviation training, presented to the Faculty of the Graduate School of The University of Texas at Austin.

Scott et al. (2017) Governance and accountability in Social Investment information release. Available at: https://treasury.govt.nz/sites/default/files/2017-07/si-governance-accountability-report.pdf)

Scherer, M.U. (2016) Regulating artificial intelligence systems: Risks, challenges, competencies and strategies. *Harvard Journal of Law and Technology* 29(2): 353.

Selbst, A.D. & Powles, J. (2017) Meaningful information and the right to explanation. *International Data Privacy Law* 7(4): 233–242.

Stanton, N.A. (2015) Responses to autonomous vehicles. *Ingenia* 62: 9.

–––(2016) Distributed situation awareness. Theoretical Issues in Ergonomics Science 17(1): 1-7.

Strauch, B. (2018) Ironies of automation: Still unresolved after all these years. *IEEE Transactions on Human-Machine Systems* 48(5): 419-433.

Susskind, J. (2018) *Future Politics*. New York: Oxford University Press.

Tan, L. (2018) Immigration NZ's data profiling "illegal" critics say. *NZ Herald* April 5.

Tatman, R. (2016) Google's speech recognition has a gender bias. *Making Noise and Hearing Things* July 12, 2016.

Tutt, A. (2017) An FDA for algorithms. *Administrative Law Review* 69(1): 83.

UN (2018) *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression.*

Van Otterlo, M. (2013) A machine learning view on profiling. In: Privacy, due process and the computational turn: *Philosophers of law meet philosophers of technology,* eds. M. Hildebrandt & K. de Vries, pp. 41–64. Abingdon: Routledge.

Villani, C. (2018) *For a meaningful artificial intelligence: Towards a French and European strategy.* Available at: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

Wachter, S. (2018) The GDPR and the Internet of Things: A three step transparency model. *Law, Innovation and Technology* 10(2): 266-294.

Wachter, S. & Mittelstadt, B. (2019) A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review* (forthcoming).

Wachter, S., Mittelstadt, B.D. & Floridi, L. (2017a) Transparent, explainable, and accountable AI for robotics. *Science Robotics* 2(6).

–––(2017b) Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law* 7(2): 76-99.

Waitangi Tribunal (2005) *The Offender Assessment Policies Report.* Available at: https://forms.justice.govt.nz/search/Documents/WT/wt_DOC_68001752/Offender%20Assessment%20Policies.pdf

Waitangi Tribunal (2017) *Tū Mai Te Rangi! The Report on the Crown and Disproportionate Reoffending Rates*

Watson, N. (2010) *Lloyd's Register: 250 years of service.* Lloyd's Register.

Walker, G.H., Stanton, N.A. & Salmon, P.M. (2015) *Human factors in automotive engineering and technology.* Surrey: Ashgate.

Walter, M. (2016) Data politics and indigenous representation in Australian statistics. In: *Indigenous data sovereignty: Towards an agenda,* eds. T. Kukutai & J. Taylor, pp. 79-97. Canberra: ANU Press.

Webb, H. et al. (2018) Multi-stakeholder dialogue for policy recommendations on algorithmic fairness. *Proceedings of the International Conference on Social Media & Society, Copenhagen, Denmark.*

Weik, M. (1961) The ENIAC story. *ORDNANCE, The Journal of the American Ordnance Association,* Jan-Feb 1961: 3-7.

Wilson, N.J., Kilgour, G. & Polaschek, D.L.L. (2013) Treating high-risk rapists in a New Zealand intensive prison programme. *Psychology, Crime and Law* 19(5-6): 527-547.

Zerilli, J., Knott, A., Maclaurin, J. & Gavaghan, C. (2018) Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy and Technology* (https://doi.org/10.1007/s13347-018-0330-6).

–––(2019) Algorithmic decision-making and the control problem (under review).

GOVERNMENT USE OF ARTIFICIAL INTELLIGENCE IN NEW ZEALAND