Exploring the Enigma: Enhancing Digital Rights in the Age of Algorithms

Jessie Malcolm

A dissertation submitted in partial fulfilment of the degree of Bachelor of Laws (Honours)
at the University of Otago – Te Whare Wānanga o Otāgo, Dunedin, New Zealand.

5 October 2018

ACKNOWLEDGEMENTS

To Colin Gavaghan, for your guidance, kindness and enthusiasm in supervising this dissertation;

To my family, particularly Heidi and John, for always being there;

And to the "colleagues", flatmates and fellow tutors, for the friendships that made law school worthwhile;

Thank you.

CONTENTS

INTRODUCTION 5

CHAPTER ONE: THE CONCERNS WITH ALGORITHMIC DECISION-MAKING 7

A) Algorithms: why should we care? 7

B) Artificial intelligence, algorithms and machine learning 8

1. Transparency and explainability 9

2. Bias 12

CHAPTER TWO: NEW ZEALAND'S CURRENT LEGISLATIVE FRAMEWORK 15

A) The plight of Malik 15

B) The Privacy Act 16

1. Breaking down the algorithmic decision-making process 17

2. The effectiveness of the Privacy Act 23

C) Requesting reasons: the OIA, LGOIMA and the common law 23

1. The OIA and LGOIMA 23

2. Common law right of reasons 25

3. Standard of reasons given 27

4. Where does this leave Malik? 27

CHAPTER THREE: FILLING THE LACUNA IN THE LAW 29

A) Where to now? 29

B) Following the European Union 29

1. "Right to an explanation" 29

2. "Human in the loop" 31

3. Summary 32

C) Strengthening existing legislation 32

1. Implementation of guidelines for "meaningful" explanations 32

2. Mandating public disclosure 34

3. A technical solution: "Responsibility by design" 35

4. Self-regulation models 35

5. Summary 36

D) Oversight regulatory body 36

1. The harm of concern 37

2. A centralised body 38

3. The function of the oversight body 38

4. Hard-edged vs. soft touch regulation 40

E) Summary 41

CONCLUSION 42

BIBLIOGRAPHY 43

“We can only see a short distance ahead,
but we can see plenty there that needs to be done”.

— Alan Turing

INTRODUCTION

Algorithms represent a present-day enigma. Their inner workings are not easily deciphered even to the experts. Algorithms increasingly control our lives. From the financial markets on Wall Street[1] to prospective matches on Tinder,[2] algorithms mediate the world we live in. As the public and private sectors opt to harness this technology in their decision-making processes, this raises concerns for human rights. The lack of transparency and explainability in algorithmic decision-making is not conducive to promoting rights to natural justice when the individual is unaware and unable to understand a decision regarding them. Likewise, algorithms embedded with bias are not consistent with the right to be free from discrimination. Coined “weapons of math destruction”,[3] algorithmic decision-making has implications for our fundamental human rights applied in the digital context.

The purpose of this dissertation is to advance three objectives necessary to enhance the digital rights of those subject to algorithmic decisions in New Zealand. These objectives are to: provide meaningful explanations; increase transparency; and mitigate bias. The following three chapters look deeper into the concerns raised by algorithms to provide more clarity around what these objectives entail and how they can be fulfilled. Ultimately, I contend that implementing an oversight regulatory body is the most logical response to achieve these objectives and enhance digital rights.

Throughout this dissertation, I have three tasks. Chapter I will provide a brief explanation of algorithms and machine learning, before outlining the concerns this technology presents to rights. As such concerns are vast, it is necessary to limit the focus to the challenges relating to the lack of transparency and explainability of algorithmic systems as well as hidden bias.

In chapter II, I introduce Malik as he encounters an algorithm in use by the Ministry of Justice. Through Malik, we will survey how the current legislative landscape of New Zealand is placed to provide recourse, or at least ameliorate the concerns identified in chapter I. It should be noted, both Malik and the algorithm he is subject to, are hypothetical. The use of algorithms by New Zealand government agencies is currently under investigation by the Department of Internal Affairs.[4] Until this report is publicly released, discussion is confined to a hypothetical algorithm predicting recidivism rates for offenders — albeit not unlike those encountered in the United States.[5]

---

1 See Chris Isidore “Machines are driving Wall Street's wild ride, not humans” (6 February 2018) CNN Money <https://money.cnn.com>.

2 See Cédric Courtois and Elisabeth Timmermans “Cracking the Tinder Code: An Experience Sampling Approach to the Dynamics and Impact of Platform Governing Algorithms” (2018) 23 Journal of Computer-Mediated Communication 243.

3 See Cathy O’Neil *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (1st ed, Crown, New York, 2016).

4 "Government to undertake urgent algorithm stocktake” (23 May 2018) Beehive.govt.nz <www.beehive.govt.nz>.

5 See Jeff Larson and others “How We Analyzed the COMPAS Recidivism Algorithm” (23 May 2016) ProPublica <www.propublica.org>.

Chapter III explores the ways New Zealand can achieve the identified objectives in order to enhance digital rights. Prospective avenues include: following the European Union; enhancing existing legislative rights in New Zealand; and introducing a regulatory oversight body. It is the final option of a regulatory body to oversee the use of algorithms in the public sector that, I argue, is the most prudent approach to addressing the concerns machine learning technology presents to digital rights.

Before we begin — a word regarding the confines of this dissertation. First, while the use of algorithmic decision-making permeates both the public and private sector, this dissertation focuses on the use of algorithms by public bodies. The reasons for this are twofold. For one, the state has legislative obligations to uphold rights. For another, the public sector can be used as a test model for any potential regulatory options before expanding this further to the private sector.

Secondly, for the purposes of this dissertation, "digital rights" concern the application of existing, legislatively recognised human rights, such as that contained in the New Zealand Bill of Rights Act 1990 ("NZBORA") and the Human Rights Act 1993. The focus on digital rights is not intended to point out any shortcomings in New Zealand human rights legislation. The Human Rights Act gives a comprehensive list of protected categories and covers indirect discrimination whilst NZBORA affords rights of explainability and review.[6] Instead, the inquiry is directed towards how to give effect to rights of natural justice and freedom from discrimination in the context of algorithmic decision-making. It is acknowledged that there are other rights that should also apply in the digital sphere, such as the "right to be forgotten". Discussion of this, however, falls outside the ambit of this dissertation.

Thirdly, algorithmic decision-making refers to machine learning algorithms. Machine learning technology brings additional challenges beyond what is encountered with human made decisions. This stems from the fact that such technology does not learn nor reason like humans do, producing outputs that are difficult to predict and explain.[7] Here we encounter the first of many dilemmas; what makes machine learning algorithms valuable is also what makes them uniquely hazardous.[8]

---

[6] Human Rights Act 1993, ss 21 and 65; and New Zealand Bill of Rights Act 1990, s 27.

[7] Andrew Tutt "An FDA for Algorithms" (2016) 69 Admin L Rev 83 at 87.

[8] At 90.

CHAPTER ONE: THE CONCERNS WITH ALGORITHMIC DECISION-MAKING

A) *Algorithms: why should we care?*

Algorithms form an integral part of our lives and are increasingly set to do so with new technologies on the horizon. Algorithms determine what advertisements are recommended when scrolling through social media, to forming the basis for granting loans, setting insurance premiums, detecting tax evasion, money laundering, trafficking and terrorist activities.[9]

The use of algorithms is not confined to the private sector. New Zealand governmental bodies have been employing algorithms to target overstayers on the immigration front, to identify those most "harmful" to the economy and prioritise their deportation.[10] Similarly, algorithms have been used to predict how long clients will stay on the books of Accident Compensation Corporation ("ACC")[11] and by the Ministry of Social Development to predict the likelihood of newborns to be subject to abuse and welfare dependency.[12] In the criminal justice arena, the New Zealand Police use a tool to predict the likelihood of re-assault in intimate partner relationships[13] while the Department of Corrections is able to express the probability of an offender being re-convicted and re-imprisoned for new offending.[14] The hypothetical recidivism algorithm discussed in chapter II is not unlike the types of algorithms currently used in New Zealand. With algorithms widespread throughout the public sector,[15] their use in decision-making processes is a present-day concern, not just a future problem.

There are many benefits to using algorithms. Such systems promise unparalleled efficiency with the ability to process huge volumes of data in order to predict outcomes. There is great potential to harness mass data on hand and use this for the public good. In the health sphere, algorithms have been developed to predict heart disease by scanning retinas,[16] or detect skin cancer by analysing a picture of a mole.[17] Such technology promises to be more efficient and a less intrusive way to catch diseases early on and improve the quality of life for many. In the public sector, reducing costs to taxpayers and increasing efficiency is an enormous incentive for incorporating algorithms into government processes.

---

[9] Paul B de Laat "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" (2017) Philosophy & Technology 1.

[10] Interview with Stephen Buranyi (Simon Martin, This Way Up, Radio New Zealand, 17 March 2018).

[11] "ACC is using a predictive modelling tool to better support our clients, faster" (28 September 2017) ACC <www.acc.co.nz/about-us/news-media/latest-news/acc-is-using-a-predictive-modelling-tool-to-better-support-our-clients-faster/>.

[12] Stacey Kirk "Children 'no lab-rats' — Anne Tolley intervene in child abuse experiment" (30 July 2015) Stuff <www.stuff.co.nz>.

[13] "Police to use new family violence risk assessment tools" (13 February 2012) New Zealand Family Violence Clearinghouse <https://nzfvc.org.nz/>; and Hamish McNeilly "Cops' new tool to predict domestic violence" *The New Zealand Herald* (online ed, Auckland, 19 April 2012).

[14] Leon Bakker, James O'Malley and David Riley *Risk of Reconviction* (Department of Corrections, 1999).

[15] Further clarification on the exact algorithms in use by various government departments is contingent on the release of the Algorithm Assessment Report by Statistics New Zealand.

[16] James Vincent "Google's new AI algorithm predicts heart disease by looking at your eyes" (19 February 2018) The Verge <www.theverge.com>.

[17] Joyce Riha Linik "Skin Cancer Detection Using Artificial Intelligence" (31 January 2009) IQ Intel <https://iq.intel.com>.

Algorithms have the potential to provide a fairer and more objective means of making determinations compared to human decision-making. It is my contention that this is contingent on adequate regulation that strengthens rights to natural justice and freedom from discrimination.

Yet there is a much darker side to algorithms. Many commentators refer to the notion of a "black box".[18] This analogy is multifaceted, encapsulating the idea that technology now serves to record thousands of variables about people, not just aeroplanes — "the black box has moved out of the plane, into our daily experience".[19] Secondly, the idea of the "black box" embodies the inherent opacity of algorithms and the consequent challenges with regulation. While the input and output of algorithms can be viewed, the internal workings remain a mystery — an enigma. This does not mean the system is incapable of being inspected, rather that there is not a simple explanation offered as to how and why the system works.[20] It is this fundamental masking of the internal processing that leads to the primary concerns of algorithmic decision-making: lack of transparency; explainability; and hidden bias.

B) Artificial intelligence, algorithms and machine learning

Broadly speaking artificial intelligence refers to technology or machines performing tasks that we would characterise as involving human intelligence.[21] At the core of artificial intelligence technology, lie algorithms.[22] Algorithms are essentially recipes; a set of instructions to be followed to accomplish a task, typically executed by a computer.[23] The application and complexity of any algorithm is vast. Most algorithms are straightforward, with basic instructions and corresponding outcomes that are relatively deterministic.[24] For example, the RoC*RoI risk measurement tool, which uses regression procedures to calculate the subject's probability of re-offending, has been in use by the Department of Corrections for nearly 20 years, producing outcomes that are relatively straightforward.[25] Recent years, however, have seen a move away from specific rule based algorithms to machine learning. Such technology is capable of dealing with immense amounts of input data to develop non-linear correlations.[26]

---

[18] See Frank Pasquale *The Black Box Society: the Secret Algorithms that Control Money and Information* (1st ed, Harvard University Press, Cambridge, 2015).

[19] Interview with Frank Pasquale, Author (Steve Paikin, The Agenda, 12 May 2016).

[20] Dallas Card "The "black box" metaphor in machine learning" (5 July 2017) Towards Data Science <https://towardsdatascience.com>.

[21] Calum McClelland "The Difference Between Artificial Intelligence, Machine Learning, and Deep Learning" (5 December 2017) Medium <https://medium.com>.

[22] Lindsay Kwan "Beyond the buzzword: What "artificial intelligence" means for marketing leaders, right now" (28 November 2017) Widerfunnel <www.widerfunnel.com>.

[23] TC "What are algorithms?" (30 August 2017) The Economist <www.economist.com>.

[24] Tutt, above n 7, at 93.

[25] See Bakker, O'Malley and Riley, above n 14.

[26] Lilian Edwards and Michael Veale "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke Law & Technology Review 18 at 25.

Sixty years on from when the term was originally coined, machine learning is fast growing in the domain of artificial intelligence.[27] The prolific collection of data, partially enabled by the incessant use of technology, provides the ideal platform for machine learning to take place. Machine learning works by feeding both input and output variables into an algorithm so that it is able to "learn" from the data.[28] Algorithms use this data to iteratively build new analytical models that can solve problems without being explicitly programmed for a particular solution.[29]

There are two categories of machine learning: "supervised" and "unsupervised".[30] In supervised learning, the algorithm is provided with a set of data (input) and the possible outcomes (output).[31] For example, the input data could take the form of pictures of moles; the output data is the determination of whether or not such moles are cancerous.[32] The aim is to approximate the mapping function so as to accurately predict the output for new input data. One could show the algorithm a picture of a mole and use the learning from its database to predict whether a biopsy is necessary. Through supervised learning, the training data set is effectively overseeing the learning process. By contrast, unsupervised learning only has input data with no corresponding output variable. The goal here is to model the underlying structure or distribution of the data in order to learn more, without placing restrictions on the model.[33] The lack of supervision means the algorithm is effectively left to its own devices in finding patterns within the data set. It is this, that makes machine learning algorithms both powerful and unpredictable, and is the basis on which the concerns surrounding transparency, explainability and bias arise.

1. Transparency and explainability

Transparency and explainability are important threads throughout our legal system. Transparency is regarded as fundamental to democracy.[34] It guarantees an opportunity to see how a result is reached, together with its reasons. This matters in terms of natural justice and understanding the system as a whole. It promotes access to justice, both retrospectively with the enforcement of rights, and prospectively in knowing what one is entitled to do in the future. Explainability is important for conceptual coherency. This aligns with Bentham's ideas of giving reasons to allow one to understand and criticise a decision.[35] Transparency and explainability are important for public confidence in decision-making processes as well as law reform, wherein "bad" decisions can initiate change.

---

[27] *Artificial Intelligence Shaping a Future New Zealand* (AI Forum NZ, May 2018) at 25.

[28] Edwards and Veale, above n 26, at 25.

[29] Nick Wallace and Daniel Castro *The Impact of the EU's New Data Protection Regulation on AI* (Center For Data Innovation, March 2018) at 7.

[30] Kwan, above n 22.

[31] Kwan, above n 22.

[32] Linik, above n 17.

[33] Jason Brownlee "Supervised and Unsupervised Machine Learning Algorithms" (16 March 2016) Machine Learning Mastery <https://machinelearningmastery.com>.

[34] José Luiz Pinheiro Lisboa "The Judiciary – Principle of Transparency and the duty of information" (2018) 7 International Journal of Open Governments 107.

[35] See Oren Ben-Dor "The institutionalisation of public opinion: Bentham's proposed constitutional role for jury and judges" (2007) 27 Legal Stud 216.

In this context, transparency refers to disclosing an algorithm's code or data, while explainability refers to the notion of making algorithms interpretable or understandable to the end user.[36] At the heart of the issue of transparency and explainability is the inability to access critical features of the decision-making process. If one does not know how or what data is being used to make a decision about them or that a decision is even being made, they are unlikely to be well placed in seeking recourse.

In one sense, difficulties with explainability and transparency are not new problems. They are challenges that have long been encountered with complex technologies.[37] In the pharmaceutical context, while a drug may prove effective for its intended use, it is hard to predict the side effects or explain the result, due to the inherent complexity of the biochemistry.[38] Humans themselves are also somewhat inexplicable. Human decision-making is known to be irrational, capricious and opaque and not any more explainable than an algorithm.[39] This represents a core argument against my contention for the need to promote explainability and transparency. If human decision-making is just as inexplicable and opaque as an algorithm, it would follow that there are no new problems to address and the suggestion of regulation is superfluous. I have several responses to this point. First, the important difference between machine learning algorithms and humans is that humans have an inbuilt advantage of trying to explain human behaviour; there is no similar edge to anticipating how algorithms will behave. Secondly, an algorithm has more capacity to systemically disadvantage (or advantage) a large group of people, compared to a human decision-maker. This is demonstrated in the later discussed example of the predictive policing tool, "Predpol".[40] Finally, accepting that algorithms are just as opaque as human decision-making, if the opportunity exists to make algorithms more explainable, then this is a desirable outcome. Striving for greater coherency and transparency in decision-making processes can only be positive.

There are a myriad of considerations to take into account when contemplating how transparency can best be promoted.[41] Mandating "full transparency" opens up privacy concerns of leaking sensitive data. As algorithms are only as good as the data set they are trained on, there is a great interest in having access to this information. Publicly disclosing this data may involve the disclosure of personal information. Disclosure also opens up the threat of manipulation with people "gaming the system".[42] Algorithms could be used in perverse ways to circumvent the activity they are designed to prevent. If, for example, it is discovered the proxies or "red flags" for an algorithm detecting tax fraud are donations to charity or membership to certain societies, tax evaders may adjust their practices to avoid detection.[43]

---

[36] Joshua New and Daniel Castro *How Policymakers Can Foster Algorithmic Accountability* (Center For Data Innovation, May 2018) at 10.

[37] Tutt, above n 7, at 103.

[38] At 103.

[39] Alex P Miller "Why do we care so much about explainable algorithms? In defense of the black box" (11 January 2018) Towards Data Science <https://towardsdatascience.com>.

[40] See page 13 of this thesis.

[41] de Laat, above n 9, at 3.

[42] At 10.

[43] At 10.

The extent to which this is relevant depends on the type of input data. If the data relates to personal characteristics, a proxy cannot so easily be evaded. On the other hand, if the input consists of behavioural data or voluntarily disclosed information, there is room for the individual to modify this information to evade the system. One solution may be to only accept algorithms that are immune to this type of manipulation. Alternatively, the disclosure of training sets and code could be limited to a small group of people such as an oversight body. The major barrier to transparency is commercial sensitivity. Disclosure of an algorithm may compromise or undermine a company's competitive edge.[44] Intellectual property and trade secrets are typically cited as ways to avoid disclosure of valuable algorithms. While this concern may appear more pertinent to the private sector, many public services employ software bought on the market or outsource the development to companies.[45] The lack of transparency exacerbates the existing power imbalance between individuals and big corporations; and similarly individuals and the state. Achieving transparency is not clear-cut. Considering how to implement transparency requires some specificity about what is being disclosed and to whom.[46] There are a range of possible recipients of disclosure: intermediary bodies with an oversight role; affected individuals; or the public in general.[47]

As machine learning algorithms are difficult to interpret, the benefits from transparency may be limited.[48] Providing details that are unilluminating and uninformative will not aid a subject's understanding of how the decision was made. While developers may know the variables and weightings of the algorithm initially, as time goes on the algorithm "learns" from encountering new data and the model adjusts accordingly. As the exact inner workings can become a mystery to those who engineered the algorithm, this complex interaction soon makes it difficult to understand how a problematic result was reached. In many cases, developers lack the ability to precisely explain how their algorithm makes decisions and instead can only express the degree of confidence they have in its accuracy.[49] The challenge of how to promote transparency and explainability to give effect to rights of natural justice, when the technology itself may not allow it, is an ever-present issue further considered in chapter III.

As explored in the following chapters, rights to explainability or the reasons for a decision, do not necessarily secure substantive justice nor effective remedies.[50] The ultimate concern is how to create explainability that is meaningful and understandable to the individual. Certain measures within the European Union's General Data Protection Regulation 2018 ("GDPR") are aimed at opening up the algorithmic black box and promoting explainability. Yet, as discussed in chapter III such measures fall

[44] At 12.
[45] At 10.
[46] At 3.
[47] At 3.
[48] At 2.
[49] New and Castro, above n 36, at 5.
[50] Edwards and Veale, above n 26, at 22.

short of providing meaningful explanations. As I argue throughout, a top-down regulatory model is best placed to promote meaningful explanations through its holistic ability to establish and enforce standards.

2. Bias

There is a common misconception that computers are more objective and can offer some neutrality when it comes to making decisions. Comparatively, human decision-making is fallible with unconscious bias. Yet such objectivity is a facade — and a dangerous one at that.[51]

The term "bias", is of itself a contested term. Humans navigate the world through bias, as do machines. Bias is inherent within machine learning algorithms by virtue of "learning" based on past data. In certain respects, bias in algorithms is desirable: how else would an algorithm be able to detect a cancerous tumour if it were not prejudiced towards those that were malignant? Bias may even be healthy in some situations — such as being biased towards erring on the side of caution when diagnosing potentially cancerous tumours. In the public sector, bias may be an inbuilt policy decision. In Durham, Police have employed a Harm Assessment Risk Tool ("HART") to classify a subject's risk of re-offending as low, medium or high-risk. The tool itself has an inbuilt predisposition or bias — designed to err on the side of caution in terms of public safety. This means it is more likely to classify an offender as medium or high-risk.[52] The question of who is entitled to make such policy decisions is another concern; whether this is entrusted to the developer, the agency using the algorithm, or some overall governing body. There is also a line that must be drawn when the use of bias is unfairly prejudicial. For an algorithm shortlisting applicants for a job in a traditionally male profession, prejudice towards what a successful candidate looks like, based on past trends, may not be desirable if this is preferring males over females.[53] The difficulty is drawing the line between harmful prejudice and the biases the algorithm is being encouraged to develop.

There is an opportunity for human bias to enter the algorithm in all stages of its development. It is software engineers who construct the data sets, define the parameters of the analysis and create the clusters, links and decision trees to generate the predictive models applied.[54] This computerisation merely drives any discrimination upstream, making it hard to uncover.[55] In one instance, developers replicated their own western standards of beauty when creating an "objective beauty evaluating algorithm" to judge a beauty contest. The resulting output chose predominately "white" contestants.[56] The risk assessment algorithm of "COMPAS" (Correctional Offender Management Profiling for Alternative

---

[51] O'Neil, above n 3.

[52] Chris Baraniuk "Durham Police AI to help with custody decisions" (10 May 2017) BBC <www.bbc.com/news>.

[53] David McRaney "YANSS 115 – How we transferred our biases into our machines and what we can do about it" (20 November 2017) You Are Not So Smart <https://youarenotsosmart.com/>.

[54] Pasquale, n 18, at 35.

[55] At 35.

[56] Noel Duan "When beauty is in the eye of the (robo)beholder" (20 March 2017) Arstechnica <https://arstechnica.com>.

Sanctions) used to evaluate the likelihood of criminals committing future crime also revealed bias. “Black” defendants were twice as likely to be incorrectly classified as high-risk compared to their “white” counterparts; while “white” repeat offenders were twice as likely to be incorrectly labelled as low-risk.[57] The problem here is not the error rate of the algorithm, it is that when mistakes were made, they tended to be in the same direction, disproportionately prejudicing a minority group. Not all errors are equal. On face value an algorithm may have a lower error rate compared to human decision-making. However, if all of these errors are concentrated towards the same group in a manner that disadvantages them, then problems of bias arise. If algorithms are learning from data that reflects an unequal society, this invariably results in the automation of bias the programs are designed to eliminate.[58] If the projection of past inequalities is combined with the public perception of algorithms being impartial, we are encroaching upon a dangerous territory of disguising this prejudice. Mitigating bias from algorithmic decision-making processes is an important component in giving effect to the right of freedom from discrimination.

The question of collective harm also arises. On an individual level outcomes may be just or differ arbitrarily from one algorithm to the next, yet on a collective level the algorithm may be biased against specific groups.[59] Not only does the opacity of the internal algorithmic process make this bias difficult to uncover, but the issue of collective harm is deeper rooted in a system of law that is orientated towards the individualistic paradigm rather than recognising rights of minority groups.[60] If this is the case, then taking steps outside of the individualistic paradigm is necessary to adequately address issues of collective harm.

Reinforcing bias is another concern pertinent to algorithmic decision-making. The predictive policing tool “PredPol” is often cited in reference to this point.[61] This algorithm is used in the United States to distribute officers based on geographical crime rates. Sending more police to an area results in more arrests being made. The information is fed back into the system, allocating increased numbers of police back to the same areas. As these areas often tend to be neighbourhoods populated with racial minorities, such groups become disproportionately represented in the arrest statistics.[62] Through this feedback loop of machine-made bias, we face the risk of algorithms creating the future they have predicted.[63]

---

[57] Christina Couch “Ghosts in the Machine” (25 October 2017) NovaNext <http://www.pbs.org/wgbh/nova/next>.

[58] Stephen Buranyi “Rise of the racist robots – how AI is learning all our worst impulses” (8 August 2017) The Guardian <www.theguardian.com>.

[59] de Laat, above n 9, at 2.

[60] See Kathleen Mahoney "The Limits of Liberalism” in Richard Devlin (ed) *Canadian Perspectives on Legal Theory* (Eamon Montgomery 1991) 57.

[61] Matt Reynolds “Biased policing is made worse by errors in pre-crime algorithms” (4 October 2017) New Scientist <www.newscientist.com>.

[62] Reynolds, above n 61.

[63] McRaney, above n 35.

One option to combat bias may be to prohibit the use of discriminating variables such as ethnicity or gender. Yet in practice, this can be substituted with proxies, such as a subject's postcode or level of education. Algorithms can learn, recognise and exploit the fact that a person's education or home address may correlate with other demographic information, which can imbue them with racial and other biases.[64] Protected characteristics, such as ethnicity might correlate with variables of interest. Statistically in New Zealand, Māori make up 50% of the total prison population.[65] Yet socially and politically it may be undesirable to reflect this in predictions, especially if we want to avoid algorithms determining the future based on predictions from the past. As there is an inevitable trade-off to be made between "fairness" and "accuracy", consideration must be given to who is best positioned to make this judgement.

There are many considerations to take into account when dealing with the question of regulating algorithmic decision-making and the hidden biases that may be embedded within. While some of these questions (such as the notion of collective harm) may be deeper rooted within the law, others are more specific to the nature of algorithms, such as reinforcing bias and carrying over human attitudes into the algorithm. Ultimately the question of regulating bias is likely to involve a policy decision, in determining when an algorithm is crossing the line in becoming unfairly prejudicial or deciding where the balance between accuracy and fairness should lie. Stepping outside the individualistic paradigm and having a mechanism that enables scrutiny of such bias, lends towards the implementation of a top-down model of regulation.

In this chapter, I have outlined the key concerns of algorithmic decision-making. These are lack of transparency and explainability as well as bias. It is these concerns that impact upon digital rights, particularly an individual's right to natural justice and the right to be free from discrimination. Having established this, the next step is to survey the current legislative landscape of New Zealand to assess how these concerns may be addressed within the present framework.

---

[64] Jackie Snow "New Research Aims to Solve the Problem of AI Bias in "Black Box" Algorithms" (7 November 2017) MIT Technology Review <www.technologyreview.com>.

[65] "Prison facts and statistics" (June 2017) Department of Corrections <www.corrections.govt.nz>.

CHAPTER TWO: NEW ZEALAND'S CURRENT LEGISLATIVE FRAMEWORK

A) The plight of Malik

It is 2019. The Ministry of Justice has introduced a new algorithm, PILATE (Predictive Indexing of Lapse or Tendency to reoffend), imported from America, to determine bail decisions in the criminal justice system. PILATE promises a more accurate method of assessing the likelihood of a defendant to recidivate, pose a flight risk or endanger the safety of the community. PILATE takes a variety of inputs, including answers to a series of questions answered by the offender through a supervising officer. These questions include the subject's: stability of residence; former substance abuse; family criminality and history; social environment; and experiences of social isolation. This is supplemented by a criminal record and other undisclosed information drawn from various sources. The internal processing results in a score of 1 (low-risk) to 10 (high-risk) of the subject's likelihood of reoffending, which in turn informs the decision to grant bail. The model was trained on past bail decisions made by judges and data sets of individuals who went on to reoffend. Rigorous trials conducted on New Zealand training data show PILATE has a lower error rate and increased consistency compared to the varying leniency among judges. After more than twelve months in operation, the results are positive. Half as many defendants are being detained, with no impact on community safety. The number of defendants failing to appear at trial remains unaffected while cutting costs of needlessly detaining those posing a low-risk.

Enter Malik, a twenty-five year old courier driver from Manukau, Auckland and a second generation New Zealander of Jordanian descent. As a result of an altercation with his former partner, Malik faces a charge of assault. It is alleged Malik entered his former partner's house threatening physical harm if his one year child was not returned. A physical altercation ensued as Malik attempted to take the child from the complainant, during which it is alleged he hit the complainant with a spatula. Malik's only previous offending relates to a breach of a protection order during events that can be attributed to the same dysfunctional relationship. While Malik is not bailable as of right, he must be released by the court on reasonable terms and conditions, unless it is satisfied that there is a just cause for continued detention.[66] The court in exercising this discretion have been heavily impacted by the recidivism score by PILATE. Under PILATE, Malik has received a score of 8, been deemed a flight risk and denied bail. Despite an early guilty plea, steady employment and favourable characteristics of good behaviour, neither Malik nor his lawyer can understand how or why this decision was reached.

With two young children to care for, Malik wishes to understand whether some error has been made in his denial of bail, to enable him to make an informed decision as to whether to challenge it. This chapter looks to give an overview of the legislative framework in New Zealand as to the recourse available in

[66] Bail Act 2000, s 7(5).

helping Malik understand the decision. Such analysis shall be directed towards the Privacy Act 1993, the Official Information Act 1982 ("the OIA") and the Local Government Official Information and Meetings Act 1987 ("LGOIMA"). The Privacy Act largely governs rights of access to information, while the OIA and LGOIMA confer a right to the reasons pertaining to a decision. Combined, these Acts constitute the main legislative body concerning access to information, data protection and privacy in New Zealand. A right to access also exists in the common law and in human rights-based legislation.[67]

B) The Privacy Act

It is necessary to first examine how the Privacy Act intersects with the algorithmic decision-making process of PILATE to establish its applicability to Malik's situation. The Privacy Act covers access by "natural persons" to information about themselves. At the core are twelve information privacy principles controlling the collection, use, security and correction of personal information by "agencies". The Ministry of Justice comes within the meaning of an "agency", defined (with a series of specific exclusions) as any "person or body of persons, whether corporate or unincorporate, whether in the public sector or the private sector".[68] As it currently stands, the Privacy Act makes no specific provisions for algorithms and no further amendments are proposed in the Privacy Bill currently going through Parliament.[69]

There are two major limits to the applicability of the Privacy Act. First, it is only engaged with "personal information", that is "information about an identifiable person".[70] While the algorithm itself is not personal information, the input data used in the algorithm (i.e. Malik's criminal record) and the corresponding output (i.e. Malik's recidivism score) is information that clearly pertains to Malik. For this reason, if an algorithm is used in a decision about an individual, it contains personal information and the Privacy Act is relevant.[71] Secondly, liability under the Privacy Act requires a breach of one of the twelve information privacy principles set out in Part II as well as a finding (in the opinion of the Tribunal or Privacy Commissioner) that Malik has suffered some sort of harm.[72] This harm includes: loss, detriment, damage, or injury to Malik; an adverse effect on Malik's rights, benefits, privileges, obligations or interests; or significant humiliation, loss of dignity or injury to feelings.[73] The threshold of "significance" sets a high bar for Malik to point to injured feelings following PILATE's decision (albeit subjective to Malik's sensitivities rather than an objective standard). One form of harm is to point to an adverse effect on Malik's rights by "not being released on reasonable terms and conditions, unless there is just cause for continued

[67] New Zealand Bill of Rights, s 27.

[68] Privacy Act 1993, s 2.

[69] Privacy Bill (2018)(34–1).

[70] Privacy Act, s 2.

[71] See Vanessa Blackwood "Algorithmic transparency: what happens when the computer says "no"?" (29 November 2017) Privacy Commissioner <https://privacy.org.nz/>.

[72] Privacy Act, s 66(1)(b)(i)–(iii). There is no need to prove loss of harm upon breach of principles 6 and 7.

[73] Section 66(1)(b)(iii).

detention".[74] As this right is subject to "justified limitations"[75] and given effect to at the discretion of the court,[76] there is no assurance of establishing that Malik's rights have been adversely affected. It is this finding of harm or loss that makes it difficult, at first instance, to engage the Privacy Act as a protective measure for Malik.

1. Breaking down the algorithmic decision-making process

Provided there is a finding of harm or loss by the Privacy Commissioner or Tribunal, the question turns to establishing a breach of one of the information privacy principles. It is helpful to break down PILATE's process into its various stages to examine the intersection with the Privacy Act.

- Phase 1: the collection of data sets as input for machine learning
- Phase 2: use of the training data in machine learning to develop a model
- Phase 3: use of the model for decision-making
    a) applying the model to an individual to make a decision
    b) the resulting prediction about an individual.

i. Phase 1: Collection of data sets

Phase 1 consists of data collection to serve as the input for the machine learning process of PILATE predicting recidivism rates of offenders. The collection of this personal information is governed by information privacy principles 1–4. At the stage of data collection, Malik's information has not yet been engaged, however, the collection of other individuals' personal information has informed the decision made about him.

Principle 1 provides personal information shall not be collected by an agency unless it is collected for a lawful purpose connected with a function/activity of the agency and such collection is necessary for that purpose.[77] This purpose limitation serves as an important constraint on the collection of personal information, establishing the fundamental framework that many other principles depend on for their operation.[78] It is assumed the collection of personal information for PILATE by the Ministry of Justice was lawful and necessary to achieve such purpose (i.e. for the administration of an efficient justice system).

---

[74] New Zealand Bill of Rights, s 24(b).

[75] Section 5.

[76] See Bail Act, s 8(1).

[77] Privacy Act, s 6, Privacy Principle 1.

[78] Paul Roth "Reports on Aspects of Privacy Compliance and Practice of the NZ Post Lifestyle Survey 2009" (20 June 2011) Office of the Privacy Commissioner <https://privacy.org.nz>.

Principle 2 mandates that personal information is collected directly from the individual.[79] This is accompanied by transparency requirements under principle 3, where the individual is to be made aware that information is being collected, the purpose for collection, intended recipients and any other relevant information.[80]

Theoretically, principles 2 and 3 address some of the transparency concerns raised in chapter I. If the Ministry of Justice is collecting personal information for the purposes of training PILATE, the subject should be made aware of the collection and use of their information. Awareness of this use of information is important, both for transparency in government processes and the individual's autonomy in controlling their own personal information. These transparency requirements are, however, limited in practice through: the government already holding information; existing information matching/sharing regimes; and the exceptions to principles 2 and 3. It is likely agencies would already hold (at least) some information for the initial training set. For example, the Ministry of Justice holds criminal records, sentencing and reoffending information that can be used to train PILATE. Similarly, the proposed model by ACC is using twelve million past decisions in its database for training purposes.[81] So long as the use of this information is directly related to the purpose in connection with which the information is obtained, the agency is free to use such information.[82] Individuals would not be informed that their information is being used in this manner. Information sharing regimes within the government also allow for personal information to be shared between government agencies for the purpose of delivering public services.[83] While these regimes are published and monitored by the Privacy Commissioner, they represent one way in which individuals may not be aware that information given to one agency is used in training an algorithm in another. Furthermore, if the agency believes on reasonable grounds that compliance is not reasonably practicable in the circumstances of the particular case[84] or necessary to avoid prejudice to the maintenance of the law,[85] this serves as an out road to the applicability of principles 2 and 3.

The exception of "non-identifiability of the information used" is the obvious way for agencies to bypass any transparency obligations when training algorithms. Where an agency believes on reasonable grounds that the information will not be used in a form in which the individual is identified or used for statistical or research purposes and will not be published in a form that could reasonably be expected to identify the individual concerned, it is not necessary to comply with principles 2 and 3.[86] Unlike the GDPR, the Privacy Act is yet to grapple with the consequences of, if and when, this information is "re-identified", or propose appropriate methods of pseudonymisation or anonymisation to protect a subject's privacy.

---

[79] Privacy Act, s 6, Privacy Principle 2.

[80] Section 6, Privacy Principle 2.

[81] *Statistical modelling to support the ACC automation cover decisions and accident description August 2018* (Accident Compensation Corporation, August 2018).

[82] Section 6, Privacy Principle 10(1)(e).

[83] Section 96E.

[84] Section 6, Privacy Principles 2(2)(f) and 3(4)(e).

[85] Section 6, Privacy Principles 2(2)(d) and 3(4)(c)(i).

[86] Section 6, Privacy Principles 2(2)(g) and 3(4)(f).

Pseudonymisation or anonymisation are techniques to enhance privacy by replacing identifying fields within a data record by one or more artificial identifiers, or pseudonyms.[87] Re-identification is the process by which anonymised personal information is linked back to the individuals the information relates to, destroying the cloak of anonymity that protects an individual's privacy.[88] While there are technical measures to mitigate this risk, such as the notion of the differential privacy,[89] it is established that data sets are capable of being re-identified and attributed to specific individuals.[90] As long as the subject is "identifiable" (as opposed to identified), the data constitutes personal information.[91] In a submission to the Select Committee, the Privacy Commissioner advocates for the inclusion of a new privacy principle in the Privacy Bill to protect against this risk.[92] In any case, these measures are directed towards protecting the privacy of those individuals whose information has been used in the training set. While these are important considerations for the future of New Zealand privacy law, for Malik's purposes, this does not assist in understanding how the decision was reached.

For completeness, principle 4 provides the collection of information must be lawful or not intrude to an unreasonable extent upon the personal affairs of the individual concerned.[93] The applicability of this largely depends on the type of information collected. It is assumed the collection of information for PILATE is lawful.

ii. Phase 2: Model construction/development

Phase 2 involves model construction and development, where the data is used in training the model through machine learning. This involves the use of various techniques such as classification, decision trees, support vector machines, ensemble methods and neural networks.[94] As PILATE was trained on past bail decisions made by hams, inevitably bias will be imported into the algorithm. There are a multitude of technical measures that can be taken to mitigate against such concerns during the construction and modelling phases of PILATE. These, however, fall outside the ambit of inquiry for the purposes of this chapter.

In terms of the Privacy Act, information privacy principles 8 and 10 are most relevant to phase 2. Principle 8 puts obligations on the Ministry of Justice to not use information without taking steps to

---

87 Olenka Van Schendel "Data masking: Anonymisation or pseudonymisation?" (7 November 2017) GDPR: Report <https://gdpr.report>.

88 John Edwards "Privacy Commissioner's Submission on the Privacy Bill to the Justice and Electoral Select Committee" at [55].

89 See *Differential Privacy for Everyone* (Microsoft Corporation, 2012).

90 See Latanya Sweeney "Weaving Technology and Policy Together to Maintain Confidentiality" (1997) 25 JLM & E 98. The results of this investigation showed that in 2000, 87% of all Americans could be uniquely identified using only three bits of information: post code, birthdate and sex.

91 Privacy Act, section 2.

92 Edwards, above n 88, at [55].

93 Section 6, Privacy Principle 4.

94 de Laat, above n 9, at 7.

ensure that it is accurate, up to date, complete, relevant and not misleading.[95] From an individual standpoint, this principle is largely redundant in the context of algorithmic decision-making. At phase 2, the information relates to people other than Malik, thus Malik has no standing to challenge its accuracy. The individuals themselves are unlikely to know their personal information is being used in training an algorithm, thus it is improbable the accuracy of such information would be challenged. Under the current Privacy Act, bringing an action requires proof of harm. Showing one has suffered harm or loss when information has been used to make a decision about another is infeasible. In addition, without knowing how much weight PILATE gives to the disputed information, it is unclear how much the data contributed to the final determination.[96]

Principle 10 provides a purpose limitation in that the Ministry of Justice, in holding personal information obtained in connection with one purpose, shall not use the information for any other purpose unless they believe on reasonable grounds that a listed exception applies.[97] This, however, is subject to the exception of when "information is used in a form which the individual is not identified", thus principle 10 is unlikely to limit the use of data in training an algorithm.

iii. Phase 3: Model use in decision-making

Phase 3 is the use of PILATE in decision-making. For the purposes of the Privacy Act, this stage can be broken down into two components: (a) applying the predictive model to Malik; and (b) the resulting prediction.

a. Applying the predictive model to Malik

Applying the predictive model to Malik involves questions of the collection of Malik's personal information as input for the algorithm. As previously discussed the vast exceptions to the transparency requirements (principles 2 and 3) may mean that Malik is not aware that this information is being collected.

Principle 10 which places limits on the use of personal information is subject to an exception of public availability. Use is permitted if the agency believes on reasonable grounds that "the source of the information is publicly available and in the circumstances of the case it would not be unfair or unreasonable to use this information".[98] The definition of "publicly available" includes when "a publication will be generally available to members of the public".[99] This becomes blurred when applied to social media. If, for example, the Ministry of Justice has used information about Malik's social networking

---

95 Section 6, Privacy Principle 8.

96 See *Case Note 14290* [2001] NZPrivCmr 5 (1 June 2001) where the Privacy Commissioner formed the view disputed information of employment may not have formed the basis for ACC's determination, no breach was found.

97 Section 6, Privacy Principle 10.

98 Section 6, Privacy Principle 10(1)(a) as amended by the Harmful Digital Communications Act 2015, s 40.

99 Section 2.

behaviour in order to predict his recidivism tendency, it is unclear whether this is “publicly available” and could be challenged on the grounds it is "unfair or unreasonable”. Social networking data has been previously used in machine learning algorithms, in one case to identify individuals with depression, before they had received a formal diagnosis.[100] In New Zealand case law, there are suggestions that once a threshold is reached, such as having two hundred friends on Facebook, the information is no longer private.[101] On the flip side, social media could be regarded as a “legally walled garden to which one can have varying levels of access depending on whether or not one has an account with a given service”, and hence not "generally available to members of the public”.[102] There is also the question of metadata (for example, the length of posting, number of social connections and key phrases used) which is one step removed from content deliberately posted on social media. Presuming the metadata pertains to the individual and constitutes personal information, if the justification for “publicly available" is that the subject has consciously chosen to make this available to the world at large, then metadata would fall outside this exception. Use of such information may be considered “unfair or unreasonable” and therefore a breach of principle 10. Without knowing whether this information forms part of the “undisclosed information drawn from various sources” used in PILATE, Malik is unable to establish a breach of principle 10. While we are limited to speculation in this instance, the use of social networking data in algorithmic decision-making and its applicability under the Privacy Act is acknowledged as a broader concern.

Principle 6 affords a right of access, where an agency holds personal information in such a way that it can be readily retrieved. This right may be directly enforced in a court of law, without proof of harm.[103] At best, Malik may gain access to the input and output information of the algorithm, constituting “personal information”. Optimistically, access to the input information may indicate if any of the data used falls into protected categories under the Human Rights Act,[104] or is information that is not “publicly available”. Knowing this may provide a finger hold for Malik to challenge the decision. Yet even if Malik obtains access to the input information, the weightings given to the variables remain hidden inside the black box.

Furthermore, certain variables, such as Malik’s postcode of South Auckland, may serve as a proxy for ethnicity and indirectly discriminate. International treaties and New Zealand law accommodate for indirect discrimination. Definitions of discrimination in international treaties refer to acts or omissions which have the purpose or effect of disadvantage or adverse treatment.[105] Similarly, New Zealand provides for indirect discrimination through NZBORA[106] and the Humans Rights Act.[107] Where a decision has the

---

[100] Heike Felzmann and Rónán Kennedy “Algorithms, social media and mental health” (2016) 27 Comp & L 31.

[101] See *Hook v Stream Group (NZ) Pty Ltd* [2013] NZEMPC 188 at [29].

[102] Felzmann and Kennedy, above n 100, at 2.

[103] Privacy Act, section 11.

[104] Human Rights Act, s 21.

[105] See Convention on the Elimination of All Forms of Discrimination against Women 1249 UNTS 13 (open for signature 18 December 1979, entered into force 3 September 1985), art 1; and International Convention on the Elimination of All Forms of Racial Discrimination 660 UTNS 195 (open for signature 21 December 1965, entered into force 4 January 1969).

[106] New Zealand Bill of Rights, s 19.

[107] Human Rights Act ss 21 and 65.

effect of adverse treatment based one of the prohibited grounds, such as race, this comes within these acts. Malik's difficulty, however, is establishing that there has been discrimination as to his Jordanian descent. He cannot do so by reference to his own information. To determine whether those of Jordanian descent are more likely to be incorrectly labelled as higher risk compared to those from European ethnicities, it is necessary to have access to all the decisions made. This is not possible for Malik within the individualistic model of the Privacy Act. This is a strong indication that departing from the current framework is necessary to be able to address the concerns of bias.

Access to information through principle 6 may provide the opportunity to challenge the accuracy of the information through principle 8, if Malik finds that it is inaccurate, out of date or misleading.[108] Extending this right of access to look into the internal workings of PILATE, beyond the input and output data, is likely stretching principle 6 too far. Aside from the fact the algorithm itself is not personal information, the previously identified concerns of intellectual property arise. Without access to the algorithm's internal workings, the weightings of any protected categories of information are unknown, as are any correlations between a protected category and predicting recidivism. With the innards of PILATE a mystery, Malik is none the wiser.

b. Prediction of the individual as a result of the algorithm

So long as Malik is capable of being identified at this stage, the prediction is regarded as personal information. Under principle 7, Malik has a right to request the correction of personal information to ensure that the information used is accurate, up to date, complete and not misleading[109] and theoretically could request a correction of his bail determination. To have any foundation for this complaint requires understanding the decision in order to show how it is incorrect.

If the Ministry of Justice is not willing to correct the prediction, they are under an obligation to take steps to attach "a statement of correction sought but not made".[110] The effectiveness of this remedy is questionable. Disagreement as to future events of becoming a flight risk or committing a crime is unlikely to warrant "correction". Thus the effectiveness of Malik disagreeing with this prediction is unlikely to carry much weight. Information which has subjective content will always be harder to correct, than that which is objectively able to be verified.[111]

[108] Privacy Act, s 6, Privacy Principle 8.

[109] Section 6, Privacy Principle 7.

[110] Section 6, Privacy Principle 7(1)(b).

[111] *Macdonald v Healthcare Hawkes Bay and Morrison* [2000] NZCRT 35 (6 December 2000) as cited in Paul Roth Privacy Law and Practice (online ed, LexisNexis, accessed 30 August 2018) at [6.10(b)].

2. The effectiveness of the Privacy Act

The practical assistance the Privacy Act can provide Malik is limited. The transparency requirements informing Malik of how this information is being used and collected are largely redundant due to the broad exception where the information used does not identify the individual concerned. In general, it is unlikely that individuals will know their information is being collected or used in algorithmic decision-making, thus they are unable to seek recourse.

Even if Mailk can show there has been a breach of the privacy principles, it would be hard to bring any action due to the additional requirement of proving harm or loss. If harm is proved, the powers of the Privacy Commissioner are limited, with no legal power to determine a complaint in a binding way. It is worth noting that one of the proposed changes in the Privacy Bill is a recommendation that the Privacy Commissioner may issue a compliance notice to an agency upon breach of an information privacy principle, without having to assess whether that individual has suffered harm.[112] Compliance notices, however, may be of limited use in Malik's situation as it is still necessary to understand how the decision is reached in order to give effect to the rights afforded under many of the information privacy principles.

The major barrier for Malik obtaining effective redress for a decision made by an algorithm is his lack of understanding about why and how he received a high score from PILATE. Under principle 6, the most Malik can do is gain access to the input and output data used for PILATE. Without understanding why the decision about his recidivism rate was made, Malik is unable to verify whether or not the Ministry of Justice is complying with the privacy principles. He is unable to correct, challenge the accuracy, or contest that information from protected categories has influenced the outcome of the decision. As Malik requires a deeper understanding of the reasons pertaining to the decision, the OIA, LGOIMA and the common law "right to reasons" becomes the next port of call.

C) Requesting reasons: the OIA, LGOIMA and the common law

1. The OIA and LGOIMA

The OIA and LGOIMA confer a right to be given access to the reasons for a decision affecting that person, whether this is made by a department, Minister of the Crown, organisation, or by a local authority in respect of the LGOIMA.[113] This right is subject to exceptions[114] and deletion of parts of the information.[115] Where decisions/recommendations have been made by departments, ministers and

[112] Privacy Bill, s 124.

[113] Official Information Act 1982, s 23; and Local Government Official Information and Meetings Act 1987.

[114] Official Information Act, ss 27–32.

[115] Section 17.

specified organisations, the right of access only applies to those with standing. Such restrictions do not apply to local authorities, wherein every person is entitled access.[116]

Prima facie, Malik does have a “right to reasons” for the decision made by PILATE provided there are no good reasons to refuse access per ss 27 and 32. As the corresponding duty to provide reasons only arises once a request has been made, this is predicated on Malik making such a request.

The question then becomes: what is the substance of a "right to reasons” in the context of algorithmic decision-making? The types of explanations or reasons given for algorithmic decisions can be distilled into two types and two timeframes.[117] The first type, “system functionality”*,* refers to the logic or general functionality of an algorithmic decision-making system.[118] For PILATE, this could be the requirements, specifications, decision trees, pre-defined models, criteria and classification structures.[119] The second type, “specific decision”, focuses on the individual circumstances of the subject, Malik, in the algorithmic decision-making process. This includes the weightings of features, machine-defined case-specific decision rules and information about reference or profile groups.[120] In relation to the timing, “ex ante” is when the explanation occurs *prior* to the decision being made. Notably, ex ante decisions can only address system functionality, as any specific decision rationale is not yet known.[121] Conversely, “ex post” occurs *after* the algorithm has made a decision, thus is able to address both system functionality and the rationale of a specific decision.[122]

The OIA and LGOIMA set out three requirements as to what a statement of reasons should include, these being: the findings on material issues of fact; a reference to the information on which the findings are based; and the reasons for the decision or recommendation.[123] The difficulty lies in translating these statutory requirements into reasons sufficient to explain the decision made by PILATE. With no requirement to summarise or supply the information on which the findings are based (only that it be referenced), it would follow the data set of which the model is trained on falls outside the substantive requirements for the “reasons for a decision being made”. “Findings on material issues of fact” suggests providing information that formed the basis for the decision: in other words, an ex post explanation of the rationale and factors showing why Malik has been deemed a flight risk and a danger to the community.

Comments from various cases can be threaded together to form a general conception of the level of adequacy required for “reasons for decisions or recommendations”. Reasons must be sufficient to allow

---

[116] Local Government Official Information and Meetings Act 1987, s 22(1).

[117] Sandra Wachter, Brent Mittelstadt and Luciano Floridi “Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation” (2017) 7 IDPL 76 at 78.

[118] At 78.

[119] At 78.

[120] At 78.

[121] At 78.

[122] At 78.

[123] Official Information Act, s 23(1); and Local Government Official Information and Meetings Act, s 22(1).

anybody with a power of review to understand the process of thought to which a conclusion is reached.[124] Reasons must also allow those with affected interests to understand the basis of the decision, and allow the public to know and comprehend the standards a decision maker sees as important.[125] The importance of enabling affected persons to determine whether to challenge the decision has also been deemed an element in determining the adequacy of a statement of reasons.[126] These comments form (at least in theory), a basis that Malik should be given reasons that go towards explaining why such a decision was made about him. The notion that a "statement of reasons must be intelligible to the recipient"[127] as well as be of "sufficient precision to give him or her a clear understanding of why the decision was made"[128] denotes that Malik should be given an explanation at a level that allows him, or a review tribunal to understand and fairly assess the decision that has been made.

The true challenge lies in how best to achieve this. Comments such as "reasons should trace the steps of reasoning in sufficient detail to allow the recipient to understand how the decision was reached",[129] become difficult when the reasoning process is not necessarily traceable. It has been suggested that where weighting exercises are involved, the weight ascribed to each should be disclosed[130] and yet where weightings are constantly adjusted every time a decision is made, such information may not be useful in understanding how the algorithm works.

2. Common law right of reasons

While the duty to give reasons exists in common law,[131] there is debate as to whether access to information legislation acts as a complete or partial code block to this common law duty. The majority of the United Kingdom Supreme Court in *Kennedy v Charity Commission*[132] suggest that the common law approach acts as an alternative to legislation.[133] Even though the information requested was subject to an "absolute exemption" from disclosure under the United Kingdom's Freedom of Information Act 2000, the common law right to reasons would be available.[134] If a similar stance is taken in New Zealand, there

---

[124] *Singh v Chief Executive Officer, Department of Labour* [1999] NZAR 258 (CA) at 263.

[125] *Re Vixen Digital Limited* [2003] NZAR 418 as cited in Graham Taylor and Paul Roth *Access to Information* (2nd ed, LexisNexis, Wellington, 2011) at 186.

[126] *Elliot v Southwark London Borough Council* [1976] 2 All ER 781, [1976] 1 WLR 499 (CA) at 508 per James LJ; adopted in *Re Palmer and Minister for the Capital Territory* (1978) 23 ALR 196 at 206–207 as cited in Taylor and Roth, above n 125, at 194.

[127] *Elliot v Southwark London Borough Council,* above n 126, at 508 per James LJ; adopted in *Re Palmer and Minister for the Capital Territory,* above n 126, at 206–207 as cited in Taylor and Roth, above n 125, at 194.

[128] *Re Poyser and Mills' Arbitration* [1963] 1 ALL ER 612, [1964] 2 QB 467 at 478 per Megaw J; adopted in *Re Palmer and Minister for the Capital Territory,* above n 126, as cited in Taylor and Roth, above n 125 at 194.

[129] *Re Palmer and Minister for the Capital Territory,* above n 126, as cited in Taylor and Roth, above n 125, at 197.

[130] *Re Salazar-Arbelaez and Minister for Immigration and Ethnic Affairs* (1977) 18 ALR 36 (AAT) at 38 as cited in Taylor and Roth, above n 125, at 197.

[131] *Lewis v Wilson and Horton* [2000] 3 NZLR 546 (CA) as cited in Taylor and Roth, above n 125, at 187. The Supreme Court in *Lewis v Wilson and Horton* indicated that a common law duty to provide statement of reasons exists in New Zealand in an appropriately argued case.

[132] *Kennedy v Charity Commission* [2014] UKSC 20, [2014] 2 WLR 808, [2014] 2 All ER 847 [Kennedy].

[133] Tim Cochrane "A common law duty to disclose official information?" [2014] NZLJ 385 at 385.

[134] At 385.

is the potential for the common law duty to operate as an alternative to the OIA. This has the benefit of facilitating information requests from those exercising public power not subject to the OIA and agencies seeking to rely on a conclusive withholding ground.[135] Per Lord Toulson, the common law may produce "a more just result" because a court would be "able to exercise a broad judgement about where the public interest lies in infinitely variable circumstances".[136] For Malik, this means that his right to reasons is more deeply rooted in common law than that afforded under legislation and a potential alternative if a right to reasons was blocked under the OIA.

Theoretically, the common law duty may be wider than that afforded under the OIA and LGOIMA. Yet the content of what this duty and right is understood to mean in modern common law is largely shaped by access to information legislation.[137] This is demonstrated in *Singh v Chief Executive Officer Department of Labour*, where the court interpreted the requirement to state reasons in the Immigration Act 1986, s 36, as akin to that required under the OIA, s 23(1).[138] Articulations of the content of this duty are vague, "vary[ing] in accordance with the role of [the] tribunal and the nature of the hearing".[139] The underlying rationale of a "right to reasons" aligns with the justifications of promoting explainability within algorithmic decision-making. It is considered to overcome the "real grievance" experienced when one is not told why something affecting them has been done.[140] Further, it enables those affected to see how the decision was reached and whether some error has been made so as to enable them to make an informed decision as to whether to challenge it.[141] The justifications of assuring the wider public of openness and legitimacy with decision-making processes[142] can be drawn on for support in favour of a system functionality explanation. In one sense, making algorithmic code and training data sets available is necessary for peer review to ensure the wider public of legitimacy of the decision-making process. To have the desired impact, it would be necessary to make this information available to the wider public, rather than just to Malik, who lacks the expertise to review algorithmic code. Making algorithmic code and training sets available, however, lies outside of the legislatively prescribed requirements for reasons given under the OIA, with no requirement to supply information on which the findings are based.[143] At a high level, support can be found for explainability in algorithmic decision-making processes, yet the practical translation of this duty remains inconclusive and does not extend to making the algorithmic code or training data available to the individual, nor to society at large.

---

[135] At 388.

[136] *Kennedy,* above n 132, at [140] per Lord Toulson.

[137] Taylor and Roth*,* above n 125, at 185.

[138] *Singh v Chief Executive Officer, Department of Labour,* above n 124, at 263 per Keith J.

[139] *Television New Zealand Limited v West HC Auckland* CIV-2010-485-2007, 21 April 2011 at [82].

[140] *Re Poyser and Mills' Arbitration,* above n 128, at 477–478; adopted in *Re Palmer and Minister for the Capital Territory,* above n 126*,* as cited in Taylor and Roth, above n 125, at 185.

[141] *Singh v Chief Executive Officer, Department of Labour, above n 124*; *Iveagh v Minister of Housing and Local Government* [1964] 1 QB 395 (CA) at 410 per Lord Denning MR at 405 per Lord Russell LJ; and adopted in *Re Palmer and Minister for Capital Territory,* above n 126*,* as cited in Taylor and Roth, above n 125, at 185.

[142] *Singh v Chief Executive Officer, Department of Labour,* above n 124, at 263 per Keith J.

[143] See Official Information Act, s 23(1)(c).

3. Standard of reasons given

As algorithms are employed to replace or supplement decisions made by humans, it would follow, the same standard of reasons expected of human decision makers would apply to algorithms. According to the psychology of the reasoning process, humans tend to form an initial sense of a conclusion and justifications are formed after the fact.[144] This ability to invent post hoc rationalisations has the effect of obscuring the real (internal motivations) and processing logic.[145] Despite being artificial and not reflective of the true processes that occurred to reach the decision, we accept these reasons from decision makers. Human reasons for decisions tend to be "practical", focused on the justification of the action (as opposed to the "epistemic" or "theoretical reason" which concern the justification of beliefs).[146] If the same standard of "human reason giving" is applied to algorithms, then an explanation should focus on a practical justification, rather than giving excessively detailed and technical reasons.

There is an argument that algorithms should be held to a higher standard of explainability than humans. As I previously acknowledged, with human decision-making being imperfect, if algorithms are able to provide clearer explanations of how a decision is reached, then this may have the effect of lifting the bar for transparency across all decision-making. The major barrier to this is a technical one; "full technical transparency is difficult and even impossible for certain kinds of AI systems in use today" and may not be "appropriate or helpful in many cases".[147] Ultimately, if obtaining a "right to reasons" is predicated on the inference that decisions made by algorithms are analogous or at least comparable to those made by humans, it would follow that the same standard of reasons should apply. Explanations of algorithms should be pitched at practical or intentional stance explanations, rather than aimed at the algorithms' complex architectural innards.[148]

4. Where does this leave Malik?

Malik is afforded a "right to reasons" for the decision made about his recidivism under the OIA and LGOIMA. If this is blocked under legislation, he can look to the common law as an alternative path, provided the United Kingdom Supreme Court decision[149] is applied in New Zealand. While various threads can be pulled together to show support for an ex post, decision specific explanation, how this is carried out in practice remains unseen.

---

[144] See Jerome Frank *Law and the Modern Mind* (Peter Smith, Gloucester, 1970) at 108.

[145] John Zerilli and others "Transparency in Algorithmic and Human decision-making: Is There a Double Standard?" (2018) 31(3) Philosophy & Technology 1 at 12.

[146] At 14.

[147] House of Lords Select Committee on Artificial Intelligence *AI in the UK: ready, willing and able?* (House of Lords, HL Paper 100, April 2018) at 38.

[148] Zerilli and others, above n 145, at 39.

[149] *Kennedy,* above n 132.

What I have revealed in this chapter is that an effective explanation for an algorithmic decision must be explainable at a level that allows an individual or a review tribunal to understand and fairly assess the decision that has been made. I have argued that the standard of explanation should be pitched at a practical level. This is analogous to the level of explanation required from humans and it is more meaningful to individuals compared to lengthy technical explanations. An effective explanation is necessary to give effect to the information privacy principles under the Privacy Act, in order to correct or challenge the accuracy of the information used in the decision. The current legislative framework is very limited in promoting transparency requirements for algorithms. The broad exceptions to principles 2 and 3 of the Privacy Act mean that individuals are unlikely to be aware that information is being collected to make a decision about them. Furthermore, making the algorithmic code and training data available does not come within the substantive requirements for providing individuals with reasons under the OIA and LGOIMA. With a holistic view necessary to assess the existence of bias in algorithmic decision-making, it is evident that the current individualistic framework is unable to address issues of collective harm.

In chapter III we assess the various paths available to achieve the objectives of providing the individual with an explanation that allows them to understand the decision made, increasing transparency by promoting peer review and ways to mitigate bias. Achieving these objectives is necessary to enhance the digital rights of those subject to algorithmic decisions in New Zealand.

CHAPTER THREE: FILLING THE LACUNA IN THE LAW

A) Where to now?

Chapter I outlined the concerns algorithmic decision-making presents to digital rights. Through Malik in chapter II, it was established that the current legislative framework is unable to ameliorate these concerns. In this chapter, we reach a metaphorical fork in the road. There are multiple paths New Zealand can take to promote effective explainability of algorithmic decisions and provide meaningful recourse for Malik. The first is following in the footsteps of the European Union to implement similar measures aimed at automated decisions as found in the GDPR. The second is to strengthen existing legislative rights in New Zealand with the introduction of various policy-based objectives, self-regulation and technical solutions. Finally, there is the option of instating a regulatory oversight body to set design and performance standards and monitor the use of algorithms by public bodies.

Throughout this chapter, we take a brief tour down the various paths available to assess how the following objectives can be achieved to: (i) provide the individual with an explanation that allows them to understand the decision made; (ii) promote peer review of the algorithm and increase transparency; and (iii) address issues of bias and collective harm. I ultimately contend that the most prudent direction forward is the implementation of a hard-edged oversight body. Adopting such a model will ensure that New Zealand is in a position to anticipate future problems presented by the advance of algorithms.

B) Following the European Union

The GDPR contains several articles aimed at automated decisions, coined by commentators as a "right to an explanation" and the requirement of having a “human in the loop”. These provisions are not new to European law, formerly existing in the Data Protection Directive 1995[150] and can be traced back further to early French data protection law.[151] Nonetheless, arts 13–15 and 22 have been the subject of academic debate as to their effectiveness in providing meaningful recourse.

1. “Right to an explanation”

There is contention over whether the “right to an explanation” translates into an explanation that is meaningful to allow the subject to understand how a decision was made. Oxford scholars, Goodman and Flaxman, argue a “right to an explanation” is afforded when the plain reading of arts 13–15 are read

[150] Data Protection Directive 1995, arts 12(a) and 15.

[151] Lee A Bygrave “Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling” (2001) 17 Computer Law & Security Report 17 at 17 as cited in Michael Veale and Lilian Edwards “Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling” (2018) 34 CLS Rev 398 at 398.

in conjunction with Recital 71.[152] The "data controller" must provide the "data subject" with information of the "existence of automated decision-making, including profiling" and "meaningful information about the logic involved as well as the significance and the envisaged consequences of such processing for the data subject".[153] According to Recital 71, a data subject should have the right to "obtain an explanation of the decision reached" and challenge it after a human has reviewed it.[154] Such an explanation means the "trained model can be articulated and understood by a human" and "any adequate explanation would, at a minimum, provide an account of how input features relate to predictions".[155] This reading indicates the substance of a "right to an explanation" is one that is ex post and about a specific decision. This appears to fulfil the type of explanation desired in chapter II.

Conversely, Watcher, Mittelstadt and Floridi give a detailed account as to why the "right to an explanation" does not exist and is more akin to a "right to be informed".[156] They contend that the information and access rights afforded by s 2 of the GDPR only go as far as extending to general ex ante information about the algorithm, rather than an ex post explanation of how a decision relating to a particular individual was reached.[157] Conferring a "right to an explanation" is based on conflating the legally binding notification duties (arts 13–14) with the non-binding duty in Recital 71.[158] As notification duties exist *before* a decision is made when the data is collected for processing, it is difficult to use this to justify an ex post explanation of a specific decision.[159] The language of "envisaged consequences" in arts 13–14 also points to an ex ante explanation. If Recital 71 were legally binding, potentially this provision would require an ex post explanation of specific decisions.[160]

The guidance provided by the Article 29 Data Protection Working Party ("A29WP") on "automated individual decision-making and profiling" also advances the notion that a "right to an explanation" under the GDPR is not as meaningful as first envisaged.[161] The guidelines suggest it is unnecessary to provide information about the "innards" of the decision-making process.[162] Rather, "meaningful information" is restricted to providing: (i) input information about the data subject; (ii) relevant information provided by others (i.e. references as to Malik's social isolation tendencies) and (iii) relevant public information used in

---

[152] See Bryce Goodman and Seth Flaxman "European Union Regulations on Algorithmic Decision-making and a Right to Explanation" (2017) 38 AI Magazine 50.

[153] General Data Protection Regulation 2018, arts 13(2)(f) and 14(2)(g).

[154] Andrew D Selbst and Julia Powles "Meaningful information and the right to explanation" (2017) 7 IDPL 233.

[155] Bryce Goodman and Seth Flaxman "European Union Regulations on Algorithmic Decision-making and a Right to Explanation" (2017) 38 AI Magazine 50 at 55.

[156] Wachter, Mittelstadt and Floridi, above n 117.

[157] Veale and Edwards, above n 151, at 399.

[158] Wachter, Mittelstadt and Floridi, above n 117, at 82.

[159] At 82.

[160] At 80.

[161] Article 29 Working Party (A29WP) "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" (WP 251, 3 October 2017).

[162] Veale and Edwards, above n 151, at 400.

the decision (i.e. criminal record).[163] Following this, Malik is not afforded anything beyond what can already be requested under access to information legislation in New Zealand.

2. "Human in the loop"

Article 22 provides for "the right not to be subject to a decision based solely on automated processing".[164] The practical application of art 22 is contingent on whether the decision is "solely" made by automatic processes and whether it has "legal" or "similarly significant" effects. Read literally, "solely" implies any human involvement in the decision-making process means that art 22 is not applicable.[165] There is concern that art 22 may be subverted with the presence of a nominal human to "rubber stamp" automated decisions. The A29WP guidelines do, however, suggest that for the system to be categorised as not "solely" automated, "meaningful human input" is required. A human should have "authority and competence" to alter a decision, rather than simply being a "token gesture".[166] Yet even if a human is genuinely positioned to provide a second opinion on an automated decision, there is a tendency for both naive and expert humans to trust in automated logic.[167] The Aviation Safety Reporting System contains many reports from pilots that link their failure to monitor, to excessive trust in automated systems such as autopilots.[168] Similarly, it has been found that skilled subject matter experts tend to misplace trust in the accuracy of diagnostic expert systems.[169] Decision-making systems could effectively operate as if they were fully automated despite having a human involved in the process.[170]

Watcher et al highlight that art 22(1) can be interpreted in two distinct ways: as a prohibition or as a right to object.[171] Such ambiguity has existed since the 1995 Directive and is yet to be resolved.[172] Read as a prohibition, the Ministry of Justice would be obligated to not engage in automated decisions until showing the requirements set out in art 22(2)(a)–(c) have been met. If interpreted as a "right to object" to automated decision-making, this would not apply if one of the requirements in art 22(2)(a)–(c) have been met. The key difference in these interpretations is the action required by the individual. The latter interpretation is a weaker one in that it relies on Malik's awareness of the decision and his willingness to intervene.[173]

---

163 Veale and Edwards, above n 151, at 400.

164 General Data Protection Regulation, art 22.

165 Veale and Edwards, above n 151, at 400.

166 At 400.

167 Linda J Skita, Kathleen Mosier and Mark D Burdick "Accountability and automation bias" (2002) 52 Int J Human-Computer Studies 701 as cited in Veale and Edwards, above n 151, at 400.

168 See Indramani L. Singh, Robert Molloy and Raja Parasuraman "Individual differences in monitoring failures of automation" (1993) 120(3) The Journal of General Psychology 357 as cited in Michael Lewis, Katia Sycara and Phillip Walker "The Role of Trust in Human-Robot Interaction" in Hussein A Abbass, Jason Scholz and Darryn J Reid (eds) *Foundations of Trusted Autonomy. Studies in Systems, Decision and Control* (Springer, Cham 2018) at 136.

169 See Richard P Will "True and false dependence on technology: Evaluation with an expert system" (1993) 7(3) Computers in Human Behaviour 171 as cited in Lewis, Sycara and Walker, above n 168, at 136.

170 Skita, Mosier and Burdick, above n 167 as cited in Veale and Edwards, above n 151, at 400.

171 Wachter, Mittelstadt and Floridi, above n 117, at 94.

172 At 94.

173 At 94.

3. Summary

While the European Union appears to be taking substantive action in addressing the need for explainability in algorithmic decision-making, in practice these measures would not add much to what already exists in New Zealand. Enacting similar provisions would not guarantee an explanation that is ex post or decision specific to allow Malik to understand how a decision regarding his bail has been reached. Even if a legally binding equivalent of arts 13–15 were instated, requiring an individual-orientated explanation of the logic involved, the technological barriers of achieving this remain. Legislating rights is one thing, yet translating this into practice is another.

Notably, European law does at the very least have a "right to be informed". Notification requirements apply regardless of whether the data is collected from the data subject. This addresses one of the fundamental problems of New Zealand law highlighted in chapter II. Under the Privacy Act, transparency requirements are not applicable when the data is not collected directly from the individual. Where individuals are not aware that an algorithmic decision is being made about them, they are unable to take further action in exercising their rights. How this notification duty is carried out in practice remains to be seen.

Introducing a similar requirement to art 22 in New Zealand could serve as a protective measure for Malik. Yet the effectiveness is limited by the degree of human involvement and the innate trust placed in machines. If judges tend to rely on the risk assessment score by PILATE they may not see it fit to override this. Further, whether it is cast as a prohibition or a right of objection is an ambiguity yet to be resolved that significantly impacts the obligations of the Ministry of Justice and the action required by Malik. Ultimately, the measures enclosed within the GDPR do not fulfil the desired objectives to enhance digital rights when individuals are subject to algorithmic decisions.

C) Strengthening existing legislation

Last chapter, left Malik with a theoretical basis for an explanation of an algorithmic decision. It was unclear how this duty could be translated into practical terms to provide him with the understanding necessary to challenge the decision. Implementing policy or legislative changes is one way in which these rights can be enhanced to set standards of what reasons should be given for an algorithmic decision.

1. Implementation of guidelines for "meaningful" explanations

In chapter II, I argued that an effective explanation for an algorithmic decision requires reasons that allow an individual or a review tribunal to understand and fairly assess the decision that has been made. I also suggested that the standard of reasons given need not be higher than that expected of human decision makers. Rather, it may be pitched at a practical level or analogous to intentional stance explanations. To translate this into practical terms, guidelines may be formulated to set the standard of explainability of what reasons are sufficient to produce an explanation that is meaningful.

Various research has been conducted on how to practically achieve explanations for machine learning.[174] Despite the technological barriers, there is potential to provide explanations that fulfil policy goals of transparency, accountability and fairness.[175] One such explanation style, which is pitched at an intentional stance and conforms to the demands of practical reason, is detailed below. Hypothetical examples illustrate how this explanation might look in Malik's situation.

[Table 1]: Example of an explanation style[176]

Type of explanation	Description	Example
Input Influence	A list of input variables is provided labelled with a quantitive measure indicating how much they influence the decision (whether positively, negatively or not at all). Static (s) and dynamic (d) predictors are also differentiated. Static predictors are unchangeable by individual effort. Dynamic variables allow for the individual's risk rating to change as a result of their own efforts and/or any treatment they receive.	PILATE has assessed your information in order to predict your tendency to recidivate. On a scale of 1 (low-risk) to 10 (high-risk), you have received a score of 8. The various factors taken into account are detailed below, along with an indication of how much they negatively (—) or positively (+) impacted the outcome. Variables that made no difference to the final output are marked (/). • Age: (/) (s) • Gender: (/) (s) • Ethnicity: (/) (s) • Previous convictions: (— —) (s) • Stability of residence: (—) (d) • Substance abuse: (— —) (d) • Family criminality: (+) (s) • Separated parents: (+) (s) • Social environment: (— — —) (d) • Social isolation: (—) (d) • Employment: (++) (d)
Sensitivity based	This specifies in relation to each input variable, the extent a variable would have to differ in order to change the output.	• Better management of substance abuse would decrease the recidivism rate to > 6 • Improving social isolation tendencies would decrease the recidivism rate to > 6.
Case based	A similar example from the training data set is provided to show the analogous characteristics of another subject with the same outcome.	This decision was based on thousands of similar past cases. For example, a 26 year old male with one previous conviction of breaching a protection order and unstable living conditions was deemed high-risk with a score of 8.
Demographic based	Reveals characteristics of those who were similarly classified and the outcome for these decisions.	• 12% living in the same region received a lower score than 8. • 10% with previous convictions received a lower score than 8. • 5% of those with previous substance abuse received a lower score than 8.

[174] Reuben Binns and others "It's Reducing a Human Being to a Percentage; Perceptions of Justice in Algorithmic Decisions" *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, 26 April 2018).

[175] At 3.

[176] At 4.

By no means is this a complete answer to creating meaningful explanations of algorithmic decisions. Especially for algorithms involving hundreds of input variables, this explanation may be too long-winded and overwhelming to be useful. This does, however, represent an illustration of what standards could be considered to provide explanations pitched at a practical level, that would allow an individual to better comprehend how the decision was reached. Here, Malik can see that his social environment, previous convictions and substance abuse have appeared to be influential on his high score. The differentiation of static and dynamic predictors highlight which variables are changeable by individual effort and which are not.[177] Knowing this, may place him in better stead to challenge the accuracy or the fairness of using this input data under the Privacy Act, or more generally through paths of judicial review.

2. Mandating public disclosure

In chapter I, I outlined the interest in disclosing the training data and code to be able to assess the strength of an algorithm. Chapter II further built on this with the importance of public legitimacy and openness being a fundamental aspect of promoting explainability. It also outlined that making the algorithmic code and training data set available, fell outside of what was required for “reasons” under the OIA and LGOIMA.

The New Zealand government could require that the code and training data of all algorithms in use be publicly disclosed. Similar measures have been taken by France. The administration of Emmanuel Macron recently announced that all algorithms developed for government use will be open-source with data used made publicly available so that society at large can verify their correct application.[178] This enables those with the technical expertise to peer review the algorithm. This has the effect of ensuring the public of legitimacy and promoting openness and transparency in the decision-making process.

Making the code and training data publicly available does, however, raise issues previously encountered in chapters I and II. There are concerns about leaking personal information through the training data. As New Zealand privacy law does not anticipate the possibility of individuals being re-identified, this is an area of law that may need strengthening before mandating disclosure. Concerns relating to commercial sensitivity are also relevant, especially if algorithms are being sourced from private companies, rather than the government funding their development as in France. Concerns of “gaming the system” were also raised. As discussed in chapter I, the validity of this comes down to the type of information used and whether it can be manipulated by the individual to change the outcome of the algorithm. Policy may dictate a diluted form of openness, wherein a “national interest” exception could apply to keep algorithms pertinent to detecting tax fraud or terrorism under wraps. There is a sliding scale of transparency as to the extent algorithmic code is made open-source and whom this is made available to.

---

[177] Bakker, O’Malley and Riley, above n 14, at 10.

[178] Jackie Snow “The president of France is promoting AI, European style” (2 April 2018) Technology Review <www.technologyreview.com>.

3. A technical solution: “Responsibility by design”

The government could opt for a technical solution and only employ algorithms that are explainable by design. “Explainable Artificial Intelligence” or “XAI” is described as “any machine learning technology that can accurately explain a prediction at the individual level”.[179] While still under development, this technology may be the answer to providing explanations to individuals that enable them to understand how a decision has been reached. Numerous large technology companies of the likes of Google, IBM and Microsoft have already expressed their commitment to developing interpretable machine learning systems.[180] Choosing to only use XAI in the public sector could produce a higher market demand for explainable algorithmic systems. Developers would be incentivised to develop such algorithms with the necessary capabilities or face the risk of losing market share to competitors.[181] It is also open to the government to choose to delay the deployment of algorithms that are unable to generate explanations for their decisions — an approach suggested in the House of Lords Select Committee Report on Artificial Intelligence.[182]

4. Self-regulation models

Another option is the implementation of self-regulatory models for government agencies to ensure their use of data, particularly personal information, is consistent with individual rights. The Ministry of Social Development has been working on an initiative called the Privacy Human Rights and Ethics framework (“PHRaE”).[183] This tool ensures people’s privacy, human rights and ethics are considered from the design stage of a new initiative. The goal is to prompt discussion and identify the risk of impinging on rights early on so the design can be modified.

This is somewhat analogous to the proposed “algorithmic impact statements” (“AIS”) incorporated as part of New York’s recent legislative changes aimed at addressing the discriminatory effects of predictive policing algorithms.[184] These statements require police departments to evaluate the efficacy and potential discriminatory effects of all available choices for predictive policing technologies.[185] It means that relevant information is made available to a larger audience that may play a role in the decision-making process and the implementation of that decision.[186] Like PHRaE, this process is primarily procedural, requiring

---

179 “Explainable AI” simMachines <https://simmachines.com>.

180 See House of Lords Select Committee on Artificial Intelligence, above n 147, at 39. See also Charles Towers-Clark "Can We Make Artificial Intelligence Accountable?” (19 September 2018) Forbes <www.forbes.com>.

181 New and Castro, above n 36, at 29.

182 House of Lords Select Committee on Artificial Intelligence, above n 147, at 40.

183 “The Privacy, Human Rights and Ethics Framework” Ministry of Social Development <www.msd.govt.nz>.

184 See Dillon Reisman and others *Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability* (AINow, 2018).

185 Andrew D Selbst “Disparate Impact In Big Data Policing” (2017) 52 Ga L Rev 109 at 110.

186 At 169.

departments to consider the question and allow society to understand the scope of the problem. This is an important first step in determining whether future intervention is required.[187]

While a step in the right direction towards considering potential discriminatory and privacy effects of the proposed algorithm initiatives, such measures do not prevent discrimination. Mandating such practice may result in the process being perceived as cumbersome. If agencies are compelled to consider the discriminatory effects of their algorithm through these "self-regulatory" processes, there is a risk of agencies "going through the motions" to merely "tick the box". This is at the expense of genuine consideration being given to the potentially discriminatory effects of algorithms. Consequently, the oversight of an independent actor is necessary to ensure appropriate consideration is given and algorithmic decision-making processes comply with the relevant standards in order to more effectively enhance digital rights.

5. Summary

Policy initiatives may be well placed to strengthen the existing "right to reasons" in New Zealand law by outlining the standard expected of an explanation to an algorithmic decision. This can be supplemented by policy decisions of the government to only employ algorithms that are explainable. In this regard, the law may be enhanced to provide individuals with an explanation that allows them to understand how the decision was made. The second objective of increasing transparency can be addressed by mandating algorithmic code to be open source; however, the potential issues relating to commercial sensitivity, privacy and manipulation require further consideration. In regards to the final aim of addressing the issue of bias and collective harm, the self-regulatory measures such as PHRaE and AIS encourage the consideration of potential discriminatory effects. They do not, however, provide an ultimate solution. Issues of bias and collective harm are unable to be addressed from inside New Zealand's current individualistic framework. To address the issue of collective harm or bias in algorithms, a top-down model, such as an oversight body is necessary.

D) Oversight regulatory body

Many in the technology industry are vocal with their calls against regulation.[188] The primary concern is the stifling effect this could have on innovation.[189] While industry self-regulation and market forces are advocated as sufficient protection to mitigate the harms algorithms bring, relying on these mechanisms is not a viable option to protect rights. To address the unique threats posed by algorithmic decision-making to individual rights, stronger action is required.

---

[187] At 110.

[188] See Daniel Saraga "Opinion: Should Algorithms Be Regulated?" (3 January 2017) phys.Org <https://phys.org>. See also New and Castro, above n 36.

[189] Saraga, above n 188.

It is my contention that the objectives of creating understandable explanations, increasing transparency and mitigating bias can be achieved through top-down regulation. This model would have two main roles. First, the oversight body would act as an expert panel in forming guidelines for levels of explanation, design standards and classification of different algorithms. This aspect is aimed at promoting transparency and explainability in algorithmic decision-making processes. Second, an element of pre-market review is necessary to prevent algorithms that present an unsatisfactory level of risk to digital rights from being introduced. Through classification, varying levels of scrutiny can be applied to algorithms depending on the level of explainability and the impact the decision has on one's rights. This broad mandate lends itself towards a hard-edged regulatory approach over a soft-touch one. This allows continued oversight of approved algorithms to ensure bias does not creep in through machine learning processes in a way that compromises digital rights. With a bird's eye view of the different algorithms in use and their application to various decisions, an oversight body would be well placed to obtain the information necessary to ascertain whether bias has unlawfully influenced a decision.

There are numerous operational matters and policy decisions to be made by the government of the day in implementing such a regulatory body. A series of considerations are discussed below to indicate the shape an oversight body might take.

1. The harm of concern

The harm regulated by the oversight body must be broader than specific harm or loss to individuals, as seen under the Privacy Act. A broad focus enables issues of bias to be addressed. At the pre-market review stage, the focus is on the level of risk of subjecting individuals to algorithms that are discriminatory or poorly designed. For example, whether the algorithm has a disproportionate impact on different groups or protected categories under s 21 of the Human Rights Act.

This is distinct from the consequentialist or harm based approach advocated by some writers.[190] This proposes when an algorithm causes harm, regulators should evaluate whether the operator can demonstrate that, in deploying the algorithm, the operator was not acting with intent to harm or with negligence and to determine if an operator acted responsibly in its efforts to minimise harms from the use of its algorithm.[191] This approach fails to recognise the harms of bias. Algorithms can have destructive consequences even with good intentions. Whether intentional or unintentional, algorithms can replicate and reinforce human bias into its system to exacerbate existing inequalities. The purpose of an oversight body is to remove the constraints imposed by an individualistic framework to allow the issue of collective harm to be monitored.

---

[190] New and Castro, above n 36.
[191] At 4.

2. A centralised body

Having a centralised body of regulation is preferable to the alternative of spreading oversight between different agencies. The primary argument for the latter is the requirement of content specific knowledge about the types of decisions an algorithm is dealing with. New and Castro, from the Centre of Data Innovation, argue that all human decisions are not regulated by one government agency, thus it would be ill-advised to have one body regulating all decisions made by algorithms.[192] The crux of this argument relies on a consequentialist approach to the identification of harm; "what constitutes harm in consumer finance is dramatically different from what constitutes harm in health care".[193] It is my contention the harm of concern should be the algorithm itself, rather than the context in which it is applied. As concerns arise from the inherent complexity of machine learning algorithms, it follows that the technical aspect is what attracts expert evaluation. As an oversight body is focused on the technical aspects of the algorithm, a centralised body is preferable. The alternative of placing regulation in multiple agencies will result in tunnel vision and inconsistency in application.[194]

3. The function of the oversight body

An oversight body would act as a standards setting body, coordinating and developing classifications, design standards and best practices. Part of this may include setting standards for explanations to be "understandable" as previously discussed.

i. Classification

Classification is important to determine the extent to which regulation is required based on the risk of the algorithm in use. Low-risk decisions that are explainable and have very little impact on rights do not need to be subject to regulatory oversight simply because they are an algorithm. For example, the proposed ACC algorithm of "Cover Decision Service" filters out the claims that are straightforward and can be accepted without manual review, from the claims that are referred through to staff for processing.[195] The algorithm is purely procedural and the ultimate decision of whether cover is granted lies with a human. In contrast, higher levels of scrutiny may be justified for algorithms used in the criminal justice context where the infringement on rights is more at risk. A classification process would recognise the distinction between the majority of algorithms with deterministic outcomes and machine learning algorithms where the outputs are hard to determine. This is illustrated in the table below.

---

[192] At 14.

[193] New and Castro, above n 36, at 14.

[194] Tutt, above n 7, at 114.

[195] *Statistical modelling to support the ACC automation cover decisions and accident description August 2018* (Accident Compensation Corporation, August 2018) at 5.

[Table 2]: A Possible Qualitative Scale of Algorithmic Complexity[196]

Algorithm Type	Nickname	Description
Type 0	"White Box"	Algorithm is entirely deterministic.
Type 1	"Grey Box"	Algorithm is non-deterministic, but its non-deterministic characteristics are easily predicted and explained.
Type 2	"Black Box"	Algorithm exhibits emergent properties making it difficult or impossible to predict or explain its characteristics.
Type 3	"Sentient"	Algorithm can pass a Turing Test (i.e. has reached or exceeded human intelligence)
Type 4	"Singularity"	Algorithm is capable of recursive self-improvement (i.e. the algorithm has reached the "singularity").

ii. Design and performance standards

An oversight body would be well placed to establish guidance for design, testing and performance to ensure algorithms are developed with adequate safety and margins of error. Policy choices can be made for the allowance given for algorithms to improve over time, acknowledging that requiring an error rate of zero would stifle innovation.[197] If the algorithm is proven to have a lower error rate than a decision made by humans and not disproportionately skewed to disadvantage or advantage a particular group this may pass the set standards.

If, through testing, the model is deemed to be accurate, yet still has discriminatory effects, the oversight body is positioned to make policy choices between the accuracy and fairness of the model. There may also be a tradeoff between explainability of an algorithm and its accuracy. Typically, an algorithm's accuracy increases with its complexity. The more complex an algorithm is, the more difficult it is to explain.[198] There is the argument that explainability should take a back seat as treating algorithmic transparency or explainability as the end goal may fail to prevent potential harm and limit innovation. This is most pertinent to algorithms wherein accuracy is paramount, for example in driverless cars; any slight compromise to navigational accuracy could be enormously dangerous.[199] When it is an individual's right at issue, however, explainability may be held in a higher regard than accuracy. As previously mentioned, XAI technology could be part the design standards to ensure explainability in algorithmic design from the outset. Ultimately any trade-off between fairness and accuracy of a model is inherently political. This will

[196] Tutt, above n 7, at 107.

[197] New and Castro, above n 36, at 6.

[198] Jason Brownlee "Model Prediction Accuracy Versus Interpretation in Machine Learning" (1 August 2014) Machine Learning Mastery <https://machinelearningmastery.com>.

[199] New and Castro, above n 36, at 13.

inevitably be an ongoing issue, if and when algorithmic decision-making becomes more explainable and transparent.

4. Hard-edged vs. soft touch regulation

If we accept that a regulatory body is inevitable, opting for the strongest model will mean New Zealand is better placed to address the inevitable advancement of algorithms. Soft touch regulation is less effective in addressing the objectives to increase transparency, explainability and mitigate bias. Taking conservative, piecemeal steps will be unable to address the future concerns presented by algorithms.[200]

A soft touch regulatory approach provides a halfway point for concerns regarding stifling innovation and compromising commercial sensitivity and privacy. The oversight body would act as a third party intermediary that government departments could choose to consult whether their use of algorithms meets the relevant design and performance standards. Disclosing code and training data to a small group of experts lessens the threat to commercial sensitivity and concerns of privacy leaks of training data. This, however, limits the second objective of transparency in enabling wider peer review.

A halfway approach comes with compromise. Even if algorithms are certified and comply with standards from the outset, due to the nature of the machine learning process wherein algorithms learn and adjust their output, there is potential for bias to creep in. It would therefore, be incumbent to have regular checks and continued oversight for approval, to ensure the algorithm still complies with standards. Under the New York legislative model, there is no compulsion for governmental bodies to comply with regulatory processes. This soft touch approach has no ability to prevent harmful algorithms being implemented in the first instance. The concerns of potentially harmful algorithms imbued with bias that are inexplicable and completely opaque remain. Having an opt in regime is therefore not conducive to the objectives necessary to enhance rights.

A hard-edged regulatory model allows an oversight body to undertake a pre-market review to ensure an algorithm's efficacy, safety and transparency have been proven before being introduced.[201] Agencies are compelled to go through this process, creating uniformity in the application of algorithms. A hard-edged approach better allows an oversight body to engage in ex ante regulation wherein the types of harms caused by algorithms can be mitigated through careful efforts during the design and development stage and extensive pre-market testing. Relying on ex post judicial enforcement is unlikely to be effective in ensuring that unsafe algorithms are kept off the market. Regular inspection could be mandated to keep the approval rating, similar to that of a “warrant of fitness” for algorithms, to ensure that the system adheres to the set standards and bias has not crept in.

---

[200] Tutt, above n 7, at 119.
[201] See Tutt, above n 7.

E) Summary

This chapter has argued the best way forward to achieve the objectives of transparency, explainability and mitigating bias, is through a hard-edged oversight body. Following an approach akin to the GDPR, does not add anything substantially different to what already exists in New Zealand. The EU model is ineffective in providing a right to an explanation pertinent to the specific individual. It is, however, acknowledged that the right to be informed when an automatic decision is being made would strengthen the existing framework in New Zealand, so that individuals are aware that decisions are being made about them. It is unclear how a right to be informed is carried out in practice and whether this involves data controllers informing each individual, or if a more general notification approach is taken.

While other proposed paths work to achieve transparency and explainability and mitigate bias, they are unable to support all three. Guidelines may assist with explainability, yet do not address bias. Self-regulation draws attention to potential discrimination but does not help promote explainability. While such measures may work in tandem with an oversight body, ultimately it is a top-down model that is required to achieve the desired objectives. A hard-edged oversight model is the best move forward in order to accommodate the risk of the vast expansion of algorithms and harness the benefits machine learning can provide to the public sector.

CONCLUSION

The enigma of algorithms is not easily deciphered. The mystique of their inner workings gives rise to the concerns pertaining to the rights of natural justice and freedom from discrimination. The opacity around the architectural innards of algorithms shrouds bias and creates technical barriers to achieving meaningful transparency and explainability for an individual subject to a decision.

Through Malik, an algorithm's capacity to impact rights is highlighted, as well as the grievance suffered when one is unable to understand a decision made about them. Surveying the legislative landscape reveals that while New Zealand does, in theory, have a "right to reasons" for decisions, translating this into a meaningful explanation is not straightforward. Recourse under the Privacy Act is limited through having to prove harm. Rights afforded under the information privacy principles are not meaningful without first understanding how a decision was reached. Access to information and privacy legislation are also unable to address issues of collective harm. It is apparent that stepping outside the individualistic paradigm is necessary to achieve this.

This dissertation draws attention to the concerns posed to digital rights: lack of transparency and explainability in algorithmic decision-making; and hidden bias. To enhance rights, three objectives are proposed: providing individuals with explanations that are understandable and meaningful; increasing transparency through peer review; and confronting issues of bias. Amongst the various paths explored to achieve these objectives, I advocate that a hard-edged regulatory oversight body is the preferred way forward. Such a body, provides for pre-market intervention of harmful algorithms and continued oversight into automated decision-making to mitigate concerns of bias. Having a bird's eye view of the algorithms in use by the public sector, enables an oversight body to apply various levels of scrutiny to algorithms, depending on their level of explainability and impact on rights. A body comprising of the relevant expertise may also formulate a yardstick as to what constitutes a meaningful explanation, to give effect to rights of natural justice in understanding how a decision was reached and hold the power to enforce this. Such measures may be additionally supported by legislative or policy decisions to solely engage algorithms that are explainable by design.

The concerns identified only scratch the surface of considerations to take into account for the effective regulation of algorithmic decision-making. With the inevitable advancement of machine learning technology, further challenges and threats to digital rights will arise. It is thus timely for New Zealand to seriously consider their response to the challenges presented by algorithmic decision-making, in both the public and private sector, to ensure adequate protection is given to fundamental human rights.

BIBLIOGRAPHY

A) Cases

1. New Zealand

Case Note 14290 [2001] NZPrivCmr 5 (1 June 2001).

Casenote 794 [1987] 8 CCNO 66.

Dotcom v Crown Law Office [2018] NZHRRT 7.

Hook v Stream Group (NZ) Pty Ltd [2013] NZEMPC 188.

Lewis v Wilson and Horton [2000] 3 NZLR 546 (CA).

M A v Legal Services Agency HC Auckland CIV-2008-404-006803, 11 December 2009.

Macdonald v Healthcare Hawkes Bay and Morrison [2000] NZCRT 35 (6 December 2000).

Re Vixen Digital Limited [2003] NZAR 418.

Singh v Chief Executive Officer, Department of Labour [1999] NZAR 258 (CA).

Television New Zealand Limited v West HC Auckland CIV-2010-485-2007, 21 April 2011.

2. Australia

Re Palmer and Minister for the Capital Territory (1978) 23 ALR 196.

Re Salazar-Arbelaez and Minister for Immigration and Ethnic Affairs (1977) 18 ALR 36 (AAT)

3. England and Wales

Elliot v Southwark London Borough Council [1976] 2 All ER 781, [1976] 1 WLR 499 (CA).

Iveagh v Minister of Housing and Local Government [1964] 1 QB 395 (CA).

Kennedy v Charity Commission [2014] UKSC 20, [2014] 2 WLR 808, [2014] 2 All ER 847.

Re Poyser and Mills' Arbitration [1963] 1 ALL ER 612, [1964] 2 QB 467.

B) Legislation

1. New Zealand

Bail Act 2000.

Crimes Act 1961.

Human Rights Act 1993.

Local Government Official Information and Meetings Act 1987.

New Zealand Bill of Rights 1990.

Official Information Act 1982.

Privacy Act 1993.

Privacy Bill (2018)(34–1).

2. *European Union*

Data Protection Directive 1995.

General Data Protection Regulation 2018.

3. *United Kingdom*

Freedom of Information Act 2000.

C) *Treaties*

Convention on the Elimination of All Forms of Discrimination against Women 1249 UNTS 13 (open for signature 18 December 1979, entered into force 3 September 1985).

International Convention on the Elimination of All Forms of Racial Discrimination 660 UTNS 195 (open for signature 21 December 1965, entered into force 4 January 1969).

D) *Books and Chapters in Books*

Cathy O'Neil *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (1st ed, Crown, New York, 2016).

Frank Pasquale *The Black Box Society: the Secret Algorithms that Control Money and Information* (1st ed, Harvard University Press, Cambridge, 2015).

Graham Taylor and Paul Roth *Access to Information* (2nd ed, LexisNexis, Wellington, 2011).

Jerome Frank *Law and the Modern Mind* (Peter Smith, Gloucester, 1970).

Kathleen Mahoney "The Limits of Liberalism" in Richard Devlin (ed) *Canadian Perspectives on Legal Theory* (Eamon Montgomery 1991).

Michael Lewis, Katia Sycara and Phillip Walker "The Role of Trust in Human-Robot Interaction" in Hussein A Abbass, Jason Scholz and Darryn J Reid (eds) *Foundations of Trusted Autonomy. Studies in Systems, Decision and Control* (Springer, Cham 2018).

Paul Roth *Privacy Law and Practice Case Notes 1994–2005* (1st ed, Lexis Nexis, Wellington).

E) Journal Articles

Andrew D Selbst "Disparate Impact In Big Data Policing" (2017) 52 Ga L Rev 109.

Andrew D Selbst and Julia Powles "Meaningful information and the right to explanation" (2017) 7 IDPL 233.

Andrew Tutt "An FDA for Algorithms" (2016) 69 Admin L Rev 83.

Bryce Goodman and Seth Flaxman "European Union Regulations on Algorithmic Decision Making and a Right to Explanation" (2017) 38 AI Magazine 50.

Cédric Courtois and Elisabeth Timmermans "Cracking the Tinder Code: An Experience Sampling Approach to the Dynamics and Impact of Platform Governing Algorithms" (2018) 23 Journal of Computer-Mediated Communication 243.

Gianclaudio Malgieri and Giovanni Comandé "Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation" (2017) 7 IDPL 243.

Heike Felzmann and Rónán Kennedy "Algorithms, social media and mental health" (2016) 27 Comp & L 31.

Indramani L Singh, Robert Molloy and Raja Parasuraman "Individual differences in monitoring failures of automation" (1993) 120(3) The Journal of General Psychology 357.

John Zerilli and others "Transparency in Algorithmic and Human decision-making: Is There a Double Standard?" (2018) 31(3) Philosophy & Technology 1 at 12.

José Luiz Pinheiro Lisboa "The Judiciary – Principle of Transparency and the duty of information" (2018) 7 International Journal of Open Governments 107.

Latanya Sweeney "Weaving Technology and Policy Together to Maintain Confidentiality" (1997) 25 JLM & E 98.

Lee A Bygrave "Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling" (2001) 17 Computer Law & Security Report 17.

Lilian Edwards and Michael Veale "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke Law & Technology Review 18.

Linda J Skita, Kathleen Mosier and Mark D Burdick "Accountability and automation bias" (2002) 52 Int J Human-Computer Studies 701.

Michael Veale and Lilian Edwards "Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling" (2018) 34 CLS Rev 398.

Oren Ben-Dor "The institutionalisation of public opinion: Bentham's proposed constitutional role for jury and judges" (2007) 27 Legal Stud 216.

Paul B de Laat "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" (2017) Philosophy & Technology 1.

Richard P Will “True and false dependence on technology: Evaluation with an expert system” (1993) 7(3) Computers in Human Behaviour 171.

Rónán Kennedy “Algorithms and the rule of law” (2017) 17 Comp & L 23.

Sandra Wachter, Brent Mittelstadt and Luciano Floridi “Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation” (2017) 7 IDPL 76.

Tal Z Zarsky “Transparent Predictions” (2013) Ill U L Rev 1503.

Tim Cochrane "A common law duty to disclose official information?” [2014] NZLJ 385.

F) Papers and Reports

Article 29 Working Party *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (WP 251, 3 October 2017).

Artificial Intelligence Shaping a Future New Zealand (AI Forum NZ, May 2018).

Dillon Reisman and others *Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability* (AINow, 2018).

House of Lords Select Committee on Artificial Intelligence *AI in the UK: ready, willing and able?* (House of Lords, HL Paper 100, April 2018).

Joshua New and Daniel Castro *How Policymakers Can Foster Algorithmic Accountability* (Center For Data Innovation, May 2018).

Leon Bakker, James O’Malley and David Riley *Risk of Reconviction* (Department of Corrections, 1999).

Nick Wallace and Daniel Castro *The Impact of the EU’s New Data Protection Regulation on AI* (Center For Data Innovation, March 2018).

Statistical modelling to support the ACC automation cover decisions and accident description August 2018 (Accident Compensation Corporation, August 2018).

G) Internet resources

“ACC is using a predictive modelling tool to better support our clients, faster” (28 September 2017) ACC <www.acc.co.nz>.

AI Now Institute "Algorithmic Impact Assessments: Toward Accountable Automation in Public Agencies” (21 February 2018) Medium <https://medium.com>.

Alex P Miller “Why do we care so much about explainable algorithms? In defense of the black box” (11 January 2018) Towards Data Science <https://towardsdatascience.com>.

Calum McClelland “The Difference Between Artificial Intelligence, Machine Learning, and Deep Learning” (5 December 2017) Medium <https://medium.com>.

Camila Domonoske “Elon Musk Warns Governors: Artificial Intelligence Poses ‘Existential Risk,’” (17 July 2017) NPR <www.npr.org>.

Cathy O’Neil "Audit the algorithms that are ruling our lives” (31 July 2018) Financial Times <www.ft.com>.

Charles Towers-Clark "Can We Make Artificial Intelligence Accountable?” (19 September 2018) Forbes <www.forbes.com>.

Chris Baraniuk “Durham Police AI to help with custody decisions” (10 May 2017) BBC <www.bbc.com/news>.

Chris Isidore “Machines are driving Wall Street's wild ride, not humans” (6 February 2018) CNN Money <https://money.cnn.com>.

Christina Couch “Ghosts in the Machine” (25 October 2017) NovaNext <http://www.pbs.org/wgbh/nova/next>.

“Claims approval process documents released” (22 August 2018) ACC <www.acc.co.nz>.

Dallas Card “The “black box” metaphor in machine learning” (5 July 2017) Towards Data Science <https://towardsdatascience.com>.

Daniel Saraga “Opinion: Should Algorithms Be Regulated?” (3 January 2017) phys.Org <https://phys.org>.

David Gunning “Explainable Artificial Intelligence (XAI)” U.S. Defense Advanced Research Projects Agency <www.darpa.mil>.

David McRaney “YANSS 115 – How we transferred our biases into our machines and what we can do about it” (20 November 2017) You Are Not So Smart <https://youarenotsosmart.com/>.

“Explainable AI” simMachines <https://simmachines.com>.

"Government to undertake urgent algorithm stocktake” (23 May 2018) Beehive.govt.nz <www.beehive.govt.nz>.

Jackie Snow “New Research Aims to Solve the Problem of AI Bias in “Black Box” Algorithms” (7 November 2017) MIT Technology Review <www.technologyreview.com>.

Jackie Snow “The president of France is promoting AI, European style” (2 April 2018) Technology Review <www.technologyreview.com>.

Jacob Koshy “Supervised vs Unsupervised Machine Learning Techniques” (20 October 2017) Prompt Cloud <www.promptcloud.com>.

James Vincent “Google’s new AI algorithm predicts heart disease by looking at your eyes” (19 February 2018) The Verge <www.theverge.com>.

Jason Brownlee “Model Prediction Accuracy Versus Interpretation in Machine Learning” (1 August 2014) Machine Learning Mastery <https://machinelearningmastery.com>.

Jason Brownlee “Supervised and Unsupervised Machine Learning Algorithms” (16 March 2016) Machine Learning Mastery <https://machinelearningmastery.com>.

Jason Tashea "Risk-assessment algorithms challenged in bail, sentencing and parole decisions” (March 2017) ABA Journal <http://www.abajournal.com>.

Jeff Larson and others “How We Analyzed the COMPAS Recidivism Algorithm” (23 May 2016) ProPublica <www.propublica.org>.

Joyce Riha Linik “Skin Cancer Detection Using Artificial Intelligence” (31 January 2009) IQ Intel <https://iq.intel.com>.

Lee Rainie and Janna Anderson “Code-Dependent: Pros and Cons of the Algorithm Age” (8 February 2017) Pew Research Center <http://www.pewresearch.org/>.

Lindsay Kwan “Beyond the buzzword: What “artificial intelligence” means for marketing leaders, right now” (28 November 2017) Widerfunnel <www.widerfunnel.com>.

Matt Burgess “Holding AI to account: will algorithms ever be free from bias if they're created by humans?” (11 Jan 2016) Wired <www.wired.co.uk>.

Matt Reynolds “Biased policing is made worse by errors in pre-crime algorithms” (4 October 2017) New Scientist <www.newscientist.com>.

Matt Wes “Looking to comply with GDPR? Here's a primer on anonymization and pseudonymization” (25 April 2017) Iapp <https://iapp.org>.

Noel Duan “When beauty is in the eye of the (robo)beholder” (20 March 2017) Arstechnica <https://arstechnica.com>.

Olenka Van Schendel “Data masking: Anonymisation or pseudonymisation?” (7 November 2017) GDPR: Report <https://gdpr.report>.

Paul Roth “Reports on Aspects of Privacy Compliance and Practice of the NZ Post Lifestyle Survey 2009” (20 June 2011) Office of the Privacy Commissioner <https://privacy.org.nz>.

“Police to use new family violence risk assessment tools” (13 February 2012) New Zealand Family Violence Clearinghouse <https://nzfvc.org.nz/>.

“Prison facts and statistics” (June 2017) Department of Corrections <www.corrections.govt.nz>.

“Privacy Act & Codes” (2013) Privacy Commissioner <www.privacy.org.nz>.

Rita Heimes “The GDPR restricts “profiling” and gives data subjects significant rights to avoid profiling-based decisions” (20 January 2016) Iapp <https://iapp.org>.

Saranya Vijayakumer “Algorithmic Decision-Making” (28 June 2017) Harvard Political Review <http://harvardpolitics.com>.

Stacey Kirk “Children ‘no lab-rats’ — Anne Tolley intervene in child abuse experiment” (30 July 2015) Stuff <www.stuff.co.nz>.

Stephen Buranyi “Rise of the racist robots – how AI is learning all our worst impulses” (8 August 2017) The Guardian <www.theguardian.com>.

TC “What are algorithms?” (30 August 2017) The Economist <www.economist.com>.

“The Privacy, Human Rights and Ethics Framework” Ministry of Social Development <www.msd.govt.nz>.

Thomas Lumley “Immigration NZ and the harm model” (5 April 2018) Statschat <www.statschat.org.nz>.

Vanessa Blackwood “Algorithmic transparency: what happens when the computer says “no”?” (29 November 2017) Privacy Commissioner <https://privacy.org.nz>.

Vignesh Ramachandran "Exploring the use of algorithms in the criminal justice system” (3 May 2017) Stanford Engineering <https://engineering.stanford.edu>.

H) Other Resources

Differential Privacy for Everyone (Microsoft Corporation, 2012).

Hamish McNeilly “Cops' new tool to predict domestic violence” *The New Zealand Herald* (online ed, Auckland, 19 April 2012).

Interview with Alistair Murray (Guyon Espiner, Morning Report, Radio New Zealand, 10 April 2018).

Interview with Alistair Murray (Guyon Espiner, Morning Report, Radio New Zealand, 5 April 2018).

Interview with Frank Pasquale, Author (Steve Paikin, The Agenda, 12 May 2016).

Interview with Stephen Buranyi (Simon Martin, This Way Up, Radio New Zealand, 17 March 2018).

John Edwards “Privacy Commissioner’s Submission on the Privacy Bill to the Justice and Electoral Select Committee” at [55]

Laws of New Zealand Information (online ed).

Mahoney Turnbull “Navigating New Zealand’s Digital Future: Coding our way to Privacy in the Age of Analytics” (LLB (Hons) Dissertation, University of Otago, 2014).

Paul Roth *Introduction to the Privacy Act 1993 Mazengarb’s Employment Law (NZ*) (LexisNexis, August 2015).

Paul Roth *Privacy Law and Practice* (online ed, LexisNexis, accessed 30 August 2018).

Reuben Binns and others “It’s Reducing a Human Being to a Percentage; Perceptions of Justice in Algorithmic Decisions” *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* (New York, 26 April 2018).