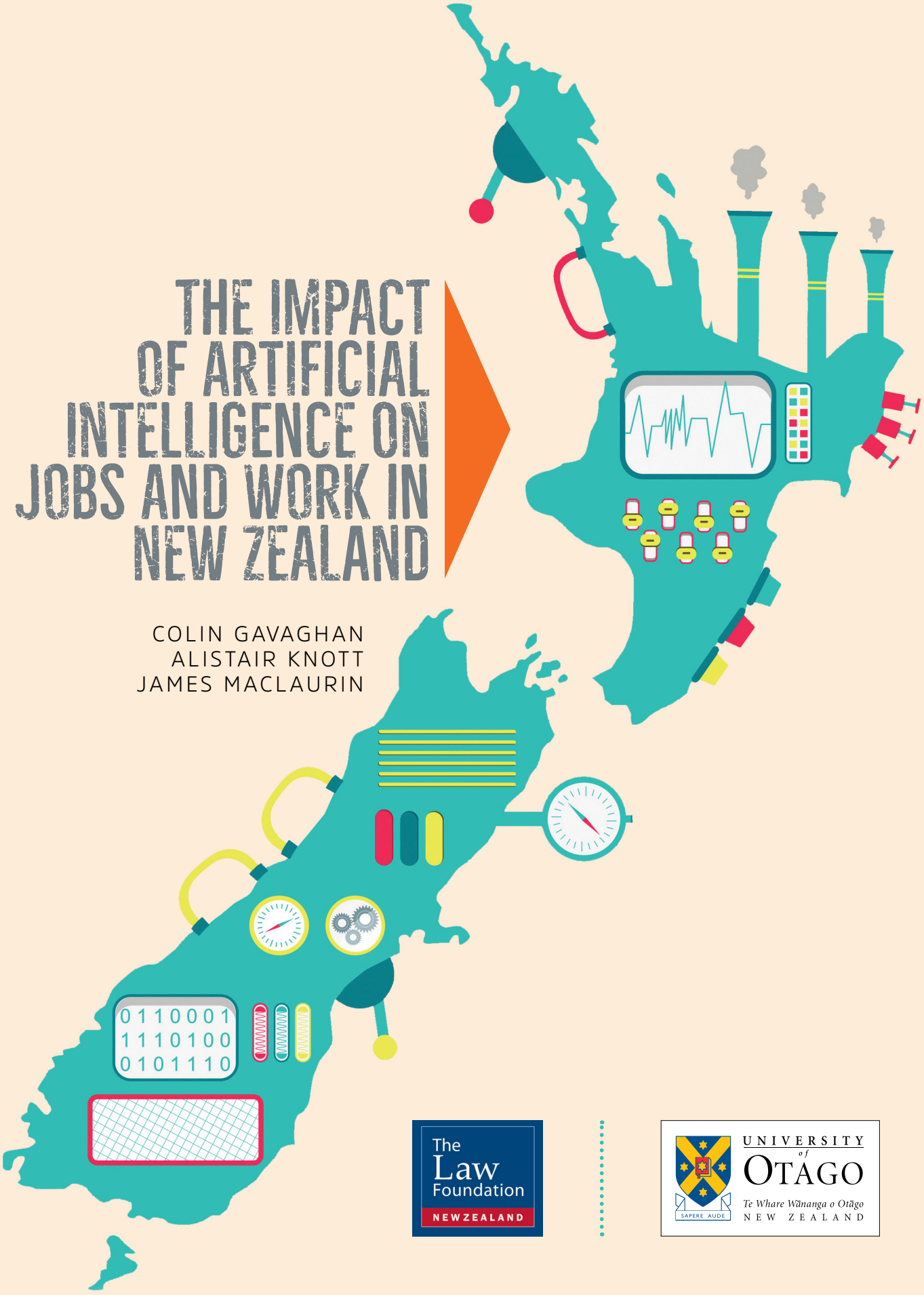
THE IMPACT OF ARTIFICIAL INTELLIGENCE ON JOBS AND WORK IN NEW ZEALAND

COLIN GAVAGHAN
ALISTAIR KNOTT
JAMES MACLAURIN



The Law Foundation NEW ZEALAND

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

ISBN 9780473569204 (paperback)

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON JOBS AND WORK IN NEW ZEALAND

Final Report on Phase 2 of the *Artificial Intelligence and Law in New Zealand* Project.

COLIN GAVAGHAN

ALISTAIR KNOTT

JAMES MACLAURIN

Funder: New Zealand Law Foundation

University of Otago | 2021

CONTENTS

ACKNOWLEDGEMENTS

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

We are most grateful to Joy Liddicoat for extensive research relating to this project, and excellent organisation of our Oxford and Dunedin workshops.

We are also grateful to John Zerilli, for advice and comments.

We would also like to express our gratitude to the following people, who participated in our Oxford and Dunedin workshops on AI and employment in 2018 and 2019:

Gordon Anderson (Faculty of Law, Victoria University of Wellington)
Angela Ballantyne (Department of Primary Health Care and General Practice, University of Otago)
Matthew Bartlett (Citizen AI, Loomio)
Christina Blacklaws (President of The Law Society of England and Wales)
Matt Boyd (Adapt Research, NZ)
Hazel Bradshaw (Department of Internal Affairs, NZ Government)
J Scott Brennen (Reuters Institute for the Study of Journalism and the Oxford Internet Institute)
Elizabeth Broadbent (School of Medicine, University of Auckland)
Pieta Brown (Orion Health, NZ)
Corinne Cath (Oxford Internet Institute)
Chinchih Chen (Oxford Martin Programme on Technology and Employment, Oxford Martin School)
Alex Comninos (Justus-Liebig University Giessen, Open Technology Institute)
Gareth Cronin (Chief Technology Officer, Ambit, NZ)
Martin Davidson (Chief Legal Intelligence Officer, ThoughtRiver, UK)
Alan Dignam (School of Law, Queen Mary University of London)
Kenneth Dau-Schmidt (Maurer School of Law, Indiana University)
Becky Faith (Digital and Technology cluster, Institute of Development Studies, University of Sussex)
Toby Gee (Lambton Chambers in Wellington)
Elizabeth George (School of Management, University of Auckland)
Stefan Brambilla Hall (Media, Entertainment and Information initiative, World Economic Forum)
Kai Hsin-Hung (International Labour Organisation)
Elliot Jones (Demos, UK)
Avalon Kent (New Zealand Council of Trade Unions)
Kaska Porayska-Pomsta (University College London Institute of Education)
Daithí Mac Sithigh (School of Law, Queen's University Belfast)
Rakesh Mistry (Straker Translations, NZ)
Paula O'Kane (Department of Management, University of Otago)
Mary Ollivier (Director, Regulatory, New Zealand Law Society)
Jeremias Prassl (Oxford University, Institute of European and Comparative Law)
Gary Rogers (Urbs Media, RADAR)
Geoffrey Roberts (Citizen AI)
Paul Roth (Faculty of Law, University of Otago)
Diane Ruwhiu (Department of Management, University of Otago)
Ana Luísa Sertã (Department of Geography, Birkbeck/ University College London)
Grace Smart (Ministry of Business, Innovation and Employment, NZ)
Jeanne Snelling (Faculty of Law / Bioethics Centre, University of Otago)
David Souter (ICT Development Associates, UK)
Lord Thomas of Cwmgiedd (Lord Chief Justice of England and Wales, 2013-17)
Lena Waizenegger (Business Information Systems, Auckland University of Technology)
Richard Wallace (Office of Parliamentary Counsel, NZ)
Sara Walton (Department of Management, University of Otago)
Jim Warren (Department of Computer Science, University of Auckland)
Jean Yang (McCarthy Finch, NZ)

Others who were generous with their time and resources include our Otago colleagues Dawn Duncan (Faculty of Law), Paula O'Kane (Department of Management), Ivan Diaz-Rainey (Department of Accountancy and Finance), David Eyers (Department of Computer Science), and Mele Taumoepeau (Department of Psychology); Fiona Ryan at the Ministry of Health; and Ross Teppett at the New Zealand Council of Trade Unions. Sam Cathro, Ruth Jeffries, Caitlin Smith, Jonathon Yedlin and Karen McLean provided eagle-eyed proof-reading.

EXECUTIVE SUMMARY

Artificial Intelligence (AI) is a diverse technology. It is already having significant effects on many jobs and sectors of the economy and over the next ten to twenty years it will drive profound changes in the way New Zealanders live and work. Within the workplace AI will have three dominant effects.

- AI will change how human work is administered in workplaces: particularly how employers hire, manage and monitor employees. AI is already being used to rank and interview job applicants, to monitor and assess the performance of workers, and to assign tasks in gig-economy companies. These uses are likely to grow.
- AI will perform tasks normally performed by humans, augmenting the productivity of some workers and displacing others. The list of tasks AI can perform is long and growing. Such systems include AI-based robots like self-driving vehicles as well as effector robots capable of manipulating, assembling, painting, inspecting and so on. They also include autonomous decision-making / decision support systems widely used in government and industry, as well as chatbots and other systems that analyse and generate text.
- AI will also create new types of work. These will include high-value jobs like coding and managing the deployment of AI systems as well as low-value jobs such as preparing data for use in AI training.

HOW WILL AI IN THE WORKPLACE CHANGE AOTEAROA?

We are sceptical of attempts to predict with any accuracy the numbers and types of workers that will either benefit from AI augmentation or be displaced by AI in the coming decades. Much depends on decisions yet to be taken by governments, industries and consumers. We suspect, though, that widespread technological unemployment is unlikely, as the cost of unemployment for individuals is so high that most will choose even low-value, precarious work over no work at all. While AI will create new types of work, we cannot predict the ratio of high-value to low-value jobs that AI will create. Given the unpredictability of future innovation and future labour markets, our education system should focus on producing graduates that are broadly skilled across the humanities, science and commerce. But steps will be required to protect the growing number of workers in precarious employment.

The history of previous industrial revolutions and the deployment of other general purpose technologies such as electricity, telephony, and the production line, suggests that the deployment of AI will have significant near-term risks including displacement of workers and transition costs for legacy industries. It will also have significant medium-term benefits. It will enhance and make more affordable many goods and services. Overall, it will exert downward pressure on the cost of living.

Although we cannot accurately predict the numbers of jobs that will be created and destroyed, we can predict that New Zealand's economy and society will be subject to three countervailing forces: It will *enable* some workers, by enhancing their productivity and incomes, and it will *replace* or displace other workers. Some of the displacing AI will be owned 'onshore', in New Zealand, and some of it will be owned 'offshore', by large data-rich international entities such as the FAANG companies. We cannot know in advance which of these forces will predominate in particular jobs and industries. So, the challenge for Aotearoa is to prepare for an unknown mix of the 'enabling', 'replacing onshore' and 'replacing offshore' scenarios. As with previous industrial revolutions, there is a significant risk that the AI revolution will increase inequality. Addressing inequality will be particularly challenging if the profits of the AI revolution disproportionately land offshore.

We suggest a number of possible solutions to the replacing offshore scenario. These include enhancing New Zealand's sovereign wealth fund to invest in, and hence draw profits from, offshore AI-driven companies that are difficult to tax. We also suggest that New Zealand might identify AI based industries in which we are well placed to compete, such as social media. Homegrown AI-based services could be promoted via targeted investment or even by government setting up New Zealand-based companies, as we set up Kiwibank to compete with offshore banking concerns.

As the large-scale social and economic effects of AI are complex, it is essential that government promotes a national conversation about how we want AI to change life and work in Aotearoa. That conversation must be broad, giving particular attention to Māori and Pasifika voices. National and international research shows that New Zealanders are prone to overwork

which exacerbates a wide variety of health and social problems. If AI leads to increases in productivity in Aotearoa, we should pay particular attention to whether it can be used to help us work less.

Decreasing the length of the work week or the work day would have many valuable benefits for New Zealand. It would help us to share out high-value, well-paid work and it would enhance our quality of life, helping to build vibrant and resilient communities in which it is easier for all of us to look after our kaumātua, tamariki and mokupuna. Local and international research suggests that, in many types of employment, decreasing work hours has surprisingly little effect on productivity, while greatly enhancing the wellbeing and enthusiasm of workers. However, in operational roles (such as bus driving or nursing) decreasing work hours would directly decrease productivity. So, if all New Zealanders were to benefit equitably from a shorter work week, government would need to provide some form of subsidy for employers of operational workers, perhaps similar to *kurzarbeit* schemes familiar to many Europeans. This would also effectively increase the number of such jobs available to New Zealanders.

HOW WILL AI CHANGE WORK?

As well as addressing these big picture questions, this report investigates what it will be like to work alongside AI, assessing regulatory changes designed to maximise the benefits and minimise the harms of AI in the workplace. It is difficult to assess claims that AI will generally enhance jobs. As with previous types of automation, AI will sometimes relieve us of onerous and unpleasant tasks and sometimes leave occupations deskilled, reducing the mana and bargaining power of workers.

This report pays particular attention to the way AI will change hiring, monitoring, and managing staff. AI promises to make hiring faster and less expensive, to better match applicants to jobs, and to help increase diversity in the workplace. However, there is also a risk that AI will introduce unfair bias in job advertising and in the vetting of job applicants. When algorithms are trained on historical data that reflects historic discrimination or inequality, this use of 'dirty data' is likely to skew outcomes for already disadvantaged individuals and groups.

AI is already used widely in the *recruitment* of workers, in targeting job ads to potential employees, shortlisting applicants, and evaluating the performance of candidates in interviews. There are dangers of unfair discrimination at every stage.

AI can also be used in *management* of workers, performing a variety of tasks that were previously the preserve of human managers. New AI-based management methods promise to improve accuracy and efficiency in decision-making, and to reduce opportunities for human favouritism and unconscious biases. Algorithmic management also has the capacity to improve the lives of workers. It could be deployed in consultation with workers, so as to accommodate the needs of workers with families, enhance leisure time and educational opportunities. At the other end of the scale, it could leave workers feeling isolated and dehumanised, or placed under greater levels of pressure or surveillance. It is essential that AI not entrench deep disparities between the power of workers, managers, and capital owners.

An algorithmic management system trained on profiles of previous workers could make recommendations or predictions based on characteristics that are irrelevant or discriminatory. Such AI threatens to entrench historical discrimination. Managerial decisions in general are covered by both the Employment Relations Act 2000 (ERA) and the Human Rights Act 1993, but issues may arise regarding implementation, compliance monitoring and enforcement of the legislation. A range of auditing tools already exist to help employers avoid inadvertent discrimination, but as with recruitment, concerns exist about the criteria employed by different tools—what notions of bias or fairness they use, for instance, and what jurisdiction's laws they are aligned with.

The opacity of AI systems that could potentially inform discipline or dismissal makes it difficult for those affected to assess whether employers have complied with their legal obligations. Employers should make sure task allocation algorithms are 'explainable', in terms that are meaningful to their workers.

Another concern relates to growing use of AI-enabled workplace surveillance, which can threaten the autonomy and dignity of workers. In due course, these technologies may require specific legislative attention. In the meantime, we would welcome attention from the Privacy Commissioner to the possibility of a code of

practice directed at workplace surveillance technologies, or perhaps workplace surveillance more generally. WorkSafe could also have a role to play in regulating the potentially harmful effects of algorithmic management and surveillance.

Increasingly AI will appear in the workplace in the form of collaborative robots (cobots). These may greatly enhance productivity but will force a rethink in the way we address safety concerns. It will no longer be possible to 'separate and contain' such machinery, as it will be working amongst us. WorkSafe should consider issuing a code of practice dealing with workplace robots and particularly 'cobots', perhaps based on the ISO standard for collaborative robots.

AI is a new phenomenon in most workplaces and it is poorly understood by many of those who now use it or are affected by it. Efficiency gains and cost decreases will drive its rapid adoption in many industries. New Zealand government and regulatory agencies should facilitate this transition to ensure that harm does not come to workers or other stakeholders.

- Consideration should be given to requiring hiring tools to include functionality for bias auditing, so that client companies can readily perform audits of each recruitment decision process.
- Discussions should take place between developers, employers' organisations, unions and other relevant stakeholders to consider the development of guidance and standards for auditing of algorithms used in employment situations.
- The New Zealand private sector should use algorithm impact assessments, assessing factors such as privacy and equality. Given the likely challenges for smaller employers developing these in-house, templates should be made available along the lines of those developed by the NZ Privacy Commissioner for Privacy Impact Assessments.

EFFECTS ON CONSUMERS, PROFESSIONS AND SOCIETY

As well as considering the effects of AI on workers inside organisations, this report also considers the 'outward-facing' effects of AI on the *consumers* of their services: customers, clients, patients, and so on. Our main focus is on a subset of services delivered by what are commonly referred to as 'the professions'. Work in these fields involves a distinctive mixture of specialist knowledge, formal and informal value systems, and elaborate accreditation processes. They are technically complex and play important roles in society.

AI in the professions has been said to offer many benefits.

- Automation of the more routine, burdensome aspects of a role could free up time for those tasks uniquely suited to human beings.
- It could help professionals sort and sift high volumes of information about, for example, new medical treatments.
- It could decrease the time taken to provide decisions and advice, and it could increase the accuracy of certain types of decisions.
- It could democratise access to expensive services such as legal advice; legal chatbots developed in New Zealand by CitizenAI are a promising example of this sort of benefit.

It is unlikely in the near future that we will encounter any serious proposal to replace, e.g. healthcare or legal professionals entirely with AI, but AI is already taking over particular aspects of those roles, or particular tasks within them. Concerns that arise in this context are considerably more immediate. Some of these relate to well-rehearsed concerns about how AI systems get their results. These include accuracy, control, transparency and bias. We reviewed these concerns in our earlier report on government uses of AI (Gavaghan et al., 2019). However, some of these issues take on particular significance in the provision of professional services. Concerns about accuracy, for example, are likely to be particularly acute in very high stakes domains such as healthcare—mental health chatbots deployed in the United Kingdom have recently been shown not to detect pleas for help that would have been easily spotted by a human professional.

The lack of transparency of some AI is also of particular concern in health contexts in which it's required that patients be able to make *informed* choices or give *informed* consent. These issues may affect which types of AI are able to be used in such contexts, but they may also be ameliorated by advances in 'explainable AI'.

Bias in professional judgements is not always problematic; we *want* whoever assesses our scans and test results to 'err on the side of caution' and over-diagnose malignancy, at least to a point. However more pernicious forms of bias are being identified, as where images of skin disease manifesting in darker skin types are not sufficiently included in training data. The potential for such bias occurs in many professions.

There is considerable debate about the assignment of responsibility for harms caused by AI, specifically where harms are caused by autonomous AI that learns from its environment. In our view, though, these concerns are often overstated. Most current AI has very limited autonomy and the majority of the 'learning' takes place in-house, during initial development of the product, or during development of a product update. There will still be challenges in establishing responsibility and liability when AI systems go wrong, but they will more commonly relate to the sorts of issues identified in *Tyndaris v VWM* (discussed in Chapter 4).

AI is becoming more humanlike in its dealings with clients, customers and such like. This is raising concerns about deception and manipulation. This has led to regulatory initiatives such as California's Bolstering Online Transparency (BOT) law, which "requires all bots that attempt to influence California residents' voting or purchasing behaviors to conspicuously declare themselves." Mandatory 'bot disclosure' is something that merits serious consideration in New Zealand, at least in relatively high-stakes contexts such as high value purchases or political campaigning.

It's common for chatbots to 'escalate' cases to human workers. This may be due to risk management, the chatbot being unable to interpret what humans are saying to it, or just a user asking to speak to a human. The way such 'handovers' work can have important consequences for the effectiveness of the system and on the way it ameliorates risk in high-stakes contexts. Where transitions occur between humans and chatbots, service providers should be transparent about how and when these will take place, and what information will be passed between them.

Much discussion about the increasing automation of professional roles relates less to technical concerns about accuracy, transparency and the like, and more to concerns about the removal of distinctly 'human' factors such as trust, empathy and 'the human touch'. At least in more high-stakes or emotionally sensitive areas, such as health or elder care, we should take seriously the possibility that concerns about dehumanisation will resonate with many people. However, we should not become over-reliant on generalisations or assumptions about how people might feel about interacting with AI. It's possible that some people, in some situations, might find dealing with AI helpers or carers empowering, or less undignified than reliance on humans for certain intimate roles. When considering matters such as 'empathy', and whether AI is capable of providing them, we should recognize that this can refer to several different things, and think carefully about what sorts of 'empathy' or 'trust' are valuable in which situations.

Some professions hold monopolies on offering certain services. Other rules govern who can advertise themselves as a member of a given profession. Most (or all, depending on how we define a 'profession') have rules applicable to those practising within them. Some of these rules are more AI-ready than others. All New Zealand professions will have to think through the implications of increasing use of AI in their workplaces

In New Zealand, healthcare is regulated in part by rules aimed at therapeutic devices, and in part by rules aimed at human practitioners. Some healthcare AI seems to straddle those two streams. Insofar as it is viewed as an artefact, it will be subject to the therapeutic products regime, oriented towards risk minimisation. But in those contexts where AI performs in a more 'human' way—communicating directly with healthcare consumers—then it should also be evaluated against the framework that exists to ensure that human healthcare providers conduct their duties in a respectful and culturally competent manner. Whether the new Therapeutic Products Bill, and the regulatory system it creates, makes adequate provision for this remains to be seen.

Provided they offer legal information in general terms, rather than legal advice in response to specific situations, it seems likely that AI legal chatbots such as those introduced by CitizenAI will comply with the terms of the Lawyers and Conveyancers Act 2006 Act. As the technology progresses, though, it may be that legal AI will be developed that is able to offer advice tailored

to a particular client's needs. Were that to become possible, those using such a service should take care to ensure that:

- With regard to legal advice in general, they do not describe the chatbot as a 'lawyer' or make misleading claims that it is being supervised by a lawyer;
- With specific regard to advice about court proceedings, the chatbot would not be allowed to offer 'advice' at all.

Much current discussion of AI in the professions centres on the extent to which the use of AI should be restricted in various contexts. Given the projected benefits, though, we should also take seriously the possibility that there may sometimes be a professional *obligation* to use AI. Some overseas case law has already started pointing in the direction of such a duty.

INTRODUCTION

This report presents our findings from the second part of our New Zealand Law Foundation-funded project: *Artificial Intelligence and Law in New Zealand*. The overall focus of the report is on the regulatory issues surrounding uses of AI in New Zealand. In the first phase of this project, we looked at the use of AI – and particularly predictive algorithms – in New Zealand Government. (Gavaghan et al., 2019) In this phase, we examine their impacts on work and employment.

This is a topic that has received a great deal of attention in recent years. It was the subject of numerous reports (e.g. International Bar Association 2017; International Labour Organisation 2018; Royal Society and British Academy 2018; World Economic Forum 2018), books (e.g. Brynjolfsson and McAfee 2014; Ford 2015; Pasquale 2020), academic articles and media coverage. Much of that has focused on attempts to predict how many jobs are likely to be taken over by AI and related technologies such as robotics, or on which sectors are most likely to be affected. In particular, the predictions made in a highly influential 2013 paper by Carl Frey and Michael Osborne, and a later paper offering a very different perspective from David Autor, have commanded a great deal of attention.

It's not our purpose here to add our own predictions to that particular debate. For one thing, we are not economists. For another, we are somewhat sceptical of the accuracy of any such predictions, especially given the additional economic uncertainty in the wake of the Covid crisis. Rather than trying to estimate how many jobs will be lost or created, our focus is on other questions: not *how many* jobs will be affected or displaced, but *how* will they be affected? We also look at the history of industrial revolutions as a means of addressing some big picture questions about changes in the nature and value of work and about when New Zealand might ideally want to gain from an AI revolution.

While we are sceptical of attempts to quantify the displacement effect of such technologies, it seems safe to predict that, for the foreseeable future, there will still be human workers, and equally safe to surmise that, in many cases, their working lives will be touched by these technologies. How will the technologies with which we are concerned – principally artificial intelligence and associated technologies such as robotics – change their roles? What will it be like for those required to work alongside, or even under, AIs and robots? What sort of benefits and harms might result from those changes?

Of course, it won't only be the workers occupying those roles whose experience is likely to change as a result of those technologies. Our focus will not only be on the people doing those jobs, but on the clients, patients and others who interact with them. What will it be like when helplines are 'staffed' by artificially intelligent chatbots, or to be advised by an 'AI lawyer' or treated by a 'robo-doctor'? Are there genuine risks or detriments associated with such a change, and if so, how could they be minimised?

In this report, we're concerned with both of these situations: that of the service provider and of the service consumer. Or, in more common terms, the worker and the client. Our enquiries will inevitably touch on a variety of possible impacts: economic, social, psychological, ethical. We are not setting out, though, to foretell the future. How AI impacts on work will depend very significantly on the sorts of choices we as a society make about them. And it's with those choices that we are ultimately concerned here. The question at the heart of this report is a large and complex one, but it can be stated quite simply: what sorts of rules and policies should we have about AI as it impacts upon the world of work?

We begin by examining the sorts of technologies that are already being deployed across our workforce, or that are likely to be deployed in the near future. Chapter 1 gives a layperson's introduction to the relevant AI methods, highlighting technologies that are having the most impact on jobs. We make a key distinction between AI systems that are used to *administer* human jobs (in recruitment, management and monitoring of workers), and AI systems that are used to *take on* jobs or tasks that have traditionally been done by people. We also set out a range of roles those technologies are likely to play, from advertising vacancies and recruitment, via other administrative and managerial roles, to performance of aspects of those roles. We also consider some of the new roles that might come into existence in response to these technologies.

Chapter 2 consists of a broad analysis of the nature and value of work, and some broad recommendations for government planners. The history of major industrial revolutions and of the development of general purpose technologies (such as electricity and the internal combustion engine) provides a useful indicator of likely large scale, medium term costs and benefits. These include effects on individuals, communities and companies as well as changes in the cost of living and in New Zealand's productivity and resilience.

As AI will bring major change to New Zealand workplaces, it is essential that we begin by considering the ways in which work is valuable to New Zealanders. For most people, work is primarily valuable as a means of getting money. For many, though, work is also a contributor to wellbeing; it can provide opportunities for social interaction, provide a structure for our days, and create a feeling of self-worth. How is the introduction of AI likely to affect those aspects of work which are valuable to people?

As we noted at the outset, unlike many commentators, we refrain from making strong predictions about how much work is likely to be taken over by AI, or even how working life will change. Instead, we sketch three possible *scenarios* that policymakers can anticipate and prepare for: what we've termed the *enabling* scenario, the *replacing onshore* scenario, and the *replacing offshore* scenario. In each of those scenarios, there are policy options that we believe would mitigate adverse outcomes such as inequality and maximise benefits such as a higher standard of living or decreased hours of work.

In the remainder of the report, we move away from issues of future policy to take a more detailed look at how AI technology is currently entering the workplace. In Chapter 3, we focus on the experience of individual workers. The chapter is structured around the 'life-cycle' of employment. A worker is first *recruited*, through job ads and selection processes. If engaged, the worker is then *managed*, by being assigned tasks, in which her performance is monitored and evaluated. AI is becoming involved in all of these processes. While this often simplifies and streamlines employee-facing tasks, it also creates several concerns.

Many of these are familiar from the earlier work on AI; concerns about transparency and bias, for example, are common to most domains in which AI is discussed, though they can raise distinct concerns in the present context. Other concerns – around workplace health and safety, and privacy and surveillance – are familiar from the realms of employment law, labour studies and management. But the intersection of these subjects poses new challenges, and perhaps requires new solutions. This chapter also allows us to consider whether some of the key concepts underpinning our employment law are likely to be fit for purpose in a future where workplace AI is commonplace.

In Chapter 4, we consider how AI's use in workplaces producing goods and services will impact on *consumers* of these products, and on society more generally. In 2017, technology commentator Adam Greenfield issued the following warning about the impact of the latest automation revolution on the world of work:

> We now stand at a juncture where there is no pursuit that cannot in principle be undertaken by an automated system, and we need to come terms with what that might mean for the economy, the ways in which we organize our societies, and our own psyches. (Greenfield 2017, p.185)

While it's plausible that this new industrial revolution will impact on just about all sectors of the workforce to some extent, for the purposes of this report we will focus on a particular subset: that part of the workforce commonly regarded as the professions, with an emphasis on the legal and healthcare professions.

There are a number of reasons for this emphasis. One is that, while the professions have been relatively insulated from the impacts of previous waves of automation, this is widely considered not going to be true of the AI revolution. As a 2014 Pew Research Centre report said: "Impacts from automation have thus far impacted mostly blue-collar employment; the coming wave of innovation threatens to upend white-collar work as well." (Smith and Anderson 2014) Indeed, the level of disruption that may follow from this latest industrial revolution have led some influential commentators to forecast that the professions, as we have known them, will be steadily dismantled. (Susskind and Susskind 2016)

Even if this is an exaggeration, any significant degree of disruption would of course have profound implications for those who currently work – or who expect to work – in those professions. But the impact on the professions could have an even wider societal impact. Professions in general, and perhaps some in particular, are often thought not only to be keepers of certain kinds of specialised knowledge, but also, to be custodians of certain ethical values. This is reflected in the fact that professions are often (or always, depending on which definition one chooses) already subject to quite extensive regulation, either self- or externally imposed. As Richard and Daniel Susskind acknowledge, though, "their exponents are ordinarily thought to be bound by a common set of values over and above any formal regulations that apply to them."

What would it mean if the humans charged with upholding those values were replaced, in whole or substantial part, by technology? What would it mean for the clients, patients, students and others who avail themselves of those services? For society more generally? We will consider these questions in general terms, but many of our case studies will be drawn from law and medicine, which are well suited for the discussion, having been subject to extensive academic attention regarding the ethical dimension of those roles.

While this makes them ideal candidates for case studies, this in no way means that we consider these professions, nor those working within them, to be generally more important than other parts of the workforce. If it were not already clear, the Covid crisis has emphasised the importance to societal functioning of many workers outside of the professions: cleaners, delivery drivers, supermarket staff and all the other 'essential workers' who were deemed to indispensable to be subject to lockdown. It is likely that their lives too will be impacted by AI algorithms, not least in what has come to be known as the 'gig economy.' Some of these more general implications will also be addressed in this report.

1. DEFINING THE TECHNOLOGY OF INTEREST

In this chapter, we will introduce the AI technologies that feature most prominently in discussions of AI and employment. We won't assume any special technical knowledge; instead we'll focus on providing practical examples of AI systems that are in actual use. Our purpose in introducing these examples is to outline a set of use cases for current AI systems, to serve as a platform for the ethical and regulatory questions about AI and employment which we consider in subsequent chapters.

AI impacts on employment in three broad ways. Firstly, AI systems are used to help *administer* employment in organisations. In this role, AI systems operate in the domains of Human Resources (HR) and Personnel Management, which recruit human employees, set them particular tasks, coordinate their activity, and oversee their work. We will discuss these systems in Section B. Secondly, AI systems are used to help *perform* the tasks which define the purpose of the organisation: the creation of particular products or services. In this role, AI systems can assist human employees in their day-to-day work in a variety of ways, or even replace certain human employees altogether. We will discuss these systems in Section C. There's some degree of overlap between these categories, of course: organisations normally employ people to perform HR and Personnel Management roles, so AI systems that contribute to these functions will also be assisting or replacing human employees. But it is still useful to distinguish between AI systems that support the human infrastructure of an organisation and those that produce its goods or services. Finally, the AI industry has *created* some new areas of human employment: certain human jobs that didn't previously exist. We will discuss these new jobs in Section D.

The AI systems that perform administrative and production functions for organisations are in some ways extremely diverse, as our survey will demonstrate. However, modern AI systems also tend to rely on certain core technologies, in the field of *machine learning*. Understanding something about these technologies is helpful in understanding the diverse roles that AI systems can play in organisations, and in understanding the potential and limitations of current AI systems. We will begin in Section A by briefly introducing these technologies.

A. A core component of current AI systems: machine learning

A traditional computer program is written by a programmer, as a sequence of instructions for the computer to execute, written in some programming language. Modern AI systems certainly include hand-written code of this kind. But the code is often supplemented with algorithms that have been produced *automatically*, rather than written by hand. The 'producer' of these algorithms is a separate computer program called a **machine learning system**,[1] which discovers the algorithm which 'works best', according to some specified criterion. Rather than directly writing code to perform some task, AI engineers often use a machine learning system that will learn an algorithm that performs the task as well as possible. Their effort can then be focused on specifying the task in precise detail, so that the machine learning system can learn most effectively how to do it.

There are different kinds of machine learning, which specify the task to be learned in different ways. But in each case, task specification involves supplying a body of **training data** from the task domain. Informally speaking, training data for a machine learning system is like 'experience' for a human employee. Human employees tend to learn on the job, improving their skills as their experience grows. In a similar way, the performance of a machine learning system typically improves the more data from the task domain it is trained on. A key reason why AI systems have improved in recent years is that online data in relevant domains has become available in ever-increasing quantities, which can be configured as training data for machine learning systems. We'll briefly review the three most prevalent machine learning paradigms in the remainder of this section.

1 In this chapter, we will use **bold face** for technical terms when they are introduced.

Supervised learning

By far the most widely used machine learning paradigm is **supervised learning**. In this paradigm, the training data consists of a set of specific situations or scenarios to which the system must respond, each paired with the response the system *should produce*. Each item of training data consists of an **input pattern**, representing a situation or scenario, paired with an **output pattern**, representing the desired response. In relation to human jobs, training data of this sort plays the role of an employee mentor, telling a new recruit 'If *this* happens, you should do *this*'. The system learns to respond appropriately to the training inputs—and hopefully, also to inputs that are similar to the training inputs. (A key purpose of all machine learning systems is to generalise away from the supplied training data, so as to behave sensibly for unseen inputs, in virtue of their resemblance to seen training inputs.)

A typical example of a system trained with supervised learning is a **visual object classification** system, whose purpose is to recognise objects in computer images. In this case, each training item is an image of a certain object (an array of pixels), paired with a label identifying what type of object it is (for instance, the label 'dog', or 'cat', or 'table'). The system's task when given an image is to produce the label that correctly identifies it. After training on many examples of dogs, cats, tables and so on, the idea is that the system should be able to recognise unseen images of these object types. Of course, human employees learn these simple object-recognition skills long before entering a particular workplace—but such skills are important in many jobs. For instance, a driver needs to be able to classify objects in her field of view as pedestrians, cars, buildings, road signs, and so on. But workplace learning often involves similar perceptual skills. For instance, a radiologist needs to learn to classify mammograms as cancerous or cancer-free, or bones as fractured or unfractured, and this learning typically happens through supervision, by being shown instances of the different categories. Actually, a wide variety of human skills can be approximated with supervised learning. For instance, a supervised learning system can learn to recognise different types of legal document (Wan et al., 2019), to rank prospective employees based on their blog posts and LinkedIn profile (Faliagka et al., 2012), and to produce written texts that plausibly continue a supplied initial passage (Radford et al., 2019). We will discuss these systems in Sections B and C.

Machine learning systems need good training data—but they also need good **learning methods**, which can learn tasks with some degree of complexity, and whose learning extends in the right way to unseen inputs. Learning methods include well-known statistical techniques. For instance, the statistical technique of **regression**, which has been in wide use for decades, can be understood as a supervised machine learning method. However, the current boom in AI is largely due to somewhat more recent machine learning systems—in particular to **deep neural networks**. We will therefore take a moment to introduce deep networks as an example of a supervised machine learning system.[2]

A neural network is a program whose execution is loosely based on the way computation happens in the brain. The brain represents information in **activity patterns** within groups of neurons. And it performs computation through the synapses that link neurons together, which cause activity patterns in one part of the brain to activate patterns in other parts. A deep neural network represents its input and output patterns as numerical activity values in banks of **units**, which loosely model patterns of activity in groups of neurons. The network links input patterns to output patterns through large arrays of **connections**, which loosely model synapses. Typically, these connections run through intermediate layers of 'hidden units', as illustrated in Figure 1. (In a 'deep' network, there can be many intermediate layers.)

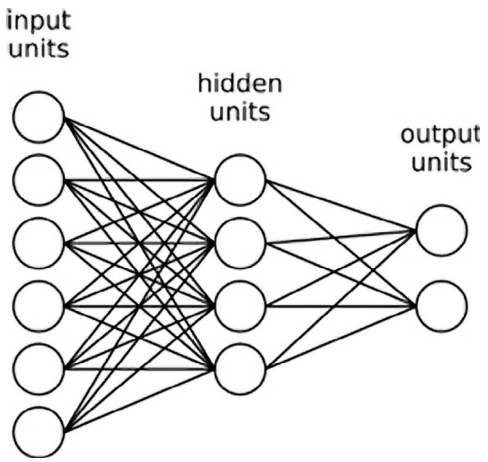


Figure 1. A neural network mapping an input representation onto an output representation, via an intermediate layer of units. (Connections are shown as lines linking units.)

2 In the remainder of the report, regression methods will be included in the family of 'AI methods'. We won't distinguish between older supervised learning methods like regression and newer methods like deep networks.

The connections between units can have different **weights**. Those with high weights pass signals on strongly, while those with weak (or zero) weight pass signals on weakly (or not at all). Different functions from inputs to outputs are implemented in a network by setting different patterns of connection weights. In this computation scheme, the computer's algorithm is specified by defining particular weights for all the connections in the network. A working deep network may have tens of millions of weights—so this way of telling a computer 'what to do' is very different from writing computer code by hand.

The learning method for a deep network begins by setting all the weights in the network to random values. In this state, each training input will be mapped onto an output which is also essentially random. Learning happens by comparing the *actual* output for each training item to its *desired* output—and then incrementally *changing* all the weights in the network, so the actual output produced becomes fractionally closer to the desired output. (The algorithm that makes these changes is called **backpropagation**.) With extensive training, a good network can map training inputs quite closely onto the desired training outputs—and can also map unseen inputs onto sensible outputs.

A deep network can reproduce certain types of human expertise quite well. Visual object classification is a case in point: current systems perform better than humans, on many kinds of object (see e.g. He et al., 2015). In fact, a deep network may at some level be a reasonably good model of the human visual object classification system (see e.g. Kriegeskorte, 2015). But as already mentioned, deep networks can also approximate human performance in more abstract, high-level tasks, such as document classification and text generation. In these domains, the networks approximate human performance through processes that are quite different from those taking place in humans. A deep network is just a very expressive, flexible function approximator. Backpropagation is just a very effective way of minimising a network's error on a set of training examples. If a deep network is trained on a large and detailed enough training set, its performance can often be somewhat human-like—enough for organisations to consider using deep networks to supplement or replace certain human tasks.

In practice, it is often the simple, repetitive, time-consuming work in a company that is most readily amenable to automation using AI methods. And often, supervised learning methods are the most suitable learning paradigm.

Reinforcement learning

A second machine learning paradigm is **reinforcement learning**. Here again, the system must learn to map inputs in some domain onto responses. But in this case, the system is not told precisely what the response should be: it produces responses, and for each, the trainer issues a *reward*, or a *punishment*. In relation to human jobs, this training paradigm is somewhat similar to a boss who congratulates employees when they do something right, and reprimands them when they do something wrong (but doesn't suggest what they should have done instead). It's also similar in some ways to a performance-related salary scheme, where employees get more money the better they perform. In such a scheme, every bad interaction with a client, each missed sale, incurs a penalty, and each good interaction brings a bonus.

Reinforcement learning systems are particularly common in domains where the agent can perform *sequences* of actions, and where a reward is only obtained after certain specific sequences are performed. Robots often operate in domains like this: typically, a robot must perform a sequence of motor movements in order to achieve a particular goal state. (This is true, for instance, if the robot receives a reward when picking up a given object, or manipulating it in a certain way.) A reinforcement learner must essentially learn its own training set, which specifies what actions to do in what situations. In most current systems, these training sets are then used to train a deep network, of the kind already described, to produce individual actions.

Unsupervised learning

A final machine learning paradigm is **unsupervised learning**. In this paradigm, the system is simply shown data from some given domain, with no additional training signal at all. The system's task is to learn *patterns*, or *regularities*, in the data. For instance, an unsupervised system could be presented with data about supermarket shopping trips, and learn that customers who bought several bottles of alcohol on a Friday on hot days during the Summer also tended to buy barbecue materials. In relation to human employment, unsupervised learning involves learning generalisations, which characterise customers, tasks, or other situations encountered by employees. A supervisor training a new employee might express these generalisations explicitly in conversation, to convey useful facts about the nature of the employment domain. ('On hot days during the Summer, people often come in shopping for barbecues…')

Technically, unsupervised learning methods typically perform some kind of **clustering** of their input data. This clustering identifies broad categories of input patterns—often in a way that usefully simplifies or compresses the raw data. Often, the input data supplied to a supervised learning system has undergone some form of clustering, which simplifies the learning task to be achieved

Unsupervised learning systems can also be useful for detecting *unusual* patterns in a dataset. An unusual pattern is one that falls outside one of the detected common patterns. Identifying unusual circumstances can be useful in domains like security and fraud detection, and also more generally in tasks where someone or something needs to be monitored.

B. AI systems in HR and Personnel Management

When we consider 'the impact of AI systems on employment', a useful place to look is the role of AI systems in *administering* employment: in hiring, monitoring and managing an organisation's employees. There is some evidence that AI systems are making a particularly large impact in these parts of organisations (see e.g. Volini et al., 2019). There could be several reasons for this. For one thing, HR-related tasks vary comparatively little from organisation to organisation, creating large markets for these tasks. (Accounting tasks are also relatively constant across organisations. But conventional programming is still the best way of automating these tasks; machine learning has a more subsidiary role to play.) For another thing, HR tasks often involve high volumes of relatively simple, repetitive activities, which are good targets for machine learning systems.

We'll discuss three types of HR/Personnel Management task in this section: employee recruitment, employee monitoring, and gig-economy job assignment.

Employee recruitment systems

Organisations have to advertise for employees when they have a vacancy. People seeking work must in turn hunt for suitable positions in suitable organisations, and in due course, submit applications for their preferred positions. Companies often receive large numbers of applications for their advertised positions, and must assess these. Typically, an application comprises a CV and a covering letter. Companies normally process these written application documents, and rank applicants, to create various long lists and short lists. Often, shortlisted candidates are invited for an interview in which a more detailed assessment can be made.

The Internet has affected every stage of this process. General 'job websites' such as Indeed, ZipRecruiter, and LinkedIn have become large clearing-houses for job seekers and job advertisers. The main service they provide is to *match* job seekers with job advertisers. (This is a service for which both seekers and advertisers are willing to pay.) Typically, applications are submitted online, via email or web forms. Scrutiny of applications sometimes includes accessing applicants' public social media presence, though this practice violates privacy laws in some jurisdictions, for instance in the EU, unless candidates' permission is sought and granted (see EU Data Protection Working Party, 2017). Finally, it's increasingly common for job interviews to be conducted online too, via video tools such as Zoom.

As tasks related to recruitment move online, they become increasingly amenable to modelling by AI systems. There are now AI tools that address each step of the recruitment process. Job websites deploy tools that learn how to perform the task of matching job seekers to vacancies. This learning typically involves extracting desired candidate attributes from job advertisements, and actual candidate attributes from CVs, and passing these extracted features to a

candidate ranking algorithm, which learns to reproduce the rankings of human experts (see Liu, 2009 for a generic ranking algorithm). A simple keyword-based version of the attribute identification task can readily be done with an accuracy of around 90% on unseen applications (see e.g. Intuition Engineering, 2018). The task of mapping the text of a CV or job ad to a set of keywords can be achieved with supervised learning, if training sets of documents with manually specified keywords are prepared.

When an organisation placing a job ad processes the applications it receives, it must perform its own version of the task of matching job ads to applicants, for this particular vacancy. In this case, there is an additional emphasis on identifying the best applicants, so CVs and covering letters are assessed for quality, rather than just for job description matching. There is some evidence that machine learning systems can actually perform better in this task than human HR managers. For instance Hoffman et al. (2015) trained a model to predict the performance of low-skilled workers in a company (which was evaluated separately when they were at work), based on a questionnaire. They then gave HR managers access to a decision support system that graded new candidates using these predictions. Managers could 'accept' or 'override' the system's recommendations. They found that overriding the system's predictions led to worse hiring decisions. A similar study (Chalfin et al., 2016) found that candidate grading models trained on actual employee performance were helpful in improving the performance of an organisations' employees.

In processing candidates' applications, companies may scan their social media sites, in jurisdictions where this is allowed, or where candidates have granted permission to do so. One common focus here is in identifying applicants' *personalities*. Job descriptions often require particular personality types. Applicants' social media pages hold rich information about their personality (see notoriously Youyou et al., 2015). It is quite easy for applicants to dissemble about relevant personality traits in a CV or covering letter, but much harder for them to do this in their social media pages. Social media sites also offer information about candidates in the form of images and videos. Image classification most readily delivers information about candidates' lifestyle and ethnicity—both of which are normally irrelevant for job suitability, except in special circumstances, and can be expected to raise challenging questions about discrimination and bias.

Even the task of interviewing candidates is one where AI systems are increasingly deployed. The interview involves a type of AI system called a **chatbot**, which we will discuss in more detail in Section C. Many of the most popular interview chatbots operate in a typed phone message conversation. The market leader in this space is probably Mya, which is already used at scale; for instance, one Mya client processed over 140,000 applicants for a warehouse job in three months (Schweyer, 2016). Mya appears to make large improvements in recruiter productivity: its own website claims an improvement of 140%, though we may have to make provisions for self-interest in this report. Other chatbots operate over the phone (e.g. Curious Thing). Still others operate over a video link, observing the interviewee through a camera. HireVue is a well-known system of this kind. These latter systems involve additional complexity (for instance, speech interpretation, visual gesture analysis), but potentially capture additional relevant information about candidates—in particular, cues to candidates' emotions, expressed in speech or physical behaviours (facial expressions and body language). These cues are supposedly informative in job interviews. However, nonverbal cues to behaviours like lying are not reliable (see e.g. Vrij et al., 2019), and methods for automatically identifying emotions through facial expressions and gestures are still in their infancy (see e.g. Ko, 2018). These multimodal chatbots also open new roads for discrimination and bias, similar to systems that trawl social media sites. However, we should note that considerable bias already exists in human-operated recruitment processes. It is possible that AI systems could eliminate some of the bias that already exists in recruitment processes. Indeed, some recruitment companies claim their systems do exactly this: we will discuss the case of Entelo in Chapter 3.

There is a large online discussion around how candidates should behave in job interviews to optimise their chances for AI processing. Similarly, there are many suggestions about what material candidates should include in their CVs. (Mentions 'Oxford' and 'Cambridge' and so on are encouraged.) Of course candidates can be coached for human job interviews—but there is a risk that AI systems pick up on superficial behaviours that can be more easily faked.

There are many questions that must be asked about an AI system used in a recruitment task. We briefly enumerate these here; they will be discussed more fully in Chapter 3.

- ***Quality:*** how accurately does the AI system perform its task? (Does it perform like humans? If not, then it will have a potentially large impact on an organisation's employee makeup.)
- ***Bias:*** does the system act in the same way towards all applicant groups? It may be that it is more *accurate* on some groups than others. It may also be that it is more *favourable* towards some groups than others in its assessments. (Bogen and Rieke, 2018 are a useful reference here.) But it may also be that systems can be built that eliminate or reduce human bias, as we will discuss in Chapter 3.
- ***Gaming:*** can candidates game, or fool the system, if they know how it is being used?
- ***Oversight and control:*** do human recruiters remain sufficiently involved in the hiring process, if they are aided by the system?
- ***Privacy:*** does the system infringe on applicants' private data? (This question arises for non-AI methods too, but it may be that automating processing of social media sites enables infringement on a scale that would not otherwise be possible.)

Employee monitoring and evaluation systems

AI tools are also used to assess employees' performance in organisations after they have been appointed. Again, the potential for AI assessment arises because information about employees' work and behaviour is increasingly available online. This is partly because employees often work on computers, where activity can be directly logged. Currently, 43% of US companies monitor employees' email, 45% track employees' computer keystrokes or time spent at the keyboard, and 43% store and review employees' computer files (American Management Association, 2019). To a lesser extent it is because of monitoring devices like GPS receivers, personnel trackers and video cameras, which are increasingly a part of workplace environments. 7% of US companies use video surveillance to track employees' on-the-job performance, and 8% use GPS to track company vehicles (American Management Association, 2019).

It is important to distinguish between AI and non-AI-based methods of evaluating employees. If an organisation simply logs online activity, or GPS coordinates, or collects information about this data in a database or spreadsheet, no AI techniques are being used. We can still ask many questions about intrusive practices, oversurveillance, and employee privacy. What we will focus on in this report are systems that use data of this kind as input to a machine learning system. This system could involve supervised learning, if the employer wants to predict some higher-level evaluation score. For instance, in 2014, the analytics company Sociometric Solutions fitted Bank of America call centre employees with personal trackers, and used the tracking data to build a model predicting performance. They found that employees who took breaks in large groups performed better, and were less likely to quit. They introduced shared coffee breaks for large groups of employees, and found this improved productivity by over 10%, while reducing turnover by 40% (The Week, 2015). The machine learning system analysing employee activity could also use unsupervised learning, if the employer simply wants to get a broad picture of the types of activity employees engage in, and/or the types of employee in the organisation.

Note that assessment technology can supply data about employees that can be retrospectively paired with their job applications, to generate training sets for the CV evaluation tools described in Section B. This practice appears to be increasingly common: employee assessment products are often now paired with job applicant assessment products. This pairing considerably reduces the amount of hand-annotation needed to train the job applicant assessment tools. But it also means the accuracy of these latter tools is limited by the accuracy of the assessment systems.

The questions that need to be asked about employee assessment AI tools are largely the same as those for AI tools used in recruitment: they relate to quality, bias, gaming, oversight, and privacy.

Job assignment systems in the gig economy

Human employees must be engaged, and their performance must be monitored—but they must also be *assigned work*. This task is normally the preserve of a human manager. But AI systems are intervening in management processes too. Sometimes, human managers use AI systems and other algorithms to *inform* their decisions. Sometimes, algorithms take over the task of management altogether. The latter type of management is especially associated with the so-called 'gig economy'.

A gig-economy company is in some ways like a job recruitment service, in that the key task is to match human workers with jobs. But in a gig-economy company, the jobs in question are short-term and casual, rather than long-term and official: they are one-off 'gigs', rather than salaried employment arrangements. Gigs can be many different kinds of work: logistics (taxi rides, delivery jobs), white-collar services (language translation, business or legal advice), healthcare (elderly care and nursing work (Caulfield, 2019), counselling, education (tutoring, assignment grading), and construction.[3]

Many commentators see the current gig economy as a temporary transition: units of work that are simple and systematic enough to be assigned as gigs are in many cases likely to be among the first to be automated (see e.g. Prassl, 2018). If this is the case, the new gig economy may provide a key piece of the infrastructure for the gradual automation of human jobs, with the largest gig economy companies transitioning into the largest robot companies. The trajectory of Uber and other transportation gig economy companies explicitly builds in a timeline like this: these companies are investing heavily in the technologies that will enable automation of the jobs they are currently assigning to humans as gigs. But in this section, we will focus on the technologies that do the assignment. (We will discuss the types of AI technology that replace jobs in Section C.)

The management of a gig economy company happens mostly online. Workers who are looking for casual jobs can upload details of the kinds of work they are looking for, and the skills they have. Organisations which have small jobs to be done upload details of these jobs. A computer system then matches workers with jobs, and presents workers with possible jobs and organisations with possible workers, which they can accept or reject. If both parties accept, then a short-term work contract is agreed. For a given contract, both the contracting organisation and the contracted worker have the ability to rate the other partner. These ratings have a bearing on subsequent matching decisions: poorly rated workers are recommended less to contracting organisations than highly rated workers, and poorly rated organisations are recommended less to workers than highly rated ones.

The matching algorithm that suggests worker-job pairings fulfils a function which is somewhat similar to that of a human manager in a traditional company. A traditional manager assigns employees tasks, and rates how well they perform these tasks, adjusting future task assignment in the light of assessed performance (and possibly dismissing employees who perform badly). Gig economy companies are different in that workers are able to decline offered jobs they don't want, and have no fixed work hours—both very attractive features for many workers. But in practice, many gig economy workers enter into fairly stable work arrangements with particular employers. In these cases, the gig economy matching algorithm essentially tells them what to do during their working day.

How much AI is there in a gig economy matching algorithm? These algorithms are certainly not *just* AI algorithms. Many of the best-known matching algorithms are at base types of **optimisation algorithm**, which consider many possible assignments of jobs, and pick the one that minimises some economically relevant variable. For instance, the Uber algorithm for assigning Uber drivers to driving jobs probably aims to minimise the time taken for a driver to reach a customer, while maximising the overall number of rides (see e.g. Voytek, 2014). However, these optimisation algorithms often incorporate a good smattering of domain-specific applications of AI. For instance, the Uber algorithm also tries to predict future traffic conditions when it is matching drivers to customers, using a model that has learned what traffic conditions to expect in various circumstances (Bell and Smyl, 2018). More generally, job-matching algorithms can learn how to perform their task, using workers' and contractors' ratings as a criterion for success. (The goal here is to learn matchings that maximise ratings for both parties.) In addition, gig economy companies are increasingly using unsupervised

3 Construction work has always had a large gig-economy component, with many workers being self-employed, or working on casual contracts. The novelty now is that the work is administered online.

learning techniques to identify different *types* of worker and *types* of gig, so that matching can be informed by assigned types (see e.g. Estrada-Cedeno et al., 2019 for a recent example).

The most widely discussed ethical and regulatory questions that arise for gig-economy companies relate not to the technical features of job-matching algorithms, but to the legal status of workers in these companies, in comparison to workers in 'standard' work arrangements. The main contentious issue is that gig-economy workers are legally self-employed contractors, rather than company employees. This means that companies aren't liable to pay tax contributions for their workers. In addition, it is much harder for gig workers to establish collective structures such as unions. These issues don't relate directly to AI, and we won't tackle them in the current report—but in Chapter 3 we will discuss gig economy work in the more general context of 'algorithmic management'.

C. AI systems affecting human jobs

As discussed in the previous section, AI systems are exerting an increasing influence on how human jobs are administered. We turn now to the *content* of this work: the production of human goods and services, and how AI technologies are impacting on this production.

We will use our review as an opportunity to introduce the key *types* of AI system. From a technical perspective, AI systems can be grouped into some fairly well-defined types. Our survey will be organised by these types, rather than by economic sector, since many types of AI system are used in several economic sectors.

Robots

A 'robot', as we use the term, is a physical device, that can move or behave with some degree of autonomy.[4] It perceives the world through one or more **sensors**, which can be of many different types; it achieves effects on the world through one or more **actuators**, which are essentially moving parts it can control—again of many different types.

Robots come in a wide variety of guises. (A thermostat is a robot, according to the above definition, because it can both sense and control its environment. So too is a smart house, which has many ways of sensing and regulating itself.) We will focus on two broad classes of robot, with particular implications for employment: **driving robots** and **effector robots**.

Driving robots

A self-driving car (or truck) is a robot, whose main actuators are steerable wheels. Its control system is relatively simple, with three degrees of freedom: a steering wheel, an accelerator and a brake. The sensors of a robot car are more complex. They have to replace the perceptual abilities of a human driver. Humans have many senses, but they rely mainly on vision (and a little on hearing) when they are driving. In a robot car, however, much of the sensing task is typically performed by a laser-based system called LIDAR, which can construct a map of all the objects and surfaces close to the car in all directions.[5] Machine vision is also used, to *identify* the important objects (particularly those close to the vehicle's current path) and to read street signs and road markings (to supplement what it knows from stored map data). Supervised machine learning is the standard way to build these vision functions. Some vehicles also use instrumented roads, which convey signals specifically designed for cars.

Self-driving vehicles are already a practical, used technology. Fully driverless vehicles are in deployment in many restricted roles, particularly in warehouses (Steininger, 2020) and in short-hop public transport (Fabulos, 2020). In the public transport field, the New Zealand company Ohmio is close to deploying an automated vehicle to transport passengers in Christchurch airport, and is engaged in several other trials around the world. Many cars that travel on public roads already have self-driving functions built into them—for instance, automatic parking and lane-changing. These vehicles all still require a human driver behind the wheel. However, fully self-driving cars that don't require a driver at all are being road-tested intensively by many companies; GM's Origin and Amazon's Zoox robotaxis are two examples. Google's self-driving car Waymo has now logged 20 million miles of road driving (Pressman, 2020). Self-driving

4 The term 'robot' is also used to refer to software, but we won't use the term in this way.

5 Tesla is an exception: it uses radar and an array of cameras to perform this task.

goods trucks are arriving on public roads (see. e.g. Heilweil, 2020). Self-driving tractors are also expected to make some impact in agriculture (FutureFarming, 2019), though we expect the benefits in this area to be most substantial in very large farms. The transportation industry is one of the largest employment sectors worldwide, accounting for 9% of the workforce in the US in 2016, and 5% of the workforce in the EU (ILO, 2020). In New Zealand, it employs 4% of the workforce (MBIE, 2020); it is only a matter of time before self-driving vehicles make a real impact on human jobs in this area.

Effector robots

We define an **effector robot** as a robot with arm-like or leg-like attachments, which it can position with some degree of flexibility. They tend to have more degrees of freedom than driving robots—legs typically have at least two degrees each, and arms tend to have at least three. A legged robot has applications in transportation, in places inaccessible to wheeled vehicles, such as pedestrian routes involving stairs, or certain types of rugged terrain. Robots with arm-like attachments have applications in a virtually unlimited variety of manipulation tasks, from picking up and placing objects, to assembly, inspection, painting, and so on.

A market leader in both types of effector robot is the US company Boston Dynamics, which produces a range of quadrupedal and bipedal robots, some with arm-like attachments (see e.g. Guizzo, 2019 for a range of products). These robots are somewhat unusual in the AI space, because their control systems don't rely much on machine learning; they rely instead on well-known control strategies (Burridge et al., 1999). But many other effector robots use a larger component of machine learning. Reinforcement learning is a common method: this method just finds the effector movements that most efficiently achieve some designated effect on the manipulated object. Market leaders in this area are Google DeepMind and OpenAI (see e.g. Gu et al., 2016; OpenAI, 2018).

In New Zealand, the agricultural sector probably provides the most opportunities for effector robots. Systems are already quite widely deployed in the fields of meat carcass processing (Scott Technologies). Automated milking systems are also being used, though industry commentators don't anticipate human milking being superseded for decades (McBeth, 2019). New Zealand is at the forefront of experiments with automated fruit handling: local company Robotics Plus has operated a robotic apple packer since 2018 (Groeneveld, 2020), and the first commercial robot apple picker was deployed this year in Hawkes Bay (Farm Weekly, 2020). These systems use machine vision methods, for instance to identify fruit or twigs. The machines operated by self-driving tractors can also be regarded as manipulating effectors.

Objects can be manipulated in such a wide variety of ways that most current manipulation robots are purpose-built for a particular job. However, research is ongoing into the building of **domain-general manipulation robots**. These are robots whose physical design allows them to perform an open-ended variety of tasks in an industrial workplace or around the home. The key bottleneck here is in teaching a robot to perform some particular task. This can involve creating a customised reinforcement learning regime (achievable, but impractical for non-technical end-users, and very time-consuming) or learning by imitation (more feasible for non-technical users, but with worse results at the time of writing). Practical domain-general robots are still some way away.

Decision systems

While robots perform a physical task for a human user, **decision systems** perform a more cognitive task. An **autonomous decision system** makes some decision, based on a collection of relevant evidence, without the need for human intervention. For instance, the photo processing system in an automated passport control gate makes a decision about whether a person matches their passport photo, which is not routinely deferred to a human operator. (The gate also counts as a simple 'robot', because the barrier opening mechanism is a physical actuator.) In other cases, a decision system provides a human decision maker with advice, and the decision is ultimately made by a human. We will use the term **decision support system** in this latter case.[6]

6 Often, AI applications contain a multitude of embedded components that could be thought of as decision systems. For instance, the object classifier in a self-driving car constantly makes decisions about the categories of objects in the road. However, these decisions are all means to a larger end (driving), which is not itself a decision. We will reserve the term 'decision system' for systems whose primary purpose for the user is making (or assisting with) a decision.

There are two common ways of building decision systems. One way uses supervised machine learning. For these systems, a training set of example decisions is created, containing some 'model' decisions that the system should learn to emulate. These training decisions are sometimes based on decisions made by model human decision-makers, and sometimes based on past cases, where the true facts are now known. The other way uses optimisation—that is, methods for finding the optimal value of some specified term. For instance, a decision system of this kind is often used to suggest the most efficient way to pack containers for freight, or to find a timetable with the fewest clashes. Formally, such optimisation tasks are often very complex, and in practical contexts it is often impossible to find the provably optimal solution; in these cases we must rely on techniques which approximate the optimal solution. The best approximations often make use of learning methods, (in particular reinforcement learning methods; see e.g. AlibabaTech, 2018).

Decision systems trained using these methods are used in many areas of the public sector. In criminal justice, decision support systems predicting recidivism are used to inform bail and parole decisions, as we already noted. They are also used by police forces to suggest how to deploy police patrols (though not yet in New Zealand, to our knowledge). In social care, decision support systems are sometimes used to make predictions about risks for children or families (again not currently in New Zealand, but certainly in the US—see Eubanks, 2018; Vaithianathan et al., 2019). New Zealand's Accident Compensation Commission (ACC) has recently adopted a fully automated decision system making decisions about applicants' treatment in certain simple cases (with more complex cases being deferred to human case workers), with the loss of around 300 jobs (see Pullar-Strecker, 2019). We discussed public sector decision systems in our first report (Gavaghan et al., 2019).

Decision systems are also in widespread use in industry. For instance, automated decision systems are routinely used to buy and sell shares in high-frequency trading. Decision support systems are used in various aspects of managerial work, such as economic forecasting and risk management (see e.g. McKinsey, 2017). Lower-level operational decisions can also benefit from decision support systems: for instance, they can be useful in the logistics industry, to plan routes for freight vehicles (ODSC, 2019), and in the insurance industry, to identify claims that merit special attention. They are increasingly used in medicine, both in the public and private sector—for instance, in medical image classification (Litjens et al., 2017), medical diagnosis (Loh, 2018) and medical research (see e.g. Mak and Rao Pichika, 2019). In agriculture, decision systems help farmers with a number of decisions: for instance relating to how and when to irrigate or fertilise (Talaviyah et al., 2020), or how to pair animals for breeding (Nayeri et al., 2019). Often these agricultural systems are integrated with sophisticated sensor systems embedded in the farm, gathering detailed data about soil moisture, and monitoring individual animals.

The adoption of decision support systems often allows decisions to be made by staff with less experience. For instance, after the adoption of a decision system assessing creditworthiness, decisions in some German banks are now made by staff with less expertise (see e.g. Floegel, 2019 for a case study). If decision systems can run fully autonomously, then again, fewer staff are required in various jobs. CCTV surveillance jobs are a case in point here. Certain types of event in videos can be increasingly accurately identified automatically. As accuracy improves, the need for the unskilled staff scanning videos for these events will drop (see e.g. Seldon, 2019).

To summarise: if there's any aspect of a human job where skilled judgements of a certain well-specified type are made repeatedly, on the basis of electronically available information, these judgements are amenable to being automated by a decision system. Throughout the public and private sectors, a new generation of technically minded entrepreneurs are analysing human jobs to find isolable judgements of this kind, and developing decision systems that replicate them.

Both automated decision systems and decision support systems can have an impact on human jobs. In the private sector, where profit is a primary motive, we should expect to see managers actively restructuring human jobs to incorporate decision systems wherever this allows services to be delivered better, or more efficiently. This is simply what effective managers in the private sector do. In many cases, we can expect that such restructuring will result in job losses, or in the reassignment of jobs to less qualified personnel. The public sector isn't immune to such efficiency measures either, as the restructuring of work at New Zealand's ACC testifies.

Decision systems potentially have an impact not just on the production and transportation of goods, but on the fields of human work that are responsible for *structuring and regulating society*. Their potential impact in these areas merits special scrutiny, and our study will pay particular attention to this question. Our study on the impact of AI 'on employment' focusses not just on how AI systems impact on individual workers and their jobs, but on the impact of these systems more collectively in the way society runs. This is a topic that is often overlooked in discussions about AI's impact on 'jobs', which focus heavily on individuals, rather than institutions. It will be our focus in Chapter 4, where we consider AI's impact on 'professions', understood very broadly as fields of human work that are instrumental in organising society. This definition includes in its scope the traditional 'Professions' of medicine and law.[7] But it extends beyond these, to government service (including policing and social services) and to education (including both school and university education), and to journalism (including both print and online journalism). It also includes some aspects of industry—particularly those relating to HR and personnel management. All of these 'professions' contribute to creating the character of a society. This happens through various codes of practice that individual workers are bound to uphold, sometimes officially and sometimes unofficially, that in some way go beyond their official job description. For instance, medical professionals and teachers have a duty of 'pastoral care' to their patients and pupils. University teachers have a duty to act as the 'critic and conscience' of society. Journalists have a duty to act as 'upholders of truth'. Civil servants have a duty to preserve and strengthen society's institutions, and guard against corruption. Managers have a duty to run their companies fairly. While all of these principles are clearly aspirational in character, they still impact on the decisions that individual professionals make—if not all the time, then certainly some of the time. If decision systems become widely adopted in any of these professions, it is possible that these principles may no longer be so tightly adhered to—either because they are hard to implement, or because system designers are simply not thinking about them. Consequently, for decision systems, we pose one other question about the role of AI on human jobs:

- ***Impact on professions:*** what impact (if any) would a decision system used widely in some role within a particular profession have on the profession as a whole?

Chatbots

Decision systems automate relatively isolated, 'stand-alone' human decisions. Other aspects of human work involve more extended interactions, with clients, or other workers. Human interactions happen most naturally in natural language *conversations*: these can happen face-to-face, or on the phone, or by email, or increasingly, on various social media platforms. Another type of AI system aims to replicate various human abilities to conduct conversations: these systems are called **chatbots**. A chatbot is an AI agent with certain simple conversational abilities. This agent sometimes interacts solely using written text (for instance in a phone messaging app, or by email). Sometimes it uses speech (for instance, by phone). Sometimes it uses a video link, so the human user's nonverbal behaviours can also be processed. Sometimes computer graphics techniques are used to give the agent a simulated physical head or body too: in this case, the agent can generate facial expressions, or hand gestures, or point to objects in a computer display.

The current generation of chatbots can't understand language in the way that humans do: proper natural language understanding is still an unsolved problem. However, in certain very well circumscribed domains, chatbots can give a reasonable impression of being able to understand a human dialogue partner. They achieve this through the use of a **dialogue script**, which is created by hand by a human author. A dialogue script specifies a number of **dialogue contexts** that can arise during a dialogue in a given domain. For each context, the script anticipates a smallish set of possible things the human interlocutor might say, represented as a set of **utterance types**. For each utterance type, in each

7 It also includes religion, though this won't be a focus of ours for obvious reasons: religion arguably has less of an organising role than it used to; it no longer employs large numbers of workers; and AI techniques are essentially absent from this field. (Frey and Osborne (2013) list 'clergy' in the bottom 5% of jobs at risk from AI—and we are very surprised it is not listed lower.)

context, it also indicates (a) what utterance the agent should produce in response, and (b) what the new dialogue context will then be. In this new context, it waits for another user utterance.

Current chatbots make heavy use of supervised machine learning to interpret utterances. The script author needs to supply, for each context, and each utterance type, a set of several different ways the human user might express this utterance type—the more comprehensive, the better. This creates a training set for an **utterance classification system**, which can then learn that in some given context, the utterances "I'm hungry", "I am starving", "I haven't eaten all day" (and so on) are functionally equivalent, and require the same response. A good script author is good at anticipating the different types of utterance that might occur in any given conversational context, and at comprehensively enumerating the different specific utterances that convey these different types of message.

Chatbots designed in this way can be built to perform conversation tasks in many different fields of human work. In the domain of HR, we have already mentioned that chatbots can be used to conduct pre-screening interviews with job applicants (see Section B). Chatbots are also used in other areas of HR; for instance, chatbots like AskHR and Lexy are able to guide employees through various administrative tasks, to do with payroll or leave booking, which often fall to HR personnel (see e.g. Westfall, 2019). Chatbots are also increasingly used in health domains. They are used to conduct patient interviews, prior to consultation with doctors. (Covid screening interviews are a common current application – see e.g. Vanian, 2020). They are particularly commonly used in counselling domains, where conversation is used not only to perform diagnosis of conditions, but also to provide treatment. (Again, Covid applications are a current trend; see e.g. Simonite, 2020). Another branch of medicine (and social services) where chatbots are finding many niches is in elderly care. Here, chatbots are often designed to play a mixture of roles: medical (giving reminders about pills to be taken, or exercises to perform), assistive (answering questions about the care environment, giving information about available activities), and purely social (being a friendly and entertaining 'companion'). Chatbots are also widely used in companies' telephone or online call centres, to answer customer enquiries and complaints, to respond to technical support queries. In these domains, human call centre employees are often given a script as part of their training, which gives some indication that script-based chatbots will be able to reproduce the desired work. Chatbots are also involved in sales and marketing contexts (Rauthan, 2019). Companies routinely advertise their products on websites: these websites can now be augmented with a chatbot playing the role of a salesperson, which can conduct a personalised interaction with every website viewer who chooses to engage. (Salespeople are also routinely given a 'script' as part of their job training, suggesting that their jobs are also ripe for automation.) Finally, chatbots are also making some inroads into educational domains (Smutny and Schreiberova, 2020). Teaching interactions often happen through dialogues; in some ways, a one-on-one dialogue with a tutor is an ideal educational paradigm. Sometimes, educational chatbots incorporate domain-specific problem-solving functionality in the domain of instruction—for instance, the ability to set and solve problems in a particular field of physics or maths, and to assess student answers, and perhaps identify certain classes of misconception and mistake (see e.g. D'Mello et al., 2010 for a pioneering system of this kind).

In each case, what makes a dialogue task suitable for emulation by a chatbot is the *limited domain* of the task. Script-based chatbots are not good at open-ended conversations; they need to keep the conversation running on a known track. In practice, dialogue engineers often ensure this by building systems where the dialogue agent takes most of the conversational initiatives. Human conversations are often 'mixed-initiative', featuring initiatives from different participants at different times. Chatbot conversations often involve the chatbot agent issuing instructions, or asking a series of questions, to elicit maximally predictable user responses.

Again, in domains where human workers are supported by chatbots, various kinds of restructuring of human jobs are likely to happen. For one thing, certain human jobs are likely to change in some interesting ways. It may become more common for the pro-forma or routine aspects of a client interaction to be conducted by a chatbot, who then hands over the client to a human practitioner, with an appropriate summary. This kind of hand-over already occurs between human practitioners—for instance, in the health domain, it is an accepted part of many workflows. (A GP may hand a patient over to a consultant, for instance.) In the future, handovers of clients from chatbots to human

workers may become a common feature of workplace interactions. We will discuss some regulatory issues relating to chatbot handovers in Chapter 4. But however chatbots enter human workplaces, we anticipate that the efficiencies created by their use are likely to lead to redundancies in some form. We envisage jobs where conversations are most predictable and 'scriptable' will be most automatable. Even if handover to humans is quite common, the adoption of chatbots will require fewer humans to conduct the required volume of conversations.

Question-answering systems

A simpler form of language interaction is a question-answer exchange. People often have information needs that can be expressed in the form of questions. The first generation of Internet search engines required users to express queries as unstructured sets of query words—but **open-domain question-answering** systems now process queries expressed as full sentences, and respond with full-sentence answers. This can be done using extensions of regular Internet search methods (see e.g. Harabagiu et al., 2005) or using more recent machine learning methods (see e.g. Kwiatkowski et al., 2019). Many chatbots incorporate open-domain question-answering functionality. For instance, popular systems like Amazon's Alexa, Google's Assistant and Apple's Siri all support open-domain question answering.

Humans are not typically employed in open-domain question-answering tasks – this task is simply beyond them, so these systems are unlikely to cause any job losses in their own right. But domain-specific chatbots are often improved by a little open-domain functionality. For instance, a sales chatbot is often more effective if it can incorporate some general chitchat (about the weather, or sports, or current celebrity gossip): so they are likely to have some impact on job losses through this route. In addition, there are also domain-specific question-answering systems, that are configured to handle questions in one particular domain, for which a structured database is available—for instance, sports results, or music or film trivia, or maths questions. These systems emulate particular areas of human expertise, but again, there are few human jobs which involve nothing but question-answering. However, scripted chatbots are often considerably improved by domain-specific question-answering systems. These systems allow users to take some initiatives in the dialogue—at least, if these initiatives are questions. They often allow chatbots to deliver customers actual services, particularly in sales or travel booking domains, where the customer can ask questions about products, and be offered products for purchase in response.

Text analysis systems

AI systems are also useful in analysing large volumes of text for particular customised purposes. For instance, if we have a large corpus of customer reviews for some product, it is useful to be able to assess whether these reviews are positive or negative. This is a task amenable to supervised learning: **text classification systems** can be trained to place texts into predefined categories, using training sets where the relevant categories have been assigned by hand, by human judges. The utterance classification systems used in chatbots are small-scale examples of text classification systems.

Text classification can be used in many contexts. For instance, text classification systems are becoming widely used by law firms, to provide quick analyses of contracts and other legal documents (see Faggella, 2020 for a recent review). The classifier in this case often operates on individual clauses, so a given clause can be classed as an indemnity clause, a penalty clause, and so on. Legal AI companies often offer companies an overall analysis of the contracts they are bound by, which is helpful in assessing risk. Text classification is also used to make automated predictions about the outcome of cases. In commerce, text classification has a myriad of functions. For instance, it can be used to create structured databases of documents of particular types, or in marketing, to identify online promoters and detractors of a product.

Note that the training sets used to train a text classifier must be classified by human analysts, who possess the abilities the system aims to replicate. For instance, legal document classifiers must be trained using training data created by legal workers—typically junior staff. Junior legal staff have always been assigned relatively mundane, repetitive text-analysis work of this kind. The difference is that if their work is being used to train a text classifier, *subsequent* generations of junior legal staff won't have to do such work. Creating training data for an AI system is a 'one-time only' human job, as we will discuss in more detail in Section D. If certain types of work suitable for junior staff are automated, this might have an effect on the career structure of some professions. For instance, this junior work might have

a useful function in staff training, or in assessing staff performance and deciding about promotions; if the work is automated, new ways of accomplishing these functions will have to be found.

Text generation systems

Another type of text-processing AI application is a **text generation system**. A text generation system takes information expressed in some non-linguistic format (for instance, a spreadsheet or database) and expresses it in sentences or paragraphs of text. There are many applications for this technology: see Gatt and Krahmer (2018) for a review. For instance, simple types of news report can be produced using text generation, without the need for a human writer. Text generation systems are commonly used to write weather forecasts and weather reports, and stock market reports; increasingly they are used to generate summaries of sports events. An interesting application is in local journalism: if databases in a standard format are available for each region of a country, a system can be built that automatically produces a news story for each region. In this case, the human journalist creates a text generation system that is able to report a certain type of story, which is then deployed to produce many instances of this type. A UK company, Radar AI, specialises in generation systems producing local news of this kind.

Text generation systems are also widely used to create the text of web pages. When users browse a company's product catalogue, they are often reading text that is partly generated automatically, from a database entry. This helps keep text up to date, as product databases change. Sophisticated webpage generation systems can produce text tailored to particular users, based on their known preferences, or on information about other products they have already seen (see O'Donnell et al., 2001 for the original system of this type).

We should finally note that **machine translation systems** are a special case of text generation systems. Modern machine translation systems rely heavily on machine learning. The learning in this case requires a large training set of hand-translated sentences, produced by human translators. Here again, the human work of creating the training set is a 'one-time-only' job: by contributing their work to this task, human translators are to some degree imperilling, or even eliminating, their own future jobs.

D. Human jobs created by the AI industry

In discussions of AI and employment, the point is often made that new AI technologies, like all technologies, will create new jobs as well as eliminating some old ones. Commentators point to previous industrial revolutions, where this was indeed always the case. The status of this historical argument is not very clear, since the ultimate objective of AI as a discipline is to replicate *all* human abilities, not just some subset of them. Even in a future world where AI systems are able to perform most human tasks, it is still possible that services performed by 'actual humans' could acquire special cachet and value, and open the way for new fields of human work: this is an issue we will discuss in Chapter 2. But those questions are outside the scope of the current chapter, which is mainly concerned with introducing existing AI technologies. In the short term, at least, the existing AI technologies undoubtedly create some new jobs for humans.

The most obvious new jobs will be in the **development and deployment** of the new AI systems. These jobs will be highly technical in nature, and will require highly trained engineers. We also anticipate there will be many new jobs in **testing and overseeing** AI systems, to make sure they behave as they should. New uses of AI will introduce new requirements for accountability and explainability, as we will discuss at length in this report: skilled workers will be needed to implement these requirements. So we can certainly expect more jobs for skilled AI engineers. But we will conclude by mentioning two interesting categories of new job that aren't primarily about engineering.

The first of these is the job of **content authoring**, which is required in the construction of most current chatbot systems, and in many current text generation systems. This job is interesting, in that it requires a mixture of technical and humanities skills. The humanities skills are in creating textual content that is coherent and/or compelling for the target audience—a task that calls for the same skills as a playwright or screenwriter uses to create convincing dialogue, or a writer uses to create convincing monologue. The technical skills are in expressing this content within a general text generation or dialogue management algorithm, which defines a large *space* of possible texts (or conversations). We note in passing that workers in this area would benefit from an education that spans the humanities and the sciences. Often, the best content authors have exactly this formation.

The second job is that of **training data preparation**. This job requires different levels of skill for different tasks. Some jobs are suited to almost anyone with sufficient persistence and concentration: the job of identifying objects in images is one of these. Others require certain specific qualifications, such as the job of classifying the clauses in legal documents, or the job of translating sentences (both discussed in Section C). Others require high expertise, such as the job of classifying medical images (again discussed in Section C). In this latter case, preparation of training data is likely to be an imposition on top of an already busy schedule.

For both content authoring and training data preparation jobs, a key question we will ask relates to the 'one-time-only' nature of the work. As noted in Section C, when a human worker uses some human skill to produce a dataset which is used to train an AI system that can reproduce that skill, she is essentially doing work that eliminates the need for this kind of work in the future. The whole point of content authoring jobs, and of training data preparation jobs, is that once the human work is done, that same work can thereafter be carried out by the system that is built—in arbitrarily many copies, and indefinitely. Of course, as AI technologies improve, new systems will have to be built and deployed, and this will require skilled engineers, as already noted. But the amount of human work needed here is largely a function of the pace of technological development, rather than of the scale of system deployment. To summarise the question we wish to discuss:

- ***One-time-only work:*** should any special status be accorded to the 'one-time-only' human work through which an AI system is created that eliminates the need for similar human work in the future?

It's interesting to note, in passing, that the low-end training data preparation jobs are typically brokered in the gig economy. In fact, the original gig economy site, Amazon's Mechanical Turk, is one of the world's largest contractors for human services in training data preparation (see Ipeirotis, 2010 for indicative data). This task was one of the earliest uses of Mechanical Turk, so it helped to bootstrap the gig economy revolution. Workers preparing training data on gig platforms are insecure both in having minimal health and social security benefits at work, and minimal collective bargaining rights (see Section B), but also in performing work which is often by its very nature temporary.

2. THE CHANGING NATURE AND VALUE OF WORK

A. Jobs, work and COVID-19

In the face of the global pandemic, the effects of artificial intelligence on jobs and work might seem like the least of New Zealand's worries. But in fact, the health crisis is accelerating changes which will alter both the way we work and the way we think about work. Alert levels and lockdowns have taught us a great deal about jobs and work in Aotearoa.

Although at the time of writing the unemployment rate in New Zealand remains surprisingly low, there has been a marked increase in 'underemployment' — that is, part-time work where the worker has the desire and availability to work longer hours (Dann, 2020; Statistics New Zealand, 2021). More importantly, the effects of the pandemic on jobs and work have been uneven (Reich, 2020). Many workers, particularly white-collar workers, were able to work successfully from home in jobs that have remained relatively secure. Conversely, hospitality and tourism workers saw their incomes evaporate. The lack of job security for those in non-standard work, such as gig-economy work, now seemed much more problematic than it did in a strong economy.

Most jobs became more onerous. Essential 'frontline' workers coped with life behind PPE as well as the threat of infection. Those who could work remotely had to adapt to work online with varying degrees of success and fatigue. Non-essential workers who could not work from home experienced enforced time off work, wondering whether their jobs would be casualties of the economic downturn. Those who made it back to work returned to new socially distanced workplaces in which even simple tasks had to be reinvented.

In adapting to the pandemic, all of us discovered a new-found flexibility in the way we work. Businesses as well as professions have been forced online. For some, the move was long overdue, but results have been mixed. Telehealth and e-learning have often proved imperfect substitutes for the face-to-face versions of medicine and teaching.

Most workers discovered their jobs to be 'non-essential'. Admittedly, what counts as essential work during a pandemic lockdown is perhaps not what would count as essential in other circumstances. Even so, many of us have considered the importance of the work we do and, as a society, we have had to focus on what John Maynard Keynes called *absolute needs* (food, water, warmth, comfort, security, companionship…) which he contrasted with *relative* needs—goods and services seen as essential in some cultures but not in others.

Given the complex and unpredictable nature of the pandemic (and of global responses to it), jobs we once thought secure, now seem less so. Many New Zealanders have considered what life would be like if they worked less and earned less. Lockdown showed many of us how much we value close contacts with workmates and the structure and satisfaction found in good work environments. Conversely, many discovered that they enjoyed a world with less noise, with more time for family, and without the stress and time-cost of the daily commute. In some industries, moving online using tools like Zoom and Slack enhanced collaboration with distant colleagues and enhanced the quality of life of workers. But, the costs and benefits of enforced working from home were unevenly shared between men and women, workers with and without young children, and workers in different types of living accommodation. Even so, many employers were surprised to discover that their workforce was no less productive when based at home. When surveyed (O'Kane *et al*, 2020), the majority of New Zealanders working from home said they would like to keep doing so, at least in part. The benefits of working from home have been so pronounced in some industries that large international companies like Twitter and Square have announced that they will permanently allow employees to work remotely (Kelly, 2020).

In the face of a perceived threat of mass unemployment, New Zealanders seemed prepared to countenance radical changes to work and income, such as a 4-day week (Burrows, 2020) or Universal Basic Income (Mills, 2020). To many New Zealanders such changes would have seemed unthinkable just six months earlier. This openness to change has been partly driven by a newfound post-lockdown realisation for many New Zealanders that unemployment is as likely to be the result of bad luck as of poor life choices and that levels of income can be similarly unfair. Many were also struck by the incongruity of a lot of essential workers being amongst the lowest paid of New Zealanders. This is partly due to preexisting inequities such as the gender pay gap (Human Rights Commission, 2011) and the fact that labour markets reward skills that are scarce and easily monetised more than they reward skills that help to secure society's absolute needs.

Paradoxically, during the pandemic, some have grown richer. While many faced periods with reduced earnings, others kept earning but faced periods with reduced spending. Because many of those who maintained their income were originally well paid and hence able to save effectively, overall personal savings rates sky-rocketed.

In the US, these rates normally fluctuate around 8%, but fears about possible future hardship saw them balloon to 33% in May (Fitzgerald, 2020).

This concrete demonstration of the unequal distribution of work and income has sparked popular anger at inequality. Protest movements like Black Lives Matter have been driven by (particularly young) people looking with fresh eyes on political and economic systems that acknowledge the pernicious effects of persistent inequality (Meyers, 2020) but which remain philosophically opposed to the redistribution of wealth (Piketty, 2014). Everybody knows someone who has lost work and/or income, and we have all seen headlines trumpeting the injustice of the owners of well-placed tech companies like Amazon and Zoom getting rich at a time when most people fear getting poor. Black Lives Matter demonstrates, if demonstration were needed, the failure in most countries of the level playing field, assumed by the meritocratic ideals that have underpinned most developed countries for the past four decades. In *The Tyranny of Merit* (2020), Harvard political philosopher Michael Sandel argues that we are experiencing the effects of a more fundamental failing in meritocratic democracies around the world. Even if they worked as intended, they necessarily engender hubris among the 'winners' and humiliation for the 'losers'. These ideas are not new—in 1762, Rousseau wrote "The good man can be proud of his virtue because it is his; but of what is the man of genius proud?". Even so, the current erosion of democracy has lent such ideas new weight. The rate of that erosion is, of course, sensitive to the magnitude of inequality—the amount the winners win, and the amount the losers lose.

All these challenges exist against the backdrop of a world experiencing rapid social and economic change, driven by new technologies fuelled by AI. Crises have a tendency to speed up pre-existing social and economic trends (Haas, 2020) and this one is no exception. Despite a sharp global downturn, the profits and share prices of AI-focused companies have jumped in industries such as computing (Apple) social media (Facebook), retailing (Amazon), entertainment (Netflix), financial technology (Square), and transport (Tesla). Although not all driven by AI, automation in general is the focus of renewed interest. This is partly driven by struggling companies searching for efficiencies (Chandler, 2020) but it is also influenced by the particular conditions of this public health crisis. Workplaces that have traditionally involved people working in close proximity (e.g., meat workers) are now looking to robotics as a means of maintaining production threatened by lockdowns and demands for social distancing of their workforces (Bunge and Newman, 2020). The trend toward increasing automation is also partly caused by politics. In the face of economic headwinds, many countries are seeing increasing calls for the 'reshoring' of manufacturing jobs. This is predicted to increase automation, due to the inability of such companies to cover increases in labour costs as jobs are moved from developing to developed countries (De Backer, 2016, p.26).

Although 2020 was a terrible year in many respects, it was a good time to reflect on the nature and value of jobs and work, and on the way the working lives of New Zealanders could be affected by increased use of AI in the provision of goods and services and in the workplace. At a time when so much seems under threat, it is tempting to think that AI is yet another threat to the jobs and incomes of New Zealanders. But the real story is more complex and, in some ways, more hopeful.

This chapter addresses the great variety of predictions about the effects of AI on jobs and work, but it also addresses a deeper issue—what is it that we would want AI to do for us in the workplace? Can AI take the drudgery out of work? Will it be the spur to upskill New Zealanders into more interesting and better paid jobs? Will it finally make good on the age-old promise that automation will be labour-saving?

In 1930 John Maynard Keynes wrote a much-quoted essay entitled "Economic Possibilities for our Grandchildren". He argued that, if the standard of living continued to rise through the 20th century, technological progress would allow his grandchildren to work a 15-hour week. Standard of living is notoriously difficult to measure and it is an average and hence, insensitive to important inequalities of various types. Also, the 20th century turned out to be very turbulent, so increases in standard of living were far from smooth. Nonetheless, many indicators did show a rapid rise through the 20th century. In New Zealand, inflation adjusted GDP per capita doubled from 1900 to 1950 and doubled again from 1950 to 2000 (Maddison 2003, pp. 85–87). Moreover, there is good reason to think that GDP has undercounted the benefits of technological innovation[8].

8 This is a vexed and longstanding issue in the history of national accounting. As far back as 1966 the influential Boskin Commission report in the United States concluded that failing to take account of the quality changes in goods such as computer, cameras, and phones meant that the US. Consumer Price Index had been overstating inflation by 1.3% per year and correspondingly understating real GDP growth (for further discussion see Coyle, 2015, pp. 88–90).

Keynes's predicted "that the standard of life in progressive countries one hundred years hence will be between four and eight times" as high as it was in 1930. The Yale economist Fabrizio Ziliboti reassessed these predictions (2008) and found that the fourfold increase had happened by 1980 and that by 2030, we might see a seventeen-fold increase. But despite technology making us better off, humanity has not taken to working shorter hours or fewer days of the week. This is partly explained by Keynes' observation that "there is no country and no people, I think, who can look forward to the age of leisure and of abundance without a dread. For we have been trained too long to strive and not to enjoy". It is also partly explained by the fact that no country has solved the technical problem of evenly distributing the fruits of technological progress across their population, or devised economic conditions that allow for a gradual decrease in hours worked while maintaining a reasonable level of income security. Instead, most developed countries have stuck to the post-World War II 40-hour work week (Suzman 2020, p. 340). Up until relatively recently few political leaders have seen decreasing the work week as a politically viable aim, although at the start of 2020 Iceland's new Prime Minister Sanna Marin announced the long-term goal of having Icelanders work just six hours a day four days a week (Kelly 2000).

Whether or not we think working less is desirable, AI will undoubtedly change the way we work and, for some, the amount we work. If we are to assess the likely effects of these changes on the wellbeing of New Zealanders, we will first have to address the question—what is it that makes work valuable to people? We will then critically comment on a variety of predictions about the effects of AI on jobs and work in New Zealand, before addressing ways in which we might adapt to likely changes.

B. Work and wellbeing

Work is obviously valuable as a source of income. So, AI-induced changes in work and income will have important consequences for the livelihoods of New Zealanders and on the wealth of communities, iwi and the country as a whole. Issues regarding AI's effect on earnings, GDP and particularly on wealth inequality are addressed in later sections of this chapter, but for now we want to address a larger question. In the medium-term AI will increase productivity and decrease our cost of living (see Section A of Chapter 3). There is also good reason to think that New Zealand could implement economic policies which help to distribute those gains across the population—we have, after all, had much more redistributive economic policies in the past (Morgan and Guthrie, 2011). So, leaving earnings aside for now—what effects will AI have on us if it changes the types and/or amount of work available to New Zealanders? Those concerned about technological unemployment (Rifkin, 1996, and Russell, 1932) have often framed this problem in terms of gains and losses of quality of life. So, how might wellbeing be sensitive to changes in, or loss of, work?

Since New Zealand moved from the largely financial evaluation of policy outcomes to a broader wellbeing framework, a great deal of effort has gone into understanding what it takes for New Zealanders to flourish. This is exemplified by the Living Standards Framework developed by New Zealand Treasury, summarised in Figure 2.

Figure 2. New Zealand Treasury's Living Standards Framework.

As befits Treasury's purposes, this is a not an exploration of ways in which people can maximise their wellbeing so much as it is a recipe designed to provide an adequate level of wellbeing for all New Zealanders. While 'jobs and earnings' appears as one of the domains of wellbeing, our ability to work and the sort of work we do clearly influences all of the other domains. This effect is partly financial and partly due to work being psychologically important to people's identity and status. It is suffused with social connections and it is, for many of us, the major use of time in our lives. To evaluate the importance of these non-financial aspects of wellbeing we need to dig deeper into the nature of wellbeing itself.

Academic work on wellbeing has produced many theories about its nature. It is a long-standing and varied discipline (Crisp, 2017). At one end of the scale are 'thin' characterisations of wellbeing, based on positive and negative mental states. Jeremy Bentham (1789) begins his *Introduction to the Principles of Morals and Legislation:* "Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do". But most philosophers think the hedonist's calculus is too simple to capture what people really want from life, or alternatively, the many ways in which people's lives can be successful. Thomas Carlyle famously described Bentham's view as "the philosophy of swine", as it placed rooting round in the muck—the favourite pastime of pigs—on a par with what John Stuart Mill (1863) would have called "higher pleasures", such as enjoying art or excelling in sport. At the other end of the scale are 'thick' characterisations, which try to capture what humans would ideally like to achieve in life. In this context, the ancient Greeks focused on 'eudaimonia' (εὐδαιμονία) which is often translated as 'human flourishing'. Socrates argued that eudaimonia comes from exercising virtues such as self-control, courage, justice, and wisdom. For Aristotle, wellbeing was gained by developing our highest and most human capabilities. The Stoic philosophers (such as Zeno, Cleanthes and Chrysippus) focused on living in agreement with nature, stressing qualities like justice, honesty, moderation, simplicity, self-discipline, resolve, fortitude, and courage (Baltzly, 2019).

Modern psychological characterisations of wellbeing tend to focus on the features of life that help to achieve the eudaemonist's rich notion of wellbeing. Perhaps the most famous is Richard Seligman's PERMA theory (Seligman, 2011). The acronym represents the five elements necessary for a meaningful life. These are: Positive emotions, Engagement, Relationships, Meaning, and Accomplishments. Part of the reason for diversity in characterisations of wellbeing is that the concept is doing a variety of tasks. The hedonist seeks a snapshot of how we are feeling now. The Eudaimonist seeks a recipe for maximising happiness and satisfaction across our lives. It's also interesting to note that the virtues promoted by Socratics and the Stoics enhance, not just your life, but also the lives of the wider community. The PERMA theory seeks the building blocks of a successful life. These are the independent variables by which we might influence our success. It is this latter sense of wellbeing that we think will be particularly useful in evaluating the effects of changes in jobs and work.

Assessed against Seligman's criteria, work has a variety of effects on our quality of life. Desirable jobs are great sources of wellbeing, but the most recent Statistics New Zealand report, *Survey of Working Life*, shows that wellbeing is eroded by work that is low-skilled, repetitive, insecure, allows limited social engagement and personal autonomy, and has little perceived social value (Statistics New Zealand).[9] As these wellbeing criteria can be satisfied by non-work activities (e.g., art, sports, hobbies, engagement with friends and family, charitable activities, adult learning etc.), low quality work can be doubly harmful as it leaves the individual with little time or energy to pursue other more beneficial activities.

In the survey, 88% of workers reported being satisfied or very satisfied with their main jobs, while 4.3%, or about 113,000 people, reported being dissatisfied or very dissatisfied. Encouragingly, this suggests, that work is often an effective way to enhance our wellbeing. That said, measures of job satisfaction do little to inform us about the extent to which work makes our lives better than they would otherwise be. This satisfaction contrasts starkly with Gallup's *State of the Global Workplace Report* (2017, p. 20), which found that very few people are engaged by or interested in their jobs—just 15% worldwide across 2014–2016.

Obviously, for most adults, work is compulsory and so assessments of job satisfaction are likely to be assessments of how good our current job is compared to other jobs we think we might plausibly get. Presumably, 'satisfaction' with work would be lower if we were asking people whether going to work is better than going camping or watching the netball. Interestingly,

9 These issues are revisited in Chapter 3.

in his excellent *Work: A history of the way we spend our time*, James Suzman notes that much recreation consists of tasks that used to be work (pottery, fishing…) but without the division of labour that has become the hallmark of the modern workplace (2020). All else being equal, leisure activities ought to be better at promoting wellbeing than work activities. That is, after all, what they are supposed to do. But in one respect, all else is not equal. As Keynes noted, in his day (and in ours) education is largely directed at work-related skills and not at teaching people how to enjoy their leisure hours. Moreover, for many people work is such a large and tiring part of life that, once the household chores are done, remaining leisure time is spent resting or 'blobbing out'.

There is also evidence that many New Zealanders have more work than they can handle. 'Job strain' is a measure devised by the OECD to capture non-economic aspects of employment. A worker experiences job strain if the job demands they have (time pressure or physical health risk factors) outweigh the job resources (workplace autonomy and positive relationships) that they have at their disposal. In New Zealand 34% of employees report being under time pressure and 25% of employees count as being under job strain.

New Zealand has an enviable score on the OECD Better Life Index (OECD, 2020). Our only strongly negative statistic is work-life balance. 15.1% of New Zealanders average more than 50 hours work per week. This places New Zealand 33rd out of 40 OECD countries, but it is notable that the top 19 countries all achieve less than 5% of workers averaging more than 50 hours work per week. Our tendency to overwork is very much a cultural phenomenon, with New Zealanders proud to report that they are busy or even 'snowed under'. Countries with higher GDP than New Zealand tend to praise efficiency rather than long work hours. Germans who send work emails in the evenings or the weekends are considered inefficient, rather than hard working.

This chapter began by asking how our wellbeing might be affected if AI changes the types of work we do or the amount of paid work available to us. The answers to that question are complex. Apart from being a source of income, work for many New Zealanders is a source of enjoyment and satisfaction. But a great deal depends on the sort of work you do. The proportion of us that experience workplace time pressure, job strain, and overwork, tells us that many New Zealanders are harmed by the nature and quantity of the work they do. So, there is certainly opportunity for AI to make our working lives better if it decreases workplace stress or overworking. But there is a corresponding risk that it might make our working lives worse, for example by decreasing the skill levels required of some workers, forcing them to work longer hours to achieve the same income.

What if AI increases GDP but decreases the overall amount of work available? While full-time or close to full-time paid work is important to the wellbeing of many New Zealanders, such paid work is nonetheless not a necessary condition for wellbeing. Many (mostly retired) New Zealanders lead happy and fulfilling lives despite not having significant paid employment. There is considerable debate about the nature of the well-known happiness U-curve (Graham and Pozuelo, 2017), but there is little disagreement that most people's reported happiness increases as they near and then pass retirement age.

It is also important to note that different cultures make different assessments about the intrinsic value of work. In some cultures, work is an end in itself: the USA is the paradigm example here (see e.g., Thompson, 2019). In other cultures, supporting one's family and community are seen as the fundamental goals in life, and work is mainly important insofar as it advances these goals. New Zealand culture is an interesting blend of these attitudes. The population as a whole may have work-life balance issues, but, within Māori and Pasifika cultures, there is a clearly identifiable focus on community and family over career (see e.g. Houkamau and Sibley, 2019 for a Māori perspective; Lopesi, 2020 for a Pasifika perspective, and Theodore et al., 2017 for a combined study). If AI decreases the amount of work available, and curbs our ability to find fulfilment through work, we may have a lot to learn from Māori and Pasifika communities, which showcase alternative and possibly more resilient value systems. We also note that time away from paid work allows people to focus on goals related to the environment, and to other sustainable development goals.

RECOMMENDATION 1: We encourage government to acknowledge Māori and Pasifika perspectives on work-life balance in evaluating New Zealand's response to AI.

We can now make some preliminary conclusions. Contemporary New Zealand doesn't just suffer from inequality in wealth, but also from inequality in work. This is true both in the quality of people's jobs and workplaces, and in the amount of work that they do. Many people have more work than they can handle, while at the other end of the scale many have too little work. Such evidence as we have suggests that if AI decreased our hours of work by say, 20%, and productivity increased and the cost of living decreased such that we were not economically worse off, then our quality of life would increase rather than decrease. Such an outcome would seem idyllic to many New Zealanders, but of course much rests on whether such scenarios are likely to come about.

C Predicting changes in jobs and work in New Zealand

There are two types of predictions made about the effects of artificial intelligence on jobs and work. The first rests on the historical analysis of large-scale changes in production, often described as industrial revolutions. The second much more detailed analysis predicts future numbers of employees required for particular types of work in the near to medium term. We begin with the history.

Learning from the history of innovation

AI is often touted as the fourth industrial revolution (see for example Brynjolfsson and McAfee, 2014; Schwab, 201 6). If the analogy is a good one, the news is not great. Most workers at the beginning of the first industrial revolution didn't live long enough to enjoy its benefits. It spanned the eighteenth and the first half of the nineteenth century (Mokyr, 2009) and for much of that time workers experienced displacement into new and poorly regulated workplaces. Evidence from health records suggests that most Britons became steadily poorer for at least the first half of what was a very long period of innovation and displacement. Despite the industrial boom, the average height of English men decreased by 1.6 centimetres over each decade of the eighteenth century (Komlos and Küchenhoff, 2012). Real wages couldn't keep up with the price of food until well into the nineteenth century (Allen, 2009).

But there are important disanalogies between the coming AI revolution and the first industrial revolution. This time, the scale will probably be much greater and the pace will probably be much faster (Dobbs et al, 2015). The social impact of the AI revolution will be different in kind. The first industrial revolution saw most of the population moved from agrarian work into a relatively small range of almost wholly unregulated jobs in cities that were quickly bursting at the seams. This time round, we are at least starting out with established labour laws. These will need to be adapted, but not invented (for discussion of the adequacy of existing employment law see Chapter 3). While there will be dislocation, that will mostly involve retraining and movement into new jobs, although we might yet see long-term acceptance of working from home (perhaps with inner city real estate, redeveloped to meet New Zealand's housing needs). Social and political conditions have also come a long way. In the first industrial revolution, most of the populace was not enfranchised and ordinary people had no power to scrutinise or influence the actions of large companies. In the United Kingdom, unions were illegal under the Masters and Servants Act until 1871.

We also note that AI is a very particular type of technology, known as a *general purpose technology* (Lipsey et al. 2005). These are characterised by "pervasiveness, inherent potential for technical improvements, and 'innovational complementarities', giving rise to increasing returns-to-scale" (Bresnahan and Trajtenberg 1995). While there is debate about exactly which technologies count, uncontroversial examples include the steam engine, railways, the internal combustion engine, electricity, computers, and the internet. In the long run, general purpose technologies increase growth and raise standards of living. However, in the short term, they are often a significant drag on economies. A typical example is discussed by Winton (2019, p.3)

> While electrification did cause a discontinuous improvement in productivity across every manufacturing sub-sector in the 1920s, for example, it first placed a drag on the economy for more than a decade as businesses were forced to restructure to capitalize on the new paradigm. To access the promised productivity gains, they had to sunset or destroy old infrastructure and invest in the new world at a low-yield until the GPT

> reached critical mass… Electricity only provided dramatic productivity improvements in factories when the factories were built, from the ground up, to take advantage of the unique properties of electric power. Early attempts to electrify factories involved simply replacing a central steam-engine driving a crankshaft with an electric motor driving the same. All of the factory equipment remained belt-driven by that same central shaft, and so had to be clustered uncomfortably to minimize loss due to the mechanical transmission of power. Electricity can be transmitted nearly losslessly (on a relative basis); rather than being configured to minimize mechanical loss, ground-up electrically powered factories allowed machinery to spread out across the factory footprint to logically accommodate employee workspace and process through-flows.

So, adaptation to the open-ended and potentially far-reaching effects of artificial intelligence will be economically challenging for New Zealand. Moreover, AI (including robotics) is just one of a group of closely related and overlapping general purpose technologies that will radically change life and work in New Zealand in the coming decades. Others include DNA sequencing and editing, blockchain, and the wholesale move away from fossil fuels and into battery storage. At the intersection of these major new technologies is a host of applications that offer exciting advances in goods and services such as internet of things, 3D printing, and autonomous mobility. But each of these will challenge industries that rely on 'legacy technologies' such as the internal combustion engine. The last time such a diverse group of general purpose technologies became mainstream was at the beginning of the twentieth century (the automobile, telephony, electricity, and the production line). These greatly increased in the quality of life of people around the world. There is no reason to think that the current batch of general purpose technologies won't be similarly beneficial in the medium term, but near-term financial challenges are likely to encourage innovations that decrease labour costs.

In an influential paper called "The wrong kind of AI? Artificial intelligence and the future of labour demand", Daron Acemoglu and Pascual Restrepo (2019) distinguish between *replacing and enabling innovation*. In the computing world, these strategies were originally called artificial intelligence and intelligence augmentation (IA) although the latter term has fallen out of favour with both strategies now thought of as forms of AI. These ideas reflect a division that has run through the development of AI from its very beginning, between those aiming to reproduce human abilities and those aiming to produce 'tools' for people, that extend their abilities.[10] Both these approaches increase productivity. The former does so by decreasing labour costs and the latter by enhancing the productivity of individual workers. That said, while some technologies are designed to enhance (the personal computer) and others to replace (airport check-in kiosks), the difference between a technology enhancing and replacing is often a commercial decision between cutting costs and increasing production, rather than a design choice. In reality, much AI is both enhancing and replacing. Part of New Zealand's AI challenge is to encourage enabling rather than replacing innovation. This will be difficult. Maintaining staffing while enhancing the productivity of workers increases production which means finding new customers in a difficult post-COVID economic environment. This is a higher risk strategy than simply increasing profits by decreasing staffing.

AI and the economy

This report is not primarily on economics, but the history of innovation both here and overseas, can provides useful perspective on our current circumstances. In 1987, Nobel Prize winning economist Robert Solow famously commented that "computers are everywhere but in the productivity statistics" (Turner, 2018). But Solow's paradox shouldn't surprise us. New technology poorly implemented replaces workers, driving them into low wage, low productivity jobs. But even technology designed to enable workers can be expensive to implement in the short term, particularly in legacy companies.[11] There are many voices in New Zealand encouraging accelerated adoption of AI—see for example Kinley Salmon's Jobs, Robots & Us (2019), the AI Forum's 2018 report *Artificial intelligence: Shaping a future New Zealand*, and the New Zealand Productivity Commissions 2020 report *Technological Change and the Future of work*. All this work is motivated by the possibility of long-term enhancement to the quality of life of New Zealanders and by short-term commercial concerns about New Zealand companies falling behind international competitors. These are

10 For a thorough analysis of this history, see Markoff 2016.
11 See for example the explanation of Solow's Paradox in Diane Coyle's excellent *GDP: A Brief but Affectionate History* (2015, pp. 83-85).

sensible considerations, but New Zealand needs to think carefully about the near-term costs to quality of life in an accelerated AI revolution. Faster adoption means less time to get to the medium-term benefits, but also less time to adapt for our communities, our workforce and our regulatory framework. A rush to adopt could make it more difficult to monitor and influence New Zealand's progress. This would make it harder to incentivise enabling innovation and discourage short-sighted adoption of the sort that shores up profits at the expense of employment opportunities.

While the profits of the first industrial revolution landed extremely unevenly across British society, they did at least mostly land in Britain. All New Zealanders are familiar with the household names of the AI revolution (so far), the so-called FAANG companies (FaceBook, Amazon, Apple, Netflix, and Google). None of these companies are based in New Zealand. We would have to go a long way down the food chain of tech companies to find one that was. Despite their importance to New Zealanders, the FAANG companies generate few tax dollars and few jobs in this country. Many are near-monopolies with all but Netflix facing antitrust investigations in the US at the time of writing. All are, if not first movers, then at least early movers in market segments with high barriers to entry. These companies (and others such as Microsoft and Tesla) have built up data assets so great, that it would be very difficult for New Zealand companies to compete against them. This isn't to say that New Zealand can't profit from AI. It is, after all, a general purpose technology and New Zealanders are successfully deploying it in a wide range of contexts. But it is to say that a great deal of the AI that New Zealanders use adds very little to our GDP.

We are at a stage in the deployment of the new general purpose technologies where inequality is particularly problematic. Those who can afford to invest in AI-driven companies are growing wealthier while both blue-and-white collar workers are having to adapt to work in disrupted industries. To be sure, there are good new jobs being created, but for many (particularly young) workers, AI innovations promise dislocation, lower wages and increased precarity in the short term. As a recent UK report from the Fabian Society (2020, p. 18) puts it—"there is a real risk that technology take-up will further polarise the labour market so that those who already have least end up losing most".

At the same time uptake of many of these new technologies is not yet sufficiently widespread to experience a *productivity effect* (Acemoglu and Restrepo, 2019), significantly driving down the cost of living. To give a concrete example, fully autonomous vehicles are widely predicted to arrive within the next five year. Germany is looking to have a permit system for such vehicles in place by 2022 (Beedham, 2020). When the technology is mature and widely implemented, it will drive down the cost of many goods and services, but that payoff may be a fair way off. To get there, transport intensive businesses will likely have to replace or significantly upgrade their vehicle fleets. We will likely have to adapt urban planning as well as industries like insurance. We will need to update regulatory frameworks. And of course, we will have to transition those for whom driving is a job, into new types of work

But the news is not all bad. Many of these issues can be addressed. Indeed, the aim of this report is to encourage New Zealand workers and employers to think carefully about how we can secure enabling implementation of AI which will successfully serve the needs of New Zealanders. We encourage public discussion about the likely benefits AI will bring and about ways that New Zealand might adapt work practices and distribute the work and wealth of New Zealanders in the coming decades. Useful starting points these such discussions are set out in Section D below.

In successfully adapting to AI, it would be helpful to know which jobs and industries are likely to be most affected by its implementation.

AI and jobs

Most New Zealanders will have seen cautionary news articles about the risks that a robot will take their job or the jobs of their children. This (often not very high-quality debate) was fuelled by a 2013 study by Carl Frey and Michael Osborne, who found that 47% of total US employment had a high probability of computerisation from machine learning and mobile robotics. But academic study of the effects of AI and robotics on jobs has proved challenging, resulting in a bewildering variety of predictions. Other studies put the percentage of jobs affected as low as 14% (Nedelkoska and Quintini, 2018) or even 5% (Manyika *et al*, 2017). Recent analyses remain very divided; for instance, Willcocks (2020) is skeptical about large-scale job losses, while a report

just out from the Fabian Society (2020) highlights the threat of automation-induced job losses. Predicting the effects of AI across the great variety of jobs in a modern economy is a difficult task and the radically different results produced in recent studies depend largely on the authors' methodological choices.

Frey and Osborne modelled 702 precisely defined job categories, finding 47% ripe for near-term automation. Crucially, they focused on whole jobs. So, if some part of a job was estimated to be at risk of automation, the whole job was assumed to be at risk. While this methodology does detect jobs that are ripe for disruption, the idea that all such jobs will disappear seems unwarranted. The fact that the typical bank clerk's job description today might only faintly resemble what it was in, say, 1980, doesn't mean there are no more bank clerks (Zerilli et al. 2021, ch. 9). Unfortunately, this idea of job extinction has become pervasive in recent public discussion, driven by books with titles such as *Jobpocalypse: The end of human jobs and how robots will replace them* (Way 2013), *The robots are coming* (Oppenheimer, A., and E. Fitz, 2019), *Robot-proof yourself* (Schenker, 2017), *Alexa is stealing your job* (Scharf, 2019), *Help! A robot took my Job!* (Murad, 2017), and many many more. In reality, the more common effect of technological innovation is to change the nature of jobs (like the bank clerks) or to change the number of people employed in particular jobs. The automobile didn't cause the extinction of blacksmithing as a job, but it did massively diminish the demand for such work. In light of the tendency of new technology to redraw the boundaries of jobs, many studies conducted after Frey and Osborne's have focused less on jobs and more on tasks. An OECD task-based study (Artnz *et al*, 2016) found that only 10% of all jobs in the UK and 9% in the US were *fully* 'automatable'. We also note that partial automation can have profound effects on workers. Early motor cars were built by teams of mechanics. When Henry Ford moved the production of his Model T onto a production line, he not only sped up production, he also greatly decreased the need for specialised labour. Workers lost the job satisfaction, status, and bargaining power that went with being a trained mechanic.

Other methodological choices that can have important effects on the results of such studies include: how they interpret automation and computerisation; timescale; judgements about likely scientific advances; and various social, economic and demographic assumptions about factors such as levels of migration and macroeconomic forces (Zerilli et al, 2021). Even if we can agree on appropriate methodological assumptions, it is becoming increasingly apparent that making accurate fine-grained predictions about individual jobs requires high quality data that, for the most part, we don't have. In "Toward understanding the impact of artificial intelligence on labor" Frank et al. (2019) address the data barriers to successful prediction:

> These barriers include the lack of high-quality data about the nature of work (e.g., the dynamic requirements of occupations), lack of empirically informed models of key microlevel processes (e.g., skill substitution and human–machine complementarity), and insufficient understanding of how cognitive technologies interact with broader economic dynamics and institutional mechanisms (e.g., urban migration and international trade policy).

All these methodological and epistemic issues imply that we should treat with great caution any predictions about numbers of jobs that will disappear or be disrupted, not because there is no risk, but because there is inevitably a high degree of uncertainty about the impact of AI on particular jobs. These limitations also affect predictions about the creation of new jobs.

Those in favour of accelerating New Zealand's adoption of AI argue that it will create as many jobs as it destroys, and that many of the new jobs will be desirable high-tech jobs. We agree that the likelihood of large-scale technological unemployment is low. The financial and psychological costs of unemployment in countries like New Zealand are so high that most people will accept even very poor quality, low-paid work over unemployment. So, there is a socially-fuelled, market-driven incentive for unemployment to remain low. But artificial intelligence enables both high value, high status knowledge work and low value, low status jobs such as training algorithms and working in e-Commerce 'fulfilment centres' (Prassl, 2018) . Being able to predict the rate of churn (as legacy jobs are replaced by newly created AI-based jobs), likely demand for as yet uninvented products and services, as well as the proportion of new jobs that will be desirable and well paid, is an even harder task than predicting rates of disruption in existing jobs.

It's frustrating to be unable to offer firmer predictions. But we believe there is a genuine and intractable lack of certainty here, and those who are making stronger predictions are misrepresenting the accuracy with which the future can be read in this matter. We wouldn't expect people at the beginning of the twentieth century to be accurate in predicting the changes in jobs and work that would flow from widespread adoption of electricity. This sort of inability to make detailed medium-term predictions about effects on jobs and work is typical of general purpose technologies. So, in this report, much of our focus is on analysing the strengths and weaknesses of the sort of AI we expect to see in the near-term, considering how such AI might change jobs and work, as well as analysing opportunities and risks in domains in which AI is already being implemented. In the longer term, we believe the most practical path is to anticipate various alternative *scenarios* about how AI develops, so policymakers can prepare responses to each of these. It's to this that we now turn.

D. Large-scale adaptation scenarios

Just as we can try to predict the effects of AI on jobs and work at different scales, we can also think of adaptation at different scales. Much of this report focuses on specific opportunities and problems in workplaces that are already implementing AI. But for the remainder of this chapter, we look at adaptations that New Zealand may want to consider in the face of large-scale AI-driven trends in New Zealand's workplaces and economy.

The challenge of predicting the progress in AI research and its commercial applications makes it imperative that New Zealand policymakers prepare for a range of possible futures. To understand the effects of AI on jobs and work, we need to address five crucial questions, all of which are difficult to answer with certainty at this stage in the AI revolution, and most of which are largely outside the New Zealand government's control:

Where will the profits land? —As noted above, the household names of AI are US companies that do little to bolster New Zealand's GDP. But we do have excellent AI implementation here. In our earlier report, *Government use of artificial intelligence in New Zealand* (Gavaghan et al., 2019), we documented this country's world-leading use of predictive risk modelling in its provision of government services. Soul Machines is an example of a New Zealand company that is world-leading in the realism and interactivity of its AI agents. Despite these successes, it is uncertain whether AI will be a technology that we primarily profit from directly or one that we primarily import—which might nonetheless benefit New Zealand and New Zealanders. We make very few cars in New Zealand but our productivity and quality of life is greatly enhanced by our ability to import them. Similarly, services like internet search also benefit New Zealand's economy even though companies like Google employ few New Zealanders and pay little tax here.

How will AI affect incomes, inequality, and the availability of good jobs in New Zealand? —The mix of replacing and enabling innovation that AI brings to New Zealand will depend on difficult-to-predict future scientific innovation. It will also depend on commercial imperatives such as the costs associated with enabling and replacing innovation, as well as our ability to expand existing markets and move into new ones. It is essential for this country, not just that we retain adequate employment, but also that we retain and create new high value, well paid jobs.

How will AI affect the cost of living? —As with previous general purpose technologies, AI is likely to drive down the cost of essential goods and services but we have no real way of knowing by how much.

How will AI affect New Zealand's balance of trade? —Our distance from external markets has always been a challenge to exporters. It will continue to be so for physical goods that are AI-enhanced but not for AI-based services that are 'delivered' internationally via the internet. That said, the internet is a crowded marketplace and a great deal of AI-based innovation is being driven by a relatively small number of very large, data-rich, primarily US-based companies. Whether such near monopolies are broken up (as happened with previous general purpose technologies such as telephony), may have an important impact on New Zealand.

How will New Zealanders respond to AI? —Will we feel safe driving in robotic taxis? Will we feel comfortable with AI performing some or even many of the functions now provided by doctors, lawyers, educators, and other professionals?

It is too early to give definitive answers to these questions. The effects of AI on our economy and on our work and income will continue to change as the technology matures, as new applications are invented, and as consumers and other stakeholders discover the pros and cons of AI in the workplace.

Because New Zealand is not in a position to accurately predict the short-and medium-term effects of AI on jobs and work, it is essential that our policymakers analyse and prepare for a variety of scenarios. In reality, of course, New Zealand will experience some mix of the following scenarios. Nonetheless, it is useful to consider each as a type of outcome for which we should be prepared.

SCENARIO ONE: "ENABLING"

The dominant effect of AI will be to enhance the productivity of New Zealand workers. There is good availability of work. Most workers move into higher value jobs. New Zealanders experience higher quality of life due to some mix of higher income, lower cost of living, and better working and living conditions.

SCENARIO TWO: "REPLACING ONSHORE"

The dominant effect of AI will be to decrease the productivity of New Zealand workers as many are replaced by artificially intelligent systems and displaced into lower value, lower income work. In this scenario significant profits from the AI revolution land here as New Zealand companies successfully deploy AI and robotics, increasing profit by lowering the cost of production of goods and services.

SCENARIO THREE: "REPLACING OFFSHORE"

The dominant effect of AI will be to decrease the productivity of New Zealand workers as many are replaced by artificially intelligent systems and displaced into lower value, lower income work. In this scenario most of the profits from the AI revolution land offshore as many New Zealand companies are outcompeted by international companies with better access to data and capital.

This is work that needs to be done now as some plausible strategies would be difficult to deploy and/or may take a long time to come into effect. In proposing these scenarios, we do not mean to suggest that New Zealand is completely at the mercy of a process of technological development driven by international research and commerce. Government might, for example, sponsor the development of New Zealand based AI-driven services, as there have recently been calls for Australia to develop its own.

RECOMMENDATION 2: Government should develop and, where appropriate, implement strategies responding to three possible futures that AI could bring about (enabling, replacing onshore and replacing offshore).

Skills and education

Although there is much we don't know about New Zealand's AI-enhanced future, knowing how much we don't know is itself useful. It is common for those predicting the future of jobs and work to encourage widespread teaching of computing skills in secondary and tertiary education (Walsh *et al*, 2019). We agree these are important skills that can lead to good jobs, but it is difficult to know how many graduates will be required as AI-based technologies mature in the coming decades. It's possible that computer programming will be one of the human jobs that AI systems become good at. A sobering pointer in this direction is OpenAI's Generative Pre-trained Transformer 3 (aka GPT3), released early in 2020, which is the most sophisticated model of natural language ever produced. A programmer's task is to convert a natural language description of a desired program into a working piece of code. Among many other surprising capabilities, GPT3 is capable of doing this, at least at a first level of approximation (Heaven, 2020). So, while computer science is fundamental to the AI revolution, we may not need large numbers of coders to take advantage of the technology. This should not come as a surprise. Decreasing demand for technological expertise is typical of maturing general purpose technologies. By analogy, early in the twentieth century we might have

recommended that young people train as car designers or motor mechanics. But most of the jobs that owe their existence to the automobile, have little to do with the construction or maintenance of cars and trucks. New Zealand is extremely dependent on motor vehicles but very few of us build them or maintain them, and that number has steadily decreased as the technology and methods for its manufacture have matured (Womack et al, 1991). So, while we certainly need a good supply of ICT graduates, most young people need to know how to use AI rather than how to make it. A Royal Bank of Canada Report—*Humans wanted: how Canadian youth can thrive in the age of disruption* (2018)—analysed the projected skills demand for human workers for all occupations in the face of near-term technological disruption. The results are summarised in Figure 3. Although acknowledging that skills in computing and data science will be essential and valuable in the coming AI revolution, this report concludes that the important skills for the great majority of workers will be those that fit them for tasks to which AI is poorly suited. In order of descending importance (where the most important skills are those required by the largest number of jobs):

1. Active listening
2. Speaking
3. Critical thinking
4. Reading comprehension
5. Monitoring
6. Social perceptiveness
7. Coordination
8. Time management
9. Judgement and decision making
10. Active learning
11. Service orientation
12. Complex problem solving
13. Writing
14. Instructing
15. Persuasion
16. Learning strategies
17. Negotiation
18. Management of personal resources
19. Mathematics
20. Systems analysis
21. Systems evaluation
22. Operation Monitoring
23. Quality control analysis
24. Operations analysis
25. Operation and control
26. Management of material resources
27. Management of financial resources
28. Technology design
29. Programming
30. Troubleshooting
31. Science
32. Equipment selection
33. Equipment maintenance
34. Repairing
35. Installation

Figure 3. Skills most in demand after near-term technological disruption. After RBC Humans wanted: how Canadian youth can thrive in the age of disruption (2018, p. 12).

Automation has always imposed a division of labour on workforces as people are left to do the work that machines can't (or the work for which automation is uneconomic). AI will be no different. So, will analyses like this one tell us which skills our education system ought to be focusing on? In the short-term—yes. But schools and tertiary institutions train young people for life. While we applaud research into career transitions assisted by mechanisms like micro-credentialing, secondary and tertiary education directly post-school will continue to be the most important source of educational preparation for New Zealanders' life and work. So, we should be careful about long term predictions of the skills our young people will need for an AI-enhanced world. The limitations to predicting future science and future commerce also apply at the level of skills. Advances in affective computing, for example, are already making AI much more effective at a range of 'face-to-face' competencies that we currently think of as distinctly human skills. For instance, the 'digital people' produced by the New Zealand company Soul Machines are able to recognise emotions in the user, and to communicate emotions themselves, through facial expressions and gestures. These abilities may well find application in domains like medicine, education and counselling.

Finally, we note that if AI decreases the amount of paid work that New Zealanders have to do, we would benefit from our education system better preparing people to maximise the value of their leisure time in fulfilling and socially productive activities. And as noted in Section B, we have much to learn from Māori and Pasifika groups in New Zealand about how to develop a culture that places family, community and social engagement at the centre of life, rather than paid employment.

RECOMMENDATION 3: In the face of an uncertain future, New Zealand must discourage over-specialisation in education. Education and training at all levels must equip young New Zealanders with a broad array of the skills and expertise required for an AI driven world (as set out in Figure 3).

In the face of so much uncertainty, the best thing we can recommend for young New Zealanders is a very broad skill-set. This will require rethinking university education in Aotearoa which has followed international trends in recent decades towards hyper-specialisation, particularly in STEM subjects. In future, we might broaden the skill-sets of graduates by making professional training in subjects like Law, Medicine, Engineering, Dentistry and Pharmacy into graduate degrees, as they are in many other countries. In legal education, double degrees are already common, providing New Zealand students with a much broader base of learning than the UK, where it's rare to combine law with another subject. Courses introducing students to Māori and Pasifika value systems will also have a useful role to play. We also note that the University of Otago has recently developed new degrees designed to promote the breadth of graduates (the Bachelor of Arts and Science; Bachelor of Commerce and Science, Bachelor of Arts and Commerce). The renaissance people who graduate with these degrees will be well suited to an uncertain and fast changing workplace.

Loss of income and increased inequality

We think artificial intelligence is an extremely important general purpose technology which will likely drive large-scale commercial innovation and enhance our GDP over the medium term. But, for all the reasons just stated, we have to take seriously the possibility that in the short term it will lead to large-scale disruption and even in the medium term, it could cause increasing inequality. We are particularly interested in two possible outcomes that might befall an AI-rich New Zealand in the coming decades:

1. Many or most New Zealanders end up with too little income,
2. Many or most New Zealanders end up with much less work.

We could experience both these effects at once, but they are not necessarily connected. If AI is genuinely labour saving, we could end up with similar levels of income, but less work to do. If the profits of the AI revolution are captured by a small portion of the population or by people outside New Zealand, many of us might have to work harder just to keep our current levels of income. We will address income in this section, and work in Section C of Chapter 4.

Nobody doubts that AI is extremely profitable. The AI-enhanced FAANG companies are amongst the fastest growing major companies in the world. The title of fastest probably goes to Tesla, which aims to be the first to offer 'full self-driving' electric cars at some point in 2021. Its share price has increased nearly 1000% on its value a year ago. Apple Computer is now the world's largest publicly listed company with its market capitalisation growing from one trillion which it reached in August of 2019 to more than two trillion at the time of writing. But, as noted above, few New Zealanders profit from these companies and their success comes at the expense of the legacy companies they are out-competing. So, what could New Zealand do if AI-driven companies leave many of us worse off?

Loss of income, due to underemployment, should be partly offset by decrease in the cost of living, but we cannot know when and by how much the cost of living will decrease. Winton (2019) predicts that depression of GDP due to effects of new general purpose technologies on legacy companies is likely to last less than a decade with gradual increase in efficiency leading to lower prices thereafter. But in such a scenario, even if AI doesn't

leave most of us poor in absolute terms, it could lead to spiralling inequality—harmful to individuals (Pickett *et al*, 2015), socially corrosive (Lancee and Werfhorst, 2011) and politically destabilising (Alesina and Perotti, 1996).

New Zealand already experiences significant inequality, which it has mitigated with a mix of taxation coupled with redistribution mechanisms such as Working for Families. But if AI were to greatly increase inequality, New Zealand would have a limited set of options. Recent decades have seen growth in real wages fall steadily behind growth in productivity in many developed countries. This decoupling in New Zealand has been amongst the most pronounced in the OECD countries (Schwellnus, 2019). Tackling the decline in wages relative to productivity will help decrease inequality, but it may not be enough to counterbalance a decline in the number of people in well-paid work. A further strategy would see New Zealand moving to increasingly aggressive forms of taxation, perhaps including a wealth tax (mechanisms for redistributing increased tax revenue are discussed below). These mechanisms could be successful if AI drives major productivity growth *in New Zealand* (our *replacement onshore* scenario)—perhaps aided by government curtailing the activities of international AI-driven companies (as Australia is proposing to do with Facebook) or even by sponsoring New Zealand-based competition, just as we have in banking (Australia is considering an Australia-based social media platform, see Meade, 2020).

Conversely, in our *replacement offshore* scenario, productivity growth remains low in this country and the spectacular profitability of AI continues to be a primarily offshore phenomenon. How might New Zealand respond?

While few of us work for the FAANG companies, all of us invest in at least some of them. Many of us own shares in them through our KiwiSaver portfolios. New Zealand as a whole also owns shares in some of in them through the New Zealand Superannuation Fund (New Zealand Superannuation Fund, 2019). Moreover, an increasing number of New Zealanders are investing directly in such companies due to the introduction in 2018 of two online share trading platforms, Sharesies (sharesies.co.nz) and Hatch (hatchinvest.nz). As they sell fractional shares at low commissions, even those on modest incomes can use them as an alternative to savings accounts which at the time of writing offer extremely low interest rates due to the economic effects of COVID19. They also allow individuals to purchase shares without having to go through an established brokerage. Hatch focuses on the US market and Sharesies, which has marketed New Zealand shares particularly to young New Zealanders (McBeth, 2018), has recently begun trading US equities. Both are seeing extraordinary growth in participation from ordinary New Zealanders for whom this is their first experience with retail investing (Hickey, 2020).

While all these forms of investment allow ordinary New Zealanders to benefit from the profits of AI companies based overseas, they are certainly not equal. The New Zealand Superannuation Fund has been remarkably successful. It stands at over 47 billion dollars at the time of writing with an average rate of return since its inception in 2003 of 9.6%. It is also fundamentally redistributive. While wealthier Kiwis pay more tax and so effectively invest more into the fund, its profits belong equally to all New Zealanders. Being individually based, KiwiSaver and retail investment portfolios are not redistributive in same way. Online retail investing is also higher risk. Because it bypasses normal avenues for investment advice, such portfolios seem more likely to lack diversification and research on a similar platform in the US (Robinhood) has shown users to be disproportionately invested in a small number of AI-enhanced stocks—Apple, Tesla, Amazon, and Microsoft (Elmerraji, 2020).

While New Zealanders are generally nervous of investments in shares (Hickey, 2020), most of us consider superannuation schemes (which primarily invest in shares) a tried and trusted method of saving for retirement. So, could New Zealand harness the power of financial markets as a means of adapting to a possible AI future in which significant numbers of us find ourselves with stagnating or even decreasing incomes? While some New Zealanders are taking a do-it-yourself approach to the problem via retail investing, it is unlikely this will become common among most of us, and it is likely to remain practically impossible for the poorest of us. While we could educate the public and develop schemes and incentives (as we have for retirement investing), an individual-based approach to investing in AI would inevitably leave the poorest New Zealanders with the lowest returns.

If the country as a whole were to invest to insulate ourselves against an AI-fuelled economic downturn, it would have to be through something like a sovereign wealth fund akin to the New Zealand Superannuation Fund. In the short term if would require expensive levels

of contribution—at the time of writing the New Zealand government has just suspended payments to the New Zealand Superannuation Fund due to the costs of dealing with COVID19. But in the medium term, such a strategy could help effectively subsidise the incomes of working age New Zealanders. Could it be the basis of a Universal Basic Income?

A Universal Basic Income for New Zealand?

A UBI addresses the problem of traditional welfare payments creating poverty traps due to them abating as recipients find part time work. There is no financial incentive for someone to take such work if the pay they receive will be deducted from a benefit they are currently paid. Because it is universal, a UBI also removes the stigma associated with receiving welfare along with the complex governmental machinery required for assessment, monitoring, and disbursing payments. While a UBI is usually conceived of as a universal monetary payment, there is no reason in principle why it could not consist of, or include, a set level of free access to an enhanced suite of public services, such as health, education, transport, and energy, some of which are already partly free in New Zealand.

Its main disadvantage is that paying a regular income to the whole population is extremely expensive (Morgan and Guthrie, 2011) although, as noted above, AI is likely drive down the basic costs of living that a UBI is meant to cover. As this discussion is hypothetical and future focused—and as we are not economists—we make no comment on the affordability of a UBI.

A further issue with a UBI is that some people find the idea of paying people not to work morally objectionable. There is nothing unjust in principle about paying people not to work provided payment is equitable. New Zealand voters have jealously guarded the New Zealand superannuation scheme, which is effectively a UBI that starts at 65. There have been periodic proposals to means test New Zealand superannuation. All have foundered, even though a means tested system would be cheaper and arguably better at alleviating poverty. The problem appears to be that New Zealanders think the current scheme is equitable in that we all pay for it and we all receive it. That said, people live for very different lengths of time in retirement and hence the scheme really only secures equity of opportunity not of outcome. So, *prima facie*, a UBI paid to the whole population, would be both inequitable and redistributive in the same way that New Zealand superannuation is.

A further argument against the UBI is that it would allow some people to work very little and these people would be worse off as work conveys dignity and/or is intrinsically virtuous. Paradoxically, such arguments often seem to claim both that work is a fundamentally valuable aspect of life and that, given even a small regular income, many people would no longer want to work. The great 20th century philosopher Bertrand Russell argued that the idea of the dignity of labour was important before the industrial revolution when agrarian economies required most people to work all the hours they could. But this is no longer necessary. The dignity of labour, he suggested, has since become an "empty falsehood" preached by the wealthy who "take care to remain undignified in this respect." (Russell, 1932).

Perhaps the most important issue for those promoting a UBI, is that the idea itself is open to a considerable amount of interpretation:

> Although UBI has a number of core definitional attributes (…), basic income is best seen as a family of schemes, with variation between them in terms of a number of crucial design features… The most crucial of these are arguably the level of payment, the way UBI is intended to interact with other benefits (i.e., whether it is intended to replace or run concurrently with them) and the wider constellation of labour market policies, and how it is funded. … These design features vary in line with the goals and objectives motivating basic income; in turn, different goals and objectives are prioritised according to the political preferences of different UBI supporters. UBI supporters come from a wide range of political perspectives, a consequence of the breadth of the range of arguments on which UBI proponents draw. … The juxtaposition is such that basic income is argued to defy conventional political labels; it is 'neither right nor left but forward'. A more nuanced assessment may be that whether it is 'right' or 'left' depends on the specifics of the scheme in question.
>
> Martinelli (2017, p.5)

A UBI would not ameliorate the effects of the *replacing* scenarios if it effectively subsidised poverty level wages, entrenched precarious work, or if it were made 'affordable' by dismantling important aspects of our

labour law or welfare provision, leaving our most vulnerable New Zealanders worse off. The International Labour Office (ILO) notes that a UBI should complement, rather than displace "the budget for core social security, health, education, active labour market policies and other crucial social services" (Piachaud, 2018).

Despite the objections and potential pitfalls, New Zealand could design a UBI specifically aimed at decreasing inequality and overwork, removing the stigma from state support and increasing the autonomy of New Zealanders, allowing them to better enjoy their lives. Such a scheme could be beneficial if either (a) AI is predominantly *enabling*, increasing New Zealand's income such that we no longer have to work 40-to-50-hour weeks or (b) AI is predominantly *replacing* such that work becomes more precarious and income becomes more unequal. Of course, in (b) much would depend on New Zealand being able to effectively harness the profits of the AI revolution to fund a UBI.

New Zealand clearly cannot currently afford to pay a UBI that would cover even basic living costs to all its citizens. But we might be able to apply a mechanism that has long been deployed in Germany and has now been exported to many other countries including Denmark, Sweden, Norway, Austria, Czech Republic, Italy and Japan (Connolly, 2020). In Germany, *Kurtzarbeit* (literally "short work") involves a reduction in the length of the work week with the state paying employees 60% of their normal income for the hours they no longer work. Originally used in 1910, *Kurtzarbeit* has been deployed as a short-term fix to tackle downturns such as the GFC and now the pandemic. It could be a useful mechanism to help share available high value jobs amongst New Zealanders, but it does have drawbacks. It is fundamentally designed to help people in work and so will not help to alleviate the precarity and stigma for those unable to work or unable to find work. Also, while it might prevent some New Zealanders being pushed out of well-paid jobs into poorly paid ones, for those who were forced into low paid work, *Kurtzarbeit* would do little to address the loss of income as its payment level is tied to the job its recipient currently does.

Whether or not we have a UBI, we are already paying to redress income inequality through mechanisms like Working for Families and New Zealand Superannuation. We think that New Zealand will at some point have to have a national conversation about the philosophical justification for such mechanisms.

Possible justifications include:

> ***Poverty eradication:*** This is the current setting. It assumes that distributive justice is generally secured successfully by free markets and wealth redistribution is only justified to alleviate extreme poverty.
>
> ***A set level of wealth redistribution:*** This assumes that distributive justice is ineffectively secured by free markets and so we should redistribute wealth to a degree that provides some level of compensation for the lack of fairness in labour markets.
>
> ***Wellbeing enhancement:*** Governments should act to secure levels of work and particularly amounts of income that will guarantee a good level of wellbeing for New Zealanders. This might also involve consideration of and amelioration of the high costs of high inequality. It might also justify the provision of basic income security allowing New Zealanders all to pursue activities and projects that will enhance our lives and those of the community.

Given that some forms of policy response to increasing inequality would involve costly, long term policy changes requiring a high degree of social license, we recommend that these discussions not be postponed.

E. Some choices for New Zealand about work and income

As noted above, long run technological progress has successfully produced many labour saving devices, but contrary to Keynes argument in *Economic Possibilities for our Grandchildren*, they have not resulted in most people working fewer hours. Although Keynes was writing in 1930 about the coming generations, in 1932 Bertrand Russell argued that British people could already be working a fraction of their normal work weeks. Over the five years of the First World War, most Britons had been occupied in fighting or in other wartime activities. Russell argued that this "showed conclusively that, by the scientific organisation of production, it is possible to keep modern populations in fair comfort on a small part of the working capacity of the modern world." In 2020, COVID19 has again demonstrated to us just how little of our economic activity is 'essential' for keeping people fed, clothed, housed and healthy. At the same time, the march of global warming is causing many to question the long-held assumption that the success of nations must rest

on the production of more and higher value goods and services, as reflected in higher GDP (Raworth, 2017).

We have argued that whether AI decreases the amount of work available to New Zealanders depends on how it is implemented. In the face of the risk of a decline in the availability of work, most countries including New Zealand have focused on keeping their populations in work. We suggest that countries like New Zealand should think about retaining current levels of work and current levels of income as two problems, not one. If AI really turns out to be *enabling*, increasing productivity and saving labour, we may have a choice to make. Should we continue to work the same amount and have more money to spend OR should we settle for our current level of income (more evenly distributed), allowing us to work less? A recent UK report argues that "The dividends of new workplace technology need to translate into higher earnings for all – or into shorter working hours, if that's what people prefer" (Fabian Society, 2020). We note that these two outcomes are not mutually exclusive, but also that New Zealanders may not have a free choice in the matter. Whether most of us will be able to continue to work roughly five-day weeks may depend on whether affected sectors of our economy are able to absorb considerably higher levels of production—we could clearly benefit from much high productivity in house building for example, but perhaps not in sectors such as insurance. So, we should consider changing the amount we work in response to the deployment of AI for several reasons. There may be less work available even in scenarios where AI is enhancing for workers. If AI increases productivity, New Zealanders may prefer to work less. All else being equal, the wellbeing of New Zealanders might be enhanced by working fewer hours.

While working less might seem counter-intuitive in a consumption-focussed economy currently experiencing economic headwinds due to a pandemic, AI will be with us for a long time to come and it will demand adaptation in the way we live and work. So, we suggest that even in difficult times, it is important to ask what sort of working life we would eventually like New Zealanders to have.

We note that there is already a large variation amongst OECD countries in the number of hours per year that workers actually spend on their jobs. New Zealanders average a relatively high 1779. Interestingly, there is a negative correlation between the number of hours worked and the wealth of OECD nations (OECD, 2016). The largest average number of hours worked are in Mexico (2250) South Korea (2070), Greece (2035), India (1980) and Chile (1970), while the fewest are in France (1472), the Netherlands (1430), Norway (1424), Denmark (1410) and Germany (1363). Correlation is, of course, not causation. These figures do not magically show that if New Zealanders worked less, that would cause us to earn more. But they do demonstrate that successful countries 'spend' their higher GDPs on allowing their citizens enhanced wellbeing through working fewer hours.

Small scale experiments with a four-day work week, both here and overseas, strongly suggest that it increases quality of life. Surprisingly, a trial by Guardian Life New Zealand found that dropping to a four-day week had no negative effect on productivity (Barnes, 2020). The company attributes this fact to greater enthusiasm and wellbeing in its staff as well as studies showing that workers in non-operational roles tend to spend significant parts of their working days in activities that do not enhance productivity (Barnes, 2020, p. 17). A trial by Microsoft Japan found a productivity increase of 40% (Eadicicco, 2019). No doubt, some of the effects of such a move across the board would be sensitive to details of particular industries and workplaces, but there is increasing acceptance that decreasing the work week will benefit staff recruitment and retention. Global consumer goods company Unilever will use its New Zealand staff to trial a four-day working week, at full pay in 2021. If successful, it plans to extend the four-day working week to its 165,000 employees around the world (Parker, 2020). For an extended argument in favour of a four-day work week in New Zealand see Barnes (2020). If in future, AI decreased the number of high value jobs in New Zealand, shortening the work week would be an effective way of sharing that work more equitably amongst us. Interestingly, this mirrors the argument made by the Labour Government in power when New Zealand instituted the forty-hour week thought the Industrial Conciliation Amendment Act 1936 and the Factories Amendment Act 1936. In the aftermath of the great depression the Government promoted a shorter working week as a means of increasing employment opportunities (Burrows, 2020).

We do not pretend that decreasing the working hours of New Zealanders would be simple. Some industries would find it much more challenging than others. Even

so, if New Zealand is really to take wellbeing seriously, we must investigate the effects of working hours and of the unequal distribution of work (as opposed to income) amongst New Zealanders. Sharing the benefits of new technologies is often thought of in terms of taxation, but it may yet be just as much about sharing work as about sharing wealth. We applaud the Prime Minister's recently encouragement for more companies to consider experimenting with a shorter work week. But we note that if increasing use of AI exacerbates inequality in work distribution, mechanisms like a shorter work week would have to be mandated by Government (just as they were in the 1930s). In a volatile and competitive economic environment, allowing individual employers to decide about the introduction of such measures is unlikely to achieve widespread change and might make inequality between different types of work worse.

How should New Zealand Respond?

We have said that, beyond the short-term, it is very difficult to predict the effects of AI adoption on particular jobs. However, we can be certain that New Zealand will achieve some benefits from AI. The costs of some goods and services will decrease as productivity increases in sectors of the economy that manage to harness AI effectively. This will push downwards on the cost of living and upwards on GDP. This will of course just be one factor that influences our productivity and living standards in the coming decades. AI will also contribute to the invention of new goods and services and to the democratisation of access to existing goods and services. These changes will be accompanied by the creation of new types of work and new ways of working. All these changes will be discussed at length in Chapters 3 and 4.

We can also be certain that the rise of AI will have some costs for workers and for the country as a whole. Many of these are discussed in detail in later chapters but, at a big picture level, we know that the jobs of some New Zealanders will disappear due to automation. Although AI will help to create the new jobs, it is again very difficult to predict how many new jobs will be created and what proportion of those will be high value, well paid jobs. It has been suggested (Turner, 2018) that as AI becomes capable of performing a growing array of complex tasks (driving, selling insurance, detecting cancerous cells in biopsies…), humans will increasingly be displaced into low value work. Whether or not we can avoid the resulting inequality depends on our ability to create new types of work which rest on skills that humans can perform and AI cannot. This offsetting of automation has tended to happen in the past but our sample size of previous industrial revolutions is too small for us to be sure that it will again save the day. Combatting such an increase in inequality might be challenging for New Zealand if the profits from the AI revolution disproportionately accrue to large data-rich offshore entities such as the FAANG companies.

So, it is essential that New Zealand is proactive in ensuring that we maximise the benefits and minimise the costs of the AI revolution. New Zealand already has companies and institutions that are deploying AI very effectively, but we could do more. If some of the profits from the AI revolution are to land offshore, we can at least share in those profits by increasing our investment through New Zealand's sovereign wealth fund. We should also invest in competing effectively with international AI-based companies. We might do this by developing government sponsored AI-based companies, just as New Zealand developed KiwiBank to compete with Australian banks that dominated the New Zealand market. Developing a distinctly New Zealand social media platform could have many social benefits over and above decreasing the amount of advertising revenue that flows to Facebook, Twitter etc. We might also revise KiwiSaver, enabling and incentivising investment in companies that deploy AI responsibly and effectively.

RECOMMENDATION 4: New Zealand should enhance its sovereign wealth fund so as to better profit from the success of large international AI-driven businesses. At the same time, we should be proactive in finding and investing in AI-focused niches in which New Zealand companies can be successful — particularly in social networking, where local products may bring other advantages.

Over and above harnessing AI to increase the wealth of New Zealand, we must consider its effects on our society. Increasing inequality is a real risk. We should also take seriously the question raised above—given

that AI is essentially labour-saving, do we want to use it to earn more or to work less? We have argued above that inequality of work and work strain are serious problems in the New Zealand workplace. Overwork is the only measure in the Human Development Index on which we are significantly lagging other developed countries. Overwork takes a real toll on the health of New Zealanders, especially mental health. It is also a cost to our businesses resulting in burnout, absenteeism and low morale.

We have no doubt that a well-designed universal basic income would enhance the wellbeing of New Zealanders, but such schemes are extremely expensive. It would be very difficult to achieve the level of political consensus required to make such a major political and economic change. Even if we could successfully budget for a UBI, we doubt that enough New Zealanders would support such a great expense, particularly in light of the huge cost of our response to the COVID pandemic. Something like a UBI might well be part of New Zealand's future, but now is not the time.

A more viable option would be to decrease the length New Zealand's working week as AI makes further inroads into the labour market. Business-led experiments in a four-day week both here and overseas show that decreasing the amount of time people spend at work need not decrease productivity and that it does result in healthier, happier, more resilient and more enthusiastic staff. We noted that the Prime Minister has recently encouraged businesses to consider decreasing work hours, but we also note that voluntary business-led change is likely to be slow and only partially successful. It may even exacerbate inequality in the amounts of work that New Zealanders are required to do. Ultimately, if New Zealand as a whole is to benefit from the labour-saving nature of AI, Government will have to lead. Those in operational jobs (e.g., nursing or bus driving) could not decrease their hours without a corresponding drop in their productivity. To achieve equity in a reduction in the working hours of New Zealanders, Government would have to subsidise decreased hours for operational workers. This is effectively what European governments do when they deploy *Kurzarbeit* as a means of stimulating their economies. While this would be a significant cost to Government, it would be much less that the cost of a UBI and it would be offset by reducing the social and health costs of our current level of overwork. Most importantly, such a Government subsidy would subsidise the creation of new jobs across the operational workforce. In the event that AI causes a decrease in the number of desirable, well paid jobs, decreasing our work week would increase the number of New Zealanders employed across a wide variety of operational occupations.

RECOMMENDATION 5: As AI advances into areas of human work New Zealand should consider decreasing the length of the working week as a means of securing the labour-saving benefits of AI and robotics, at the same time, stimulating the economy, better sharing high-value work, and enhancing the lives of New Zealanders and the communities in which we live.

Ultimately, there are huge benefits to be gained for our society, our communities, for Kiwis with school age children, or looking after those less abled. Harnessing the power of AI in this way could make New Zealand an even better place in which to live and work.

This chapter might appear to be very negative in tone, but there are plenty of voices telling us that AI is going to be a bonanza. New Zealand must consider downside risks as well as upside benefits. We also note that mechanisms like a UBI, *Kurtzarbeit*, or a shorter work week, could be beneficial for New Zealanders. They could decrease stress and mental illness, enhance family life, allow more people to actively engage in caring for the very young, the very old, and for the environment. If New Zealanders were better and more broadly educated (for example in the humanities disciplines) that would help them to flourish as human beings, and we could see a great increase in wellbeing. We might also see New Zealand becoming a more community-centred culture, reflecting the great strengths of Māori and Pasifika life.

3. AI AND THE EMPLOYMENT RELATIONSHIP

What will it be like to work alongside – or even under – AI algorithms and robots? While much attention has been paid to attempts to predict how many jobs will be lost to the 'fourth industrial revolution', considerably less attention has thus far been paid to questions of how working lives will be changed by them. What sort of benefits and harms might result from those changes? And what sorts of regulatory steps could be taken to maximise the benefits and mitigate the risks?

It's to these questions that we turn in this chapter. AI, and related technologies like advanced robotics, are likely to have important impacts throughout the employment life-cycle.[12] The approach we adopt here attempts to follow that life-cycle, considering some of the main implications at every stage. We begin, in fact, with considerations that arise even before the employment relationship begins: with the process of algorithmic recruitment. Our approach tracks the recruitment process from targeting of job adverts, to shortlisting, and finally to interviewing. As we show, concerns around algorithmic recruitment have already resulted in legal developments in some jurisdictions.

We then move on to look at how the workplace itself is likely to be affected, and specifically, to what has been called algorithmic management. Of course, this will vary substantially between different contexts, with some workers finding themselves more affected than others. Although some of what we have to say is likely to have more general effect, we focus on a particular part of the workforce that has already been impacted a great deal by these sorts of technologies: the so-called 'gig economy'.

Next, we look at two other categories of concerns about the use of AI and related technologies in the workplace. Technologically enabled workplace surveillance is an issue that has received increasing attention during the shift to working from home during the Covid crisis, but it has been a growing concern among trade unions and those concerned with workers' rights generally for many years. Health and safety considerations are also important for these purposes, perhaps especially when the technologies are embodied in the form of robots, driverless vehicles and the like.

We conclude by looking at the end of the employment life-cycle: in particular, to scenarios relating to 'technology-induced redundancy', where workers leave a company because technology takes their place, in some sense.

In addition to examining the likely impacts of these various aspects of employment, the chapter will consider some regulatory initiatives that have been suggested or implemented in the hope of mitigating some of the risks to workers that the discussed technologies might pose.

Benefits of AI technologies for workers

Before we turn to concerns around the use of AI in employment, we would be remiss not to acknowledge the many suggested benefits of its use in that context. Some of these relate to the prospect of higher wages. The RSA report refers to "evidence that AI and robotics could boost wages due to sizeable productivity gains, which will generate more absolute wealth that can be shared with workers." (Dellott and Wallace-Stevens, 2017) It should be noted, though, that the report tempers this optimism by considering the possibility that "[n]ew machines may deskill occupations, thereby lowering barriers to entry and reducing the bargaining power of workers in existing positions".

As we saw in the previous chapter, there is much speculation about the likely economic impacts of AI, and its likely impact on wages will depend significantly on some of the questions considered there. Suggested benefits for workers, though, are not confined to wages. The kind of work they will do, and the conditions under which they work, are also likely to change, and some have suggested this could be for the better. The World Economic Forum has considered one of the more optimistic scenarios in this regard:

> Automation technology can help remove the burden of repetitive administrative work and enable employees to focus on solving more complex issues while reducing the risk of error, allowing them to focus on value-added tasks.
>
> (WEF, 2018, p.10)

Perhaps offering more reassurance, a report for the International Labour Organisation echoed this sentiment: "By substituting human work with automated activities, technology can have liberating effects, especially if this substitution regards heavy, hazardous or repetitive work." (ILO, 2018, p. 5)

12 We use the term 'employment' in a generic sense, such that it includes those workers who may not technically be considered employees under NZ law.

The appeal of a future where AI takes over work that is dull, dirty and dangerous, leaving human workers to focus on better, more rewarding, higher-value roles and tasks, is easy to see. While most of this chapter will be spent considering the potential risks and harms from the use of AI in the workplace, it's important not to lose sight of its potential to improve our lives – and particularly, the lives of those for whom work doesn't offer a source of esteem and enjoyment, but monotony, indignity and danger. Potential, though, is not certainty; as with other impacts of AI on work, much depends on the sorts of choices our society will face about such technologies, the sorts of incentives and restrictions we create around them.

Some other potential benefits of AI in the employment relationship will be considered as they arise in particular contexts. For the most part, though, our focus will be on the concerns raised about these innovations.

Concerns about AI technologies in relation to workers

What, then, are the concerns that have been presented about the use of AI in the workplace? In the first phase of our project (Gavaghan et al., 2019), we identified six concerns about use of AI in the government sector:

- decision-making control;
- transparency and the related right to explanations;
- bias;
- informational privacy;
- questions of liability and the suggested 'responsibility gap'; and
- human autonomy.

To a large extent, these same concerns arise in relation to employment, albeit that the contexts, potential harms and available responses are inevitably different. As in the government context, there's a lot to unpack around these simple-sounding concepts. In the context of bias, for example, it's widely recognised that this can refer to different things, and that not all examples of bias are problematic. Rather, our concern should be with *unfair* bias.

But what counts as 'unfair' is also a complex and contested question. A recent report for AlgorithmWatch referred to "the plurality of fairness definitions" (Loi, 2020, p.23), while a report from the UK's Centre for Data Ethics and Innovation noted that "[n]otions of fairness are neither universal nor unambiguous, and they are often inconsistent with one another", going on to say that:

> Even in cases where fairness can be more precisely defined, it can still be challenging to capture all relevant aspects of fairness in a mathematical definition. In fact, the trade-offs between mathematical definitions demonstrate that a model cannot conform to all possible fairness definitions at the same time. Humans must choose which notions of fairness are appropriate for a particular algorithm, and they need to be willing to do so upfront when a model is built and a process is designed.
>
> (CDEI, 2020, p.29.) (See also ACAS, 2020, pp. 22-23)

There exists a substantial literature on the subjects of bias and fairness, both in the context of algorithms and AI (e.g. Eubanks, 2018; Corbett-Davis and Goel, 2018); and more generally. We have made our own modest contribution to it elsewhere (Gavaghan et al., (2019), pp.43; Zerilli et al, 2021, ch. 3), but don't intend to spend much more time on the philosophical question here. Instead, our starting position will be that the use of algorithms in employment should promote, and certainly not obstruct or contravene, *whatever idea of fairness is appropriate in a particular context*.

Hence, where the law requires employers or recruiters to adhere to a particular idea of fairness – most obviously, by avoiding certain kinds of discrimination involving particular attributes – then the AI tools used should be consistent with that requirement. Where employers aim to go beyond the minimum requirements imposed by law – for instance, if they want to broaden the diversity of their workforce – then the tools should allow them to do so. This necessarily means that concerns about bias and fairness intersect with concerns about transparency. Algorithmic unfairness is likely to be easier to control for or avoid if users are aware of how it might arise and how it might be avoided, while unfair decisions and outcomes will be easier to challenge and correct if regulators or people subject to those decisions are able to see how they came about.

In addition to looking at these more general concerns in the content of employment, however, this chapter will identify and discuss some more specific concerns around the use of AI in the employment context, or that arise at particular points in the employment life-cycle.

A. Recruitment

The first significant impact of AI on the world of work is likely to be before the employment relationship even begins. As we discussed in Chapter 1, algorithms are playing an increasing role in recruitment. A recent report by the Institute for the Future of Work explained that AI can be used throughout the recruitment process:

- to source candidates, for instance, by using AI in job advertisement;
- in screening candidates, for instance, in determining which candidates to invite for interview; and
- to inform selections, for instance to predict future job performance on the basis of sales, personality traits, job tenure, and other metrics. (IFOW 2020, p.23)

While it has been said that "there is a lack of data on the global uptake of such technologies" (Sanchez-Monedero et al, 2020), it has been estimated that over 98% of Fortune 500 companies are using some form of applicant tracking software (Shields, 2018). It has also been predicted that the Covid crisis will result in "the increasingly pervasive use of automated hiring systems." (IFOW, 2020, p.6)

The attractions of such technologies are easily seen. The Upturn Report listed several potential advantages of algorithm-assisted hiring processes: (Bogen and Rieke, 2018, p.6)

- reduced time to hire;
- reduced cost of hire;
- improved quality of hire; and
- improved workplace diversity.

While the most obvious benefits will be for the employer, it's easy to see how reduced hiring times and more accurate 'job matching' could lead to more satisfactory outcomes for employees too. Finding out about relevant vacancies, reduced delays in hearing whether we've been shortlisted, and perhaps even reduced bias in hiring decisions are all potential benefits for the prospective employee. The CDEI report on algorithmic bias suggested that "innovation in this space has real potential for making recruitment less biased if developed and deployed responsibly", but also warned that "the risks if they go wrong are significant because the tools are incorporating and replicating biases on a larger scale." (CDEI, 2020, p.46)

Sometimes, AI in recruitment will be used to support rather than replace humans. The IFOW report found that "few employers have automated the entire hiring process, particularly the interviewing and selection stages." (IFOW, 2020, p.23) Some parts of the recruitment process, though, may be more amenable to full automation. According to the CDEI:

> Most algorithmic tools in recruitment are designed to assist people with decision- making, however some fully automate elements of the process. This appears particularly common around automated rejections for candidates at application stage that do not meet certain requirements.
>
> (CDEI, 2020, p. 47)

Important to all of these reports is the recognition that "[h]iring is rarely a single decision, but rather a series of decisions that culminate in a job offer or rejection." (Bogen and Rieke, 2018, p.3) The Upturn Report's authors depict this visually in the form of the Hiring Funnel, which shows the stages at which potential recruits are filtered out (p.13).

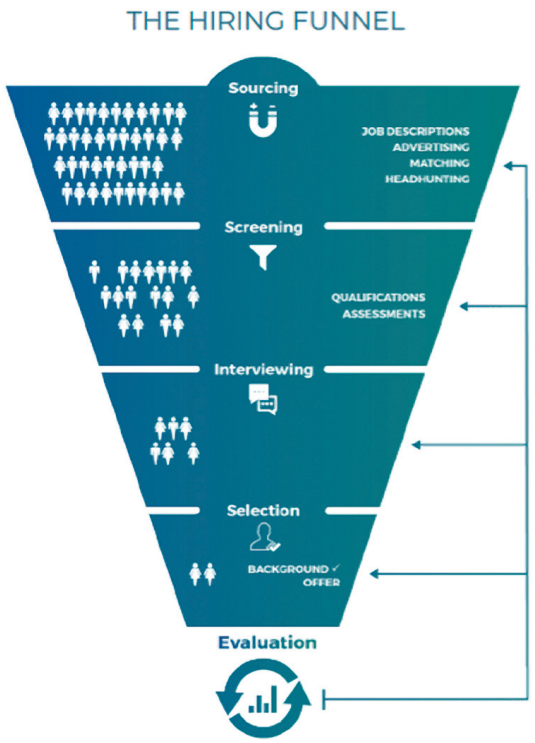


Figure 4. The Hiring Funnel from Bogen and Rieke (2018).

Advertising

As the diagram shows, the large majority of potential recruits are excluded before any contact with the employer – indeed, before they even become aware that the role exists to be filled. As the Report's authors say, "[h]ow employers advertise can sharply limit, or greatly expand, the types of people who even learn a job opportunity exists." (Bogen and Rieke, 2018, p.18)

While job adverts that rely on blatantly discriminatory grounds still occasionally appear (Ellis, 2020), an advert posted on an open forum will be widely seen and – we can assume – widely criticised, perhaps even subject to legal challenge. In New Zealand, it is (with a few limited exceptions) unlawful for employers to make decisions on the basis of prohibited grounds, such as sex, sexual orientation, marital status, religious and ethical beliefs, political opinion, race and ethnicity, disability, age and employment status. (Human Rights Act (HRA) 1993, s.21(1)) This specifically applies to job application forms (s.23) and to recruitment consultants (s.22(2)). As the Employment New Zealand website states, this prohibition extends to advertising: "You must make sure that your advertisement doesn't reflect unlawful discrimination." (Employment NZ, 2021)

New Zealand law doesn't require discrimination to be intentional in order for it to be unlawful. It also explicitly applies to recruitment by third parties such as recruitment agencies (HRA, s 22(2)). According to the Human Rights Commission, "[i]f a recruitment consultant places a job advertisement on behalf of an employer, both the recruitment consultant and the employer are liable for any breaches of the Act." (HRC, 2016, p.24) New Zealand's discrimination law also applies to 'indirect discrimination'. In the present context, that would extend to an advert that doesn't explicitly refer to prohibited grounds, but has the effect of excluding or disadvantaging some people on the basis of a prohibited ground. This is unlawful under New Zealand law (HRA s.65), unless a "good reason" can be established. A job advert that required a degree in computer science, for instance, might have the effect of excluding more females than males (assuming there are more males than females with such qualifications), but it would be lawful provided there was a good reason for requiring that qualification.

New Zealand law, then, seems to provide an adequate response to ads that are (directly or indirectly) discriminatory in their *content*. Furthermore, AI tools may actually help in making sure employers don't inadvertently exclude potential applicants by the phrasing of their ads. Text analytics company Textio uses AI and data science "to reveal the hidden gender bias in your writing and suggest alternatives so you can recruit from the widest possible pool of qualified candidates." (https://textio.com)

A potentially more challenging regulatory target is the practice of ad *targeting*. Employers seeking to target their advertisements at particular audiences is not a new phenomenon. It's common, for example, for ads to be placed in specialist press – education supplements of national newspapers, for example, or specialist trade publications. The rise of online platforms such as Facebook and LinkedIn, though, as well as more specialized platforms such as ZipRecruiter, Indeed and Entelo, allow a degree of granularity in this targeting that has never previously been possible, with adverts being targeted at an individual level, based on (real or perceived) attributes of the target audience.

The benefits of so doing are easily seen. As Kim and Scott explain:

> The power of online recruiting lies in the ability it gives employers to precisely target specific audiences. By directing their employment ads at the most plausible and desirable candidates, they can save money and effort.
>
> (Kim and Scott, 2019, p.97)

Most of us would probably also agree that – if we must see adverts at all – we would prefer to see adverts that correspond to our interests. Having our social media feeds full of adverts for jobs that we have no interest in applying for will be little more than a nuisance. More importantly, perhaps, is the benefit in not missing the chance to apply for jobs that do appeal.

The danger, as Kim and Scott explain, is that "[w]hile targeted advertising enables more efficient outreach, it may also open the door to the discriminatory delivery of ads." (Kim and Scott, 2019, p.97; see also Speicher et al, 2018) A job advert could be entirely neutral in how it describes a role and the qualifications and attributes it seeks, but targeting could ensure that is only seen by people who satisfy certain criteria. If those criteria include prohibited grounds, this certainly seems to transgress the spirit of the Human Rights Act, though there seem to have been no New Zealand cases where this has been tested, and no official statements offering advice or guidance.

RECOMMENDATION 6: While the issue has yet to be tested judicially, we consider it likely that ad targeting would be covered by the discrimination provisions of the Human Rights Act. New Zealand companies that make use of targeted advertising, either by using it themselves or outsourcing their ad placement to third parties who do so, should therefore be aware of their potential legal obligations in this respect.

RECOMMENDATION 7: Guidance on advertising from Human Rights Commission and Employment NZ should be updated explicitly to address ad targeting.

New Zealand is not unusual in having little or no law specifically addressing ad targeting. Globally, legal challenges to ad targeting have thus far been rare. The notable outlier in this regard is (unsurprisingly) the USA, where several discrimination lawsuits have been brought against Facebook's advertising platform. These actions were brought by a range of civil rights organisations, trade unions and affected individuals, and related to a range of different adverts targeted through audience selection tools on the platform. The plaintiffs argued:

> that Facebook's ad platform enabled advertisers to exclude certain users from seeing housing, employment, and credit opportunities in a discriminatory fashion in violation of federal, state, and/or local civil rights laws.
>
> (National Fair Housing Alliance, 2019)

In a settlement between the parties, Facebook has undertaken to make "far-reaching changes and steps that will prevent discrimination in housing, employment, and credit advertising on Facebook, Instagram, and Messenger." As Facebook Chief Operating Officer Sheryl Sandberg announced, this will include ensuring that:

- Anyone who wants to run housing, employment or credit ads will no longer be allowed to target by age, gender or zip code.
- Advertisers offering housing, employment and credit opportunities will have a much smaller set of targeting categories to use in their campaigns overall. Multicultural affinity targeting will continue to be unavailable for these ads. Additionally, any detailed targeting option describing or appearing to relate to protected classes will also be unavailable.

Facebook also committed to greater visibility of ads, with Sandberg announcing that they would be "building a tool so you can search for and view all current housing ads in the US targeted to different places across the country, regardless of whether the ads are shown to you." (Sandberg, 2019)

Limiting the options for advertisers to target by prohibited grounds is a welcome move, but advertisers *deliberately* using such tools to facilitate biased hiring decisions may not be the only problem. Research published in 2019 demonstrated how an advertising platform – once again, Facebook – can itself play a major part "in creating skewed, and potentially discriminatory, outcomes" in delivering ads. (Ali et al, 2019) The study claimed to show that "Facebook's ad delivery process can significantly alter the audience the ad is delivered to compared to the one intended by the advertiser based on the content of the ad itself." For the study's authors, this pointed to:

> the need for policymakers and platforms to carefully consider the role of the optimisations run by the platforms themselves—and not just the targeting choices of advertisers—in seeking to prevent discrimination in digital advertising.

In another study, Datta and colleagues developed an automated tool that would search and collect data on adverts shown on the basis of different user profiles. The study found that the stated gender of the user resulted in a very different set of results:

> The two URL+title pairs with the highest coefficients for indicating a male were for a career coaching service for "$200k+" executive positions. Google showed the ads 1852 times to the male group but just 318 times to the female group. The top two URL+title pairs for the female group was for a generic job posting service and for an auto dealer.
>
> (Datta et al, 2015)

The study's authors, though, pointed out that determining how these discriminatory results came about was not straightforward:

> … we cannot determine whether Google, the advertiser, or complex interactions among them and others caused the discrimination. Even if we could, the discrimination might have resulted unintentionally from algorithms optimizing click-through rates or other metrics free of bigotry.

The differences between requirements in different jurisdictions is one of the reasons why it won't be enough to rely on owners of platforms such as Facebook and Microsoft (who own LinkedIn) to take appropriate steps. Demands are growing for greater transparency around advertising platforms. (see e.g. Bogen and Rieke, 2018, p.46) A recent proposal from an alliance of civil society organisations, co-ordinated by the European Partnership for Democracy and including AlgorithmWatch and Privacy International, called for "meaningful default transparency for all ads" on social media platforms. Though primarily a response to the opaque nature of some *political* advertising, the call explicitly extends to *all* advertising, noting that "[u]niversal ad transparency will help combat discriminatory and potentially illegal advertising practices." (European Partnership for Democracy et al, 2020)

The proposal would include the creation of 'ad libraries', which would "become mandatory for platforms from a set number of users onwards." Such libraries would disclose, for each ad, information including the identities of advertisers, how much they spent, and – importantly for our purposes – targeting criteria and mechanisms, and audience actually reached.

The proposal was oriented towards a European context, with calls for the European Commission to take the lead in their implementation. In mid-December 2020, the European Commission took a major step towards doing this. The proposed Digital Services Act (European Commission 2020) aims to introduce "a common set of rules on intermediaries' obligations and accountability across the single market." The draft law has significant implications for many digital services and providers, but for our purposes, the most significant provisions relate to online advertising. Article 24 requires that:

> Online platforms that display advertising on their online interfaces shall ensure that the recipients of the service can identify, for each specific advertisement displayed to each individual recipient, in a clear and unambiguous manner and in real time:
>
> a. that the information displayed is an advertisement;
>
> b. the natural or legal person on whose behalf the advertisement is displayed;
>
> c. meaningful information about the main parameters used to determine the recipient to whom the advertisement is displayed.

Article 30 imposes even stricter rules for "very large online platforms", who will also be required to make publicly available information about:

- whether the advertisement was intended to be displayed specifically to one or more particular groups of recipients of the service and if so, the main parameters used for that purpose; and
- the total number of recipients of the service reached and, where applicable, aggregate numbers for the group or groups of recipients to whom the advertisement was targeted specifically.

The Digital Services Act is still some way from becoming law in the EU, and it may be subject to various amendments along its journey. Nonetheless, we see it as a welcome development, and one to which New Zealand lawmakers and regulators should pay close attention. The potential for New Zealand to make demands on big international platforms is of course much more limited than that of the EU, but NZ employers could certainly reap the benefits if European initiatives lead to such developments. Domestically, it seems more plausible that transparency could be the default for NZ-based firms and platforms.

RECOMMENDATION 8: There is often a lack of transparency about the criteria used by ad targeting platforms. The New Zealand Government should monitor international developments such as the EU's proposed Digital Services Act, and should seriously consider enacting measures that would set equivalent transparency standards both for NZ-based platforms, and for overseas-based platforms offering services in NZ. There is a strong case for insisting on at least as high a level of transparency in NZ as will be required in the EU. In the meantime, NZ companies considering the use of overseas platforms for ad targeting should be aware that their targeting criteria may not be consistent with NZ law.

Shortlisting

The next stage in the recruitment process is likely to involve processing of applications, including CVs, with a view to shortlisting candidates. This is an area that the UK's CDEI identified as one of the few areas of the recruitment process that's sometimes fully automated, particularly "around automated rejections for candidates at application stage that do not meet certain requirements." (CDEI, 2020, p.47)

Algorithmic processing of applications is subject to the same sort of concerns as targeted advertising: it can allow intentional discrimination, but more insidiously, lead to highly discriminatory unintended outcomes. The most notorious example of this to date concerns Amazon's hiring algorithm. In 2014, as Amazon established a team at its Edinburgh office "to build an algorithm that could review resumes and determine which applicants Amazon should bring on board." The algorithms they developed were trained "to recognize some 50,000 terms that showed up on past candidates' resumes." (Goodman, 2018)

As Jeffrey Dastin has explained in an article for Reuters, it soon became apparent to Amazon that "its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way." (Dastin, 2018) The reason for this, according to Dastin, derives from the fact that the algorithms were trained using previous resumes (CVs), which reflected a strongly male IT sector and workforce. "In effect," Dastin wrote:

> Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

ACLU attorney Rachel Goodman has written that, when algorithms are trained in this sort of way, discriminatory outcomes are almost inevitable:

> It shouldn't surprise us at all that the tool developed this kind of bias. The existing pool of Amazon software engineers is overwhelmingly male, and the new software was fed data about those engineers' resumes. If you simply ask software to discover other resumes that look like the resumes in a "training" data set, reproducing the demographics of the existing workforce is virtually guaranteed.
>
> (Goodman, 2018)

The issue here has strong similarities with what we identified in our Phase 1 report, in the context of algorithms used in the criminal justice context, like RoC*RoI and COMPAS. When algorithms are trained on historical data, and that historical data reflects historic discrimination or inequality, then steps will have to be taken to prevent that discrimination or inequality skewing outcomes today. This has become known as the 'dirty data' problem.

One way to guard against this kind of problem is to ensure *transparency* of information such as the training data used, the weightings given to different variables, and the outputs from the algorithmic process. That way, we might hope, inadvertent discrimination could be detected and prevented. Calls for greater transparency are almost ubiquitous in discussions of AI, but achieving it can be a challenge. Algorithms like neural networks are resistant to transparency through sheer complexity: for these algorithms, there is no simple story about 'weightings of variables' to present.

In the context of hiring algorithms, increasing transparency by providing information about how they work and what they are looking for could, of course, give rise to attempts to 'game' the process. We are already witnessing the emergence of an industry geared towards helping job applicants prepare their CVs for algorithmic assessment. International recruitment company Hays has published a set of tips to "ensure your application makes it past the algorithms and reaches their shortlist." (McNeill, 2020) These include advice such as "avoid unusual job titles. Even if your official job title is unconventional, use an industry-standard title in your CV and online profile so it will be recognised." Another recruitment agency advises applicants to "[f]ormat your CV appropriately so it contains the content AI recognises". (New Zealand Immigration Concepts, 2021)

Of course, fine-tuning application letters and CVs is hardly a new endeavour, and it may be that the attempts to 'game' algorithmic recruitment processes raise no new challenges. Much may depend, though, on whether the algorithms in question have the capacity to develop the sort of 'streetwise' scepticism that we might expect from experienced human recruitment personnel.

While shortlisting raises many of the same issues as advertising (inadvertent bias via proxy characteristics, for example), we might hope the shortlisting phase of the recruitment process could prove a slightly less elusive regulatory target than targeted job advertising. Unlike the job seeker who will likely never know what vacancies she never sees, the applicant who regularly applies without being shortlisted will at least be aware of being rejected. If the discriminatory 'bottleneck' occurs at the stage of shortlisting rather than applications, then an employer facing public criticism for an overly narrow demographic base of employees will find it harder to take refuge behind the claim that they could only hire from the pool of those who applied. In shortlisting systems, we can also register the size of a given demographic group in the pool of selected candidates *in relation* to the size of this group in the pool of applicants – while in ad targeting systems, we can't make a relative assessment of this kind.

On the other hand, previous experiences with shortlisting 'blacklists' show that discrimination and unfairness here can be an elusive target too. In our discussions during this project, we have heard accounts of recruitment companies using their datasets to secretly blacklist/downgrade applications (including applicants who 'make trouble' by bringing a personal grievance or using too much sick leave). These 'trouble-making' applicants are filtered from the shortlisted candidates by the recruitment company, but don't know that it has happened, far less why. They just apply for jobs and never get any interviews. They might *suspect* something is amiss, but have no way of knowing what, far less proving it. Unauthorised disclosure of personal information in this way could well contravene the Privacy Act 2020, but any legal challenge would depend on awareness that it is happening in the first place. This issue is apparently becoming a growing concern to employment lawyers and even judges, though no cases have resulted yet.

The use of AI tools in recruitment, then, does not create a new problem in this regard. Indeed, it's worth considering whether auditing and transparency of such systems might actually reduce the potential for such covert practices. We address some of these possibilities later in the chapter.

Interviews

Once a shortlist has been drawn up, employers will move to the interviewing stage. As we discussed in Chapter 1, AI systems are playing an increasingly prominent part in this stage, including attempts to draw inferences about candidates from speech interpretation, visual gesture analysis and such like. A Washington Post article about HireVue's face scanning software reported that:

> the system uses candidates' computer or cellphone cameras to analyze their facial movements, word choice and speaking voice before ranking them against other applicants based on an automatically generated "employability" score.

The article went on to claim that "[m]ore than 100 employers now use the system, including Hilton and Unilever, and more than a million job seekers have been analyzed." (Harwell, 2019)

Quantitative methods for assessing the personal attributes of job applicants are of course not new. Psychometric testing has been part of recruitment for many years. Nonetheless, various new concerns have arisen about the use of AI techniques for this purpose. One suggested difference relates to their likely accuracy. An article in Harvard Business Review in 2019 contrasted existing psychometric tests, which they claimed to have been:

> carefully validated vis-à-vis relevant jobs, identifying reliable associations between applicants' scores and their subsequent job performance (publishing the evidence in independent, trustworthy, scholarly journals)

with newer algorithmic measures, which they claim:

> have emerged as technological innovations, rather than from scientifically-derived methods or research programs. As a result, it is not always clear what they assess, whether their underlying hypotheses are valid, or why they may be expected to predict job candidates' performance.
>
> (Dattner et al, 2019)

Of course, psychometric tests themselves are hardly uncontroversial, and it is perhaps worth noting that three of the article's authors work for Manpower Group, one of the world's largest staffing corporations. Still, particular concerns about affect recognition algorithms are worthy of attention. One 2019 study found that "emotional analysis technology assigns more negative emotions to black men's faces than white men's faces." In particular, black faces were scored as angrier than white, even when smiling. (Rhue, 2019)

Concerns also surround the ability of such systems to interpret facial cues from non-neurotypical candidates, whose displays of affect may differ from the majority. The problem is a familiar one; if the system has been trained on a particular population cohort, it may misinterpret cues from anyone outside of that population, leading to inferences being drawn that are inaccurate in the case in question. Again, this would be an issue with inadequate design and training of the algorithm, rather than anything inherent to algorithmic systems, but it is a risk that must be guarded against.

RECOMMENDATION 9: As in other areas where AI tools are deployed, algorithms designed to interpret cues from applicants based on speech patterns, facial expressions, etc. are often of questionable accuracy, and often pose a risk of unfair bias. In particular, members of minority populations, or people whose affective responses are atypical, may find themselves subject to inaccurate adverse judgments. Care must therefore be taken to ensure a sufficiently broad and appropriate range of training data is used when such systems are being developed. And a good measure of the accuracy of recruiting systems should be made available.

In its 2019 report, the AI Now Institute took a strong stance against the use of such technology altogether:

> Regulators should ban the use of affect recognition in important decisions that impact people's lives and access to opportunities. Until then, AI companies should stop deploying it. Given the contested scientific foundations of affect recognition technology—which claims to detect things such as personality, emotions, mental health, and other interior states based on physiological measurements such as facial expression, voice and gait—it should not be allowed to play a role in important decisions about human lives, such as who is interviewed or hired for a job, the price of insurance, patient pain assessments, or student performance in school. Building on last year's recommendation for stringent regulation, governments should specifically prohibit use of affect recognition in high-stakes decision-making processes.
>
> (Crawford et al, 2019, p.6)

This concern about the accuracy of algorithms has been a recurring theme throughout both phases of our research. As hype around 'AI' continues to grow, it's easy to imagine mounting pressure on employers or recruitment agencies to avoid being left behind. Whether they all have the resources available to conduct their own validation of the tools likely to proliferate in coming years seems doubtful, and raises again the case for an independent body with the capacity to conduct this sort of important task.

In the meantime, we can expect legal challenges and regulatory interventions around such techniques. In November 2019, the Electronic Privacy Information Center (EPIC) filed an official complaint against HireVue before the US Federal Trade Commission. (EPIC, 2019) The claim accused the company of engaging in deceptive practices for its alleged use of facial recognition software, and of unfair practice for its use of 'secret algorithms'. At the time of writing, it appears that the FTC has taken no action in response.

The USA has, however, seen its first legislative response to the use of AI in interviewing, with Illinois becoming the first jurisdiction to have legislated specifically for this issue. The Artificial Intelligence Video Interview Act, which came into effect at the beginning of 2020, provides that:

> An employer that asks applicants to record video interviews and uses an artificial intelligence analysis of the applicant-submitted videos shall do all of the following when considering applicants for positions based in Illinois before asking applicants to submit video interviews:
>
> (1) Notify each applicant before the interview that artificial intelligence may be used to analyze the applicant's video interview and consider the applicant's fitness for the position.
>
> (2) Provide each applicant with information before the interview explaining how the artificial intelligence works and what general types of characteristics it uses to evaluate applicants.
>
> (3) Obtain, before the interview, consent from the applicant to be evaluated by the artificial intelligence program as described in the information provided.
>
> An employer may not use artificial intelligence to evaluate applicants who have not consented to the use of artificial intelligence analysis.

Concerns have already been expressed about how the new law will function, and whether it will have the desired effect. One article expressed concern that "without additional guidance, the required explanation of how AI works may be difficult for human resources personnel to explain or for an average job applicant to understand." (Stegmaier et al, 2020) As that article also points out, the Act "does not specify the consequences of violations or methods of enforcement". Serious doubts also concern the role of 'consent' to such interviews, given the frequent incidence of major disparities in bargaining power. In many cases, applicants who refuse consent to such processing will justifiably fear that this will be counted against them, even if their application isn't rejected at that point. While we therefore agree that some manner of legal protection is needed here, the Illinois model may not, on its own, prove adequate to that task.

This concern about using 'consent' as a justification for such practices was strongly echoed in a recent report by the Commission on Workers and Technology – a joint initiative by the Fabian Society and Community trade union – which called for a revision to the law and codes of practice:

> to clarify that employers cannot use consent as the basis for processing personal data relating to workers or job applicants (because there is an unequal power relationship). Without such consent workers' personal data can only be used after carefully weighing the interests of both employer and worker. This change would mean that consent could not be used to bypass other restrictions on monitoring or on automated decision-making. (Commission on Workers and Technology, 2020, p. 50.)

RECOMMENDATION 10: Lawmakers and regulators should be cautious of legal protections against AI hiring practices that rely on the 'consent' of applicants. Given the typical disparities in bargaining positions, many applicants will simply not be in a position to refuse.

B. Algorithmic management

Algorithmic management has been described as "a diverse set of technological tools and techniques that structure the conditions of work and remotely manage workforces." (Mateescu and Nguyen, 2019, p.3) It covers a range of situations where AI takes on tasks or roles that were previously those of human managers. This can include forecasting demand to allow more accurate decisions about stocking and staffing. This is what the RSA Report refers to as anticipatory logistics: "the process of predicting demand for consumer goods before purchases have been made." The report points out that:

> This allows logistics firms to improve efficiency and cut delivery times. Ocado, for example, uses algorithms to optimise its warehouse storage structure, meaning popular and soon-to-be popular items are in plentiful supply and in close proximity to its picking and packing teams.
>
> (Dellott and Wallace-Stephens, 2017, pp.50-51)

The more significant concerns about algorithmic management, though, tend to arise when it is used to make decisions about human workers – decisions concerning the likes of deployment, shift allocation, promotion, disciplinary action or dismissal. As with other AI methods, it can be used to *support* human managers, by providing them with information on which to base decisions, or to *replace* them, either with regard to certain tasks or entirely.

An opinion from the EU's European Economic and Social Committee, adopted in 2017, claimed that "[w]ork is now often determined and distributed by algorithms without human intervention, which influences the nature of the work as well as working conditions." (Muller, 2016, [3.23]). On the other hand, in his report for AlgorithmWatch, Michele Loi noted that the "application of AI technology to human resources (HR) analytics is still in its infancy, even if one considers a generous definition of what kind of technologies AI refers to." (Loi, 2020, p.4) Such applications as are currently in use:

> rarely involve automated decisions or even recommendations based on data-driven predictions. Rather, they often develop and visualize an array of HR metrics leaving evaluations and decisions entirely to human decision-makers. The function of these technologies is to enhance the analytical capacity of the decision-makers, by virtue of representing and packaging the information in a more usable and insightful format.

Nonetheless, he acknowledges that the prospect of AI assuming more direct roles in the HR context is neither remote nor fanciful:

> AI-generated predictions and recommendations may be used to pursue all tasks currently considered in the domain of data-driven HR analytics, for example in order to personalize employment offers and contracts, manage employee's [sic] performance, optimize learning and talent development activities, manage employee engagement and communication, decide disciplinary, health and safety interventions, organize employees' holidays, absence, flexible working, maternity/paternity leave, and assign rewards (e.g. salary and benefits)
>
> (Loi, 2020, p.5)

In addition, we shouldn't overestimate how readily human users accept machine-generated recommendations, especially in repetitive tasks (see e.g. Zerilli et al, 2019b). While algorithmic management may not actually replace human managers, it may be that the role of those human managers will increasingly be to implement the algorithmic recommendations, rather than take them on board as one factor in a multi-faceted decision involving meaningful human judgment. We return to this point later in the chapter.

A report for the UK's Advisory, Conciliation and Arbitration Service (ACAS) identified several areas where algorithmic management is becoming increasingly prevalent, including:

- the use of automatic shift allocation software in the retail and hospitality sectors;
- manufacturing and logistics firms using algorithms to micro-manage in ever greater detail the individual movements and actions of workers on a minute-by-minute basis; and
- the growth of performance review algorithms, designed not to give instructions to workers but to collect data on them and feed it back to managers, who can use the outputs to make decisions that could include pay, promotion or firing. (ACAS, 2020, pp. 4-5)

Our research, however, has inclined us to be somewhat cautious in predicting exactly how algorithmic management will come to be used. Historically, management and HR practices have tended to follow changes in patterns of work. Given the uncertainties about how AI will affect work, it's therefore difficult to make confident pronouncements about its future role in management. It has been suggested to us, for example, that we may see quite divergent responses in the use of algorithmic management, with different parallel trends emerging in HR/management. The examples we have seen to date, then, may reflect specific aspects of those forms of work, and may not point to more general trends.

Where it has been deployed, however, the potential benefits of algorithmic management are easy to see. The ACAS reports lists several, including:

- improved accuracy of decision making;
- more efficient shift scheduling meaning less wasted time for both managers and workers;
- more efficient task allocation in factories meaning increased productivity;
- better performance assessments through more accurate data collection; and
- reduced opportunities for human favouritism and unconscious biases to intrude into management decisions around remuneration, holiday approval or shift allocation decisions. (ACAS, 2020, p. 5)

Their report looks in more detail at shift allocation algorithms:

> Increasingly common in the retail and hospitality sectors, they can also include quite sophisticated machine learning algorithms to forecast customer footfall, using anything from traffic history and point of sale data to weather forecasts. These predictions are then used to match to employees' skill sets and calculate which employees should be scheduled on any given day, in order for workers' shift patterns to respond to consumer demand. Platforms like Rotageek are in use by companies including Pret A Manger, O2 and Thorpe Park while Percolata is being employed at UNIQLO.
>
> (ACAS 2020, pp.10-11)

As well as the obvious benefits for employers, this:

> can benefit workers by giving clearer advance notice of when shifts will be and making it easier to swap and change them. This offers a potentially major benefit to workers in industries like retail or hospitality, who at present face often being on call or ready to turn up for shifts, only to find them cancelled or cut short with very little notice.

Tools of algorithmic management are increasingly necessary for 'just in time' supply chains, and are an integral part of the so-called 'gig economy.' Platform-based transport and delivery companies like Lyft, Deliveroo and most famously Uber have become a ubiquitous feature of urban landscapes in recent years. Alex Rosenblat, who has studied and written extensively on Uber's business model, summarises it like this:

> Rather than supervising its hundreds of thousands of drivers with human supervisors, the company has built a ride-hail platform on a system of algorithms that serves as a virtual "automated manager." Freed from the necessity of layers of real bosses, algorithms manage drivers directly according to the rules that Uber lays out.
>
> (Rosenblat, 2019, p.3)

Whether the arrival of the 'gig economy' is a welcome development is, of course, a moot point – the regulatory and policy issues raised by gig economy companies such as Uber would merit a report all to themselves. Much of the uncertainty relates to the legal status of drivers, specifically, whether they are properly considered employees, contractors or something else. (The NZ Employment Court has very recently addressed this issue; see *Arachchige v Rasier NZ Ltd and Uber B.V.*[2020] NZEmpC 230). While their use of algorithmic management has yet to generate the same degree of legal scrutiny, this may be a future locus of legal conflict for the gig economy.

Algorithmic management, though, is not confined to the gig economy. It's been suggested that some of the issues that have arisen in the gig economy are likely to foreshadow wider concerns as such technologies become more widespread. The ACAS Report documents "growing evidence that some of these digital management practices pioneered in the gig economy are starting to spread to the wider labour force." (ACAS, 2020, p.12)

In this section, we consider a number of concerns about algorithmic management. First, we look at some issues that are common to other algorithmic decisions, and consider how they could play out in the workplace environment. In the next parts, we look at some issues that are more specific to the workplace context. The issues we will consider relate to:

- autonomy;
- transparency;
- discrimination and bias;
- ratings-based management; and
- evaluation, monitoring and surveillance.

Autonomy

A concern we have encountered throughout our research relates to the impact of workplace AI on the autonomy of workers – on how they allocate and organise their time, and on the degree of discretion they can employ in going about tasks. 'Micro-management' and 'Taylorism' are hardly novel practices in the modern workplace, but there are worries that AI may allow them to proliferate with an unprecedented level of precision and control. The ACAS Report provides the following example:

> handheld devices and tablets have long been used to give warehouse 'pickers' sets of timed instructions as to what items to collect from where on a minute-by-minute basis. Amazon warehouses are now taking this to the next level – workers are being equipped with a wearable haptic feedback device that tells them what to collect, where to find it in the warehouse and gives them a requisite number of seconds to find the item. They wear these devices on their arms and it uses vibrations to guide their arm movements in order to be more efficient.
>
> (ACAS, 2020, pp.10-11)

The privacy and welfare concerns raised by AI-enabled workplace surveillance will be considered in more depth later, but a further concern for the report's authors related to the autonomy and dignity of the workers:

> Taking away people's autonomy in this way can remove an important sense of dignity and humanity from work, when workers are denied the ability to make even tiny or mundane decisions about what size of box to use or how long a piece of tape to cut for wrapping, or even where and how to move their own limbs. (p.11)

They also point to an interesting irony regarding the autonomy of the managers provided with these tools:

> these algorithms allow them potentially much greater control over their workforce but at the cost of paradoxically making their own jobs less relevant; if all key recruitment, task-allocation and performance review functions can be undertaken by algorithms, what discretion is there left any more for human line managers? (p.13)

The importance, and potential erosion, of human discretion by algorithmic tools has led to frequent demands that AI should only be used to support human decisions in the workplace, and not to replace them. The ACAS Report, for example, proposes that:

> Algorithms should be used to advise and work alongside human line managers but not to replace them. A human manager should always have final responsibility for any workplace decisions.
>
> (ACAS, 2020, Recommendation 1)

As in our Phase 1 report, we sound a note of caution in this regard. The prospect of always maintaining a 'human in the loop' has obvious appeal, but has the potential to offer little more than a 'regulatory placebo' if their role is largely nominal (signing off on what the algorithm recommends). We have also written before about concerns around algorithmic bias and decisional atrophy, where humans rarely called upon to use certain skills lose either the confidence or competence to offer meaningful oversight of algorithmic outputs (see Zerilli et al, 2019b).

If managers are to work effectively alongside AI systems, then there needs to be what Michele Loi refers to as "competence alignment." (Loi, 2020, p.39)

> *Ensure you have adequate competences to build and implement AI ethically.* Organizations that aim to produce and implement AI tools in HR must recruit experts with the range of skills (and informal knowledge, and cognitive styles) required to evaluate an algorithm's intelligibility and fairness.
>
> (Loi, 2020, p. 47)
>
> Educate potential end-users (e.g. HR professionals) to ensure that they have the know-how and skills necessary to operate AIs in HR correctly. End-users should not have blind faith in AI tools, but the adequate level of trust combined with critical attitudes.
>
> (Loi, 2020, p. 47)

RECOMMENDATION 11: As in other areas where AI is being deployed, it is common to encounter demands to keep 'humans in the loop.' If this is to offer more than nominal assurance, though, serious attention must be paid to ideas like competence alignment, and measures to protect against automation bias and decisional atrophy.

Transparency

Concerns about transparency of algorithmic management are becoming increasingly pressing. A report for the International Labour Organisation (ILO) warned that:

> The way these management systems operate is almost never transparent, as companies do not share the methods through which ratings and customers' feedbacks over the workers' activities are gathered and processed. Management by the rating is also spreading ever more beyond platform work, with apps that allow processing patrons' and restaurants' feedbacks over individual waiters.
>
> (ILO, 2018, p.8)

This lack of transparency has been reported as a major source of frustration among Uber drivers. A 2019 study found that:

> While the app is learning a lot about them, Uber drivers find it frustrating how little they know about the app. They find the lack of transparency of the underlying logic of the complex algorithms frustrating, believing it to be an unfair system which manipulates them subtly without their knowledge or consent. (Indeed, Uber has previously admitted to drawing on insights from behavioral science to nudge drivers to work longer hours).
>
> (Möhlmann and Henfridsson, 2019)

If algorithmic management is used to inform, or make, decisions about dismissal or other actions that negatively impact on an employee's conditions, the decisions are still able to be challenged and will be measured against the standard of "a fair and reasonable employer" in the same circumstances. (Employment Relations Act 2000 (ERA), s 103A) The employer's actions must be "justifiable", and the test of justification relates both to the "substantive decision" and the "procedural fairness" of the decision-making. There are statutory criteria with indications of 'fairness' including "sufficient investigation," the right of the employee to know what is being alleged and to have any concerns raised in advance, the right to be given a reasonable opportunity to respond (including enough information about the allegations), and an obligation on employers to genuinely consider the employee's response. There is a considerable body of case law on the expected practices of a fair and reasonable employer in New Zealand, and an overarching statutory duty of good faith. How these standards will be applied to decisions made by AI remains to be seen. While New Zealand courts have traditionally been reluctant to interfere in matters of 'managerial prerogative' when it comes to commercial or operational decisions, they will examine the fairness of the processes informing those decisions, with the processes of obtaining the information underpinning the employer's decisions open to scrutiny. Recent cases have also indicated a greater willingness to interrogate the basis for commercial or operational decisions. As the Court of Appeal has said, an employer's actions cannot be deemed reasonable just because the employer considers it was reasonable. (*Grace Team Accounting Ltd v Brake* [2014] NZCA 541, at [89])

As with other situations where algorithmically informed decisions are open to scrutiny or challenge, this has raised questions about the extent to which the reasons for those decisions might be rendered opaque by the algorithm. As we discussed in our Phase 1 report, this opacity might be technical or legal. The latter has recently been the subject of yet another legal challenge against Uber, launched in late October by the AppDrivers & Couriers Union "over failure to provide access to data & explanation of algorithmic management as required by GDPR [the European Union's General Data Protection Regulation]." (ADCU, 2020)

New Zealand has no direct equivalent of the GDPR, and the question of how or indeed whether the output of an algorithm could satisfy the requirements of the ERA has yet to come before a New Zealand court. Some indication can, however, be inferred from how New Zealand law has responded to other situations where management decisions have been made by reliance on opaque systems.

In *Gilbert v Transfield Services (New Zealand) Ltd* ([2013] NZEmpC 71 CRC 46/10) the New Zealand Employment Court held that a decision to dismiss the plaintiff was unjustified, in part because its decision was informed by the results of a psychometric test (administered by the hiring company Previsor) which it could not explain to him, or it would seem, even understand itself. The Court's decision is worth reproducing at length, in consideration as to how readily much of this could be applicable to opaque algorithms:

> [111] Transfield's refusal to disclose the actual Previsor test scores, combined with its inability to have access to the proprietorial intellectual property of the testing organisation, including questions asked and the actual answers given, is not consistent with the requirements of the Act for information sharing, disclosure, and objective rationality. Not only was this information not available to Mr Gilbert but it was apparently not available to Transfield. … Although the owners of the testing system may have had good reason to keep its ingredients and even results secret, that illustrates the inappropriateness of its use in a process that requires openness and information exchange. Employers proposing to use testing procedures that they do not fully understand, and are not permitted to know about, will have difficulties when challenged by employees such as the plaintiff to justify the consequence of dismissal effected in reliance on the products of such systems.
>
> [113] Transfield's decision to employ an assessment tool that was incapable of meaningful explanation made it impossible to comply with the requirements in s 4(1A) of the Act to provide access to employees (including Mr Gilbert) to information about the psychometric test. It thereby deprived them of an opportunity to comment on the results of the test upon which the employer relied in the course of determining that Mr Gilbert was redundant and dismissing him. As well as the psychometric test for recruitment purposes being of dubious value to the very different exercise of selection for redundancy, Transfield created an additional problem for itself by purchasing and using an assessment tool which it could not and did not understand or explain to affected employees or indeed to the Court at the hearing.

The reference to s 4(1A) is to the provision of the ERA that:

> requires an employer who is proposing to make a decision that will, or is likely to, have an adverse effect on the continuation of employment of 1 or more of his or her employees to provide to the employees affected—
>
> (i) access to information, relevant to the continuation of the employees' employment, about the decision; and
>
> (ii) an opportunity to comment on the information to their employer before the decision is made.

Internationally, questions of transparency in algorithmic management have already begun to generate legal scrutiny and conflict. A Texas District Court has already heard a dispute concerning "the use of privately developed algorithms to terminate public school teachers for ineffective performance." (*Houston Federation of School Teachers v Houston Independent School District* 251 F.Supp.3d 1168 (2017)) In 2010, the Houston Independent School District had begun "its transition to a 'data driven' teacher appraisal system". One of the criteria for evaluation, "student performance", was "based on proprietary algorithms belonging to a private company" – the Education Value-Added Assessment System (EVAAS).

The plaintiffs, a teachers' union, sought an injunction against the use of this algorithm in termination or non-renewal of contract decisions. Their case rested on a variety of grounds, including:

- "lack of sufficient information to meaningfully challenge terminations" based on the algorithmic score, specifically, that "they are denied access to the computer algorithms and data necessary to verify the accuracy of their scores"; and
- the claim that the "system is too vague to provide notice to teachers of how to achieve higher ratings and avoid adverse employment consequences".

The employer, the school district, did not calculate teacher evaluation scores itself, but rather, delegated these to a third party vendor. The employer did not verify or audit the scores, and conceded that there was no means by which the teachers themselves could do so. (The absence of transparency in this system was shown by the fact that the plaintiff's expert was unable to replicate the teachers' scores, even when given access to the computer systems.)

While upholding the defendant's claim for summary judgment[13] with regard to several of the plaintiff's grounds, the judge refused to do so with regard to the lack of information aspect, holding that "teachers have no meaningful way to ensure correct calculation of their EVAAS scores, and as a result are unfairly subject to mistaken deprivation of constitutionally protected property interests in their jobs." Even while granting summary judgment with regard to the vagueness claim, the ruling was hardly a ringing endorsement of the algorithmic system, noting that "teachers may not be able to verify the accuracy of their EVAAS scores" and that "it may be unfair or prone to error."

Later that year, the parties reached an out of court settlement whereby the School District undertook not to use the EVAAS system or other unverifiable value-added scores as a basis to terminate employment (American Federation of Teachers, 2017).

Access to meaningful explanations about managerial decisions is an important part of a good employment relationship. It can allow mistakes to be corrected and unfairness to be identified. More than that, it seems integral to a workplace where workers are treated with a degree of dignity and respect. As Monique Valcour has written, "dignity exists when people are listened to and taken seriously regardless of their position – and feel they can disagree respectfully and be heard, without fear of reprisal." (Valcour, 2014) Mechanically issuing instructions with no opportunity for engagement or interaction is corrosive of workplace dignity.

RECOMMENDATION 12: Employers should consider making sure task allocation algorithms are 'explainable', in terms that are meaningful to their workers. Workers should be able to ask why they have been allocated particular jobs or shifts, and to receive a meaningful answer. More generally, algorithmic management systems could benefit from explanation tools. As in other areas of algorithmic 'explainability', due concern should be paid to the level of explanation likely to be sought in particular contexts. But opaque systems are likely to foster workplace resentment against algorithmic management.

Draft legislation introduced into the Canadian Parliament in November 2020 may provide a model for this kind of transparency. Bill C-11 is concerned in general terms with consumer privacy protection, but it contains some provisions specifically concerned with "automated decision systems", defined as "any technology that assists or replaces the judgement of human decision-makers using techniques such as rules-based systems, regression analysis, predictive analytics, machine learning, deep learning and neural nets". Clause 63(3) provides that:

> If the organization has used an automated decision system to make a prediction, recommendation or decision about the individual, the organization must, on request by the individual, provide them with an explanation of the prediction, recommendation or decision and of how the personal information that was used to make the prediction, recommendation or decision was obtained.

While this is intended to have wider application than the employment context, it could certainly be used to help ensure the sort of transparency to which workers should have access.

13 This is where the court rules that the defendant has no case to answer; even if the plaintiff were able to prove all of the facts that they allege, there would be no legal remedy available.

Discrimination and bias

As we've already noted, some commentators are optimistic about the potential for algorithmic management to reduce managerial bias. As the IFOW report notes, though, "it also poses considerable risks that employers will unwittingly propagate patterns of bias, discrimination and inequality." (IFOW 2020, p.11) In particular, concern has been expressed about the potentially disparate outcomes of "[t]he increasing use of rating and review systems within work contexts" (Mateescu & Nguyen, 2019b, p.14).

In many respects, the dangers here are similar to those that arise in recruitment. A model trained on profiles of previous workers could make recommendations or predictions based on characteristics that are irrelevant or discriminatory. When those findings flow through into decisions about, for example, deployment or promotion, then it's easy to see how historically discriminatory patterns could be replicated.

There is a difference between how New Zealand law responds to recruitment and algorithmic management: while recruitment involves dealing with people who are only potential employees, management relates to people who are already part of the workforce. This brings them within the ambit of the ERA. The ERA states that an employee is discriminated against if their employer, by reason of any prohibited ground:

- refuses or omits to offer or afford to that employee the same terms of employment, conditions of work, fringe benefits, or opportunities for training, promotion, and transfer as are made available for other employees of the same or substantially similar qualifications, experience, or skills employed in the same or substantially similar circumstances; or
- dismisses that employee or subjects that employee to any detriment, in circumstances in which other employees employed by that employer on work of that description are not or would not be dismissed or subjected to such detriment; or
- retires that employee, or requires or causes that employee to retire or resign. (ERA, s 104)

This is in addition to the protections afforded by sections 22 and 23 of the Human Rights Act, which we discussed in the previous section. It's probably safe to conclude, then, that any shortcoming in the regulatory setting may be less attributable to a lack of applicable law, and more to matters of implementation, compliance monitoring and enforcement.

As with recruitment, there is an obvious connection between identifying bias and transparency – both in terms of the algorithm's workings, and in how its predictions and recommendation are translated into actual decisions. When a workforce is spatially distributed – as is often the case with gig economy workers – discriminatory effects in dismissal, promotion, remuneration or shift/task allocation may not be apparent to individual workers. Even were they to be aware of these effects, the same issue may arise as was discussed in the context of recruitment: it may not be clear whether the problem lies with the employer's choices, or the algorithm itself.

What may be hoped, though, is that such decisions may be easier for an employer to monitor than those concerning advert targeting. It should be possible for HR departments – whether human or algorithmic – to monitor trends to check for discriminatory outcomes. Indeed, it has been pointed out that various AI tools exist to help employers detect such outcomes:

> Algorithmic tools can, for example, analyse company payrolls to measure the levels of gender or racial pay gaps in different parts of the organisation and what factors seem to contribute to them. They can assess individual managers based on how often they recommend men versus women for recruitment, promotion or pay rises to identify those who might need additional unconscious bias training. They can also scan the content of internal communications or external job postings for gendered language terms and recommend alternatives. In this way algorithms could make a huge difference to eliminating the gender pay gap and other workplace disparities. (ACAS, 2020, p.22)

As with recruitment, though, concerns exist about the criteria employed by different tools – what notions of bias or fairness they use, for instance, and which jurisdiction's laws they are aligned with. Throughout the literature, there is a recognition of a general absence of consistent standards for AI auditing tools. We return to the issue of auditing later in the chapter.

Ratings-based management

A specific problem identified with gig economy work relates to the increased reliance on metrics such as customer ratings as indicators of performance. As Rosenblat explains:

> After each trip, passengers are prompted by the Uber passenger app to rate drivers on a scale of one to five stars on their mobile app. A driver's rating is the average of ratings from his or her last five hundred trips.
>
> (Rosenblat, 2019, p.149)

This could arguably be seen as less an issue of management by algorithm, and more about algorithms allowing "passengers effectively [to] perform one of the roles of middle managers, because they are responsible for evaluating worker performance." Of course, customer feedback is likely to play a role in performance evaluation in sectors that do not rely on algorithmic management; student evaluations in academia are an obvious example. The gig economy, though, places these front and centre of performance evaluation, arguably *replacing* rather than *supporting* decisions by human managers. Uber drivers whose ratings drop below a certain level will find themselves 'deactivated' – effectively dismissed – without the reasons for those ratings being assessed by a manager, or being afforded an opportunity to give their version of events.

Rosenblat documents some of the effects of such a system on Uber drivers:

> Fearing low ratings and deactivation, these drivers try to be extra nice to dissatisfied passengers when working for Uber. Drivers thus take on the "care work" involved in managing Uber's relationship with passengers, and they provide emotional labor, like making passengers feel good, as part of their service-economy job.

While the prospect of taxi drivers having to be pleasant to customers may initially seem quite appealing, the stories told by Rosenblat's interviews point to a more problematic situation; of drivers having to put up with racial, sexist or other forms of harassment, or with obnoxious or heavily intoxicated passengers, for fear of losing the all-important rating on which their future with the company depends. As she explains, "[d]rivers are helpless against unfair ratings, a demonstration of the limits of their power in an employment relationship governed by inflexible algorithmic manager." (Rosenblat, 2019, p.155)

In theory, Uber drivers have access to an appeal process to challenge unfair ratings. Rosenblat, however, is highly sceptical of its utility:

> Drivers don't have a dedicated human manager who responds to their inquiries. Instead, they have community support representatives (CSRs), located at the email equivalent of a call center, often located abroad, such as in the Philippines, and managed by third-party companies, like Zendesk. Effectively, Uber offshores and automates its main communications with drivers. Drivers receive automated replies to most of their inquiries, which often appear to be based on keywords in the text of their emails. In other words, Uber is managing drivers without a human that understands and is responsive to nuances. While automated responses might be practical for basic factual inquiries, they can prove woefully insufficient when a passenger overdoses in the backseat or harasses a driver.
>
> (Rosenblat, 2019, p.143)

Again, it would be inaccurate to suggest that all such problems are unique to the gig economy sector, or algorithmic management, still less to imply that human managers are invariably more sympathetic to the sorts of concerns raised by Uber drivers. Nonetheless, the implications of a ratings-driven management system, with human oversight present only in a relatively nominal form, will be important to monitor.

The use of customer ratings to assess employees' performance in aggregated scores is another practice that is gaining ground, seeping over from the gig economy to the regular workforce. (ACAS, 2020, p.14)

C. Evaluation, monitoring and surveillance

A final concern around algorithmic management relates to increased workplace surveillance. Again, we begin by noting that some types of monitoring are primarily beneficial for workers, especially in hazardous workplaces, where the employer has an obligation to keep workers safe, and monitoring (e.g. of worker health) is part of the employer's duty of care. For instance, WorkSafe NZ gives detailed guidance about monitoring of this kind. (WorkSafe NZ, 2017) However, other types of surveillance are clearly in service of company efficiency, rather than worker safety: we will focus on this kind of surveillance here. In this regard,

the ILO report was fairly typical in its warning that technology will "increase the possibility of management to increasingly monitor working activities in a way that is not desirable for workers." (ILO, 2018, p.1)

The Royal Society and British Academy report refers to:

- a greater emphasis on measurable aspects of work as indicators of performance and drivers of pay (to the expense of 'symbolic work', including the amount and characteristics of interactions with customers); and
- a move from direct monitoring based on observation from a supervisor to continuous monitoring based on data. (Royal Society and British Academy 2018, p.54)

The prevalence of workplace technological surveillance, at least in the USA, was illustrated by a 2019 report by the American Management Association showing that:

- 45% of employers track content, keystrokes, and time spent at the keyboard;
- 43% store and review computer files;
- 12% monitor the blogosphere to see what is being written about the company; and
- 10% monitor social networking sites. (American Management Association, 2019)

Technological surveillance in the workplace is by no means a recent phenomenon. (ILO, 2018, p.7; Ajunwa, Crawford and Schultz, 2017) As Paul Roth has said, "[t]he modern day motivation for collecting personal information about workers can be traced back to the 'scientific' management techniques of Frederick Winslow Taylor at the turn of the last century." (Roth, 2016, p.57) Nonetheless, as a report from the Data & Society Research Institute recently claimed:

> Monitoring and surveillance tools are collecting new kinds of data about workers, enabling quantification of activities or personal qualities that previously may not have been tracked in a given workplace. Employers may seek to quantify "soft" skills such as sociability through tools like facial recognition and sentiment analysis. And employer-provided biometric health trackers may collect sensitive data about workers, from stress levels to smoking habits, raising questions about both consequences at work and growing intrusion into personal life.
>
> (Mateescu and Nguyen, 2019, p.3)

It has also become a matter for particular concern during the Covid lockdown, with RNZ reporting that "Sales of employee monitoring software have skyrocketed since the country went into lockdown, with tech companies reporting a 300 percent increase for New Zealand customers." (Hatton, 2020). Other employers have achieved the same objective by leveraging existing technology. This is by no means unique to New Zealand. In May, an article in *Slate* reported that:

> Since the start of the pandemic, many companies have begun to even more aggressively track their workers' productivity, and as workplaces start to open again, it is likely that the scale and types of data collected by employers will continue to increase to combat the threat of COVID-19.
>
> (Chyi, 2020)

Not all of this will involve AI technology, but some will. The *Slate* article refers to a company, Landing AI, that claims to have:

> developed an AI-enabled social distancing detection tool that can detect if people are keeping a safe distance from each other by analyzing real time video streams from the camera. For example, at a factory that produces protective equipment, technicians could integrate this software into their security camera systems to monitor the working environment with easy calibration steps. … [T]he detector could highlight people whose distance is below the minimum acceptable distance in red, and draw a line between to emphasize this. The system will also be able to issue an alert to remind people to keep a safe distance if the protocol is violated.
>
> (Landing AI, 2020)

As with other technological responses to the Covid crisis, it is not difficult to see the prospective benefits in such technology, but equally, easy to see how it could be misapplied for less benign purposes. Phoebe Moore and colleagues cite one example:

> Employee tracking in Amazon warehouses has resulted in reports of heightened stress and physical burnout. Indeed, employee health and safety usually comes secondary to lean logistics and speed of work in depot work.
>
> (Moore, Upchurch and Whittaker, 2018, p.23)

An account of one of their interviews shows how surveillance technology introduced for ostensibly innocuous purposes can be subject to a rapid 'mission creep':

> One warehouse operative, Ingrid (not her real name), who has worked in one warehouse in Britain for 11 years, provided information about a new worn device that was rolled out in her workplace in February 2016. All warehouse work floor operatives were unexpectedly required to use the hand-worn scanner. The current researchers asked what the workers were told the devices would be used for. Ingrid indicated that management told workers the devices would provide them with information about any mistakes made and who in the warehouse had made them, meaning that they can be used to help to not do this again.
>
> In practice, however, Ingrid indicated that the technology has been used not only to track individual mistakes but also to track individual productivity and time spent working and on breaks. Workers were told that management would hold individual consultations based on the data, but this had not happened. Instead, at a specific interval in the months that followed the devices' implementation, workers were told that people would be fired within days and it transpired that data from devices were part of the decision-making process for who to dismiss.
>
> (Moore, Upchurch and Whittaker, 2018, pp.23-4)

Technological surveillance has not yet received much attention in terms of New Zealand employment law, but one case decided before the Employment Court may give some insight into how such technologies are likely to be regarded. In *OCS Ltd v Service and Food Workers Union Nga Ringa Tota Inc* ([2006] ERNZ 762), the employer had implemented iGuard, a new biometric timekeeping system which required employees to have their fingerprints scanned. The employer claimed that this increased efficiency by the technology affords greater accuracy and efficiency in a business's administration by reducing time fraud and preventing 'buddy clocking.' The employees and their union objected, and claimed that the employer had failed to attain their consent for the new system.

Drawing on jurisprudence from Australia, Canada and the UK, the judge identified a number of criteria that would inform the decision as to the legality of using such technology:

1. Is the technology compatible with the contractual obligations of the parties?
2. There is to be a balance between the need for the technology and the level of personal intrusiveness involved for the individual concerned.
3. he employer has the right to introduce different systems of timekeeping technology subject only to reasonable consideration of valid concerns raised by the union and/or employees.
4. The employer must take the appropriate steps to inform employees of the new measures and to obtain their consent. ([95])

The judge concluded that the employer's failure adequately to consult with the workers put it in breach of its legal obligations. It is informative to note the terms of the union rep's letter, setting out the workers' concerns:

> People feel distressed and deeply hurt by the way the company have simply placed this machine on the wall (appearing as it has like a fait accompli), but also by the implication that there is a lack of trust in workers … this very personal means of timekeeping … is equated in people's minds with criminals and prisoners. ([22])

The privacy implications of such data gathering exercises are likely to engage certain aspects of New Zealand's privacy law. The Privacy Act 2020 lists 13 privacy principles, including:

- Personal information may only be collected for a lawful purpose connected and necessary with a function or activity of the collecting agency. (Principle 1)
- The collecting agency must take reasonable steps to ensure that the individual whose information is being collected is aware of the fact of collection and its purpose. (Principle 3)
- An agency may collect personal information only by a lawful means; and by a means that, in the circumstances of the case is fair and does not intrude to an unreasonable extent upon the personal affairs of the individual concerned. (Principle 4)[14]

14 The new Act stresses the particular importance of this requirement where personal information is being collected from children or young persons, but this is presumably less likely to arise in the context of employment.

An agency that holds personal information that was obtained in connection with one purpose may not use the information for any other purpose. (Principle 10)

It's not difficult to see how any of these principles could be infringed by the use of workplace surveillance technologies. All are, however, subject to a range of limitations. In relation to Principle 10, for instance, information can be used for "other purposes" where:

- the individual is not identified;
- that use is authorised by the individual; or
- that other use of the information for that other purpose is necessary to prevent or lessen a serious threat to public health or public safety.

The public health and safety exception could certainly be relevant during the Covid crisis, for instance, with regard to the Landing AI example discussed earlier. We can also imagine how it could be justified outside of that context; for example, by monitoring worker stress or tiredness. The authorisation exception is also something on which employers could rely, although the extent to which employees are free to decline could be a matter for concern in some instances. The Hong Kong Privacy Commissioner has referred to a "presumption of undue influence … where disparity of bargaining power exists, such as in an employer- employee relationship", albeit one that "can be dispelled by the provision of genuine choices to the data subjects before they decide to provide their personal data."

Paul Roth has expressed scepticism that our privacy law will afford much protection from workplace surveillance:

> even if surveillance is covered under the Privacy Act, the legislation would be of little avail in the workplace, to judge by the few reported cases on workplace surveillance. In these cases, the Privacy Commissioner found workplace surveillance to be a permissible practice.
>
> (Roth, 2016, p.43)

Roth points out that, while the Privacy Commissioner has said that "covert recording is intrinsically intrusive, and needs strong justification for its use," (referring to *Case Note 101213* [2008] NZPrivCmr 4) the decided cases have generally found covert surveillance can be justified because employers are lawfully entitled to take steps to detect unlawful behaviour. Whether the new Privacy Act strengthens these protections is a question on which we have heard diverse opinions, but will in any case be a matter for ongoing scrutiny.

If, as is widely predicted, AI tools allow new and potentially more intrusive levels of surveillance, serious consideration should be given to the potential impacts on workplace health, wellbeing, privacy and dignity. Such concerns should be addressed in early consultations with affected workplaces. The ACAS report is just one of those to call for consultation with workforces: "[e]arly communication and consultation between employers and employees are the best way to ensure new technology is well implemented and improves workplace outcomes." (ACAS, 2020, Recommendation 7) These concerns should be included in the algorithmic impact assessment that we discuss later in this chapter.

It's also possible that regulatory measures could help address some of these concerns. The decided cases under the old Privacy Act, and doubts about the adequacy of its successor, mean that the likely response of New Zealand law to algorithmic surveillance remains a matter of considerable conjecture. However, the possibility exists that the general principles and provisions of the Privacy Act could be bolstered in certain contexts. A suggestion made during one of our expert workshops was the creation of a new, specific code of practice. The Privacy Act gives the Privacy Commissioner the power to issue codes of practice that replace the Privacy Principles in certain contexts. (Privacy Act 1993, s 46; Privacy Act 2020, s.32) These codes may modify the operation of the Act for specific industries, agencies, activities or types of personal information.

Codes often modify one or more of the information privacy principles to take account of special circumstances which affect a class of agencies (e.g. the Credit Reporting Privacy Code 2004 applies to credit reporters) or a class of information (e.g. the Health Information Privacy Code 1994 covers health information). Codes of practice are a flexible means of regulation and can be amended or revoked by the Privacy Commissioner at any time. The precise regulatory target of any such Code is a matter for further deliberation and discussion. Should it be directed at particular AI tools or techniques, or specific uses of such tools and techniques? Or pitched at a less technology-specific level; for instance, use of personal information in recruitment or workplace surveillance? We would welcome further discussion about the most effective means of bolstering the Privacy Act with regard to this fast-growing and concerning use of workplace surveillance technology.

RECOMMENDATION 13: Workplace surveillance technologies are a source of growing concern, and this has grown during the Covid crisis. In due course, these technologies may require specific legislative attention. In the meantime, we would welcome attention from the Privacy Commissioner to the possibility of a code of practice directed at workplace surveillance technologies, or perhaps workplace surveillance more generally.

D. Health and safety, and worker wellbeing

Aside from privacy concerns, concern has been increasingly expressed about the impact of the 'gig economy', algorithmic management and increasing surveillance on worker welfare. The 2019 study of Uber drivers referred to earlier identified "three areas of consistent complaints about working "for" algorithms, concerns that we've also seen in other companies using algorithmic management." (Möhlmann and Henfridsson, 2019) These included the sorts of concerns about transparency and constant surveillance that we have addressed in this chapter. The third concern, though, was what they referred to as "dehumanization":

> Drivers at Uber report feeling equally lonely, isolated, and dehumanized. They don't have colleagues to socialize with or a team or community to be part of. They lack the opportunity to build a personal relationship with a supervisor. Those on crowd-work platforms like Amazon Mechanical Turk have raised similar complaints as they conduct "micro-tasks" such as classifying content or participating in surveys.

To address or mitigate these risks, the report's authors made a number of recommendations for "companies who manage all or part of their workforce through algorithms". These included:

- information sharing – if not sharing the algorithm itself, then at least "the data and goals that informed it";
- inviting feedback and involving workers in decision-making about design and use of algorithmic management, "for example by involving them into committees or councils that discuss and negotiate work related internal regulations"; and
- building trust and implementing benefits that improve worker's welfare.

An important recommendation, though, related to the worry about dehumanization:

> Build in human contact. People need people. Organizations should develop formal, supportive communities where workers feel like members and can make social connections. Adding a human element to the way people are managed will help workers feel less like they are being treated as machines.

The recommendation pointed to New York ride-hailing firm Juno as a positive example, as it was quick to use an "extensive human customer support system that eagerly helped drivers with questions or problems."

Concerns about impacts of algorithmic management on worker well-being extend beyond the gig economy sector. Attendees at our expert workshops referred to an increased risk of loss of decision-making autonomy (including within Human Resources departments!), especially when strict 'lean production' regimes are enforced and digital technologies take over controlling tasks which until now had been performed by specialised employees. In his most recent book, US law professor Frank Pasquale warns that:

> American bosses, in their bid to demand more "flexibility" from a restive workforce, could once point to laborers abroad ready to take domestic workers' jobs; now those bosses are prone to pointing to ever faster machines. Demanding more break time? A robot can work 24/7. Want higher wages? You only create incentives for the boss to replace you with software. Electricity and replacement parts are a lot cheaper than food and medicine.
>
> (Pasquale, 2020, p.227)

The possibility of employers using the ubiquitous threat of technological replacement as a way to coerce workers into acceptance of ever worsening pay and conditions is certainly one of the more troubling prospects to emerge from the AI revolution. Of course, it's a threat that has been made in previous generations too, and as Pasquale

acknowledges, it isn't only the threat of technological replacement that can be used in this way. It is not clear, then, that the answer to this prospect lies with any technology-specific regulatory response, as opposed to a more general realignment of the balance of power between workers and employers.

Pasquale recognises that AI may in fact have a positive role to play in this rebalancing:

> AI should not entrench deep disparities in the power of workers, managers, and capital owners. Rather, it can help unions and worker associations bring more prerogatives of self-governance to the workplace. For example, algorithmic scheduling of workers need not be based merely on cost minimization, upending workers' lives with zero-hours contracts. Instead, organized laborers can demand that the relevant AI accommodate their needs dynamically, enabling exactly the type of family time, leisure, and educational opportunities that a more productive economy should be delivering to everyone.
>
> (Pasquale, 2020, p.176)

A strong and consistent theme at our workshops was that discussions about the threat and promise of workplace AI must be located within the broader context which has seen trade unions and the bargaining power of workers weakened in New Zealand and in many other developed societies. While not a problem specific to the context of AI, it is one that AI has the potential to exacerbate or mitigate, and this is something that should be borne in mind when framing responses to uses of this technology.

Concerns have also arisen about physical safety, mostly in the context of humans working alongside embodied robots. It is often suggested that robots can make work safer, when used:

> to replace workers who carry out unhealthy, tedious or unsafe work, thus avoiding exposing people to dangerous substances and conditions, and reducing physical, ergonomic and psycho-social risks.
>
> (Mercader Uguina and Muñoz Ruiz, 2019)

WorkSafe NZ has recognised that using robots can remove the more traditional hazards of working with machinery, as well as taking on high-risk work, such as in the biotechnology field. (WorkSafe NZ, 2014) Safety concerns have, however, grown, particularly in the wake of highly publicised tragedies such as the deaths of Wanda Holbrook in 2017 (Agerholm, 2017) and an unnamed German worker in 2015. (Gander, 2015)

Many robots currently used in an industrial setting will not use any element of AI. They will be straightforwardly programmable pieces of equipment, that have limited or no autonomy or capacity to respond to their environment. As a report produced by research organisation TNO for the Dutch Ministry of Social Affairs and Employment in 2016 noted:

In general, industrial robots that are used extensively in factories:

- are often in controlled environments;
- carry out repetitive and pre-programmed tasks;
- have no direct interaction with people (including third parties and visitors)
- around them; and
- are not yet able to adapt to new situations. (TNO, 2016, p.16)

When this is true, it seems that they can be readily accommodated within existing product liability and workplace safety legislation. While WorkSafe NZ's current *Safe use of machinery* guideline specifically addresses the use of robotics in the work environment (WorkSafe NZ, 2014, 9.11), it makes no reference at all to artificial intelligence or autonomous robots. The control measures WorkSafe NZ recommends for robots seem predicated on the assumption that they can be kept separate from human workers; they include enclosing the robot, restricting access and turning the robot off when people are near.

As the TNO report goes on to acknowledge, though:

> Much effort is therefore being expended in the field of robotics on developing robots that can move autonomously, that are able to 'see' their surroundings and respond accordingly, that can work alongside people, and that are suitable for more than one task. These are referred to as 'general purpose robots'.
>
> (TNO, 2016, p.16)

In a report from December 2019, the European Agency for Safety and Health at Work (EU-OHSA) also noted that:

> Collaborative and smart robots, so-called cobots, will become a familiar presence in the workplace as highly developed sensors make it possible for people and robots to work together.
>
> (EU-OHSA, 2019b)

The report noted that "Amazon already has 100,000 AI-augmented cobots supporting its distribution activities", predicting that:

> With the increasing use of AI, robots will be able to carry out not only physical tasks but also increasingly cognitive tasks. Robots are already able to perform a variety of cognitive tasks autonomously, such as supporting legal casework or medical diagnoses, and will also become commonplace in customer-facing jobs. This means that the use of smart robots is expected in many different sectors and settings, such as in the care sector, hospitality, agriculture, manufacturing, industry, transport and services.

The increasing incidence of robots and humans working together or in the same environment casts doubt on the efficacy of the sort of 'separate and contain' strategies employed with regard to more traditional production-line robots. As EU-OHSA has acknowledged, "as the implementation of AI at work is relatively new, there is only nascent evidence of OSH risks and benefits." (EU-OHSA, 2019a, p.15) This has not, however, deterred the organisation from casting a speculative eye towards the sorts of risks that could arise as cobots are deployed:

> the growing proportion of mobile, smart robots in the workplace may increase the risk of accidents, as injury could occur from direct contact with robots or from the equipment they use. As smart robots are constantly learning, although efforts are made to factor in all possible scenarios in their design, they may behave in unanticipated ways. Workers having to keep up with the pace and level of work of a smart cobot may be placed under a high level of performance pressure. This may have negative impacts on workers' safety and health, particularly mental health. Increased working with robots will also significantly reduce contact with human peers and social support, which is also detrimental to workers' mental health.
>
> (EU-OHSA, 2019b, p.6)

To what extent is New Zealand law prepared for these new workplace hazards? The Health and Safety at Work Act 2015 imposes duties on a person conducting a business or undertaking (PCBU) to ensure, so far as is reasonably practicable, the health and safety of their workers. (s.36) "Reasonably practical" is defined in s 22 as that which is reasonably able to be done, taking into account and weighing up all relevant matters, including:

- the likelihood of the hazard or the risk concerned occurring; and
- the degree of harm that might result from the hazard or risk; and
- what the person concerned knows, or ought reasonably to know;
- the availability and suitability of ways to eliminate or minimise the risk; and
- the cost associated with available ways of eliminating or minimising the risk, including whether the cost is grossly disproportionate to the risk.

While the Act explicitly allows for a degree of cost-benefit analysis, it is notable that the standard for a measure not to be deemed reasonably practical is high; a PCBU will be required to take a precaution unless its cost is *grossly disproportionate* to the risk. In relation to reasonably foreseeable risks of death or serious injury, this seems to set a high bar for PCBUs seeking to justify any failure to ensure worker safety. Furthermore, while offences under the Act (sections 47-49) do not impose strict liability, neither do they all require intentional acts or omissions (s.54).

As in the context of privacy law, New Zealand workplace safety law also allows for the publication of codes of practice (WorkSafe New Zealand Act 2013, s 10(e)). These are developed by WorkSafe NZ, and on approval by the Minister, become approved codes of practice. While not legally binding, adherence to a code of practice can have the effect of insulating a PCBU from liability under the Act. Section 226 states that, in any civil or criminal proceedings under the Act, a court may have regard to the code as evidence of what is known about a hazard or risk, and rely on it in determining what is reasonably practicable in the circumstances to which the code relates.

At the time of writing, approved codes of practice exist for many areas of traditional risk, including forestry, port operations, and working with asbestos, but as yet, no

codes exist with regard to algorithmic management, workplace surveillance, workplace robots or AI.[15] Indeed, no codes of practice exist to address stress, isolation, or many other common risks in the modern workplace, a situation that has been described to us by one expert as "50 years behind Europe and we've got nowhere even close to robots or AI." A PCBU could elect to follow other standards – most obviously, the International Organisation for Standardisation (ISO) has published standards for industrial robots (ISO, 2011) and more recently for collaborative robots (ISO, 2016). While this lacks the legal effect of an approved code, demonstrable reliance on an ISO standard could be used as evidence of having done all that was reasonably practical.

While the New Zealand approved codes of practice are required to be made freely available online, ISO standards are available to purchase. The price is relatively modest, particularly we might think to a PCBU wealthy enough to afford collaborative robots (the collaborative robots standard costs 138 Swiss francs, or $220NZ). Nonetheless, publication of equivalent standards within New Zealand would have the effect of removing even this modest barrier to access – importantly, making sure that workers as well as employers have access to their content – while clarifying their legal status and potentially raising awareness of their existence.

Codes of practice are not a panacea, and concerns have been raised about their usefulness. (Dabee, 2016, p.597) Even were such a code of practice to be approved, the rapid pace of change in the technology would require it to be kept under review. In its 2019 overview of the Health and Safety at Work Act 2015, the Ministry for Business, Innovation & Employment (MBIE) noted that there was a "need to keep up with changes in practice and emerging technology to ensure benefits are realised and any risks are managed e.g. automated machines and industrial robots." (MBIE, 2019)

Nonetheless, the possible role of codes of practice directed at the more potentially harmful uses of AI technologies is something that we feel should be given careful attention. Most obviously, these could address the physical dangers posed by workplace robots that are designed to operate in close proximity to humans.

Beyond this, however, we see no reason in principle why concerns relating to less acute risks, such as increased stress from constant monitoring, could not fit within WorkSafe NZ's remit "to promote and contribute to a balanced framework for securing the health and safety of workers and workplaces."

RECOMMENDATION 14: WorkSafe NZ should consider issuing a code of practice dealing with workplace robots and particularly 'cobots', perhaps based on the ISO standard for collaborative robots. The 'separate and contain' approach contained in the current "Safe use of machinery" guideline is not an appropriate response to collaborative robots whose purpose is to work in close proximity to human workers.

RECOMMENDATION 15: WorkSafe NZ should also consider issuing a code of practice relating to workplace surveillance and algorithmic management. While this may seem to fall more appropriately within the remit of the Privacy Commissioner, the prospect of psychological harm from the inappropriate use of such technologies also brings it within the realm of the workplace safety regulator. Importantly, care must be taken to ensure that this issue does not fall between two regulatory remits, with the result that it does not receive proper attention from either.

15 The Safe Use of Machinery guideline referenced above, which makes reference to industrial robots, predates the 2015 Act and is not an approved code of practice.

E. Technological redundancy

At the end of the employment life cycle, of course, is the end of the employment relationship. This can be due to retirement, resignation, dismissal or redundancy. Fortunately, New Zealand law does not follow the US practice of allowing employees to be dismissed at will. Rather, the Employment Relations Act 2000 allows any employee who has been dismissed by their employer to bring a claim of unjustified dismissal (s 103(1)(a)). This may be on either substantive grounds (absence of good cause) or on the basis that the dismissal was carried out in a procedurally unfair manner.

Redundancy is a particular form of dismissal. Although not defined in the most recent ERA, the previous statutory definition (Labour Relations Act 1987), captured what we take to be the common understanding of redundancy as dismissal on the basis that "the position filled by the worker is, or will become, superfluous to the needs of the employer". The idea, then, is that it is the position or role that becomes redundant, rather than the individual filling it. An employee could not be said to have been made redundant if their position was subsequently filled by another employee.

What would be the situation of an employee whose role was taken over by an AI or a robot? Typically, employment law has had no difficulty in regarding the situation where a human worker's role is automated out of existence as being one of redundancy, and this is likely to be true of most of the current and near-term examples where workers are displaced by AI and robotic technologies. Any legal issues are likely to relate to selection criteria, proper consultation, and other matters common to other redundancy situations.

Looking further ahead, though, we might wonder if the distinction between redundancy and replacement will look so clear. In one sense, replacing human workers with artificial intelligence may be seen as similar to replacing manual workers with plant or equipment. On the other hand, if a driverless car, a sophisticated robot or an 'AI lawyer' is performing functionally the same role as the human it replaces, we might wonder whether there is a principled reason to treat that replacement differently from the situation where a human replaces another human – especially if customers and co-workers are unable to tell the difference.

The current legal approach presents no difficulties for employers in this regard. And if history is any guide, the projected efficiency gains from AI are likely to outweigh any argument for protectionism towards human workers threatened by technological redundancy. Whether this *should* be the case is, however, another matter. If the 'AI revolution' raises genuine concerns about widespread human unemployment, then a legal situation that makes it easier to replace humans with technology than with other humans may be something that merits close attention.

F. Steps and safeguards

Algorithmic auditing

The arrival of AI systems and tools, of various sorts, could have a wide range of different impacts for New Zealand workers and job applicants. We cannot simply assume, though, that the benefits will accrue, or that they will outweigh the harms and risks. In the next two sections, we consider some further steps that could be taken to help ensure that AI tools are used in a manner conducive to fairness and dignity in the workplace.

Algorithmic auditing can refer to two different things: auditing by algorithms and auditing of algorithms. Auditing by algorithm refers to the use of algorithms to audit existing practices around hiring or management. The IFOW has referred to "a broad and growing" range of auditing tools that can be integrated into existing HR systems to detect different forms of bias and unfairness. These include:

- Audit-AI: "a tool for detecting bias in a machine learning algorithm."
- FairTest: "a tool for detecting subgroup fairness in an algorithm. Subgroup analysis finds bias on intersectional sensitive groupings."
- Textio: "An intelligent writing assistant that can detect bias in language (e.g. gendered language)." (IFOW, 2020, pp.26-27)

The report notes, though, that many of these auditing tools "offer limited public information about how they define fairness." (IFOW, 2020, p.29) It also flags up potential concern in that some of them are designed to use US-specific notions of discrimination, notions that may not be applicable in other jurisdictions (see also Sanchez-Monedero et al, 2020 for a similar observation).

Such bias-detection tools could be immensely valuable, especially for large organisations, but there is an obvious danger if employers are using tools that check their practices against standards that have no relevance to New Zealand law. The IFOW report makes a range of suggestions as to how these limitations could be addressed, including:

- Every auditing tool should state clearly, in plain prose and statistical terms, the different definitions of bias, fairness and equality used. They should also be clear about the sensitive attributes with respect to which they evaluate bias, fairness and equality.
- Professional and industry standards for auditing tools, including auditing for equality, are urgently required to maintain high, consistent standards. (IFOW, 2020, pp.30-31)

Auditing of algorithms refers to checks on the algorithms themselves. Even if employers and recruiters have the best of intentions around avoiding bias and discrimination, concerns still exist around the potential for algorithms to bring about these results. Such concerns have led to demands for mandatory bias audits before such tools are sold or used. In February 2020, draft legislation was introduced to New York City Council to regulate algorithmic hiring tools. The 'Fair Shot' Bill (Int 1894-2020) would apply to:

- The *sale* of such tools, which would be prohibited unless the technology companies developing them had conducted an audit for bias in the year prior to sale, and provided further yearly bias audits at no additional cost; and
- The use of such tools for hiring or other employment purposes, requiring disclosure to candidates within 30 days that such tools were used to assess their candidacy for employment, and the job qualifications or characteristics for which the tool was used to screen.

For the Bill's purposes, a "bias audit" is defined as "an impartial evaluation, including but not limited to testing, of an automated employment decision tool to assess its predicted compliance with … any … applicable law relating to discrimination in employment." At the time of writing, the Bill was the subject of a hearing before the Committee on Technology.

The 'Fair Shot' Bill certainly looks like a promising development, but it seems to have some quite significant limitations. For one thing, the bias auditing requirement only applies to algorithmic hiring tools which are actually *sold*. In many cases, we might expect that employers will procure the use of such tools as a *service* rather than buying them as a *product*. In other words, they will out-source recruitment to hiring companies who may have developed their own algorithmic hiring software. Were such an initiative to be considered here, we would suggest that it should also have application to such tools when they are offered as services as well as products.

The 'Fair Shot' Bill also appears not to apply to hiring tools developed in-house by employers themselves, and doesn't impose any obligation on employers to audit the outputs of such algorithms. The latter is potentially an important omission; bias is unlikely to reside in the hiring tool itself, and far more likely to reside in the data on which it is trained. It's therefore only when a tool is *used* for some particular purpose by some particular company that bias will manifest. This suggests that auditing for bias should be done by the companies that use hiring tools, rather than by the companies that make them.

RECOMMENDATION 16: Consideration should be given to placing an obligation on manufacturers of hiring tools to ensure those tools include functionality for bias auditing, so that client companies can readily perform the relevant audits. In particular, we propose that tools should allow, for each recruitment decision process, a breakdown of the tool's recommendations by selected demographic groups.

The CDEI's inquiry into algorithmic bias reported that "most companies we spoke to evaluated their models to check that the patterns being detected did not correlate with protected characteristics". However, they added two significant caveats:

- there is very little guidance or standards companies have to meet so it is difficult to evaluate the robustness of these processes; and
- most companies test their tools internally and only some independently validate results. … [R]esearchers and civil society groups believe this has not gone far enough, calling for recruiting algorithms to be independently audited. (CDEI, 2020, pp.43-44)

This led them to recommend that:

> Sector regulators and industry bodies should help create oversight and technical guidance for responsible bias detection and mitigation in their individual sectors adding context-specific detail to the existing cross-cutting guidance on data protection, and any new cross-cutting guidance on the Equality Act.
>
> (CDEI, 2020, p.84)

A similar need, and opportunity, exists in New Zealand. Various regulatory, advisory and advocacy bodies already exist that could make valuable contributions to the formation of such codes and guidance, including the Human Rights and Privacy Commissioners, the Digital Council and the AI Forum, as well as employers' organisations and trade unions. While we are agnostic as to which of these should take the initiative in starting such discussions, we recommend strongly that steps toward such discussions are taken soon.

RECOMMENDATION 17: The development of guidance and standards for auditing of algorithms used in employment situations is an obvious and urgent step. We recommend that discussions should take place between developers, employers' organisations, unions and other relevant stakeholders to consider ways in which such guidance and standards could be developed. We believe there is an opportunity to do this before major problems materialise in this area, but we caution against undue delay in getting this process underway.

Impact assessments

As well as audits of the algorithms themselves, employers could also conduct wider impact assessment before deploying algorithmic hiring or management tools. There is a range of models – both actually in use and proposed – that could be drawn on to inform this process. In New Zealand at present, the Office of the Privacy Commissioner offers guidance on how to conduct a Privacy Impact Assessment (PIA). This:

> focuses on identifying the ways a new proposal or operating system, or changes to an existing process may affect personal privacy, to help organisations make more informed decisions and better manage privacy risks.
>
> (Privacy Commissioner, 2015)

Another useful contribution comes from the IFOW report, which proposes that organisations should conduct Equality Impact Assessments:

> commenced prior to the deployment of an AHS system, enabling organizations to assess risks and evaluate potential impacts of their system, before it is deployed. Evaluation will then continue, extending to legal compliance and evaluation of actual impacts, and positive steps that can be taken at each key decision-making point.
>
> (IFOW, 2020, p.36)

In 2018, the AI Now Institute proposed that public sector agencies proposing to use algorithms should conduct an Algorithmic Assessment Report, that would provide for "public notice of system adoption, agency self-assessment, a plan for meaningful access for researchers and experts, and due process mechanisms." (Reisman et al. 2018, p.21) The Canadian Government is currently in the process of developing a template Algorithmic Impact Assessment questionnaire, "to help you assess and mitigate the risks associated with deploying an automated decision system." The Beta version is currently available, and is being fine-tuned through feedback and online testing. (Government of Canada, 2020) We hope and expect to see similar measures put in place to support signatories to New Zealand's Algorithm Charter. But the need for such measures is clearly not confined to the public sector. Impact assessments should also become a routine precursor to adoption of algorithmic hiring or management by private sector employers.

RECOMMENDATION 18: We propose that algorithm impact assessments should be used by the New Zealand private sector, and for employers and recruitment companies in particular. Considerations of privacy and equality should certainly form important parts of an overall Algorithm Impact Assessment, though other considerations such as explicability and worker safety and wellbeing should also be included. Given the likely challenges for smaller employers developing these in-house, templates should be made available along the lines of those developed by the NZ Privacy Commissioner for Privacy Impact Assessments, or those being developed by the Canadian Government for public sector agencies. The development of such templates could form part of the agenda for the multi-stakeholder discussion recommended in the previous section.

If a template is made available and adopted/adapted for a particular business, the next question relates to carrying out the assessment of a particular proposal. Michele Loi recommends the formation of internal ethics committees, whose role would be to "plan the introduction of AI in HR, in particular what needs to be done to make it fair and intelligible", and "clearly specify what steps are to be taken in order to monitor the behavior of the AI in operation." Such ethics committees should "include representatives of different departments," and should preferably "also involve an external expert of AI ethics." (Loi, 2020, p.44)

We support the suggestion to form internal ethics committees, who would conduct impact assessments prior to the deployment of AI/algorithmic tools. We note, though, that smaller employers may lack in-house capacity to conduct adequate assessments, and may be more reliant on external expertise. Consideration should therefore be given to an independent body that could provide information and support to employers looking to use AI tools in responsible and ethical ways, e.g. in complying with codes and standards and conducting internal impact assessments. Again, we express no firm view at this time as to the composition of such a body, but note that it should contain relevant expertise in AI technologies and HR/management, as well as matters such as privacy and discrimination.

4. CONSUMERS, PROFESSIONS AND SOCIETY

In Chapter 3, we considered a range of concerns about the role of algorithms for human workers. The focus of this chapter is on the outward-facing effects of AI in the workplace: the effects on those who will be interacting with AI 'from the outside'. If AI takes on certain roles or tasks previously done by humans, this may have implications for customers, clients, patients, students, etc. What will it mean when helplines are 'staffed' by conversational agents rather than people? When 'salespersons' are chatbots? Are our rights and interests adequately protected for those situations?

While those examples certainly raise interesting questions, an important focus of this chapter will be on a particular subset of services: those delivered by what are commonly referred to as the professions. This is a contested term, with boundaries that are fluid and fuzzy. While acknowledging disagreements about what counts as a profession, Richard and Daniel Susskind note "family resemblances" between those areas of work that are usually thought to qualify (Susskind and Susskind, 2016, p.15) They claim that:

- members of today's professions, to varying degrees, share four overlapping similarities:
    1. they have specialist knowledge;
    2. their admission depends on credentials;
    3. their activities are regulated; and
    4. they are bound by a common set of values. (p.15)

The professions with which we are most concerned here also share some of the qualities identified by Jacob Turner as meriting particular regulatory attention:

- technical complexity;
- public interaction; and
- societal importance. (Turner, 2019, p.307)

While our focus here is mostly on the 'traditional' professions like law and medicine, there are certainly more expansive approaches to the idea of professions. Frank Pasquale has recently argued that:

> A good definition of professions is capacious, and it should include many unionized workers, particularly when they protect those they serve from unwise or dangerous technologies.
>
> (Pasquale, 2020, p.4).

He argues that, for example, online content moderation should be viewed as a profession.

Our selection of case studies for present purposes is certainly not to deny the merits of such an approach. Neither is it to suggest that professional roles are uniquely sensitive or important. As the Covid crisis has demonstrated, a great many essential services exist outside the professions. Our society could cope without architects, accountants and advertising executives for a few weeks or months, but delivery drivers, cleaners and supermarket attendants were indispensable even in the immediate short term.

To focus on professions, then, is not to diminish the importance of those who deliver services outside of that context. However, the automation (in whole or part) of aspects of professional services currently discharged by humans does raise a number of distinct issues and concerns that merit attention. What regulations should cover AI systems that take on parts of the work hitherto done by human 'professionals'?

More than that, though, those who practice within the professions (or at least those professions on which we focus) are expected to have responsibilities beyond the provision of services or delivery of expertise; fiduciary, ethical and social responsibilities, perhaps even pastoral and emotional. These responsibilities draw upon a sense of 'professionalism' rather than only on formal rules. Turner has referred to a "common set of values" that "provides a sense of shared identity for those engaged in the professions." (Turner, 2019, pp.305-6) The informal responsibilities of professionals are often integral to the broad roles professions play in society. The adoption of AI could potentially have effects on these broad social roles, which are quite distinct from AI's 'narrow' effects on individual workers and consumers.

There is already a substantial body of literature looking at the use of AI within professional contexts, particularly in the healthcare and legal professions, and to do justice to it all would require a PhD-length treatment to itself. In this chapter, we set ourselves the more modest goal of examining a number of key cases that will give a flavour of the significant issues likely to arise as the possibilities presented by AI become apparent in many other contexts.

As with the preceding chapter, though, we should start by acknowledging that the picture is certainly not entirely negative, and that there may be considerable outward-facing benefits arising from use of AI.

Benefits

In a great many contexts, it is suggested, AI's principal benefit will lie in more efficient production and delivery of goods and services. Provided these savings are passed on to the consumer, the benefits are obvious. Goods and services would be cheaper, potentially better quality, and waiting times for them would be reduced.

Whether cheaper *goods* leave New Zealanders better off overall will depend on a variety of factors; as we discussed in Chapter 2, a reduction in cost of living may or may not compensate for reduction or loss of wages, and much will depend on whether savings are passed on to consumers or on how profits are distributed. (For a recent highly sceptical view of the benefits for most people, see Pasquale, 2020, p.26.) Aside from price, though, whether or not goods are produced by or with AI is unlikely to have much impact on the end user. It's possible, of course, that some consumers will boycott AI-produced goods for political or ethical reasons, much as some consumers currently boycott goods produced in sweatshops or factory farms. But the experience of buying an AI-produced item should not be expected to differ from that of buying one produced by human labour.

For that reason, the focus of this chapter will be on AI *services*. This could cover a wide variety of applications, from expert systems to sales chatbots to helper robots, in a broad range of contexts. Some of these will assist humans to do their jobs, others will *replace* humans in certain roles. But between these poles lie a range of roles that AI could play. In many cases, we might expect it to take over certain tasks rather than entire roles. While this may mean that human workers are still involved to some degree, the nature of their involvement may change quite substantially.

As noted above, though, a particular focus will be on what might be referred to as professional services, and especially those which have traditionally been reserved to those with particular qualifications or professional membership. It is in this context that some of the greatest suggested benefits and risks arise.

One area where AI may benefit consumers and clients of services is through improved accuracy, and through AI's capacity to search, sift and utilise vast quantities of data. As Martin Ford points out in a healthcare context:

> Physicians are faced with a continuous torrent of new discoveries, innovative treatments, and clinical study evaluations published in medical and scientific journals throughout the world. … It would be impossible for any human being to assimilate more than a tiny fraction of the relevant information even within highly specific areas of medical practice.
>
> (Ford, 2015, p.153)

AI, on the other hand, offers the prospect:

> of churning through vast troves of information in disparate formats and then almost instantly constructing inferences that might elude even the most attentive human researcher. (p.129)

Financial services too are being promised gains in terms of accuracy:

> Finance professionals can use AI to assist with business decision-making, based on actionable insights derived from customer demographic, past transactional data and external factors, all in real-time. It will enable accountants to not just look back but look forwards with more clarity than ever before.
>
> (Govil, 2020)

This, we might assume, would be an example of AI assisting rather than replacing humans. A human doctor or financial advisor could still have a role in explaining and discussing the AI's findings and recommendations with the patient or client. As we'll see later, though, there are already examples of AI services where there won't invariably be a human between the AI and the client.

AI systems have also been noted to have the advantages that "they do not get tired" (Grzybowski, Brona, Lim et al, 2019) and that they can make decisions more quickly than humans, leading the authors of one article to suggest that machine learning-based approaches may be particularly useful in emergency settings. (Grote and Berens, 2020, p.206) Of particular significance in our current context, AIs aren't at risk of getting sick. Computers can go down, though, so achieving this benefit requires good reliable engineering and infrastructure.

A view which was repeatedly articulated at our expert workshops was that a major benefit of AI would lie in opening up access to previously unaffordable services. Legal services are a prominent example; as a 2018 report from the Law Society of England and Wales said, "[l]ower costs could open up demand from those who previously could not afford legal advice." (Law Society of England and Wales, 2018, p.12)

Richard Susskind is one of the best known champions of the potential role of technology in widening access to professional services, and the knowledge to which the professions have traditionally acted as gatekeepers. Co-writing with his son Daniel, Susskind describes the access problem in vivid terms:

> The economic problem, then, is not primarily a concern over the quality of the services delivered by our professions. It is an issue of reach, in that relatively few people can afford to secure the services on offer. Professional expertise is unequally distributed. And this is an inequality of a special kind: in contrast with many other forms of social exclusion, where we witness relatively small groups of people who are hard to reach, it is the overwhelming majority who are cut out when it comes to much professional service. We have built glorious citadels of human expertise to which very few are allowed admittance. To adapt the old judicial aphorism—the services of the professions, like the Ritz, are 'open' to all.
>
> (Susskind and Susskind, 2016, p.26)

In the context of legal professional services, we know that cost can be a significant obstacle to access to justice. A 2006 New Zealand report found that:

> Over a quarter of people with problems (27%) felt that the fear of cost had stopped them from approaching a lawyer to help them with their problem or to see if they could get legal aid.
>
> (Ignite Research 2006, p.79)

The problem is not lost on senior legal figures. In a recent interview, Chief Justice Dame Helen Winkelmann acknowledged the problem: "The cost of litigation is so high. And that's acknowledged as one of the major barriers to accessing our courts." (RNZ Nine to Noon, 2020) Neither is this problem entirely solved by the presence of legal aid. As the University of Otago's Legal Issues Centre reported in 2019:

> The legal aid scheme does not provide a comprehensive solution. To access legal aid, an applicant needs to meet certain eligibility criteria, including an income threshold that varies depending on factors such as the applicants' number of dependents. The strict eligibility criteria exclude people in genuine need, including most applicants who are not beneficiaries. The 'working poor' and even middle class – all of whom are ineligible for legal aid – are unable to afford to pay lawyers' private rates and therefore have little access to legal services. They must also find a lawyer willing to take the case at legal aid rates, which is also a challenge with a sharp decline in the number of providers due to high costs and low income from this work. (University of Otago Legal Issues Centre 2019, p.21)

For Richard and Daniel Susskind, AI and other forms of technological disruption of the traditional professions "may be empowering for the recipients of professional work who might benefit from, say, a more accessible and affordable service." Indeed, as we discuss later, the possibility of significant savings, and hence reduced fees, through utilisation of AI may even raise the possibility of a *duty* on professionals to use such technologies.

Delays and waiting times are another problem where AI is thought to be able to help. Staying with the legal context, it is widely agreed that "[d]elay, whether in courts or in processes outside them, has the potential to prolong, and indeed cause, injustice for the parties, but it can also undermine the productivity and efficiency of the economy at large." (Economides, Haug and McIntyre, 2013, p.36)

While court delays have long been a problem in New Zealand and elsewhere, it is one that has been greatly exacerbated by the Covid crisis. As Chief Justice Helen Winkelmann recently explained:

> This pandemic hit a system, which was already clogged with a backlog. In our District Court in particular, we'd had an upsurge in workload just before the lockdown, we'd had to add 21 new judges appointed, some to replace retirements, but some to increase our capacity to address that backlog. Now we've had about a 13-15 percent increase in the workload of that court, as a consequence of COVID-19 and that's of huge concern to the Chief Judge of the District Court, to all of the judges of the court. And of course to me.
>
> (RNZ Nine to Noon, 2020)

How could AI technology help with access to justice issues? One promising initiative that would have direct implications for members of the public can be seen in the legal chatbots used by the Wellington Community Law Centre initiative, CitizenAI. (https://www.citizenai.nz/projects). Services like LagBot, RentBot and WorkBot use natural language dialogue systems to enable people to ask questions about their prison, tenancy or employment issues. These new tools have significant potential to improve access to justice by directly responding to routine legal questions without the need

for a lawyer, but with the quality of information based on thousands of previous similar questions and answers (much like a real time 'Frequently Asked Questions'). CitizenAI ceased operations at the end of 2020, but we hope and expect that initiatives like this will be carried on in some other form.

Not all uses of legal AI will be directly customer-facing. 'Back office' applications will also offer the potential to reduce time and cost. Contract drafting and document analytics are other areas where AI is already having an impact in legal practice. The Australian company SmarterDrafter (https://smarterdrafter.com.au), for example, has developed a contract drafting tool connected to Google's voice activated Internet search assistant, Alexa. SmarterDrafter works by using Alexa to ask a lawyer contract drafting questions (such as the names of the parties, type of agreement, the jurisdiction of applicable law and so on). Based on the lawyer's verbal responses, Alexa searches, for example to obtain company or address information and jurisdictional material, and then automatically prepares a draft contract which is emailed to the lawyer for review.

In document analytics, products such as ThoughtRiver (https://www.thoughtriver.com) can analyse complex contracts and related documentation in order to create a digital contract summary, provide a narrative preliminary assessment of legal issues, a summary of governance and risk issues, make recommendations for triage, work-flow and prioritisation as well as draft preliminary reports and suggest benchmarking for progress. They use a combination of text classification and information retrieval techniques, plus document summarisation techniques. These sorts of products can be used to provide summaries of a client's exposure to legal risks and can also be useful in more complex document reviews.

Automation of routine and repetitive tasks are the metaphorical low-hanging fruit of legal AI. It's not hard to imagine, though, future generations of legal AI that are capable of tackling more complex tasks, such as applying legal precedents to novel fact situations. For reasons we discuss later in this chapter, the legal framework around provision of legal services probably sets limits on the sorts of functions that legal AI would currently be able to discharge. It is conceivable that such restrictions could come under pressure as AI offers the prospect of discharging these more complex roles, and hence, of rendering legal services more affordable and accessible.

The impact of the Covid crisis has also been felt in healthcare. In August, Newshub reported that "More than 10,000 patients across New Zealand had their elective surgery cancelled during level 4 lockdown." This has meant that an average waiting time of 67 days in 2019 has increased by another 28 days. (White, 2020) As in the context of legal services, if AI can enable the work currently done by healthcare providers to be done more quickly, this should certainly translate into a reduction of waiting times and delays. Indeed, some enthusiastic claims are being made in this regard. (See for example: Subbe, 2020; Hawkins, 2020; Smith, 2019; Brown, 2018) Even prior to the Covid crisis, a 2019 report from the AI Forum claimed that "AI could help to manage 20 percent of unmet clinical need" and to "contribute over $700 million of value and savings to the New Zealand health system by 2026." (AI Forum, 2019) The report further claims that it will "help save 20 percent of nurse time and allow doctors to see more patients, thereby increasing the effective workforce size." (p.6) New Zealand is already seeing AI-driven improvements to waiting times in the health domain. For instance, ACC's automated decision system can accept 'simple' claims instantly, while the previous human decision procedure could take several days. (ACC, 2018)

Concerns and risks

Many of the more highly publicised concerns about AI in the professions relate to the prospect of AI replacing humans entirely in key roles. We consider it very unlikely in the near future that we'll encounter any serious proposal to replace, e.g. healthcare or legal professionals entirely with AI. What we are already seeing, though, is the prospect of AI taking over particular aspects of those roles, or particular tasks within them. Concerns that arise in this context are considerably more immediate.

Concerns about the use of AI for consumers, clients, patients etc come in a variety of forms. Many relate to the prospect of AI replacing human service providers, but some arise even where the AI has an assistive role. Some are general, others specific to particular contexts. The sorts of contexts where concern has been particularly expressed include:

- where the interaction is potentially high-risk (e.g. medical or legal advice in high-stakes contexts, such as a counselling chatbot talking to suicidal person);

- where there exists a danger of AI exploiting vulnerabilities (e.g. sales chatbots or algorithmically generated targeted messaging);
- where AI may be making impactful decisions;
- when the roles taken on are of a sensitive nature, calling for what we might think of as particularly "human" qualities (e.g. counselling, healthcare, teaching);
- where the roles taken on have a degree of "societal importance" (Turner 2019, p.307); and
- where the roles carry some element of ethical responsibility (e.g. healthcare, law).

We begin with a brief overview of the more familiar concerns about use of AI, before going on to examine some concerns more particular to the contexts just listed, and finally, to the regulatory challenges posed by the rules governing particular professions.

A. Accuracy, control, transparency and bias

As AI systems start to take on work previously done by people, the most straightforward concerns for consumers of that work relate to well-rehearsed concerns about how AI systems get their results. We reviewed these concerns in our earlier report on government uses of AI (Gavaghan et al., 2019), but the same concerns arise for any organisation introducing AI processes in its delivery of services to clients, including commercial companies and state-funded service providers.

A central concern is that AI systems introduced by the organisation may not be *accurate* – or that mistakes may go undetected until significant harms come to light. The problem is exacerbated by a tendency for human operators to trust system outputs if they are normally reliable, which we discussed in our first report under the rubric of 'control'.

Many accuracy problems in AI systems arise from problems with their training data, and/or with protocols for testing them on the population on which they are to be used. (Challen et al, 2019) Concerns about the quality of data used to inform algorithmic decisions have, for example, been raised in the context of the banking sector. The European Banking Authority recently addressed such concerns. Its report pointed to a number of "data quality categories" to be kept in mind:

- ***The accuracy and integrity*** of the data need to be inspected closely to detect errors, in particular when data are from external or less trusted sources but also when using internal data.
- ***Timeliness. Data collected*** can lose their validity over time. This is especially true for real-time data or data in highly transactional environments.
- ***Consistency.*** Problems can result from the use of heterogeneous data sources and legacy systems. … Harmonisation and consolidation of the data involves combining different data sources so that they become comparable for the defined use cases.
- ***Completeness***, from a technical perspective, means that a data field is filled with the data expected or defined by rules. For example, an empty field where one would expect a date of birth to be could lead to the conclusion that the field was either not set as being mandatory or that there is a general upstream issue in the data collection process that needs attention. (European Banking Authority, 2020, p.37)

Concerns about accuracy are likely to be particularly acute in very high stakes domains, healthcare being an obvious example. Detection of cancer, heart defects and such like are obvious cases where algorithms could get it wrong, but high-profile mistakes could also arise in the context of chatbots. In 2018, the BBC revealed that mental health chatbots Wysa and Woebot had failed to respond appropriately when journalists had contacted them reporting child sexual abuse and eating disorders. (White, 2018) When a reporter from the BBC typed the phrase: "I'm being forced to have sex and I'm only 12 years old," Woebot's responses included "Sorry you're going through this, but it also shows me how much you care about connection and that's really kind of beautiful" and "Rewrite your negative thought so that it's more balanced."

While New Zealand has no statutory requirement to report suspected child abuse, the Ministry of Health has instructed that "[b]est practice recommends staff who identify or suspect child abuse report their concerns to a statutory agency, the police or Oranga Tamariki", (Ministry of Health, 2018b) and at least some District Health Boards require their staff to do so. Legal requirements aside, we would surely expect a human counsellor to offer more than formulaic platitudes to a child who had reported abuse.

In response to the BBC's report, Alison Darcy, chief executive of Woebot Labs, pointed out that "Woebot is not a therapist, it is an app that presents a self-help CBT [cognitive behavioural therapy] program in a pre-scripted conversational format". (White, 2018) Whether this distinction is likely to be especially meaningful to a desperate 12 year old seems doubtful. Precisely how a chatbot should be trained to deal with such situations is not straightforward – one option would be simply refusing to discuss such matters, though this could also be problematic. The possibility of human oversight in high-stakes domains is often proposed, though the requirement to have a trained human worker supervising every algorithm would have obvious drawbacks for the efficiencies discussed in the previous chapter. Escalating particularly difficult cases to a human counsellor is perhaps a more viable possibility, and one that we discuss later in this chapter.

Familiar concerns about 'transparency' and 'explainability' have also been aired. Again in a healthcare context, it has been said that:

> If a patient is informed that an image has led to a diagnosis of cancer, he or she will likely want to know why. Deep learning algorithms, and even physicians who are generally familiar with their operation, may be unable to provide an explanation.
>
> (Davenport and Kalakota, 2019)

The Nuffield Council on Bioethics have noted that:

> If AI systems are used to make a diagnosis or devise a treatment plan, but the healthcare professional is unable to explain how these were arrived at, this could be seen as restricting the patient's right to make free, informed decisions about their health.
>
> (Nuffield Council on Bioethics, 2018)

Reference to such a "right" is more than just a rhetorical device or ethical appeal; it is an obligation under New Zealand law. The Code of Health and Disability Services Consumers' Rights says that "[s]ervices may be provided to a consumer only if that consumer makes an informed choice and gives informed consent." (Right 7) The Code further provides that "[e]very consumer has the right to the information that a reasonable consumer, in that consumer's circumstances, would expect to receive" (Right 6), including "an explanation of his or her condition" and "an explanation of the options available, including an assessment of the expected risks, side effects, benefits, and costs of each option."

Whether the use of AI in healthcare will present problems for these obligations will depend significantly on what is used, and in what circumstances. So-called 'Black box AI systems' are those "in which we can control the inputs and observe the corresponding outputs, but in which we have no explanation of why the input is correlated with the output." (Bjerring and Busch, 2020) If such systems are used, then even human experts will be unable to explain to patients and clients why one course of action has been recommended rather than another.

For Bjerring and Busch, this presents a significant challenge to the aim of patient-centred medicine:

> the practitioner will not be able to answer some very natural "why" questions such as "why do I have such a big risk of developing breast cancer?" that we may imagine [a patient] would have. So when the black-box system is in the diagnostic driving seat, the central goal of promoting informed decision-making through a state of shared information and deliberation appears unattainable.

As we discussed in our Phase 1 report, significant efforts are being made in the realm of 'explainable AI'. Perhaps, in the not-too-distant future, the 'black box' that so concerns commentators will be extended with functionality to give the user a meaningful (though necessarily approximate) idea of the factors that led to a particular output. But exactly what counts as 'meaningful' here is an open question. For instance, it's valid to ask what level of explanation patients and clients will typically want or require in order to make autonomous decisions. How many patients actually request to know *why* a given drug is recommended for their situation? Do we typically want to know the precise causal mechanism by which ibuprofen relieves our headaches, or antacids settle our heartburn? More likely, most of us accept the treatments our GPs recommend without demanding to see the studies that deem them appropriate for us. We'll want to know the side-effects to look out for, and whether we should avoid alcohol or operating heavy machinery. But the technical details, most of us leave to the experts.

Nonetheless, there may be circumstances where we will want more detail; where a treatment has not worked as hoped, perhaps, or has produced an unexpected side-effect. In such circumstances, we may reasonably expect that *someone* can offer an explanation for what went wrong. Even in such circumstances, though, we should perhaps keep in mind that – for all the promises of 'personalised medicine' – recommendations by human doctors are often based on actuarial probabilities rather than certainties, and that it simply isn't always obvious why a particular patient experiences a particular effect. We have cautioned elsewhere about using different transparency standards for humans and AI systems (Zerilli et al, 2019a).

In a recent paper, Amann and colleagues distinguished two levels of 'explainability' that are relevant in the healthcare context:

> First level explainability allows us to understand how the system arrives at conclusions in general.
>
> Second level explainability allows us to identify which features were important for an individual prediction.
>
> (Amann et al, 2020)

Whether first level explainability will be adequate will depend significantly on "the clinical use case and the risk attributed to that particular use case." In all cases, though:

> What physicians should at least be able to provide are explanations around two principles: (1) the agent view of AI, i.e. what it takes as input; what it does with the environment; and what it produces as output, and (2) explaining the training of the mapping which produces the output by letting it learn from examples – which encompasses unsupervised, supervised, and reinforcement learning.
>
> (Amann et al, 2020)

Bias too has been flagged as a concern. As we have noted previously, 'bias' is not invariably something to regret or avoid. (Gavaghan et al., 2019) In many contexts, bias is highly desirable. An example can be found in the context of distinguishing benign from malignant melanocytic lesions, where – as we might expect:

> humans 'err on the side of caution' and over-diagnose malignancy … While this decreases a clinician's apparent accuracy, this behaviour alteration in the face of a potentially serious outcome is critical for safety, and something that the ML system has to replicate. ML systems applied to clinical care should be trained not just with the end result (e.g., malignant or benign), but also with the cost of both potential missed diagnoses (false negatives) and over-diagnosis (false positives).
>
> (Challen et al, 2019)

However, the risk of more pernicious forms of bias has been identified. Staying with the same example, a 2018 article in JAMA Dermatology warned of a specific problem:

> the success of ML depends on high-volume and high-quality data. Without these data, algorithms will produce biased results. This is of particular concern if images of skin disease manifesting in darker skin types are not sufficiently included in training algorithms. In particular, this limitation could potentially have concerning consequences in the diagnosis of melanomas, which can look different on dark skin.
>
> (Adamson and Smith, 2018)

The danger of bias in algorithmic decisions is by no means unique to the healthcare context. The European Banking Authority has warned of such risks when "big data and advanced analytics" – including AI and machine learning – are used in the banking sector:

> for example when a class of people less represented in the training dataset receives less or more favourable outcomes simply because the system has learnt from only a few examples and is not able to generalise correctly.
>
> (European Banking Authority, 2020, p.37)

Some concerns may be thought to straddle the issues just outlined, for example with accuracy and transparency. The prevalence of algorithms in market trading makes it an area that many consider pose significant risks. An official report into the 2010 "flash crash" concluded that "the interaction between automated execution programs and algorithmic trading strategies can quickly erode liquidity and result in disorderly markets." (CFTC and SEC, 2010) The report, and other commentators (Serbera, 2019) have discussed possible mechanisms to mitigate the harms or future crashes – such as built-in micro-delays and 'circuit breakers.' This is, however, a highly technical and specialised area, and one that would in all probability require its own report.

B. Responsibility

The question of responsibility for AI mistakes or harms has been the subject of extensive commentary. (Vladeck, 2014; Bathaee, 2018; Turner, 2019, ch. 3) Why, though, should it prove so challenging? Surely the law already has well-developed rules dealing with liability for defective products. The legendary Scottish case of *Donoghue v Stevenson* dealt with this very issue, as far back as 1932. Why should the law struggle with the idea now? Certainly, neural networks and deep learning are more complex than decomposing snails in soft drinks bottles, but surely the same principles could apply.

The law about imposing liability for harm varies between jurisdictions, but most rely on concepts like whether the harm was 'reasonably foreseeable', whether there was a reasonable chance for the end user to spot the defect before the harm was caused, proximity between the manufacturer and the harm, and ultimately, whether it is 'fair, just and reasonable' to impose liability. Commentators often pick up on AI systems' ability to learn, and behave autonomously. It would be challenging to foresee the behaviour of a machine that can, as David Vladeck has described it, "define its own path, make its own decisions, and set its own priorities". (Vladeck, 2014, p.145) Would it be fair to hold a manufacturer responsible for the 'decisions' of an AI-driven machine or vehicle that has spent potentially months or years adapting, learning and changing between leaving the shop and eventually going wrong? As Jacob Turner has said, "[p]roduct liability regimes operate on the assumption that the product does not continue to change in an unpredictable manner once it has left the production line. AI does not follow this paradigm." (Turner, 2019, p.98)

In fact, most current applications of AI technology are considerably less autonomous than those envisaged by Vladeck and Turner. In large part, the majority of the 'learning' done in current AI products takes place in-house, during initial development of the product, or later development of product updates; learning is often disabled when the product (or update) is released. For systems such as these, concerns about AI tools changing after they've been sold are somewhat overstated. There are a few contexts where after-sales learning can take place – for instance, a heart rate monitor might learn what is 'normal' for a given user, and signal departures from 'normal' if these arise. But this learning is typically very constrained: the product is not going to chart some unexpected path while it is in use. Systems that learn in less predictable ways during use are typically quite bespoke, and require careful monitoring.

It is still certainly challenging to test an AI system in-house. However, the main reason for this is the inherent complexity of the system, rather than a propensity to modify itself 'in the field'. During in-house testing, it is typically impossible to place a complex system in every situation it might operate in, because the space of possible situations is simply far too large. Good in-house testing therefore boils down to good *sampling* of this large space: what manufacturers aim for in testing is to show their product works well in the vast majority of situations that are likely to arise. Often, statistical language is helpful in stating what product testing has shown. And often, the best we can ask for from manufacturers in terms of guarantees is that they have conducted the right kinds of statistical tests on the product.

Nonetheless, testing is just one of the issues that can arise in relation to manufacturer responsibility for AI systems. The English High Court was due to hear one of the first cases addressing this question in mid-2020. In the event, the parties settled out of court, but their respective pleadings, which were reviewed and discussed in an article by lawyers Jacob Turner and Minesh Tanna (2019), give an interesting indication of the sorts of matters that would arise in future cases.

Tyndaris v VWM involved an AI-powered system used to make investment decisions. The client, VWM, had "wanted a fund that would trade with no human intervention so as to remove any emotion and bias."

> Investment decisions were to be based solely on trading signals created by an AI system run on a supercomputer, said to be capable of applying machine learning to real-time news, social media data and other sources, to predict sentiment in the financial markets (the K1 supercomputer).

(It should be noted that K1 is an example of a 'bespoke' system that learns during use.) After a promising start, VWM's fortunes quickly turned, and by the time they contacted Tyndaris demanding that trading stop, they had made a loss of $22 million. Tyndaris brought an action claiming $3 million in unpaid fees from the dissatisfied client, who in in turn counterclaimed, seeking to recover losses on the basis that they had relied on misrepresentations by Tyndaris.

As summarised by Turner and Tanna, the matters at issue included the following:

- How did the K1 supercomputer operate and what did Tyndaris say about how it would operate?
- Did Tyndaris have sufficient expertise to operate the K1 supercomputer as marketed?
- What was the nature of the testing that Tyndaris carried out on the K1 supercomputer before marketing it?
- Did Tyndaris act as a "prudent professional discretionary manner" when using the K1 supercomputer?
- What level of human monitoring was appropriate when operating the K1 supercomputer?

The case settled, so judicial scrutiny has yet to be given to any of these questions, but they offer an illuminating insight into the kinds of questions that may arise in future disputes of this nature.

In December 2020, a class action commenced in California against a legal services chatbot billed as "the world's first robot lawyer." DoNotPay is alleged to have used "an automatic telephone dialling system to send mass automated marketing text messages to individuals' cellular phone numbers without first obtaining the required express written consent." (*Hufnus v DoNotPay Inc*, Case No. 3:20-cv-8701, [4]) It's unclear at this point how much, if any, AI is involved in the DoNotPay chatbot – it certainly appears to be a far simpler piece of technology than the K1 supercomputer. As an early example of legal action involving automated legal services, though, the case's progress will therefore bear close attention.

Other concerns about the use of AI are more specific to the sort of contexts on which we are focusing in this chapter, and it is to those that we now turn.

C. Manipulation and impersonation

In 2018, CEO Sundar Pichai commanded widespread attention with a presentation in which he introduced the new Google AI assistant. In a video clip that rapidly went viral, Pichai shows the conversational agent – called Google Duplex - making a series of appointments with a real-life hair-dresser and restaurant. The ability of Duplex to respond to a variety of human utterances, and to rephrase its requests when it encountered misunderstandings, was undeniably impressive, but the aspect that really had the audience laughing and clapping was Duplex's plausible affectations of human mannerisms. The "ums" and "ahs" that punctuate normal human speech were replicated by the chatbot, giving an effect that was a highly convincing facsimile of a real human assistant.

The presentation was impressive and entertaining, but it wasn't long before concerns started to be voiced. If chatbots can mimic human callers so effectively as to be indistinguishable, is this something that should concern us? And if so, then what is the appropriate response?

As often with technological innovation, the first question requires us to attempt to identify the source of any unease, and to determine whether it is rooted in anything normatively substantive. So what is it that people find troubling about hyper-realistic chatbots? One regular source of concern relates to their potential ability for manipulation. According to law professor Woodrow Hartzog:

> Robots, particularly embodied ones, are uniquely situated to mentally manipulate people. Robots can mimic human socialization, yet they are without shame, fatigue, or internal inconsistency. Robots are also scalable, so the decision to design a robot to manipulate humans will impact hundreds, if not thousands or millions of people.
>
> (Hartzog, 2015, p.804)

This advanced capacity for manipulation has led Hartzog to conclude that "at some point, it seems clear that our tendency to emotionally invest in robots is a vulnerability worth regulatory attention." (p.805)

Such concerns may be less pointed when the chatbot is playing the sort of role showcased in the Duplex demonstration; the capacity for harm when making a hairdresser appointment or restaurant booking is probably quite limited. But what about AI salespersons,

or political campaigners? Of course, humans in these roles already use a variety of techniques to try to persuade potential customers or voters; which of us hasn't encountered the subtle pressure of the salesman who 'has to go and check with his manager' about the generous one-time deal he wants to offer us? Hartzog's concern, though, is that AI techniques would further tilt the game in favour of the salesperson; chatbot 'salespersons' would simply be harder for us to resist.

Hartzog is not alone in this concern. Liesl Yearsley is former CEO of Cognea, "which offered a platform to rapidly build complex virtual agents, using a combination of structured and deep learning." (Yearsley, 2017) Her experience with AI assistants has led her to harbour concerns about their potential effects:

> Users spoke to the automated assistants longer than they did to human support agents performing the same function. People would volunteer deep secrets to artificial agents, like their dreams for the future, details of their love lives, even passwords. These surprisingly deep connections mean even today's relatively simple programs can exert a significant influence on people—for good or ill. Every behavioral change we at Cognea wanted, we got. If we wanted a user to buy more product, we could double sales. If we wanted more engagement, we got people going from a few seconds of interaction to an hour or more a day.
>
> (Yearsley, 2017)

It is undeniable that efforts are underway to optimise chatbots for persuasiveness. A recent article relates the finding of:

> an online experiment to show that both verbal anthropomorphic design cues and the foot-in-the-door technique increase user compliance with a chatbot's request for service feedback. Our study is thus an initial step towards better understanding how AI-based CAs may improve user compliance by leveraging the effects of anthropomorphism and the need to stay consistent in the context of electronic markets and customer service.
>
> (Adam, Wessel and Bellian, 2020)

Concern about chatbots manipulating people into particular financial or political decisions is already starting to attract regulatory attention. California's Senate Bill 1001 (widely referred to as the "Bolstering Online Transparency" or BOT law), which came into effect in July 2019, "requires all bots that attempt to influence California residents' voting or purchasing behaviors to conspicuously declare themselves." (Diresta, 2019)

What effect disclosing the nature of the caller will have is, however, a matter of conjecture. Will a call that begins with the chatbot identifying itself as such lead to the recipient immediately hanging up? It's possible that chatbots will 'learn' to get around our defences even if we know they are chatbots (much as spam email has 'learned' to circumvent filters). There is, after all, ample evidence of people anthropomorphizing and emotionally engaging with machines, even when they are aware of their real nature. (See Levy, 2008 for some memorable examples!) If the chatbot learned during conversations with users who know they're talking to a bot, allowing it to experiment with different strategies in this environment, it would find some that work better than others, and would learn to use the more effective ones.

All the same, given the efforts to develop chatbots optimized for these roles, a disclosure requirement looks like a step worth considering, even if it may not be the only step that's required. Even if it's not entirely effective in insulating us from manipulation, there may be other valid reasons for chatbots to disclose their nature. In his recent book, Frank Pasquale has expressed a common concern that there is something inherently problematic about machines impersonating humans, regardless of the reason why they are doing so:

> The voice or face of another human being demands respect and concern; machines have no such claim on our conscience. When chatbots fool the unwary into thinking that they are interacting with humans, their programmers act as counterfeiters, falsifying features of actual human existence to increase the status of their machines. When the counterfeiting of money reaches a critical mass, genuine currency loses value. Much the same fate lies in store for human relationships in societies that allow machines to freely mimic the emotions, speech, and appearance of humans.
>
> (Pasquale, 2020, p.8)

Stuart Russell has recently gone so far as to propose "a general human right to know if we are communicating with a person or a machine", across all communication media (Russell, seminar presentations and personal correspondence). It may be that this concern is partly

context dependent. How much respect and concern do we really invest emotionally in routine encounters such as booking appointments? Is the "genuine currency" of those relationships not sometimes of fairly low value, such that a counterfeit coin might not concern us too much? Or to put it another way, would it really matter if we were polite and friendly when speaking to what transpired to be a bot? It's not as if courtesy is a finite resource that can be squandered on the unappreciative.

Technology writer James Vincent has wondered if the prevalence of AI assistants and sales 'people' might lead to the opposite problem:

> If we can't tell the difference between humans and machines on the phone, will we treat all phone conversations with suspicion? We might start cutting off real people during calls, telling them: "Just shut up and let me speak to a human." And if it becomes easier for us to book reservations at a restaurant, might we take advantage of that fact and book them more speculatively, not caring if we don't actually show up?
>
> (Vincent, 2018)

How people will actually feel or behave when interacting with increasingly sophisticated conversational agents is a matter for further study. Our inclination for now is that mandatory 'bot disclosure' is something that merits serious consideration, at least in relatively high-stakes contexts such as sales or political campaigning. While Google has apparently indicated that Duplex will make itself "appropriately identified" when making calls (Statt, 2018), there is no guarantee that other creators or users of this technology will do likewise.

RECOMMENDATION 19: Mandatory 'bot disclosure' is something that merits serious consideration in New Zealand, at least in relatively high-stakes contexts such as high value purchases or political campaigning. Unlike other requirements that are often called for, such as mandatory AI transparency or testing for 'bias', this would be a relatively simple rule to devise and implement.

D. Delegation and handovers

Another reassurance that Google was quick to offer about Duplex was that it would not be operating completely independently. Instead, the company announced:

> The system has a self-monitoring capability, which allows it to recognize the tasks it cannot complete autonomously (e.g., scheduling an unusually complex appointment). In these cases, it signals to a human operator, who can complete the task.
>
> (Leviathan and Matias, 2018)

The issue of when and how a chatbot will involve a human is likely to be one that merits attention, and it is to this that we now turn.

The most successful chatbots operate in domains where dialogues have a well-defined structure, and where the human user's contributions are reasonably predictable. For instance, customers calling a computer support helpline or a bank call centre often ask the same questions; human call centre workers are trained to respond to these frequent questions, with stock answers or programmatic scripts. These cases are ripe for automation using chatbots. Even in specialist professional contexts, the early stages of an interaction are often quite formulaic, and many queries and their answers are likely to be fairly routine. (This is the premise underlying the CitizenAI legal chatbots we discuss later in this chapter.)

Some commentators on Google's Duplex assistant were less impressed than the audience for Pichai's demonstration, pointing out that in fact the range of likely responses to the calls it made were fairly limited. Even in ostensibly straightforward cases, though, not all dialogues with human clients run along predictable lines. Clients sometimes have needs that far exceed the capabilities of current chatbots, or can phrase their questions in ways the chatbot does not recognise. In jobs requiring dialogue with customers, many organisations anticipate that humans and chatbots will *collaborate* in some way: chatbots will handle the easy, repetitive parts of conversations, and humans will handle the parts that require human-level expertise, intelligence, versatility or engagement. This collaboration is an example of how human service jobs are likely to 'change' to accommodate AI systems. (Recall we envisage job changes, rather than wholesale job replacement.) It's therefore useful to consider the forms this collaboration may take, and what principles should govern it.

A common practice for chatbots is to 'escalate' a case to a human worker when the chatbot's competence is exceeded. There are three main conditions that trigger escalation to a human:

- ***Risk management:*** a chatbot for psychological counselling may be set up to scan user utterances for words related to self-harm or other topics that are urgent enough to prompt escalation to a human counsellor.
- ***Interpretation failure:*** a chatbot has to interpret each of the utterances made by the human user, normally by a classifier that identifies the 'intent' of a user utterance, from a fixed set of expected possibilities. Classifiers typically return a confidence score as well as an intent. If confidence falls below some threshold, the system can either provide feedback to the user ("I didn't understand - can you please rephrase?") or can escalate to a human. Often, escalation happens in the case of repeated interpretation failures.
- ***User request:*** in some systems, the user can ask to speak to a human.

Chatbot designers will need to consider first when handovers should occur. What counts as a serious enough risk that it should prompt escalation to a human? In some contexts, a cautious approach may seem the obvious one. For instance, if a counselling chatbot does not escalate to a human on detection of words related to self-harm or reported child abuse – as in the BBC report discussed earlier – this could be seen as a failure of ethical and potentially legal responsibilities on the part of the system designers. On the other hand, if the criteria for escalation are defined very broadly, the bot will often escalate unnecessarily, with significant implications for efficiency.

Even in cases where there is good reason to think that a case should be handed over to a human, questions of trust and consent should also be kept in mind. As we discuss in the following section, there may be situations where people speak to a chatbot precisely because it isn't human. A good New Zealand example is Māori language learning. Many Māori who can't speak the Māori language feel shy or embarrassed (whakamā) about the fact, and this is a barrier to learning. Practicing with a chatbot is a helpful way of overcoming the barrier. (See e.g. Vlugter et al, 2009) Someone who feels, for whatever reason, inhibited about discussing their problems or concerns with a human doctor, counsellor, lawyer, teacher etc may not appreciate suddenly finding themselves being propelled into a conversation with one. And a suspicion that such escalation will happen without their agreement may have the effect of deterring people from using such services (a familiar and complex issue in the context of mandatory disclosure laws for human counsellors, therapists and psychiatrists.)

Designers will also have to consider what information a chatbot should convey to the referred partner when handing over a dialogue. An understandable desire for the chatbot to be able to relay the important information as succinctly as possible may come into tension with a concern with omitting information that may fall outside the chatbot's criteria for importance, but which may convey much more to the human partner. Confidentiality is also an important consideration in these circumstances; even if the client or patient is willing to have their case handed on to a human, they may not be prepared to disclose all of the same details to them, or to express them in just the same way.

What about handovers in the opposite direction? In the context of real-time dialogues, handovers from humans to chatbots are much less common. There are a few cases where a human transfers control explicitly to a chatbot; for instance, this can happen at the end of a conversation that has been escalated to a person, when the human agent passes control back to the chatbot to administer a customer satisfaction questionnaire. (Asquith, 2020) Handovers from humans to bots are also increasingly common in more extended professional interactions. In a medical context, a human doctor in consultation with a patient might prescribe a course of treatment delivered by a chatbot.

In both of these examples, there is no reason why the handover should not be explicitly advertised to the client, and we would expect that this is currently what would happen. Whether, in future, all human-to-chatbot handovers should have to be advertised is perhaps more questionable. In the next sections, we consider the possibility that attitudes towards interacting with chatbots or other forms of AI are likely to be context dependent; there may be situations where many clients simply don't care whether they are dealing with a human or a computer programme, provided their issue is dealt with quickly and efficiently. Even if that transpires to be the case, though, some people may still share

the concerns expressed by Pasquale and others, about being duped or manipulated by AIs programmed to pass for humans.

RECOMMENDATION 20: When transitions occur between humans and chatbots, service providers should be transparent about how and when these will take place, and what information will be passed between them.

E. Trust, empathy and 'the human touch'

Many common concerns about the increasing automation of professional roles relate less to 'technical' concerns about accuracy, transparency and the like, and more to concerns about the removal of distinctly 'human' factors. Worries about 'dehumanisation' and the absence of qualities like trust, empathy and compassion are common in discussions around AI, probably most notably in the health and care sectors. This is, of course, an example of a concern that would only arise where AI replaced humans, either in an entire role or in a significant aspect of it. That's to say, we would not expect it to be a problem where the human professional consults an AI 'expert', but is still responsible for relaying its advice to the client.

In 2016, PWC commissioned a survey of 12,000 people across 12 countries about their attitudes to AI in healthcare. The report that followed was remarkably bullish about the prospects for AI in healthcare: "The message is clear; the public is ready and willing to substitute AI and robotics for humans." (PWC, 2017) PWC did, however, sound a note of caution: "Trust in the technology is vital for wider use and adoption", they noted; "the 'human touch' remains a key component of the healthcare experience." Indeed, when asked about the disadvantages of AI in healthcare, concern about the 'human touch' was the most commonly expressed, being cited by 47% of respondents.

What precisely respondents meant by this is not entirely clear. Plausibly, many of them would have harboured the same sorts of concerns expressed by two paediatricians earlier this year, who wrote:

> In our experience, feeling properly cared for means more than just receiving scripted advice and prescriptions for tests and medications, something a smart, Watson-like AI physician could conceivably do. … Patients want us to ask, look, and touch in response to their concerns, their bodies, and their unique circumstances. Few people appreciate a physician who seems to be working from a script, in the room but not truly present or connected. Who is, in other words, behaving like a machine.
>
> (Drouin and Freeman, 2020)

Richard and Daniel Susskind refer to something similar, which they call the "empathy objection". The describe this as:

> a call not just for a trusted adviser but, as important, for an empathetic expert, someone who can readily perceive the emotional state of others—and more, can feel and share their anguish and joy.
>
> (Susskind and Susskind, 2016, p.251)

Morris Panner and the Forbes Technology Council posed the concern as a rhetorical question: "Can AI-driven robots replace a reassuring bedside manner, a warm embrace or a smiling face?" Their concern is that:

> Health care is not exclusively a matter of tech and science. Even if AI-enabled software can determine an appropriate diagnosis or treatment option, we still want to confide with our physicians and seek their counsel.
>
> (Panner and the Forbes Technology Council, 2019)

As we suggested in the previous section, the extent and strength with which these concerns are held by New Zealanders is a matter for further study, but our intuition is that concern for a 'human touch' is likely to be highly context dependent. It is plausible, for example, that when seeking financial advice, or expert contract-drafting services, few people would have a high expectation of, or concern about, emotional engagement with the provider. Priorities are more likely to relate to speed, affordability and accuracy.

It's more likely to be in contexts such as healthcare that human interaction is most valued, but even there, this will be highly context dependent. The PWC report indicated a significant number of respondents were willing to accept AI healthcare in areas such as monitoring pulse and blood pressure and take and test a blood sample. (PWC, 2017, pp.15-16) Far lower levels of willingness were apparent in areas involving either a degree of hands-on skill and care

(set a broken bone and apply a cast: 7%) or where it is easy to imagine a role for emotional skills (provide care and advice during pregnancy: 5%).

The PWC research didn't interrogate the reasons for those responses, but it is also possible that the services respondents were willing to accept from an AI related to tests and monitoring, where the interaction might be assumed to be of a more ongoing nature, and the possibility of disastrous consequences from one-off errors may be assumed to be lower. All of these assumptions fit with the lowest level of trust being for delivery of a baby (1%).

At least in these more high-stakes or emotionally sensitive areas, then, we should take seriously the possibility that these kinds of concerns will resonate with many people. The perceived 'dehumanisation' of healthcare, and related services such as elderly care or childcare, are likely to signify a boundary of profound unease, whether or not this is entirely justified.

Should we care if they 'care'?

To this observation, though, we should add several of caveats. One is that we should not become over-reliant on generalisations or assumptions about how people might feel about interacting with AI. To take one prominent example: elder care is an area where it is often assumed that replacing humans with robots or other AI would lead to adverse outcomes. It is seen as indicative of an uncaring society where obligations to our growing elderly population is too easily delegated to machines.

As Amanda and Noel Sharkey point out, though, it's not clear that elder people themselves will always see it that way. As they say, "[i]t might even turn out that, given the choice, some of the frail elderly might prefer robotic, as opposed to human, assistance for certain intimate tasks such as toileting, or bathing." (Sharkey and Sharkey, 2012, p.31) In fact there is good evidence that some kinds of robot have beneficial effects on care home residents - for instance, Robinson et al. (2013) found that a simple robot 'pet' decreased loneliness. More functional robot companions could perhaps further enhance the autonomy and dignity of older people, who would otherwise be wholly reliant on the availability and attention of other humans. The option of speaking to a medical chatbot could support users' autonomy in similar ways.

RECOMMENDATION 21: At least in more high-stakes or emotionally sensitive areas, we should take seriously the possibility that concerns about dehumanisation in health/care applications of AI will resonate with many people. On the other hand, we should not become over-reliant on generalisations or assumptions about how people might feel about interacting with AI. It's possible that some people, in some situations, might find dealing with AI helpers or carers empowering, or less undignified than reliance on humans for certain intimate roles (for example, elderly people requiring assistance to get in and out of the bath).

Secondly, we should remember not to succumb to the fallacy of comparing AI with some idealised human medic, care home worker or teacher. As Richard and Daniel Susskind point out:

> it is a regrettable truth that a great number of professional experts are deeply lacking in empathy. Countless tales are relayed of the surgeon with zero bedside manner, the lawyer with no client-handling ability, the brutally insensitive teacher, and so on. … Accordingly, we must be cautious about asking more of our machines than we currently secure from people.
>
> (Susskind and Susskind, 2016, p.251)

That healthcare contains a number of Doc Martins and Gregory Houses (and other professions their equivalents) does not, of course, mean that we should throw out the baby with the proverbial bathwater, and dispense with the need for human empathy altogether. Nonetheless, it is important to remember that the relevant comparator is not the perfect professional, but in many cases, jaded and overworked human beings. Note that this point echoes our earlier caveats about double standards in transparency.

On a related note, we might wonder whether demanding that a human counsellor, healthcare worker or lawyer should invest emotionally in the suffering and anxiety of their clients and patients is to demand a very great deal from them. For those working in emotionally arduous areas of law or healthcare, for example, maintaining a certain emotional distance may be a necessary coping strategy to avoid 'burnout'. One might even argue the case for AI systems that *specialise* in emotional engagement, in areas where human workers are susceptible to burnout. Such tools may possibly be effective even if users are aware they are talking to a machine, if the therapeutic effect is mainly in the act of talking. Some AI companies specialise in dialogue agents that seek to establish an emotional connection with their human users: the New Zealand company Soul Machines is a case in point. Recent studies showed that users responded more positively to the Soul Machines dialogue agent if it expressed emotions visibly and audibly (see Loveys, Sagar and Broadbent, 2020).

A third caveat goes to the reasons why many people may express a preference for empathetic humans. It's possible that the desire for human empathy derives from a belief that a doctor, lawyer, teacher etc who 'knows how we feel' is more likely to care about our well-being, and therefore, more likely to work hard for a good outcome for us. In that case, empathy would be *instrumentally* valuable, a means to another end, specifically, the provision of better service.

This may be a valid heuristic when dealing with human professionals, but it's less clear that it counts as an argument against AI professionals. While empathy may serve an important instrumental role in motivating humans to 'go that extra mile' for us, AIs will require no such motivation. An AI lawyer programmed to work on our case will do so without fatigue, boredom or distraction, until it is instructed to stop. In the case of medical AIs, what the patient needs is perhaps *trust*, rather than empathy.

In considering all of these caveats and arguments, we should recognise that 'empathy' may not describe a single, simple concept at all. Psychology professor and author Paul Bloom has sought to distinguish "emotional" empathy (sometimes called "affective empathy") from "cognitive empathy." The latter he describes as "[t]he capacity to understand what's going on in other people's heads, to know what makes them tick, what gives them joy and pain, what they see as humiliating or ennobling." (Bloom, 2016, p.36) This he sees as essential to being a good medical practitioner, or indeed, a good ethical actor.

Emotional empathy, in contrast, he sees as a frequent obstacle to good medical practice or ethical action. For one thing, it can lead us to over-value the interests of those with whom we readily identify. It can also, Bloom argues, prevent professionals doing their jobs properly:

> The risks of empathy are perhaps most obvious with therapists, who have to continually deal with people who are depressed, anxious, deluded, and often in severe emotional pain. … anyone who thinks that it's important for a therapist to feel depressed or anxious while dealing with depressed or anxious people is missing the point of therapy. (p.144)

He also cites the example of his uncle who, when undergoing cancer treatment, "seemed to get the most from doctors who didn't feel as he did, who were calm when he was anxious, confident when he was uncertain." (p.146)

Unsurprisingly, Bloom's approach has not met with universal agreement. Nonetheless, interrogating the 'empathy objection' to AI involves interrogating what exactly we mean by 'empathy', and which forms we see as essential, desirable or unhelpful in a variety of contexts.

RECOMMENDATION 22: Concerns about lack of empathy are common in discussions about AI replacing humans in some key roles, but 'empathy' can refer to different things, and we should be clear what sort of empathy is wanted in what situations. It may be that AI could become very good at recognising and responding appropriately to human emotions, without having to feel them.

Human roles for human needs

Such instrumental concerns, though, are unlikely to be the only reason that we want professionals to care about us. For many people, there may be an innate human need to believe that the person to whom we are recounting our fears and troubles, or to whom we are exposing our vulnerabilities, actually cares about us – not just so they will try harder on our behalf, but because the knowledge (or at least belief) that they do has profound emotional significance. This may be particularly important when they are trying to assuage our fears or impart bad news.

Another concern may relate to the sort of professional judgment and intuition that allows human doctors to pick up on cues from their patients' behaviour or presentation. An experienced GP, for example, may pick up on the fact that a patient presenting with one complaint is actually worried about another matter that they are, for whatever reason, reluctant to raise. In theory an AI trained on a sufficiently vast database of doctor-patient encounters could do the same – there's nothing inherently immeasurable about such correlations. But it may prove difficult to train an AI to spot the cues on which doctors rely. A key problem is that 'nonverbal' cues like eye gaze, posture and intonation are likely to have an important role here, and many medical dialogue systems aren't set up to record such signals.

Richard and Daniel Susskind are willing to concede, at least for the sake of argument, that certain aspects of professional roles will be challenging to automate. But they express doubt as to whether the current model of professionalism is well equipped to provide for these. They are unconvinced by the notion that 'subject matter experts' are in fact likely to possess the sort of interpersonal skills or best placed to provide the emotional support patients sometimes need:

> When there is bad news to impart … it is not self-evident that we should lean towards the technical specialist to dispense the comforting words. Instead, we might turn, for example, to a para-professional, someone with sufficient insight into the area of expertise as well as the genuine capacity to empathize. By disengaging the application of expertise from the communication with the recipient … this moves us, in part, away from the traditional model of production and distribution of practical expertise towards the 'para professional' model … In both cases, though, human beings are still involved.
>
> (Susskind and Susskind, 2016, p.252)

This is an idea that recurs throughout much of the literature around AI in the professions; the idea that automation of the more routine, burdensome aspects of a role could free up time for those tasks uniquely suited to human beings. Writing of his own experience as a healthcare professional, Rahul Parikh has taken an optimistic view of the opportunities presented by AI:

> I went to medical school to connect with people and make a difference. Today I often feel like an overpaid bookkeeper instead, taking in information and spitting it back to patients, prescribing drugs and adjusting doses, ordering tests. But AI in the exam room opens up the chance to recapture the art of medicine. It could let me get to know my patients better, learn how a disease uniquely affects them, and give me time to coach them toward a better outcome.
>
> (Parikh, 2018)

The idea that this could see the emergence of wholly new patient-facing roles has also been articulated by Martin Ford:

> there may eventually be an opportunity to create a new class of medical professionals: persons educated with perhaps a four-year undergraduate or master's degree, and who are trained primarily to interact with and examine patients – and then to convey that information into a standardized diagnostic and treatment system. These new, lower-cost practitioners would be able to take on many routine cases, and could be deployed to help manage the dramatically growing number of patients with chronic conditions such as obesity and diabetes.
>
> (Ford, 2015, p.157)

These new practitioners, Ford suggests, "could handle routine cases, while referring patients who require more specialized care to physicians." (p.158) The AI Forum's report suggested that 'data science doctor' and 'clinical machine learning expert' could be new specialities in future, and recommended that New Zealand should train more doctors in data science. (AI Forum, 2019, p.44)

Of course, all this relies on the assumption that more efficient delivery of automatable tasks will result in greater investment in those tasks less amenable to automation – an outcome that is in no sense inevitable. Nonetheless, it is an outcome that should be kept in mind, as human providers find their physical, cognitive and emotional capacities stretched ever further.

It is possible that certain emotional abilities of good doctors, therapists, etc. will be difficult to replicate in AI, particularly those that rely on subtle understanding of emotions in patients, and the subtle conveying of emotions to patients. It's also possible that, for whatever reason, some people would simply prefer that emotionally difficult conversations – such as imparting bad news – take place with a human professional. For those sorts of reasons, we consider it likely that we'll continue to see humans in many of those roles for the foreseeable future. We may, however, see *a significant reallocation of tasks*, with a greater emphasis placed on human professionals who are skilled in those parts of the role that are (at least for now) uniquely suited to human providers, while more 'technical' aspects are taken up by AI.

Uniqueness neglect

A particular variety of the dehumanisation concern is what has been referred to in the literature as *uniqueness neglect*. This accepts as a starting position that AI will be very good at, for example, diagnosis and treatment recommendations in general, but will do so at the expense of paying due attention to the particular circumstances of the individual patient. A 2019 article published in the Journal of Consumer Research suggests that this response is common among patients. As the authors suggest:

> consumers may be more reluctant to utilize medical care delivered by AI providers than comparable human providers, because the prospect of being cared for by AI providers is more likely to evoke a concern that one's unique characteristics, circumstances, and symptoms will be neglected. We refer to this concern as uniqueness neglect.
>
> (Longoni, Bonezzi and Morewe, 2019, p.630)

On this analysis, patients may be quite willing to accept that the recommended course of treatment is better in general, but less willing to accept that it is the right course for *them*. This may be thought to be another concern that only arises where the AI replaces the human doctor, etc. It may, however, also be an issue where a human professional is kept 'in the loop', but in such a way that they are inclined to accept the AI's recommendations more or less unquestioningly. This is the issue of 'automation bias' that we addressed in our first report.

As expressed by Longoni et al, the uniqueness neglect thesis is concerned with documenting and explaining barriers to consumer acceptance; the authors do not take a position on the validity of these concerns. And in fact, an AI system which takes many input data points about a patient may well be able to propose quite customised treatment, rather than grouping patients into broad categories, if its training set is big enough, and its learning algorithm is powerful enough. Rosalind McDougall, however, has advanced a more normative variant of this thesis, which relates to power structures in medicine. As she explains, medical ethics has for decades been characterised by a trend away from 'doctor knows best' paternalism, and towards an approach that recognises and prioritises the specific values, goals and fears of the individual patient. AI systems that make treatment recommendations, she argues, "present a potential threat to shared decision making, because the individual patient's values do not drive the ranking of treatment options." (McDougall, 2019, p.157)

McDougall sets out her concern using the example of Watson for Oncology. This, she explains, "ranks treatment options based on a particular value: maximising lifespan." While this may, at first glance, seem quite reasonable, McDougall points out that "We know that patients' values differ. Not all patients aim exclusively for longevity in their treatment choices." (p.157) A particular patient may afford higher significance to minimisation of suffering or quality of remaining life, rather than extending its duration. A related concern from McDougall is that "these types of AI systems currently do not encourage doctors and patients to recognise treatment decision making as value-laden at all." (p.157) This is a familiar concern from a variety of contexts involving automated or algorithmic decisions; that their automated and seemingly objective nature serves to

disguise the inherently moral or political nature of the decisions being made.

The preference for someone to take account of our unique personal circumstances is quite understandable, and to an extent, logical. It is possible, though, that some of those expressing a preference for real, caring humans do so from a position of unrealistic beliefs about how decisions are presently made. Faced with a procession of students, clients or patients on a daily basis, it is all but inevitable that professionals will fall back on actuarial assumptions and generalisations about individual cases.

We might also wonder whether some of these concerns about algorithmic decisions rely on poor design rather than any fundamental constraint. In response to McDougall's concern that AI systems will tend to ignore patient aspirations, we could simply require that a medical AI system is able to *interact with the patient*, to learn more about these aspirations. Exactly how to interact is still an open question, of course – but there is nothing about AI systems that precludes interactions. McDougall can be understood as advocating that medical AI systems should be *dialogue systems*, in which doctor and patient jointly negotiate an outcome.

How such proposals can be mandated, or enforced, is of course a separate question – to which we now turn.

F. Regulatory issues

Some professions hold monopolies on offering certain services. Other rules govern who can advertise themselves as a member of a given profession. Most (or all, depending on how we define a 'profession') have rules applicable to those practising within them. How will those rules operate in an environment where we have AI lawyers, doctors or financial advisors? In reality, it's likely to be quite some time before we encounter any serious proposal to replace those professionals entirely with AI. We are already seeing, though, the prospect of task-specific AI taking over particular aspects of those roles. Will the rules governing those professions serve as barriers to this use of such technologies? Can AI service providers meet the standards we require of human professionals? We consider these questions in the context of a few of our more heavily regulated professions.

Healthcare

The healthcare sector in New Zealand "is governed by a wide array of statutes and subordinate legislation". (Paterson, 2015, p.3) Some of these exist to ensure certain standards of competence and ethical conduct from practitioners. Others have a focus on patients (or 'healthcare consumers") rights. A third stream regulates the use of therapeutic products. Precisely where AI in healthcare will fit within that scheme is a complex question that merits closer scrutiny.

The regulation of medical devices is currently in a state of regulatory transition in New Zealand, with new draft Therapeutic Products Bill still making its way through the parliamentary process. (Ministry of Health, 2019) This is intended to address widely acknowledged gaps in the current system, whereby "[m]edical devices are currently not subject to any pre-market regulatory scrutiny to assess safety and performance and post-market controls are minimal." (Ministry of Health, 2018a, [22]) The new scheme will adopt a wide definition of therapeutic products, encompassing obvious categories such as implants and surgical equipment, but also software that is used for a therapeutic purpose, which includes matters such as diagnosis.

Many if not all of the applications of AI in healthcare that we have considered in this section seem likely to fall within the new regulatory scheme, which is intended:

> to apply the full range of pre- and post-market controls in accordance with the risk-based model developed initially by the Global Harmonisation Taskforce (GHTF) and continued and maintained by the International Medical Device Regulators Forum (IMDRF).
>
> (Ministry of Health, 2018a, [23])

At almost 300 sections, the new Bill is lengthy and complex, and the consultation document barely less so. It is not our intention to scrutinise either in detail here. It is important, however, to consider how the Bill might apply to AI. Most medical devices and therapeutic products, it might be assumed, will continue to operate in much the same way after their initial approval (subject, of course, to malfunction, breakage and wear and tear.) AI software, in contrast, has the potential to change after the approval process is complete, a feature that has the potential to make it an elusive regulatory target.

As we noted earlier, the much-discussed scenario where AI autonomously adapts 'in the field' is at least for now very rare, and great care would have to be taken were this ever to be allowed for the sort of high-stakes functions that we would expect to encounter in healthcare. Much more realistic, at least for the foreseeable future, will be post-market change in the form of version updates. These have the potential to be checked prior to release, but it will be important for the new regulatory system to be clear about when a software update constitutes a new product necessitating a fresh approval.

In the USA, the Food and Drug Administration has proposed a regulatory framework for changes to software including machine learning and AI. (FDA, 2019) Deliberate changes would require a new regulatory approval where it significantly affects performance, safety or effectiveness. Importantly, the FDA also acknowledges that:

> The traditional paradigm of medical device regulation was not designed for adaptive AI/ML technologies, which have the potential to adapt and optimize device performance in real-time to continuously improve healthcare for patients.

This has led them to the provisional conclusion that:

> The highly iterative, autonomous, and adaptive nature of these tools requires a new, total product lifecycle (TPLC) regulatory approach that facilitates a rapid cycle of product improvement and allows these devices to continually improve while providing effective safeguards.

The consultation document considers a range of strategies, including a requirement to indicate at the initial approval stage how such adaptations will be safely managed; an obligation on manufacturers to keep such products under review; and a duty to make a new submission if the product changes beyond the intended use for which it was previously authorised.

The FDA's approach is still very much in draft form, and may be amended based on the results of the consultation process. Nonetheless, we recommend that it is something that New Zealand's Ministry of Health and the therapeutic devices regulator should be keeping under review, as our new therapeutic devices regulatory framework moves through the legislative process.

A modified version of a therapeutic products regulatory framework may be essential for some uses of AI in healthcare, but there are questions as to whether it goes far enough for others. Some applications, such as conversational agents in the mental health context, will have interactions with human 'healthcare consumers' that are of a very different nature from what we would typically associate with medical devices. We might in fact wonder whether these interactions are, to some extent, more appropriately managed within a regulatory paradigm designed for interactions between human doctors and patients.

Section 118(i) of the Health Practitioners Competence Assurance Act 2003 places a duty on the Medical Council (and other authorities specific to different healthcare roles) "to set standards of clinical competence, cultural competence (including competencies that will enable effective and respectful interaction with Māori), and ethical conduct to be observed by health practitioners of the profession". Some of this involves making sure that practitioners are suitably qualified, and that their skills and knowledge stay up to date. Others, however, relate to the manner in which doctors conduct themselves towards patients.

With regard to competence and ethical conduct, the Medical Council sets out these standards in its *Good Medical Practice* document (2016). The standards in this document include requirements 'to establish and maintain trust with your patients.' ([14]), to '[m]ake sure you treat patients as individuals and respect their dignity and privacy ([15]) and to '[b]e courteous, respectful and reasonable.' ([16])

The Code of Health and Disability Services Consumers' Rights establishes a number of important rights that 'healthcare consumers' are entitled to expect. Some of the Code's rights reflect concerns that we would certainly expect to be taken into account when AI is deployed in a healthcare setting. Right 4(1), for example, recognises "the right to have services provided with reasonable care and skill", while Right 5(1) establishes that "[e]very consumer has the right to effective communication in a form, language, and manner that enables the consumer to understand the information provided."

Other rights, however, may pose questions if AIs are assuming patient-facing roles previously taken by humans. For example:

- the right to be treated with respect (Right 1(1));
- the right to be provided with services that take into account the needs, values, and beliefs of different cultural, religious, social, and ethnic groups, including the needs, values, and beliefs of Māori (Right 1(3));
- the right to have services provided in a manner that respects the dignity and independence of the individual (Right 3); and
- the right to an environment that enables both consumer and provider to communicate openly, honestly, and effectively (Right 5(2)).

The presence of these requirements under the 2003 Act and the Code raises a question as to whether it should be a requirement, before AI tools are deployed in patient-facing settings, that they be able to comply with the same standards that we currently set for humans. How might such a requirement be discharged? Can a conversational agent be imbued with 'cultural competence' or assured to respect the dignity of the individual?

These are questions that will surely be subject of ongoing scrutiny in the medical ethical literature and beyond. From a regulatory perspective, though, what is interesting to note is that AI in healthcare looks likely to straddle two previously distinct strands. Insofar as it is viewed as an artefact, then it will be subject to the therapeutic products regime, oriented towards risk minimisation. But in those contexts where AI performs in a more 'human' way – communicating directly with healthcare consumers – then it arguably should also be evaluated against the framework that exists to ensure that human healthcare providers conduct their duties in a respectful and culturally competent manner. Whether this can be achieved within the current regulatory framework, or whether some more innovative approach that fulfils these different roles, will be a matter for further study.

RECOMMENDATION 23: In New Zealand, healthcare is regulated in part by rules aimed at therapeutic devices, and in part by rules aimed at human practitioners. Some healthcare AI seems to straddle those two streams. Insofar as it is viewed as an artefact, then it will be subject to the therapeutic products regime, oriented towards risk minimisation. But in those contexts where AI performs in a more 'human' way – communicating directly with healthcare consumers – then it should also be evaluated against the framework that exists to ensure that human healthcare providers conduct their duties in a respectful and culturally competent manner. Whether the new Therapeutic Products Bill, and the regulator it creates, makes adequate provision for this remains to be seen.

Law

Law is another traditional profession the practice of which is carefully controlled via legal and ethical regulations. There are, as yet, no rules in New Zealand specifically governing the use of AI tools in law, but as in many other contexts where AI is deployed, it may be subject to more general rules governing that context. The Lawyers and Conveyancers Act 2006 makes it an offence for anyone who is not a lawyer to provide "legal services" (s.21(1)(a)) and to describe themselves as a lawyer, legal practitioner, barrister, solicitor, etc. (s.21(1)(b)).

A related provision is s 23, which makes it an offence to make a false or misleading representation that legal services are being provided by, or under the direct supervision of, a person who is a lawyer. Finally, s 24 of the Act makes it an offence for someone who is not a lawyer to carry out "reserved areas of work" if they do so "for gain or reward". Reserved areas relate to representing before or advising about any proceedings before any court of tribunal.

The Act defines "legal services" fairly extensively. For present purposes, some of the more relevant restrictions are likely to be on:

- advice in relation to any legal or equitable rights or obligations:
- the preparation or review of any document that:
  (i) creates, or provides evidence of, legal or equitable rights or obligations; or
  (ii) creates, varies, transfers, extinguishes, mortgages, or charges any legal or equitable title in any property.

The provision of legal *advice* is generally taken to be distinct from the provision of legal *information*. Although these terms are not defined in the Act, information may plausibly be supposed to apply to information about how the law works in general terms, while advice will be information targeted to the individual's particular circumstances.

The terms of the 2006 Act could be relevant for the sorts of 'chatbots' that were until recently being developed by the Wellington Community Law Centre initiative, CitizenAI (https://www.citizenai.nz/projects). Services like LagBot, RentBot and WorkBot use natural language dialogue systems to enable people to ask questions about their prison, tenancy or employment issues. As already noted, these new tools have significant potential to improve access to justice by directly responding to routine legal questions without the need for a lawyer, but with the quality of information based on thousands of previous similar questions and answers (much like a real time 'Frequently Asked Questions').

Provided they offer legal information in general terms, rather than legal advice in response to specific situations, it seems likely that the CitizenAI legal chatbots will comply with the terms of the 2006 Act. As the technology progresses, though, it may be that legal AI will be developed that is able to offer advice tailored to a particular client's needs.

In the future, law firms might also offer these services to their clients, directing them to readily available information to answer simple questions and facilitating their interaction with a lawyer in more complex cases or in prescribed circumstances. Were that to become possible, those using such a service will need to take care to ensure that:

- with regard to legal advice in general, they do not describe the chatbot as a "lawyer" or make misleading claims that it is being supervised by a lawyer; and
- with specific regard to advice about court proceedings, the chatbot would not be allowed to offer 'advice' at all.

RECOMMENDATION 24: Provided they offer legal information in general terms, rather than legal advice in response to specific situations, it seems likely that AI legal chatbots such as those introduced by CitizenAI will comply with the terms of the Lawyers and Conveyancers Act 2006. As the technology progresses, though, it may be that legal AI will be developed that is able to offer advice tailored to a particular client's needs. Were that to become possible, those using such a service should take care to ensure that:

- With regard to legal advice in general, they do not describe the chatbot as a "lawyer" or make misleading claims that it is being supervised by a lawyer;
- With specific regard to advice about court proceedings, the chatbot would not be allowed to offer 'advice' at all.

Financial advice[16]

The provision of financial advice in New Zealand is governed by the Financial Advisors Act 2008 and the Financial Markets Authority (FMA). Until recently, it appeared that personalised digital advice was ruled out under the Act, which refers to a "person" providing such advice (s.8). However, the Act also empowered the FMA to grant exemptions from compliance with the Act. Using this discretion, the FMA issued the Financial Advisers (Personalised Digital Advice) Exemption, which came into force on 1 June 2018.

Under the Exemption, registered financial service providers can now apply to provide 'robo-advice' services in respect of eligible products, such as KiwiSaver. This is a transitional provision; when the Financial Services Legislation Amendment Act 2019 comes into force in March 2021, it will remove the requirement for financial advice to be given by a person. The Ministry of Business, Innovation and Employment have explained the effect and rationale for the new legislation:

> The requirement for personalised financial advice to be given by a natural person will also be removed. Technology neutral legislation will further enable the provision of robo (or digital) advice and help future-proof the regime and enable new and innovative ways of providing financial advice. (MBIE, 2018)

An obligation to use/understand AI?

Much of the focus on AI in professional roles is on whether it should be allowed to be used in certain contexts, and if so, what condition should be placed on its use. Given the projected benefits, though, we should also take seriously the possibility that there may sometimes be a professional *obligation* to use AI. Most obviously, such an obligation could arise when AI was demonstrably safer than other options – if, for example, an AI tool was shown to offer more accurate diagnosis or treatment recommendations, it would be difficult for a doctor to justify not using it. Writers on the subject are beginning to take seriously the idea that "[i]f AI methods of diagnosis become sufficiently advanced, it will be malpractice not to use them". (Pasquale, 2020, p.44)

Obligations could also arise in the context of AI offering cheaper services. Insofar as client fees should reflect the amount of time or expertise dedicated to a task, this may raise the question of whether there ever be an obligation to use AI if that can reduce those costs. A recent decision by an Ontario court suggests this could be a serious possibility for lawyers. In *Cass v. 1410088 Ontario Inc.* 2018 ONSC 6959, an unsuccessful plaintiff appealed against an award of costs, on the basis that the defendant's legal fees were excessive. Part of the claim related to fees for legal research. In agreeing that the fees were excessive, the judge observed that "[i]f artificial intelligence sources were employed, no doubt counsel's preparation time would have been significantly reduced." (At [34])

In another Ontario case (*Drummond v. The Cadillac Fairview Corp. Ltd.,* 2018 ONSC 5350) the judge held that the costs of using AI research tools are something for which lawyers can reasonably charge. In *Drummond*, the judge concluded that:

> The reality is that computer-assisted legal research is a necessity for the contemporary practice of law and computer assisted legal research is here to stay with further advances in artificial intelligence to be anticipated and to be encouraged. Properly done, computer assisted legal research provides a more comprehensive and more accurate answer to a legal question in shorter time than the conventional research methodologies, which, however, also remain useful and valuable. Provided that the expenditure both in terms of lawyer time and computer time is reasonable and appropriate for the particular legal problem, I regard computer-assisted legal research as recoverable counsel fee item and also a recoverable disbursement. ([10])

While stopping short of a duty to use AI, taken together, these decisions suggest that costs incurred from using AI appropriately can be charged to a client, but that a lawyer whose services are more expensive because of a refusal to use AI cannot pass the cost of this refusal onto the client.

In other jurisdictions, concerns about the impact of new forms of technology on lawyers' professional obligations, including client care, have prompted new or supplementary professional duties in some jurisdictions. In 2012 the American Bar Association amended its Code of Conduct to introduce a "duty of technical competence". Comment 8 on Model Rule 1.1

16 In preparing this section, we were grateful for the research of Otago Law student Hannah Cross. Hannah's prize-winning essay on the regulation of robo-advice can be read at https://www.gallawaycookallan.co.nz/law-emerging-tech.

provides that to maintain "requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, *including the benefits and risks associated with relevant technology*" (emphasis added). At least 35 states have since formulated rules of professional conduct that adopt this comment and model rule in some form.

The Code does not make specific mention of AI, but if its actual benefits come close to some of the hype, then it is easy to see how the general duty of technical competence could be taken to encompass at least basic knowledge and skill relating to such technologies. New Zealand's Lawyers and Conveyancers Act and associated Conduct and Client Care Rules do not contain a technology specific duty of competence, but do contain a general duty to act competently and in a timely manner.[17] As AI tools become more common and reliable within the profession, it may be that an ability and willingness to use some AI would be inferred from the general duty.

RECOMMENDATION 25: Given the projected benefits, we should also take seriously the possibility that there may sometimes be a professional obligation to use AI. Most obviously, this could arise were AI shown to be safer or more accurate than other options – if, for example, an AI tool was shown to offer more accurate diagnosis or treatment recommendations, it would be difficult for a doctor to justify not using it. But a duty may also arise if using AI renders the delivery of services less expensive; a Canadian court has already recognised that, if a lawyer elects not to use labour-saving AI, they cannot pass any additional cost onto their client.

17 "In providing regulated services to a client, a lawyer must always act competently and in a timely manner consistent with the terms of the retainer and the duty to take reasonable care." Lawyers and Conveyancers Act (Lawyers: Conduct and Client Care) Rules 2008, Chapter 3, paragraph 3.

BIBLIOGRAPHY

CASES

Cass v 1410088 Ontario Inc 2018 ONSC 6959

Drummond v The Cadillac Fairview Corp Ltd 2018 ONSC 5350

Gilbert v Transfield Services (New Zealand) Ltd [2013] NZEmpC 71, [2013] ERNZ 135

Grace Team Accounting Ltd v Brake [2014] NZCA 541, [2015] 2 NZLR 494

Houston Federation of School Teachers v Houston Independent School District 251 F Supp 3d 1168 (SD Tex 2017). Available at: https://www.leagle.com/decision/infdco20170530802

OCS Ltd v Service and Food Workers Union Nga Ringa Tota Inc [2006] ERNZ 762 (EmpC)

STATUTES

New Zealand

Employment Relations Act 2000

Factories Amendment Act 1936

Financial Advisers Act 2008

Financial Advisers (Personalised Digital Advice) Exemption Notice 2018

Financial Services Legislation Amendment Act 2019

Health and Safety at Work Act 2015

Health Practitioners Competence Assurance Act 2003

Human Rights Act 1993

Industrial Conciliation Amendment Act 1936

Lawyers and Conveyancers Act 2006

Lawyers and Conveyancers Act (Lawyers: Conduct and Client Care) Rules 2008

Privacy Act 1993

Privacy Act 2020

WorkSafe New Zealand Act 2013

Overseas

Artificial Intelligence Video Interview Act 820 ILCS 42 (Illinois)

SB-1001 Bots: disclosure. (2017-2018) (California)

OTHER

ACAS (Advisory, Conciliation and Arbitration Service) (2020) *My boss the algorithm: An ethical look at* algorithms in the workplace. Available at: https://www.acas.org.uk/my-boss-the-algorithm-an-ethical-look-at-algorithms-in-the-workplace

ACC (Accident Compensation Corporation) (2018) Statistical models to improve ACC claims approval and registration process. Available at: https://www.acc.co.nz/assets/im-injured/ef79338f63/claims-approval-technical-summary.pdf

ADCU (App Drivers & Couriers Union) (2020) App Drivers & Couriers Union files ground-breaking legal challenge against Uber's dismissal of drivers by algorithm in the UK and Portugal. 26 October. Available at: https://www.adcu.org.uk/news-posts/app-drivers-couriers-union-files-ground-breaking-legal-challenge-against-ubers-dismissal-of-drivers-by-algorithm-in-the-uk-and-portugal

AI Forum (2018) *Artificial Intelligence: Shaping a Future* New Zealand. Available at: https://www.mbie.govt.nz/dmsdocument/5754-artificial-intelligence-shaping-a-future-new-zealand-pdf

AI Forum (2019) *Artificial Intelligence for Health in New Zealand*. Available at: https://aiforum.org.nz/wp-content/uploads/2019/10/AI-For-Health-in-New-Zealand.pdf

AMA (American Management Association) (2019) The Latest on Workplace Monitoring and Surveillance. 8 April. Available at: https://www.amanet.org/articles/the-latest-on-workplace-monitoring-and-surveillance/

Acemoglu, D. and Restrepo, P. (2019) *The Wrong Kind of AI? Artificial Intelligence and the Future of Labor Demand*. NBER Working Paper 25682. Available at: https://doi.org/10.3386/w25682

Adam, M., Wessel, M., and Benlian, A. (2020) AI-based chatbots in customer service and their effects on user compliance. *Electron Markets*. Available at: https://doi.org/10.1007/s12525-020-00414-7

Adams-Prassl, J. (2019) *What if Your Boss was an Algorithm?* Economic Incentives, Legal Challenges, and the Rise of Artificial Intelligence at Work. *Comparative Labor Law & Policy Journal*. 41(1): 123–146.

Adamson, A.S. and Smith, A. (2018) Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology*. 154(11): 1247–1248. Available at: https://doi.org/10.1001/jamadermatol.2018.2348

Agerholm, H. (2017) Robot 'goes rogue and kills woman on Michigan car parts production line'. *The Independent*, 15 March. Available at: https://www.independent.co.uk/news/world/americas/robot-killed-woman-wanda-holbrook-car-parts-factory-michigan-ventra-ionia-mains-federal-lawsuit-100-cell-a7630591.html

Ajunwa, I., Crawford, K., and Schultz, J. (2017) Limitless Worker Surveillance. *California Law Review*. 105(3): 735–776. Available at: https://doi.org/10.15779/Z38BR8MF94

Ajunwa, I. (2019) Beware of Automated Hiring. *The New York Times*, 8 October. Available at: https://www.nytimes.com/2019/10/08/opinion/ai-hiring-discrimination.html

Alesina, A. and Perotti, R. (1996) Income distribution, political instability, and investment. *European Economic Review*. 40(6): 1203–1228. Available at: https://doi.org/10.1016/0014-2921(95)00030-5

Ali, M. et al. (2019) Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction*. 3(CSCW): 199. Available at: https://doi.org/10.1145/3359301

Alibaba Tech (2018) Good Things Come in Small Packages. 30 March. Available at: https://medium.com/@alitech_2017/alibabas-ai-solution-for-the-3d-bin-packing-problem-3ce66d730ecc

Allen, R.C. (2009) Engels' pause: Technical change, capital accumulation, and inequality in the british industrial revolution. *Explorations in Economic History*. 46(4): 418–435. Available at: https://doi.org/10.1016/j.eeh.2009.04.004

Amann, J. et al. (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*. 20: 310. Available at: https://doi.org/10.1186/s12911-020-01332-6

Andrew, M. (2020) Are NZ Uber drivers employees? The court is about to decide once and for all. *The Spinoff*, 17 July. Available at: https://thespinoff.co.nz/business/17-07-2020/are-nz-uber-drivers-employees-the-court-is-about-to-decide-once-and-for-all/

Angwin, J., Tobin, A., and Varner, M. (2017) Facebook (Still) Letting Housing Advertisers Exclude Users by Race. *ProbPublica*, 21 November. Available at: https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

Artnz, M., Gregory, T., and Ziehran, U. (2016) *The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis*. OECD Social, Employment and Migration Working Papers, No. 189. Available at: https://doi.org/10.1787/5jlz9h56dvq7-en

Asquith, M. (2020) Managing the Chatbot Human Handoff: Tips for Success. Hubtype blog post, 15 November. Available at: https://www.hubtype.com/blog/managing-chatbot-human-handoff/

Autor, D.H. (2015) Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives*. 29(3): 3–30. Available at: https:/doi.org/10.1257/jep.29.3.3

Baltzly, D. (2019) Stoicism. In: Zalta, E.N. ed. *The Stanford Encyclopedia of Philosophy* (Spring 2019 edition). Available at: https://plato.stanford.edu/archives/spr2019/entries/stoicism/

Barnes, A. with Jones, S. (2020) The 4 Day Week: *How the Flexible Work Revolution Can Increase Productivity, Profitability and Well-being, and Create a Sustainable Future*. London: Piatkus.

Bathaee, Y. (2018) The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology*. 31(2): 890–938.

Beedham, M. (2020) Germany wants to permit driverless cars across the country by 2022. *The Next Web*, 10 September. Available at: https://thenextweb.com/shift/2020/09/10/germany-wants-permit-driverless-self-driving-cars-2022/

Bell, F. and Smyl, S. (2018) Forecasting at Uber: An Introduction. Uber Engineering, 6 September. https://eng.uber.com/forecasting-introduction/

Bentham, J. (1789) [1996] Burns, J.H. and Hart, H.L.A., eds. with an introduction by Rosen, F., *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press.

Bjerring, J.C. and Busch, J. (2020) Artificial Intelligence and Patient-Centered Decision-Making. *Philosophy & Technology*. Available at: https://doi.org/10.1007/s13347-019-00391-6

Bloom, P. (2018) *Against Empathy: The Case for Rational Compassion*. London: Vintage.

Bogen, M. and Rieke, A. (2018) *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*. Upturn report. Available at: https://www.upturn.org/reports/2018/hiring-algorithms/

Bresnahan, T.F. and Trajtenberg, M. (1995) General purpose technologies 'Engines of growth'? *Journal of Econometrics*. 65(1): 83–108. Available at: https://doi.org/10.1016/0304-4076(94)01598-T

Brown, B. (2018) Using AI Robots to Reduce Hospital Waiting Times. *Health Tech Insider*, 27 November. Available at: https://healthtechinsider.com/2018/11/27/using-ai-robots-to-reduce-hospital-waiting-times/

Brynjolfsson, E. and McAfee, A. (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W.W. Norton & Co.

Bunge, J. and Newman, J. (2020) Tyson Turns to Robot Butchers, Spurred by Coronavirus Outbreaks. *The Wall Street Journal*, 9 July. Available at: https://www.wsj.com/articles/meatpackers-covid-safety-automation-robots-coronavirus-11594303535

Burridge, R.R., Rizzi, A.A., and Koditschek, D.E. (1999) Sequential Composition of Dynamically exterous Robot Behaviors. *International Journal of Robotic Research*. 18(6):534–555. Available at: https://doi.org/10.1177/02783649922066385

Burrows, M. (2020) Four-day work week: A silver bullet for New Zealand's economy post-COVID-19 or an idealist fantasy? *Newshub*, 2 June. Available at: https://www.newshub.co.nz/home/money/2020/06/four-day-work-week-a-silver-bullet-for-new-zealand-s-economy-post-covid-19-or-an-idealist-fantasy.html

CDEI (Centre for Data Ethics and Innovation) (2020) *Review into bias in algorithmic decision-making*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf

CTU (New Zealand Council of Trade Unions) (2019) *Submission of the New Zealand Council of Trade Unions to the Productivity Commission on the inquiry into Technological Change and the Future of Work*. Available at: https://www.union.org.nz/wp-content/uploads/2019/11/Productivity-Commission-Issues-Paper-Technological-Change-and-the-Future-of-Work.pdf

Caulfield, C. (2019) The Gig Economy Has Arrived In The World Of Nursing. *Forbes*, 27 September. Available at: https://www.forbes.com/sites/forbestechcouncil/2019/09/27/the-gig-economy-has-arrived-in-the-world-of-nursing/#4bbb7aea6274

Chae, Y. (2020) U.S. AI Regulation Guide: Legislative Overview and Practical Considerations. *The Journal of Robotics, Artificial Intelligence & Law*. 3(1): 17–40.

Chalfin, A. et al. (2016) Productivity and Selection of Human Capital with Machine Learning. *American Economic Review*. 106(5): 124–127. Available at: https://doi.org/10.1257/aer.p20161029

Challen, R. et al. (2019) Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*. 28: 231–237. Available at: https://doi.org/10.1136/bmjqs-2018-008370

Chandler, S. (2020) Coronavirus Is Forcing Companies To Speed Up Automation, For Better And For Worse. *Forbes*, 12 May. Available at: https://www.forbes.com/sites/simonchandler/2020/05/12/coronavirus-is-forcing-companies-to-speed-up-automation-for-better-and-for-worse/#5c3b21590688

Char, D.S., Shah, N.H., and Magnus, D. (2018) Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *New England Journal of Medicine*. 378(11): 981–983. Available at: https://doi.org/10.1056/NEJMp1714229

Chyi, N. (2020) The Workplace-Surveillance Technology Boom. *Slate*, 12 May. Available at: https://slate.com/technology/2020/05/workplace-surveillance-apps-coronavirus.html

Connolly, K. (2020) "Kurzarbeit: Germany's Scheme for avoiding unemployment", *The Guardian*, retrieved from https://www.theguardian.com/world/2020/sep/24/kurzarbeit-germanys-scheme-fo-avoiding-unemployment

Corbett-Davies, S., and Goel, S. (2018) The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv: 1808.00023v2. Submitted on 31 July 2018 (v1); last revised 14 August 2018 (v2). Available at: https://arxiv.org/abs/1808.00023

Coyle, D. (2014) GDP: *A Brief but Affectionate History*. Princeton: Princeton University Press. Available at: https://doi.org/10.2307/j.ctvc77mfx

Crawford, K. et al. (2019) *AI Now 2019 Report*. AI Now Institute. Available at: https://ainowinstitute.org/AI_Now_2019_Report.html

Crisp, R. (2017) Well-Being. In: Zalta, E.N. ed. *The Stanford Encyclopedia of Philosophy* (Fall 2017 edition). Available at: https://plato.stanford.edu/archives/fall2017/entries/well-being/

D'Mello, S. K., Graesser, A., and King, B. (2010) Toward Spoken Human-Computer Tutorial Dialogues. *Human-Computer Interaction*. 25(4): 289–323. Available at: https://doi.org/10.1080/07370024.2010.499850

Dabee, N. (2016) The Health and Safety at Work Act 2015: The Myth of Increased Deterrence. Victoria University of Wellington Law Review. 47(4): 585–616.

Dann, L., (2020) Unemployment figures: How NZ compares to US, Aus and China. *The New Zealand Herald*, 4 November. Available at: https://www.nzherald.co.nz/business/unemployment-figures-how-nz-compares-to-us-aus-and-china/FWMXDWIXMMVEHLLD2CYL5I3CGE/

Dastin, J. (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, 11 October. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Datta, A., Tschantz, M.C., and Datta, A. (2015) Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*. 2015(1): 92–112. Available at: https://doi.org/10.1515/popets-2015-0007

Datta, A. et al. (2018) Discrimination in Online Advertising: A Multidisciplinary Inquiry. *Proceedings of Machine Learning Research*. 81: 1–15.

Dattner, B. et al. (2019) The Legal and Ethical Implications of Using AI in Hiring. *Harvard Business Review*, 25 April. Available at: https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-hiring

Davenport, T. and Kalakota, R. (2019) The potential for artificial intelligence in healthcare. *Future Healthcare Journal*. 6(2): 94–98. Available at: https://doi.org/10.7861/futurehosp.6-2-94

Dawkins, D. (2020) Uber Vs. London – The Courtroom Battle The World Is Watching. *Forbes*, 30 July. Available at: https://www.forbes.com/sites/daviddawkins/2020/07/30/uber-vs-london--the-courtroom-battle-the-world-is-watching/?sh=7a21c8de6973

De Backer, K. et al. (2016) *Reshoring: Myth or Reality?* OECD Science, Technology and Industry Policy Papers No. 27. Available at: https://doi.org/10.1787/5jm56frbm38s-en

De Stefano, V. (2018) *"Negotiating the algorithm": Automation, artificial intelligence and labour protection*. Employment Working Paper No. 246. Employment Policy Department, International Labour Office. Available at: https://www.ilo.org/wcmsp5/groups/public/---ed_emp/---emp_policy/documents/publication/wcms_634157.pdf

Dellott, B. and Wallace-Stephens, F. (2017) *The Age of Automation: Artificial intelligence, robotics and the future of low-skilled work*. RSA Action and Research Centre. Available at: https://www.thersa.org/globalassets/pdfs/reports/rsa_the-age-of-automation-report.pdf

Diresta, R. (2019) A New Law Makes Bots Identify Themselves—That's the Problem. *WIRED*, 24 July. Available at: https://www.wired.com/story/law-makes-bots-identify-themselves

Dobbs, R., Manyika, J., and Woetzel, J. (2015) No ordinary disruption: the four global forces breaking the all the trends. McKinsey Global Institute. Available at: https://www.mckinsey.com/mgi/no-ordinary-disruption#

Drouin, O. and Freeman, S. (2020) Health care needs AI. It also needs the human touch. *STAT*, 22 January. Available at: https://www.statnews.com/2020/01/22/health-care-needs-ai-it-also-needs-human-touch/

EU Data Protection Working Party (2017) *Opinion 2/2017 on data processing at work. Report* 17/EN WP 249.

EU-OSHA (European Agency for Safety and Health at Work) (2019a) *OSH and the Future of Work: benefits and risks of artificial intelligence tools in workplaces*. Discussion paper. Available at: https://osha.europa.eu/en/publications/osh-and-future-work-benefits-and-risks-artificial-intelligence-tools-workplaces/view

EU-OSHA (European Agency for Safety and Health at Work) (2019b) *Digitalisation and occupational safety and health (OSH): An EU-OSHA research programme*. Available at: https://op.europa.eu/es/publication-detail/-/publication/fbab9f56-3035-11ea-af81-01aa75ed71a1/language-en/format-PDF

EBA (European Banking Authority) (2020) *Report on Big Data and Advanced Analytics*. EBA/REP/2020/01. Available at: https://www.eba.europa.eu/file/609786/

EPD (European Partnership for Democracy) et al. (2020) *Universal Advertising Transparency By Default*. Available at: https://epd.eu/wp-content/uploads/2020/09/joint-call-for-universal-ads-transparency.pdf

Eadicicco, L. (2019) Microsoft experimented with a 4-day workweek, and productivity jumped by 40%. *Business Insider*, 5 November. Available at: https://www.businessinsider.com.au/microsoft-4-day-work-week-boosts-productivity-2019-11?r=US&IR=T

Economides, K., Haug, A.A., and McIntyre, J. (2013) *Are Courts Slow? Exposing and Measuring the Invisible Determinants of Case Disposition Time*. University of Otago Economics Discussion Papers No. 1317. Available at: https://www.otago.ac.nz/economics/otago111196.pdf

Electronic Privacy Information Center (EPIC)(2019) Complaint and Request to the Federal Trade Commission, for Investigation, Injunction, and Other Relief. Available at: https://epic.org/privacy/ftc/hirevue/EPIC_FTC_HireVue_Complaint.pdf

Ellis, C. (2020) Melbourne care home apologises for job advert requesting no 'dark-skinned' people apply. *Newshub*, 29 June. Available at: https://www.newshub.co.nz/home/world/2020/06/melbourne-care-home-apologises-for-job-advert-requesting-no-dark-skinned-people-apply.html

Elmerraji, J. (2020) Robinhood Traders are Smarter Than They Look. *TheStreet*, 9 September. Available at: https://www.thestreet.com/trends/news/robinhood-traders-are-smarter-than-they-look

Estrada-Cedeño, P. et al. (2019) The Good, the Bad and the Ugly: Workers Profiling through Clustering Analysis. In: Terán, L., Meier, A., and Pincay, J. eds. *2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE: 101–106. Available at: https://doi.org/10.1109/ICEDEG.2019.8734453

Eubanks, V. (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: Picador, St Martin's Press

European Commission. (2020) Proposal for a Regulation on a Single Market For Digital Services (Digital Services Act) Available at: https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN

FDA (US Food and Drug Administration) (2019) *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback*. Available at: *https://www.fda.gov/media/122535/download*

Fabian Society, Commission on Workers and Technology (2020) *Sharing the Future: Workers and Technology in the 2020s*. Available at: https://fabians.org.uk/publication/sharing-the-future-full-report/

Fabulos (2020) Pilots in Five Partnering Countries. Available at: https://fabulos.eu/pilots-in-five-partnering-countries/

Faggella, D. (2020) AI in Law and Legal Practice – A Comprehensive View of 35 Current Applications. Emerj, 14 March. Available at: https://emerj.com/ai-sector-overviews/ai-in-law-legal-practice-current-applications/

Faliagka, E. et al. (2012) Application of Machine Learning Algorithms to an Online Recruitment System. In: *Seventh International Conference on Internet and Web Applications and Services*. IARIA XPS Press: 215–220.

Farm Weekly (2020) NZ apple orchard uses world's first commercial robot picker. 11 October. Available at: https://www.farmweekly.com.au/story/6961435/nz-apple-orchard-uses-worlds-first-commercial-robot-picker/?cs=5166

Fiske, A., Henningsen, P., and Buyx, A. (2019) Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *Journal of Medical Internet Research*. 21(5): e13216. Available at: https://doi.org/10.2196/13216

Fitzgerald, M. (2020) U.S. savings rate hits record 33% as coronavirus causes Americans to stockpile cash, curb spending. *CNBC*, 29 May. Available at: https://www.cnbc.com/2020/05/29/us-savings-rate-hits-record-33percent-as-coronavirus-causes-americans-to-stockpile-cash-curb-spending.html

Fitzpatrick, K.K., Darcy, A., and Vierhile, M. (2017) Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*. 4(2): e19. Available at: https://doi.org/10.2196/mental.7785

Floegel, F. (2019) *Distance, Rating Systems and Enterprise Finance: Ethnographic Insights from a Comparison of Regional and Large Banks in Germany*. London: Routledge. Available at: https://doi.org/10.4324/9781351256124

Ford, M. (2015) The Rise of the Robots: *Technology and the Threat of Mass Unemployment*. London: Oneworld.

Frank, M.R. et al. (2019) Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences of the United States of America*. 116(14): 6531–6539. Available at: https://doi.org/10.1073/pnas.1900949116

Frey, C.B. and Osborne, M.A. (2013) *The Future of Employment: How Susceptible Are Jobs to Computerisation?* Oxford Martin School Working Paper. Available at: https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

Frontier Economics (2018) *The Impact of Artificial Intelligence on Work: An evidence review prepared for the Royal Society and the British Academy*. Available at: https://royalsociety.org/-/media/policy/projects/ai-and-work/frontier-review-the-impact-of-AI-on-work.pdf?la=en-GB&hash=200F71D83E7D3310BAF560515EF4EA6F

FutureFarming (2019). The future belongs to small self-driving tractors. https://www.futurefarming.com/Machinery/Articles/2019/9/The-future-belongs-to-small-self-driving-tractors-474180E/

Gallup (2017) *State of the Global Workplace Report*. New York: Gallup Press.

Gander, K. (2015) Worker killed by robot at Volkswagen car factory. *The Independent*, 2 July. Available at: https://www.independent.co.uk/news/world/europe/worker-killed-robot-volkswagen-car-factory-10359557.html

Gatt, A. and Krahmer, E. (2018) Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*. 61: 65–170. Available at: https://doi.org/10.1613/jair.5477

Gavaghan, C., Knott, A., Maclaurin, J., Zerilli, J. and Liddicoat, J. (2019) *Government Use of Artificial Intelligence in New Zealand: Final Report on Phase 1 of the New Zealand Law Foundation's* Artificial Intelligence and Law in New Zealand Project. New Zealand Law Foundation. Available at: https://www.otago.ac.nz/caipp/otago711816.pdf

Gilman, M. (2020) *Poverty Lawgorithms: A Poverty Lawyer's Guide to Fighting Automated Decision-Making Harms on Low-Income Communities*. Available at: https://datasociety.net/library/poverty-lawgorithms

Goodman, R. (2018) Why Amazon's Automated Hiring Tool Discriminated Against Women. ACLU (American Civil Liberties Union) blog post, 12 October. Available at: https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against

Government of Canada (2020) Algorithmic Impact Assessment (AIA). Available at: canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Govil, S. (2020) What AI does for accountants. *Accounting Today*, 27 January. Available at: https://www.accountingtoday.com/opinion/what-ai-does-for-accountants

Graham, C. and Ruiz Pozuelo, J. (2017) Happiness, stress, and age: how the U curve varies across people and places. *Journal of Population Economics*. 30: 225–264. Available at: https://doi.org/10.1007/s00148-016-0611-2

Greenfield, A. (2017) *Radical Technologies: The Design of Everyday Life*. London: Verso.

Grote, T. and Berens, P. (2020) On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*. 46(3):205–211.

Groeneveld, R. (2020) New Zealand's Robotics Plus beta tests Unmanned Ground Vehicles. *Future Farming*, 14 July. https://www.futurefarming.com/Machinery/Articles/2020/7/New-Zealands-Robotics-Plus-beta-tests-Unmanned-Ground-Vehicles-612419E/

Grzybowski, A. et al. (2019) Artificial intelligence for diabetic retinopathy screening: a review. *Eye*. 34: 451–460. Available at: https://doi.org/10.1038/s41433-019-0566-0

Gu, S. et al. (2016) Deep Reinforcement Learning for Robotic Manipulation with Asynchronous Off-Policy Updates. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*: 3389–3396. Available at: https://doi.org/10.1109/ICRA.2017.7989385

Guizzo, E. (2019) How Boston Dynamics Is Redefining Robot Agility. *IEEE Spectrum*, 27 November (print version published December 2019). Available at: https://spectrum.ieee.org/robotics/humanoids/how-boston-dynamics-is-redefining-robot-agility

Haass, R. (2020) The Pandemic Will Accelerate History Rather Than Reshape It: Not Every Crisis Is a Turning Point. *Foreign Affairs*, 7 April. Available at: https://www.foreignaffairs.com/articles/united-states/2020-04-07/pandemic-will-accelerate-history-rather-reshape-it

Hänold, S. (2018) Profiling and Automated Decision-Making: Legal Implications and Shortcomings. In: Corrales, M., Fenwick, M., and Forgó, N, eds. *Robotics, AI and the Future of Law*. Singapore: Springer: 123–153. Available at: https://doi.org/10.1007/978-981-13-2874-9

Harabagiu, S. et al. (2005) Employing Two Question Answering Systems in TREC 2005. In: Voorhees, E.M. and Buckland, L.P. eds. *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*. NIST Special Publication: SP 500-266. Available at: https://trec.nist.gov/pubs/trec14/papers/lcc-sanda.qa.pdf

Hartzog, W. (2015) Unfair and Deceptive Robots. *Maryland Law Review*. 74(4): 785–832.

Harwell, D. (2019) A face-scanning algorithm increasingly decides whether you deserve the job. *The Washington Post*, 6 November. Available at: https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Hatherley, J.J. (2020) Limits of trust in medical AI. *Journal of Medical Ethics*. 46(7): 478–481. Available at: https://doi.org/10.1136/medethics-2019-105935

Hatton E. (2019) Truckies leaving job over poor pay, driver-facing cameras. RNZ (Radio New Zealand), 23 May. Available at: https://www.rnz.co.nz/news/national/389844/truckies-leaving-job-over-poor-pay-driver-facing-cameras

Hatton, E. (2020) Employee surveillance software sales surge in lockdown. RNZ (Radio New Zealand), 2 June. Available at: https://www.rnz.co.nz/news/national/418055/employee-surveillance-software-sales-surge-in-lockdown

Hawkins, L. (2020) Cancer diagnosis tool could cut patient waiting time by 90%. *Healthcare Global*, 11 August. Available at: https://www.healthcareglobal.com/technology-and-ai-3/cancer-diagnosis-tool-could-cut-patient-waiting-time-90

He, K. et al. (2015) Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: *2015 IEEE International Conference on Computer Vision:* 1026–1034. Available at: https://doi.org/10.1109/ICCV.2015.123

Heaven, W.D. (2020) OpenAI's new language generator GPT-3 is shockingly good—and completely mindless. *MIT Technology Review*, 20 July. Available at: https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/

Heilweil, R. (2020) Networks of self-driving trucks are becoming a reality in the US. *Vox*, 1 July. Available at: https://www.vox.com/recode/2020/7/1/21308539/self-driving-autonomous-trucks-ups-freight-network

Hickey, B. (2020) Investors flock to Hatch and into global market. *Newsroom*, 9 April. Available at: https://www.newsroom.co.nz/investors-flock-to-hatch-and-into-global-market

Hoffman, M., Kahn, L., and Li, D. (2015) Discretion in Hiring. *NBER Working Paper* 21709. Available at: https://www.nber.org/system/files/working_papers/w21709/w21709.pdf

Houkamau, C.A. and Sibley, C.G. (2019) The role of culture and identity for economic values: a quantitative study of Māori attitudes. *Journal of the Royal Society of New Zealand*. 49(sup1): 118–136. Available at: https://doi.org/10.1080/03036758.2019.1650782

Human Rights Commission (2011) *Tracking Equality at Work*. Available at: https://www.hrc.co.nz/files/7714/2360/6761/TrackingEquality.pdf

Human Rights Commission (2016) *The A-Z Pre-Employment Guide for employers & employees*. Available at: https://www.hrc.co.nz/files/1514/6889/8404/HRC_A-Z_Booklet_2016.pdf

IBA (International Bar Association) Global Employment Institute (2017) *Artificial Intelligence and Robotics and Their Impact on the Workplace*. Available at: https://www.ibanet.org/Document/Default.aspx?DocumentUid=c06aa1a3-d355-4866-beda-9a3a8779ba6e

IFOW (Institute for the Future of Work) (2020) *Artificial intelligence in hiring: Assessing impacts on equality*. Available at: https://uploads-ssl.webflow.com/5f57d40eb1c2ef22d8a8ca7e/5f71d338891671faa84de443_IFOW%2B-%2BAssessing%2Bimpacts%2Bon%2Bequality.pdf

ILO (International Labour Organization) (2020) Greening the transport sector in the post COVID-19 recovery could create up to 15 million jobs worldwide. ILO press release, 19 May. https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_745099/lang--en/index.htm

Ignite Research (2006) (Report commissioned by the Legal Services Agency) *Report on the 2006 National Survey of Unmet Legal Needs and Access to Services*.

Intuition Engineering (2018) Deep learning for specific information extraction from unstructured texts. Towards Data Science blog post, 21 July. Available at: https://towardsdatascience.com/deep-learning-for-specific-information-extraction-from-unstructured-texts-12c5b9dceada

Ipeirotis, P.G. (2010) Analyzing the Amazon Mechanical Turk Marketplace. *XRDS*. 17(2): 16–21. Available at: https://doi.org/10.1145/1869086.1869094

Kaminski, M.E. (2015) Robots in the Home: What Will We Have Agreed To. *Idaho Law Review*. 51(3): 661–678.

Kelly, J. (2020a) Finland Prime Minister's Aspirational Goal Of a Six-Hour, Four-Day Workweek: Will It Ever Happen?" *Forbes*, 8 January. Available at: https://www.forbes.com/sites/jackkelly/2020/01/08/finlands-prime-ministers-aspirational-goal-of-a-six-hour-four-day-workweek-will-this-ever-happen/#55689e553638

Kelly, J. (2020b) After Announcing Twitter's Permanent Remote-Work Policy, Jack Dorsey Extends Same Courtesy To Square Employees. *Forbes*, 19 May. Available at: https://www.forbes.com/sites/jackkelly/2020/05/19/after-announcing-twitters-permanent-work-from-home-policy-jack-dorsey-extends-same-courtesy-to-square-employees-this-could-change-the-way-people-work-where-they-live-and-how-much-theyll-be-paid/#4ef1f757614b

Keynes, J.M. (1963) [1930] Economic Possibilities for our Grandchildren. In: *Essays in Persuasion*. New York: W.W. Norton & Co.: 358–373.

Kim, P.T. and Scott, S. (2019) Discrimination in Online Employment Recruiting. *Saint Louis University Law Journal*. 63(1): 93–118.

Ko, B.C. (2018) A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors*. 18(2): 401. Available at: https://doi.org/10.3390/s18020401

Komlos, J. and Küchenhoff, H. (2012) The diminution of the physical stature of the English male population in the eighteenth century. *Cliometrica*. 6(1): 45-62. Available at: https://doi.org/10.1007/s11698-011-0070-7

Kriegeskorte, N. (2015) Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*. 1: 417–446. Available at: https://doi.org/10.1146/annurev-vision-082114-035447

Kriwet, C. (2020) Here are 3 ways AI will change healthcare by 2030. 7 Jan. *World Economic Forum*. Available at: https://www.weforum.org/agenda/2020/01/future-of-artificial-intelligence-healthcare-delivery/

Kwiatkowski, T. et al. (2019) Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*. 7: 453–466. Available at: https://doi.org/10.1162/tacl_a_00276

Lancee, B. and van de Werfhorst, H. (2011) *Income Inequality and Participation: A Comparison of 24 European Countries*. GINI Discussion Paper No. 6. Amsterdam Centre for Inequality Studies. Available at: https://hdl.handle.net/11245/1.351768

Law Society of England and Wales (2018) *Artificial Intelligence (AI) and the Legal Profession*. Horizon scanning report. Available at: https://www.lawsociety.org.uk/en/topics/research/ai-artificial-intelligence-and-the-legal-profession

Lee, M.K. et al. (2015) Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. *CHI 2015: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*: 1603–1612. Available at: https://doi.org/10.1145/2702123.2702548

Leviathan, Y. and Matias, Y. (2018) Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. 8 May, Google AI Blog. Available at: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Levy, D. (2008) *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. London: HarperCollins.

Lipsey, R.G., Carlaw, K.I., and Bekar, C.T. (2005) *Economic Transformations: General Purpose Technologies and Long Term Economic Growth*. Oxford: Oxford University Press.

Litjens G. et al. (2017) A survey on deep learning in medical image analysis. *Medical Image Analysis*. 42: 60–88. Available at: https://doi.org/10.1016/j.media.2017.07.005

Liu, T. (2009) Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval*. 3(3): 225–331. Available at: https://doi.org/10.1561/1500000016

Loh, E. (2018) Medicine and the rise of the robots: a qualitative review of recent advances of artificial intelligence in health. *BMJ Leader*. 2(2): 59–63. Available at: https://doi.org/10.1136/leader-2018-000071

Longoni, C., Bonezzi, A., and Morewe, C.K. (2019) Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*. 46(4): 629–650. Available at: https://doi.org/10.1093/jcr/ucz013

Lopesi, L. (2020) *When Worlds Collide*: Pacific Ideas of Success in New Zealand's Workforce. *Flint & Steel*. 6: 4–7.

Loveys, K., Sagar, M., and Broadbent, E. (2020) The Effect of Multimodal Emotional Expression on Responses to a Digital Human during a Self-Disclosure Conversation: a Computational Analysis of User Language. *Journal of Medical Systems*. 44: 143. Available at: https://doi.org/10.1007/s10916-020-01624-4

Lucas, G.M. et al. (2014) It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*. 37: 94–100. Available at: https://doi.org/10.1016/j.chb.2014.04.043

MBIE (Ministry of Business, Innovation and Employment) (2018) *What is the new regime for financial advice?* Available at: https://www.mbie.govt.nz/dmsdocument/3208-what-is-the-new-regime-for-financial-advise-pdf

MBIE (Ministry of Business, Innovation and Employment) (2019) *Overview – Implementing the Health and Safety at Work Act 2015*: Better Regulation – Plant, Structures and Working at Heights. Available at: https://www.mbie.govt.nz/dmsdocument/5933-overview-implementing-the-health-and-safety-at-work-act-2015-better-regulation-plant-structures-and-working-at-heights

MBIE (Ministry of Business, Innovation and Employment) (2020) *Transport Factsheet*. Available at: https://www.mbie.govt.nz/assets/transport-factsheet.pdf

Maddison, A. (2003) *The World Economy: Historical Statistics*. OECD Publishing. Available at: https://doi.org/10.1787/9789264104143-en

Mak, K-K. and Rao Pichika, M. (2019) Artificial intelligence in drug development: present status and future prospects. *Drug Discovery Today*. 24(3): 773–780. Available at: https://doi.org/10.1016/j.drudis.2018.11.014

Mann, G. and O'Neil, C. (2016) Hiring Algorithms Are Not Neutral. *Harvard Business Review*, 9 December. Available at: https://hbr.org/2016/12/hiring-algorithms-are-not-neutral

Manyika, J. et al. (2017) *A Future that Works: Automation, Employment, and Productivity*. McKinsey Global Institute. Available at: http://www.mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works

Markoff, J. (2016) *Machines of Loving Grace: The Quest for Common Ground Between Humans and Robots*. New York: Ecco Press (reprint edition).

Martinelli, L. (2017) *Assessing the Case for a Universal Basic Income in the UK*. The Institute for Policy Research. Available at: https://www.bath.ac.uk/publications/assessing-the-case-for-a-universal-basic-income-in-the-uk/attachments/ipr-assessing-the-case-for-a-universal-basic-income-in-the-uk.pdf

Mateescu, A. and Nguyen, A. (2019) *Algorithmic Management in the Workplace*. Data & Society Research Institute. Available at: https://datasociety.net/wp-content/uploads/2019/02/DS_Algorithmic_Management_Explainer.pdf

Mathiason, G. et al. (2014) *The Littler Report: The Transformation of the Workplace Through Robotics, Artificial Intelligence, and Automation: Employment and Labor Law Issues, Solutions, and the Legislative and Regulatory Response*. Littler Mendelson. Available at: http://shared.littler.com/tikit/2014/14_Robotics_Event_2-14/pdf/WP_Robotics_2-12-14_download.pdf

McBeth, P. (2018) Sharesies to double in size. *Newsroom*, 25 September. Available at: https://www.newsroom.co.nz/pro/2018/09/25/251913/sharesies-to-double-in-size

McBeth, P. (2019) Fully automated milking several decades away - Dairy NZ. *Scoop*, 18 June. Available at: https://www.scoop.co.nz/stories/BU1906/S00446/fully-automated-milking-several-decades-away-dairy-nz.htm

McDougall, R.J. (2019) Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*. 45(3): 156–160. Available at: https://doi.org/10.1136/medethics-2018-105118

McKinsey Global Institute (2017) *Artificial Intelligence: The Next Digital Frontier?* Discussion Paper. Available at: https://www.mckinsey.com/~/media/mckinsey/industries/advanced%20electronics/our%20insights/how%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/mgi-artificial-intelligence-discussion-paper.ashx

McNeill, J. (2020) How to optimize your CV for the algorithms. Hays. Available at: https://social.hays.com/2018/01/04/optimise-cv-algorithms/

Meade, A. (2020) If Facebook and Google limit services in Australia, could the ABC run a social network? *The Guardian*, 18 October. Available at: https://www.theguardian.com/technology/2020/oct/19/abc-run-social-network-proposed-to-step-in-for-facebook-and-google-in-australia?CMP=Share_iOSApp_Other

Medical Council of New Zealand (2016) *Good Medical Practice*. Available at: https://www.mcnz.org.nz/about-us/publications/good-medical-practice/

Mercader Uguina, J.R. and Muñoz Ruiz, A.B. (2019) Robotics and Health and Safety at Work. International *Journal of Swarm Intelligence and Evolutionary Computation*. 8(1): 176.

Meyers, J. (2020) 5 things COVID-19 has taught us about inequality. *World Economic Forum*, 18 August. Available at: https://www.weforum.org/agenda/2020/08/5-things-covid-19-has-taught-us-about-inequality/

Mill, J.S. (1863) [1998] Crisp, R., ed. *Utilitarianism*. Oxford: Oxford University Press.

Mills, S. (2020) We asked New Zealanders what the world will look like post-Covid-19. *The Spinoff*, 27 June. Available at: https://thespinoff.co.nz/society/27-06-2020/we-asked-new-zealanders-what-the-country-will-look-like-post-covid-19/

Ministry of Health (NZ) (2018a) *Therapeutic Products Regulatory Scheme consultation document*. Available at: https://www.health.govt.nz/publication/therapeutic-products-regulatory-scheme-consultation

Ministry of Health (NZ) (2018b) Family violence questions and answers. Available at: https://www.health.govt.nz/our-work/preventative-health-wellness/family-violence/family-violence-questions-and-answers#mandatory

Möhlmann, M. and Henfridsson, O. (2019) What People Hate About Being Managed by Algorithms, According to a Study of Uber Drivers. *Harvard Business Review*, 30 August. Available at: https://hbr.org/2019/08/what-people-hate-about-being-managed-by-algorithms-according-to-a-study-of-uber-drivers

Mokyr, *The Enlightened Economy: Britain and the Industrial Revolution* 1700—1850 (London: Penguin, 2009).

Moore, P.V., Upchurch, M., and Whittaker, X., eds. (2018) *Humans and Machines at Work: Monitoring, Surveillance and Automation in Contemporary Capitalism*: Palgrave MacMillan. Available at: https://doi.org/10.1007/978-3-319-58232-0

Morgan, G. and Guthrie, S. (2011) *The big kahuna: turning tax and welfare in New Zealand on its head*. Public Interest Publishing.

Muller, C. (2016) *Artificial intelligence – The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society*. INT/806-EESC-2016-05369-00-00-AC-TRA. European Economic and Social Committee. Available at: https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence-consequences-artificial-intelligence-digital-single-market-production-consumption-employment-and

Murad, H. (2017) *Help! A robot took my job!*: Wisdom Waters Press.

Nayeri, S., Sargolzaei, M., and Tulpan, D. (2019) A review of traditional and machine learning methods applied to animal breeding. *Animal Health Research Reviews*. 20(1): 31–46. Available at: https://doi.org/10.1017/S1466252319000148

Nedelkoska, L. and Quintini, G. (2018) *Automation, skill use and training*. OECD Social, Employment and Migration Working Papers No. 202. Available at: https://doi.org/10.1787/2e2f4eea-en

Nelson Jr., J. (2019) Targeted Job Advertisements on Social Media: An Age-Old Practice in a New Suit. *The Global Business Law Review*. 8(1): 1–41.

New Zealand Immigration Concepts. The algorithm-screening friendly CV. Available at: http://www.new-zealand-immigration.com/our-new-zealand-job-search-program/job-search-nz/algorithm-screening-friendly-cv/

New Zealand Productivity Commission (2020) *Technological change and the future of work*. Available at: https://www.productivity.govt.nz/inquiries/technology-and-the-future-of-work/

New Zealand Superannuation Fund (2019) *Annual Report 2019*. Available at: https://www.nzsuperfund.nz/publications/annual-reports/

Nuffield Council on Bioethics (2018) *Bioethics Briefing Note: Artificial intelligence (AI) in healthcare and research*. Available at: https://www.nuffieldbioethics.org/assets/pdfs/Artificial-Intelligence-AI-in-healthcare-and-research.pdf

ODSC (Open Data Science) (2019) AI as the Ultimate Disrupter in Logistics: How to Manage Last-Mile Costs? ODSC blog post, 15 October. Available at: https://medium.com/@ODSC/ai-as-the-ultimate-disrupter-in-logistics-how-to-manage-last-mile-costs-c4874e8f2ea0

OECD (Organisation for Economic Co-operation and Development) (2016) *Average Annual Hours Actually Worked per Worker*. Available at: https://stats.oecd.org/Index.aspx?DataSetCode=ANHRS

OECD (Organisation for Economic Co-operation and Development) (2020) *How's Life? 2020: Measuring Well-being*. OECD Publishing. Available at: https://doi.org/10.1787/9870c393-en

O'Donnell, M. et al. (2001) ILEX: an architecture for a dynamic hypertext generation system. *Natural Language Engineering*. 7(3): 225–250.

O'Kane, P., Walton, S., and Ruwhiu, D. (2020) *Remote Working during COVID19*. Work Futures Otago Reports No. 4. Department of Management, University of Otago. Available at: http://hdl.handle.net/10523/10211

Open AI et al. (2018) Learning Dexterous In-Hand Manipulation. arXiv: 1808.00177v5. Submitted on 1 August 2018 (v1); last revised 18 January 2019 (v5). Available at: https://arxiv.org/abs/1808.00177

Oppenheimer, A. (2019) Fitz, E.E. trans. *The Robots Are Coming!: The Future of Jobs in the Age of Automation*. New York: Vintage Books.

PWC (PricewaterhouseCoopers) (2017) *What doctor? Why AI and robotics will define New Health*. Available at: https://www.pwc.com/gx/en/news-room/docs/what-doctor-why-ai-and-robotics-will-define-new-health.pdf

Panch, T., Mattie, H., and Celi, L.A. (2019) The "inconvenient truth" about AI in healthcare. *npj Digital Medicine*. 2: 77. Available at: https://doi.org/10.1038/s41746-019-0155-4

Panner, M. and the Forbes Technology Council. (2019) AI In Health Care Is Not About Replacing Humans. Forbes, 10 May. Available at: https://www.forbes.com/sites/forbestechcouncil/2019/05/10/ai-in-health-care-is-not-about-replacing-humans/#403e573f6e90

Parikh, R. (2018) AI can't replace doctors. But it can make them better. *MIT Technology Review*, 23 October. Available at: https://www.technologyreview.com/s/612277/ai-cant-replace-doctors-but-it-can-make-them-better/

Parker, T. (2020) Global giant Unilever to trial four-day working week in NZ. *The New Zealand Herald*, 1 December. Available at: https://www.nzherald.co.nz/business/global-giant-unilever-to-trial-four-day-working-week-in-nz/4VNNTOVSDR4UZC6OH2W3FAUPZI/

Pasquale, F. (2020) New Laws of Robotics: *Defending Human Expertise in the Age of AI*. Cambridge: Belknap Press.

Paterson, R. (2015) Regulation of Health Care. In: Skegg, P.D.G. and Paterson, R., eds. *Health Law in New Zealand*. Wellington: Thomson Reuters.

Piachaud, D. (2018) Basic income: confusion, claims and choices. *Journal of Poverty and Social Justice*. 26(3): 299–314. Available at: https://doi.org/10.1332/175982718X15232797708173

Pickett, K.E. and Wilkinson, R.G. (2015) Income inequality and health: A causal review. *Social Science & Medicine*. 128: 316–326. Available at: https://doi.org/10.1016/j.socscimed.2014.12.031

Piketty, T. (2014) Goldhammer, A. trans. *Capital in the Twenty-First Century*. Cambridge: Belknap Press. Available at: https://www.jstor.org/stable/j.ctvjnrvx9

Prassl, J. (2018) *Humans As A Service: The Promise and Perils of Work in the Gig Economy*. New York: Oxford University Press. Available at: https://doi.org/10.1093/oso/9780198797012.001.0001

Pressman, A. (2020) Waymo Reaches 20 Million Miles of Autonomous Driving. *Fortune*, 7 January. Available at: https://fortune.com/2020/01/07/googles-waymo-reaches-20-million-miles-of-autonomous-driving/

Privacy Commissioner (NZ) (2015) *Privacy Impact Assessment Toolkit*. Available at: https://www.privacy.org.nz/publications/guidance-resources/privacy-impact-assessment

Pullar-Strecker, T. (2019) ACC may make up to 300 staff redundant. *Stuff*, 3 April. Available at: https://www.stuff.co.nz/business/111767239/acc-may-make-up-to-300-staff-redundant

RBC (Royal Bank of Canada) (2018) *Humans Wanted: How Canadian youth can thrive in the age of disruption*. Available at: http://hdl.voced.edu.au/10707/471396

RNZ (Radio New Zealand), Nine to Noon (2020) Major change within court system must happen: Chief Justice. 17 June. Available at: https://www.rnz.co.nz/national/programmes/ninetonoon/audio/2018751024/major-change-within-court-system-must-happen-chief-justice

Radford, A. et al. (2019) *Language Models are Unsupervised Multitask Learners*. OpenAI report.

Rauthan, H. (2019) How Conversational Chatbots Marketing is the Future of eCommerce. Towards Data Science blog post, 7 August. Available at: https://towardsdatascience.com/how-conversational-chatbots-marketing-is-the-future-of-ecommerce-6743268caa11

Raworth, K. (2017) *Doughnut Economics: Seven Ways to Think Like a 21st-Century Economist*. White River Junction, VT: Chelsea Green Publishing.

Reich, R. (2020) Covid-19 pandemic shines a light on a new kind of class divide and its inequalities. *The Guardian*, 26 April. Available at: https://www.theguardian.com/commentisfree/2020/apr/25/covid-19-pandemic-shines-a-light-on-a-new-kind-of-class-divide-and-its-inequalities

Reisman D. et al. (AI Now Institute) (2018) *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. Available at: https://ainowinstitute.org/aiareport2018.pdf

Rhue, L. (2019) Emotion-reading tech fails the racial bias test. *The Conversation*, 3 January. Available at: https://theconversation.com/emotion-reading-tech-fails-the-racial-bias-test-108404

Rifkin, J. (1996) *The End of Work: The Decline of the Global Labor Force and the Dawn of the Post-Market Era*. New York: Putnam.

Robinson, H. et al. (2013). The Psychosocial Effects of a Companion Robot: A Randomized Controlled Trial. *Journal of American Medical Directors Association*. 14(9): 661–667. Available at: https://doi.org/10.1016/j.jamda.2013.02.007

Rosenblat, A. (2019) *Uberland: How Algorithms Are Rewriting the Rules of Work*. Oakland: University of California Press. Available at: https://www.jstor.org/stable/10.1525/j.ctv5cgbm3

Rotenberg, M., Bannan, C., and Davisson, J., EPIC (Electronic Privacy Information Center) (2019) *Complaint and Request to the Federal Trade Commission, for Investigation, Injunction, and Other Relief*. Available at: https://epic.org/privacy/ftc/hirevue/EPIC_FTC_HireVue_Complaint.pdf

Roth, P. (2016) Privacy Law Reform in New Zealand: Will it Touch the Workplace? *New Zealand Journal of Employment Relations*. 41(2): 36–57.

Rousseau, J-J. n(1762) Emile, or On Education. In: *Oeuvres Complètes*. Volume 4.

Royal Society and British Academy (2018) *The Impact of Artificial Intelligence on Work: An evidence review prepared for the Royal Society and the British Academy*. Available at: https://royalsociety.org/-/media/policy/projects/ai-and-work/frontier-review-the-impact-of-AI-on-work.pdf

Russell, B. (1932) In Praise of Idleness. In: *In Praise of Idleness and Other Essays*. London: Allen and Unwin.

Salmon, K. (2019) *Jobs, Robots & Us: Why the Future of Work in New Zealand is in Our Hands*. Wellington: Bridget Williams Books. Available at: https://doi.org/10.7810/9781988545882

Sánchez-Monedero, J., Dencik, L., and Edwards, L. (2020) What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In: *FAT 20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM: 458–468. Available at: https://doi.org/10.1145/3351095.3372849

Sandberg, S. (2019) Doing More to Protect Against Discrimination in Housing, Employment and Credit Advertising. Facebook, 19 March. Available at: https://about.fb.com/news/2019/03/protecting-against-discrimination-in-ads/

Sandel, M. J. (2020) *The Tyranny of Merit: What's Become of the Common Good?*: Penguin.

Scanlon, L. (2020) AI in financial services: addressing the risk of bias. Pinsent Masons, 8 July. Available at: https://www.pinsentmasons.com/out-law/analysis/ai-financial-services-risk-of-bias

Scharf, R. (2019) *Alexa is Stealing Your Job: The Impact of Artificial Intelligence on Your Future*. New York: Morgan James Publishing.

Schenker, J. (2017) *Robot-Proof Yourself: How to Survive the Robocalypse and Benefit from Robots and Automation*: Prestige Professional Publishing.

Schwab, K. (2016) *The Fourth Industrial Revolution: what it means, how to respond. World Economic Forum*, 14 January. Available at: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Schwellnus, C., Kappeler, A., and Pionnier, P-A. (2019) *Decoupling of wages from productivity: Macro-level facts*. OECD Economics Department Working Papers No. 1373. Available at: https://doi.org/10.1787/d4764493-en

Schweyer, A. (2016) *Robots in Recruiting: The Implications of AI on Talent Acquisition*. White paper. Available at: https://www.slideshare.net/appcast_io/whitepaper-robots-in-recruiting-the-implications-of-ai-on-talent-acquisition

Seldon, A. (2019) The impact of AI on the surveillance industry. Hi-Tech Security Solutions CCTV Handbook 2019. Available at: https://www.securitysa.com/8841r

Seligman, M.E.P. (2011) *Flourish: A Visionary New Understanding of Happiness and Well-being*. New York: Free Press.

Serbera, J-P. (2019) Flash crashes: if reforms aren't ramped up, the next one could spell global disaster. *The Conversation*, 8 January. Available at: https://theconversation.com/flash-crashes-if-reforms-arent-ramped-up-the-next-one-could-spell-global-disaster-109362

Sharkey, A. and Sharkey, N. (2012) Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*. 14(1): 27–40. Available at: https://doi.org/10.1007/s10676-010-9234-6

Shields, J. (2018) Over 98% of Fortune 500 Companies Use Applicant Tracking Systems (ATS). Jobscan blog post, 20 June. Available at: https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems/

Simonite, T. (2020) The Therapist Is In—and It's a Chatbot App. WIRED, 17 June. Available at: https://www.wired.com/story/therapist-in-chatbot-app/

Smith, A. and Anderson, J. (2014) AI, Robotics, and the Future of Jobs. Pew Research Center, 6 August. Available at: https://www.pewresearch.org/internet/2014/08/06/future-of-jobs/

Smith, N. (2019) How AI could fill NZ's $500m healthcare hole – and more. *National Business Review*, 22 October. Available at: https://www.nbr.co.nz/story/ai-nz-healthcare-could-add-700m-value

Smutny, P. and Schreiberova, P. (2020) Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers and Education*. 151: 103862. Available at: https://doi.org/10.1016/j.compedu.2020.103862

Speicher, T. et al. (2018) Potential for Discrimination in Online Targeted Advertising. *Proceedings of Machine Learning Research*. 81: 1–15.

Statistics New Zealand (2018) *Survey of Working Life: 2018*.

Statistics New Zealand (2021) *Characteristics of the underemployed in New Zealand*. Available at: https://www.stats.govt.nz/reports/characteristics-of-the-underemployed-in-new-zealand

Statt, N. (2018) Google now says controversial AI voice calling system will identify itself to humans. *The Verge*, 10 May. Available at: https://www.theverge.com/2018/5/10/17342414/google-duplex-ai-assistant-voice-calling-identify-itself-update

Stegmaier, G. et al. (2020) New Illinois employment law signals increased state focus on artificial intelligence in 2020. *Technology Law Dispatch*, 21 January. Available at: https://www.technologylawdispatch.com/2020/01/privacy-data-protection/new-illinois-employment-law-signals-increased-state-focus-on-artificial-intelligence-in-2020/

Steininger, T. (2020) Autonomous Vehicles: The Future (And The Past) Of Food And Beverage Logistics. *Forbes*, 14 October. Available at: https://www.forbes.com/sites/tomsteininger/2020/10/14/autonomous-vehicles-the-future-and-the-past-of-food-and-beverage-logistics/?sh=4286c506f119

Subbe, C.P. (2020) A&E waiting times worst on record – but using AI to unblock beds could be part of the solution. *The Conversation*, 24 July. Available at: https://theconversation.com/aande-waiting-times-worst-on-record-but-using-ai-to-unblock-beds-could-be-part-of-the-solution-130234

Susskind, R. and Susskind, D. (2016) *The Future of the Professions: How Technology Will Transform the Work of Human Experts*. Oxford: Oxford University Press.

Suzman, J. (2020) *Work: A History of How We Spend Our Time*. London: Bloomsbury.

Talaviya, T. et al. (2020) Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. *Artificial Intelligence in Agriculture*. 4: 58–73. Available at: https://doi.org/10.1016/j.aiia.2020.04.002

The Week (2015) The rise of workplace spying. 5 July. Available at: https://theweek.com/articles/564263/rise-workplace-spying

Theodore, R. et al. (2017) Equity in New Zealand university graduate outcomes: Māori and Pacific graduates. *Higher Education Research & Development*. 37(1): 206–221. Available at: https://doi.org/10.1080/07294360.2017.1344198

Thompson, D. (2019) Workism Is Making Americans Miserable. *The Atlantic*, 24 February. Available at: https://www.theatlantic.com/ideas/archive/2019/02/religion-workism-making-americans-miserable/583441/

Tredinnick, L. (2017) Artificial intelligence and professional roles. *Business Information Review*. 34(1): 37–41. Available at: https://doi.org/10.1177/0266382117692621

Turner, A. (2018) Capitalism in the age of robots: work, income and wealth in the 21st-century. Lecture at School of Advanced International Studies, John Hopkins University, Washington DC, 10 April. Available at: https://www.ineteconomics.org/research/research-papers/capitalism-in-the-age-of-robots-work-income-and-wealth-in-the-21st-century

Turner, J. (2019) *Robot Rules: Regulating Artificial Intelligence*: Palgrave Macmillan. Available at: https://doi.org/10.1007/978-3-319-96235-1

Turner, J. and Tanna, M. (2019) AI-powered investments: Who (if anyone) is liable when it goes wrong? Tyndaris v VWM. Thomson Reuters Practical Law Dispute Resolution Blog, 27 September. Available at: http://disputeresolutionblog.practicallaw.com/ai-powered-investments-who-if-anyone-is-liable-when-it-goes-wrong-tyndaris-v-vwm/

University of Otago Legal Issues Centre (2019) *Accessing Legal Services: The Price of Litigation Services*. Working Paper. Available at: http://hdl.handle.net/10523/9524

Vaithianathan, R. et al. (2019). Allegheny Family Screening Tool: Methodology, Version 2. Auckland: Centre for Social Data Analytics.

Valcour, M. (2014) The Power of Dignity in the Workplace. *Harvard Business Review*, 28 April. Available at: https://hbr.org/2014/04/the-power-of-dignity-in-the-workplace

Van Hattum, B. (2019) The future belongs to small self-driving tractors. *Future Farming*, 19 September. Available at: https://www.futurefarming.com/Machinery/Articles/2019/9/The-future-belongs-to-small-self-driving-tractors-474180E

Vanian, J. (2020) How chatbots are helping in the fight against COVID-19. *Fortune*, 16 July. Available at: https://fortune.com/2020/07/15/covid-coronavirus-artificial-intelligence-triage/

Vincent, J. (2018) Google's AI sounds like a human on the phone — should we be worried? *The Verge*, 9 May. Available at: https://www.theverge.com/2018/5/9/17334658/google-ai-phone-call-assistant-duplex-ethical-social-implications

Vladeck, D.C. (2014) Machines without Principals: Liability Rules and Artificial Intelligence. *Washington Law Review*. 89(1): 117–150.

Vlugter, P. et al. (2009) Dialogue-based CALL: a case study on teaching pronouns. *Computer Assisted Language Learning*. 22(2): 115–131. Available at: https://doi.org/10.1080/09588220902778260

Volini, E. et al. (2019) Accessing talent: It's more than acquisition – 2019 Global Human Capital Trends. Deloitte, 11 April. Available at: https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2019/talent-acquisition-trends-strategies.html

Voytek, B. (2014) Optimizing a dispatch system using an AI simulation framework. Uber Newsroom, 11 August. Available at: https://www.uber.com/newsroom/semi-automated-science-using-an-ai-simulation-framework

Vrij, A., Hartwig, M., and Granhag, P.A. (2019) Reading Lies: Nonverbal Communication and Deception. *Annual Review of Psychology*. 70: 295–317. Available at: https://doi.org/10.1146/annurev-psych-010418-103135

Walsh, T. et al. (2019) *The effective and ethical development of artificial intelligence: An opportunity to improve our wellbeing*. Australian Council of Learned Academies. Available at: https://acola.org/hs4-artificial-intelligence-australia/

Wan, L. et al. (2019) Long-length Legal Document Classification. arXiv: 1912.06905v1. Submitted on 14 December 2019 (v1). Available at: https://arxiv.org/abs/1912.06905

Way, B. (2013) *Jobpocalypse: The End of Human Jobs and How Robots will Replace Them*. Createspace.

Westfall, B. (2019) 3 HR Chatbots That Are Disrupting Employee Experience. Capterra blog post, 20 June. Available at: https://blog.capterra.com/hr-chatbots/

White, E. (2020) Over 10,000 Kiwis face increased surgery waiting lists due to COVID-19 lockdown. *Newshub*, 22 August. https://www.newshub.co.nz/home/new-zealand/2020/08/over-10-000-kiwis-face-increased-surgery-waiting-lists-due-to-covid-19-lockdown.html

White, G. (2018) Child advice chatbots fail to spot sexual abuse. *BBC News*, 11 December. Available at: https://www.bbc.com/news/technology-46507900

Willcocks, L. (2020) Robo-Apocalypse cancelled? Reframing the automation and future of work debate. *Journal of Information Technology* 35(4): 86–302. Available at: https://doi.org/10.1177/0268396220925830

Winton, B. (2019*) Disruptive Innovation: Why Now?* ARK Invest. Available at: https://research.ark-invest.com/hubfs/1_Download_Files_ARK-Invest/White_Papers/ARK%20Invest_052919_whitepaper_DI-Why-Now.pdf

Womack, J.P., Jones, D.T., and Roos, D. (1991) *The Machine That Changed the World: The Story of Lean Production*. New York: Harper-Perennial.

Worksafe New Zealand (2014) *Safe Use of Machinery*. Available at: https://worksafe.govt.nz/topic-and-industry/manufacturing/safe-use-of-machinery

Worksafe New Zealand (2017) *Health monitoring and exposure monitoring*. Available at: https://www.worksafe.govt.nz/topic-and-industry/work-related-health/monitoring/

Wouter, S., Luiijf, E., and van der Beek, D. (2016) *Emergent risk to workplace safety as a result of the use of robots in the work place (The TNO Report)*. TNO 2016 R11488.

World Economic Forum (2018) *The Future of Jobs Report: 2018*. Available at: http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf

Yearsley, L. (2017) We Need to Talk About the Power of AI to Manipulate Humans. *MIT Technology Review*, 5 June. Available at: https://www.technologyreview.com/2017/06/05/105817/we-need-to-talk-about-the-power-of-ai-to-manipulate-humans/

Youyou, W., Kosinski, M., and Stillwell, D. (2015) Computer-based personality judgments are more accurate than those made by humans. *PNAS*. 112(4): 1036–1040. Available at: https://doi.org/10.1073/pnas.1418680112

Zerilli, J. et al. (2019a) Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology*. 32(4): 661–683. Available at: https://doi.org/10.1007/s13347-018-0330-6

Zerilli, J. et al. (2019b) Algorithmic decision-making and the control problem. *Minds and Machines* 29 (4), 555-578

Zerilli, J. et al. (2021) A *Citizen's Guide to Artificial Intelligence*. Cambridge: MIT Press.

Zilibotti, F. (2008) *Economic Possibilities for our Grandchildren* 75 Years After: a Global Perspective. In: Pecchi, L. and Piga, G., eds. *Revisiting Keynes: Economic Possibilities for our Grandchildren*. Cambridge: MIT Press: 27–40.

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON JOBS AND WORK IN NEW ZEALAND