Research insights

Digital Council of Aotearoa New Zealand: trust and automated decision-making

November 2020

ISSUED BY

Brainbox Institute (Ltd)

REPRESENTATIVES

Tom Barraclough, Curtis Barnes
tom@brainbox.institute / curtis@brainbox.institute
+6421900848 / +64212477739

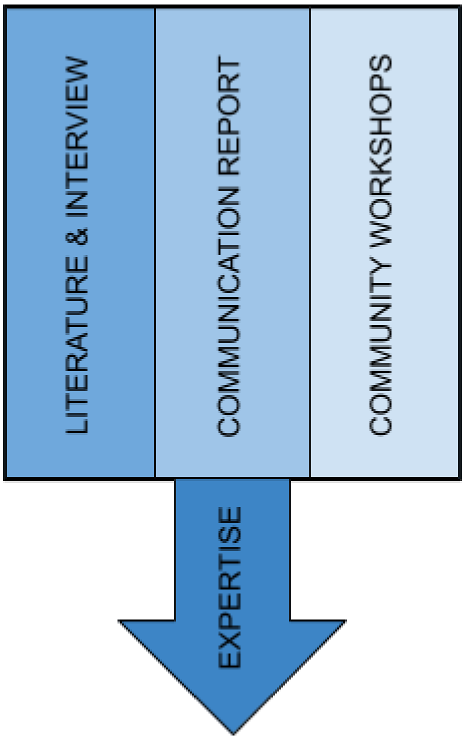
PREFACE

Background to this document

The information contained in this document is intended to provide insights developed from reviewing literature, talking to experts, and conducting our own analysis. This process primarily took place between 4 May and 10 July.

At the outset, it became clear to us that neither "trust" nor "ADM" are the subject of a single distinct body of literature. Rather, each are broad concepts that overlap and envelop numerous other literatures, each very large and complex in its own right. These bodies of literature also incorporate the massive emerging literature on artificial intelligence and ethics, which was simply not possible to cover in the time available. As a result, our aim was to distill key elements of these broad topic areas into insights that were accessible, communicable, and manageable, primarily for informing the next stage of the Council's research stream.

This document and its analysis form part of a wider research collaboration. For its part, the document provides information and guidance that supports and shapes other aspects of the research stream (depicted in the diagram shown). We hope that this document will:



1) Inform and be informed by the qualitative work by Toi Āria, which through a programme of workshops, gathered local perspectives on trust and ADM from a range of New Zealand citizens.

2) Inform the report by AntiStatic, which aims to communicate a clear framing of trust and ADM for a New Zealand audience by incorporating what is learned across the research stream.

3) Inform the Council downstream at the point that they write and talk about the subject, particularly to key government officials and stakeholders.

4) Synthesise our pre-existing expertise with the subject matter to provide a nuanced perspective.

5) Provide a record of resources in the reference list for interested people to consult.

To that end, the document contains three main features, broadly split into three sections:

1) Our summary analysis.

2) Analysis of key ideas raised in the literature.

3) Our reference list, which serves as a record for further research

The state of the literature

As a phenomenon, ADM is inseparable from the literature around digital privacy and the trusted use of data-driven technologies. Because of this, the subject area is saturated with work, both globally and locally. Internationally there is a high degree of public scrutiny, particularly from human and civil rights-based perspectives. Locally, there is increasing appreciation of the value of incorporating perspectives from different stakeholders into the use and development process of certain ADM systems.

The emphasis on human and civil rights in the use of digital technologies naturally overlaps with the literature and advocacy around Māori data sovereignty, informed by values from te Ao Māori like tikanga, tino rangatiratanga, kaitiakitanga, whanaungatanga and kotahitanga. In light of this, we advised the Council to seek a dedicated literature review on this topic, to be conducted by researchers with standing, experience, and expertise in mātauranga Māori, noting that New Zealand researchers lead the world on this topic. That has led to a report by Te Kotahi research institute which is a valuable and important resource for the Council and for senior policy makers.

There is a dense theoretical literature around "trust", "automation", and "automated decision-making", to the extent that it is inseparable from contemporary discussion around ethics and artificial intelligence. In practice, the Council's focus has also tended toward government and public applications of such technologies, which also incorporates a whole range of literature around human and civil rights and methods of organising government intervention. Based on our reading of the literature on these topics, participation in expert discussions, and prior knowledge of digital technologies, what follows is a summary of the relationship between trust and ADM.

There is a well-developed literature focused on the working relationship between automated systems and human operators. This provides a range of insights relating to trust and its components, with a view to improving the adoption and proper use of ADM systems. We suggest any future work focus on this literature because it provides the best practical insights into the ways that trust is generated in automated digital systems by humans who use them, with associated impacts for people subject to them.

TABLE OF CONTENTS

SUMMARY OF ANALYSIS

Understanding Automated Decisions Making systems

"Automated decision making" encompasses most digital or data-driven technologies

"ADM" is not really a distinct technology. Rather, it encompasses a broad array of digital and data-driven technology groups. Each of these is the subject of its own vast literature. For instance, ADM certainly includes AI, which is itself a clustering concept for a wide variety of technologies. As such, "ADM" is a difficult subject to investigate concisely, and "trust in ADM" covers most of what humans do with computers, especially when "trust" is taken to also include the technical and contextual effectiveness and reliability of the system in question.

ADM systems are many-layered

While it is tempting to talk about ADM systems as single units, such systems have many different levels and layers operating in conjunction to deliver a workload or output. The layers are both social and institutional, as well as digital and technical. As such, these are inherently complex systems that require nuanced analysis specific to each system in its context.

ADM systems contain interactions between people + automation at most layers

All ADM systems use a combination of humans and computers to deliver their outputs or workload, though the ratios of each may vary for different systems and at different points in the system. It is incorrect to talk about ADM systems as operating without human input. Moreover, this requires that analysis advance beyond merely stipulating that systems should have humans "on" or "in" the "loop", and so on. At each level of an ADM system, both humans and computers are most likely present.

All ADM is a product of human agency

There is no ADM system which is entirely 'free floating', or without a human touchpoint. For instance, even highly autonomous systems were designed by humans and deployed by humans. This could (and should) mean that there is always a person responsible for part or whole of an ADM system. This responsibility encourages further discussion about legal accountability.

Understanding trust

Trust is being vulnerable to another party because of an expectation of positive actions by that party, in situations of risk and interdependence

We found several papers that attempted a definition of trust across a range of disciplines.[1] By way of summary, the literature proposes a succinct theoretical concept for trust, which is that it is a psychological state where someone is willing to be vulnerable to another's power over them, based on positive expectations of that person's actions. Added to this is that trust is assessed in a situation of some interdependence between the actors and a risk that either party may be negatively affected by the other's actions.[2] Building on these concepts we can conclude that being *distrusting* is where a person is unwilling to be vulnerable to the power wielded by another party, or where they have negative expectations about how another person's actions may affect them. In some situations, trust and distrust may co-exist, for example in a business transaction between competitors.

Trust is related to, but different, from "trustworthiness"

It is possible to trust someone or something based on a misperception. The literature distinguishes the concept of trustworthiness from trust, and relates that particularly to whether something or someone justifies another's trust or not, depending on factors like predictability and reliability or other components of trust. In an interpersonal or institutional setting, it is also possible for trust to be misplaced, due to incorrect perception of the likelihood of positive outcomes, or the result of manipulating perception.

Social, historical, and psychological reasons may make a person unwilling to be vulnerable, or to have negative expectations of how someone else's actions may affect them

The literature refers to the role of trust propensity, which is someone's willingness to be vulnerable even where they might have had no opportunity to test the past conduct of the other party or have access to other means of testing trustworthiness.

There are many reasons why somebody might have negative expectations about the risk to them from a person's actions based on an uneven power dynamic, and they include historical experiences, personal experiences, and psychological states. Groups and demographics who have previously been harmed in relationships of uneven power are likely to have an expectation of negative outcomes from those relationships and, subsequently, will experience low trust. This is made clear by various international perspectives on the risks of specific ADM systems to disempowered groups with experiences of discrimination, as well as the insights drawn from the Toi Āria workshop participants and the te Kotahi wananga.

[1] Rousseau, D.M., S.B. Sitkin, R.S. Burt and C. Camerer (1998), 'Not so different after all: a cross-discipline view of trust', Academy of Management Review, 23(3), 292–404. D. Harrison McKnight and Norman L. Chervany "Trust and Distrust Definitions: One Bite at a Time" in R. Falcone, M. Singh, and Y.-H. Tan (Eds.): Trust in Cyber-societies, LNAI 2246, pp. 27–54, 2001: "An analysis of the word trust in three unabridged dictionaries (Websters, Random House, and Oxford) showed that trust had far more definitions (9, 24, and 18, respectively) than did the terms cooperation (3, 2, 6), confidence (6, 8, 13), and predictable (1, 2, 1). On average, trust had 17.0 definitions, while the others had an average of 4.7. Trust had close to as many definitions as did the very vague terms 'love' and 'like.'
[2] Rousseau et al (ibid).

Trust is inherently related to power dynamics and control

In theoretical conceptualisations, and in practice, trust is an issue that arises in relation to power relations between interdependent individuals or groups. As a result, resolving low levels of trust can occur in two broad ways. Either one can attempt to improve the perception of positive effect between the parties (i.e. make Person X believe Person Y will use a power to Person X's benefit), or to re-balance power between the individuals. The literature speaks of an interrelationship between trust and control: "Trust and control are complements as well as substitutes ... Complete trust, that is, unconditional or blind trust, is ill advised, and where trust ends one needs control. Vice versa, complete control is impossible, and trust is needed where control ends. At the same time, more trust allows for less control."[3]

In reality, changing the perception of a more vulnerable party can be extremely difficult, particularly when considering groups of people, rather than individuals. For instance, regarding relationships between Māori and the Crown, improving Māori perception of how Crown powers will be used works against centuries of negative interactions. This sits alongside the fact that making somebody believe you will use your power to benefit them does not mean you actually will do so - the matter of trustworthiness. On the other hand, it can also be possible to change the perception of another party easily in the right situations, for example through deliberate messaging and communications where trust propensity is already high, for example in groups of people who have no significant history of government abuses of power.

If power is redistributed between the parties to a relationship, to some extent, this actually eliminates the need for trust at all. In other words, by granting power to Party X within its relationship to Party Y, and incorporating accountability and oversight mechanisms within systems, there is a lesser emphasis placed on trust, because the power imbalance has been rendered neutral by greater control, limiting vulnerability and risk. In other words, if low trust is to be remedied by incorporating within a system a raft of new accountability mechanisms, oversight mechanisms, and punitive mechanisms, then the problem has been resolved not through enhancing trust between the parties, but by making breach of trust much more risky and difficult for the party that holds a power. We noted from work by Toi Āria and te Kotahi, that people must still have confidence in those mechanisms for them to be effective.

Understanding trust in ADM systems

ADM systems rely on trust 'internally' between people and automation

As above, ADM systems are multi-layered interactions between people and computers to deliver a workload. In many instances, the people working 'inside' an ADM system must have trust in the computer systems in order for the system to be effective. Inadequate or excessive trust in these computer systems can result in over or under reliance, with subsequent ineffectiveness. Equally, people must trust other people - those that build, maintain, and monitor the ADM system. For example, if senior decision-makers do not have trust in the people building ADM systems, they will not trust the system itself.

[3] Bart Nooteboom "Trust and innovation" in Handbook of Advances in Trust Research (Reinhard Bachmann and Akbar Zaheer, eds) Edward Elgar 2013.

ADM systems rely on trust 'externally' between the system and the people who are impacted

ADM systems also tend to have outputs that affect people who are external to the system. For example, people outside an ADM system will be impacted in some way by what the ADM system delivers in many cases. Frequently, these external people must have trust in the ADM system for it to effectively deliver the output of its workload. For instance, where a system is intended to deliver a service as an output, it will be unable to do so where people refuse to engage with those outputs due to low levels of trust in the system that delivers them.

Trust is relevant in every relationship within an ADM system

As above, there are many different people situated internally and externally to an ADM system, and trust matters between all of them. For example, a commercial pilot must trust aircraft technicians who she has never met, which means she must also trust the people who trained the technicians (who she also never met), and so on. There are many more links in this chain, all of which must be present to ensure the pilot is confident to fly the plane (and indeed, rely on its automated systems). Without this trust, she would never even start the plane's engines.

The same is true for people external to the ADM system. Not only must they have trust towards the system 'as a whole', but also towards each person within that ADM system: trust in the designers, the builders, the communicators, the data collectors, the procurers, the decision-makers, the contractors, etc. A distrust towards even one person among the hundred that comprise an ADM system may undermine trust in the entire ADM system. We've since observed this phenomenon corroborated by participants in the Toi Āria workshops.

Trust is relevant at every level of the ADM system

Moreover, this holds true for every level of a multi-layered ADM system. There is also an 'upstream, downstream' element. For example, if a person does not trust the data collection upstream, they are unlikely to trust the decisions made downstream based on that data. If a person does not trust a government or a company that procures or builds an ADM system in the first instance, then they probably will not trust any subsequent element of that system - even if it is a perfectly good system in terms of its intended purpose and use.

Improving trust in ADM systems

Trust is always a question of "the appropriate level"

Trust shouldn't be thought of as a binary - i.e. I trust the ADM system, or I don't trust it. Rather, trust in ADM is always a question of "appropriate level" depending on the capabilities of the system and its role in an overall human-computer workload. According to the literature, trust is also dynamic and changes over time. Whether levels of trust are "appropriate" becomes a contextual and applied question, rather than being able to be assessed in the abstract. This in turn compels us to think about how, where, and why the ADM system is being deployed, what it does well, and what it does poorly. Communicating these facts to operators and outsiders is an important step in developing appropriate levels of trust.

Degree of trust in an ADM system may be shaped by a range of consistent factors:

The literature assessing interactions between human operators and automated computer systems suggests some consistent factors that affect the level of trust those humans have in the automation. Ensuring that human trust is at an appropriate level is extremely important to get the most out of the systems, especially in high-stakes, high-pace operation environments. As such, the following are possible factors that a human assesses to determine their level of trust in a system:[4]

- "Competence" – the system's ability to do the prescribed task
- "Predictability" – the extent to which the operator can predict the system's behaviour
- "Dependability" – the system's ability to perform in all anticipated situations
- "Consistency" – low variability in the system's behaviour
- "Confidence – the extent to which the operator can act on the system's outputs

Perception of ADM systems may be shaped by a range of consistent factors:

Some of the literature describing empirical research into how ADM systems are perceived concluded that there is variation around the following factors:[5]

- "Fairness" – the system is perceived to be not unfair, or not used unfairly
- "Usefulness" –the system is perceived as making a positive contribution to society
- "Riskiness" – the systems is perceived as potentially socially harmful or undesirable

Depending on how these factors were perceived, people had different attitudes towards ADM systems, either more enthusiastic or less enthusiastic.

We've since seen these factors confirmed by participants in the Toi Āria workshops.

Good communication is important for establishing appropriate levels of trust

It is important to communicate honestly and effectively about what an ADM system does and what it is capable of, including its limitations and flaws. Sometimes, this information can be very complex. The literature shows that adverse outcomes are more likely where people are instilled with false confidence in ADM systems, or alternatively, not enough confidence in them. In conjunction with finding "the appropriate level" of trust, we can see that communication serves the purpose of equipping people (internal and external to an ADM system) with accurate information about the specific things the system can do well, and vice versa.

[4] Miller, E.J., Perkins, L. "Development of Metrics for Trust in Automation", Proceedings of the 15th International Command and Control Research and Technology Symposium(ICCRTS '10), Santa Monica, CA, June 22-24, 2010

[5] Araujo, T., Helberger, N., Kruikemeier, S. et al. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Soc (2020).

State of the literature

The international literature is vast

Depending on how you scope the topic of “trust and/in ADM”, it is easy to reach the conclusion that there is an enormous amount of attention being given to the subject worldwide. For example, to the extent that “AI” and “data science” (particularly the discourse around ethics in these subjects) fall under the umbrella of ADM, the amount of information around this subject is vast and growing more every day.

There are many frameworks and set of principles around AI ethics, data use, and ADM

Across both the private and public sectors globally, there are many different attempts at frameworks, principles, and high-level covenants around ADM-related technologies. These, in turn, are subjects of analysis (both individually and in the aggregate). A common criticism of these is that they are abstract so as to be broadly agreeable to the widest possible audience. However, as a result, they lack the specificity and contextualisation necessary for them to guide real-world systems in their operational context.

Need for real case studies and use cases

The literature has reached a point where the limitations of abstract norms and guidance has become widely accepted.[6] Frequently, stakeholders now advocate a need for applied thinking – either through assessing existing algorithmic systems, or as part of the development of new algorithms within their domain-specific contexts. We strongly recommend that any future work by the Council focus on real testing of specific systems in actual use cases, progressing the work that it has already done with Toi Āria’s participatory research to the next level.

Trust literature extends across many disciplines

Trust is a complex phenomenon, the analysis of which is the subject of a variety of disciplinary approaches. This includes but is not limited to psychological and sociological applications and cross-disciplinary meta-analyses.[7]

The New Zealand perspective

New Zealand already deploys a range of private and public sector ADM systems

Like most countries, ADM is an important part of service and product delivery in contemporary New Zealand. As a nation and economy, we are subject to the same sorts of

[6] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter and Luciano Floridi, "The ethics of algorithms: Mapping the debate" (2016) Big Data & Society July-December 1-21.

[7] D. Harrison McKnight and Norman L. Chervany "Trust and Distrust Definitions: One Bite at a Time" in R. Falcone, M. Singh, and Y.-H. Tan (Eds.): Trust in Cyber-societies, LNAI 2246, pp. 27–54, 2001. Rousseau, D.M., S.B. Sitkin, R.S. Burt and C. Camerer (1998), ‘Not so different after all: a cross-discipline view of trust’, Academy of Management Review, 23(3), 292–404. Jason A. Colquitt, Brent A. Scott, and Jeffery A. LePine "Trust, Trustworthiness, and Trust Propensity: A Meta-Analytic Test of Their Unique Relationships With Risk Taking and Job Performance" (2007) 92(4) Journal of Applied Psychology 909-927

conditions, pressures, and resource constraints which encourage the use of ADM systems. As a result, there are ADM systems currently in operation which should be used as case studies for further analysis.[8] Some of these have been touched upon in the course of the Toi Āria workshops. We also note that there will be many systems we are not aware of, some of which will be commercially sensitive. Among the many ADM systems operating in New Zealand, we expect the importance and implications of trust to vary on a case by case basis, reflecting variations in the data, method, effect, and context in which the system operates. In short, an ADM system for sorting and packaging kiwifruit will differ from one for processing votes in the general election, or for determining the availability of credit, or for delivering a government entitlement, or for flying an aeroplane. These are all ADM systems, vastly different in their workload, outputs, technologies, risks, and effects.

There have been attempts to address ADM and data governance in New Zealand

The risks of data use and ADM have been considered within the New Zealand context.[9] Concern around these risks has catalysed several attempts to develop best practice, which in turn has yielded outputs like the Statistics New Zealand Algorithm Charter – a voluntary piece of guidance. The Charter is new and needs a chance to be operationalised, however, the Charter has already been rejected by some government agencies and we predict it will struggle to make inroads into influencing the use of ADM within the public sector for a number of practical reasons, many of which cannot be addressed by high-level, generalised norms such as those included in the charter.

Existing scholarship on data sovereignty overlaps with the question of trust and ADM

Again, “trust and ADM” encompasses many other concepts and phenomena, including data use. As such, Māori data sovereignty is a significant topic. In New Zealand, Māori data sovereignty is subject to extensive academic research, and is still growing. In particular, this considers data sovereignty rights in light of the Treaty of Waitangi and its implications.

[8] For example, we discuss the SmartGate system later in this report and include a range of publicly available resources in the reference list about “automated electronic systems” in New Zealand law for making decisions.

[9] See for example Office of the Privacy Commissioner "Submission to the Government Administration Committee on the Border (Customs, Excise and Tariff) Processing Bill 2009" (30 October 2019).

UNDERSTANDING AUTOMATED DECISION-MAKING

What is automated decision-making?

> Automated decision-making can be defined in several ways. Narrowly, it may be described as decisions by technological means without human involvement. Broadly, it may be described as the increasingly common process by which personal data is processed by computers and used to help make data-driven decisions.

It is all too easy to become bogged down while attempting to precisely define what ADM is. At the same time, we heard from interviews that it is essential to have a robust working concept of ADM to inform accurate investigation and avoid directionless discussions about ambiguous topics. Ultimately, our analysis leads us to the following summary:

> Automation is a means to augment human activity. These human activities invariably require or involve decision-making. Automating parts of that activity forms part of a human-machine workload. In this sense, most automated decision-making is a "socio-technical" process, because the process involves both human and technological elements to deliver the work.[10] The ratios by which this workload is divided can lean more heavily towards the machine or towards the human. Either way, the use of the automation augments normal human work capacity, and in doing so, extends the range, speed, and volume of many human activities.

One of our key findings is that ADM is actually a conceptual group that encompasses a vast variety of digital and data-driven technologies. This is the cause of significant confusion, and it makes ADM a challenging subject for a literature review. For example, ADM extensively incorporates elements of the following technical subjects, all of which are subject to their own public, technical, and legal discourses:

- Conventional computing
- Machine learning
- Algorithms
- Personal data collection and use
- Data science
- Artificial Intelligence

A definition we found that mirrors these conclusions is adopted by Canada in its Directive on Automated Decision-making, which defines an automated decision-system as follows:[11]

> *Automated Decision System includes any technology that either assists or replaces the judgement of human decision-makers. These systems draw from fields like statistics, linguistics, and computer science, and use techniques such as rules-based systems, regression, predictive analytics, machine learning, deep learning, and neural nets.*

A similar scope is adopted in New Zealand's algorithm charter and the policy documents that

[10] Aaron Rieke, Miranda Bogen, David G Robinson "Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods" (February 2018) <https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods>

[11] Canada Treasury Board "Directive on Automated Decision-Making" <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592&section=html>.

informed its development.[12] Notably, we also found that there is a standard definition of an "automated electronic system" in more than one piece of legislation in New Zealand, which also empowers chief executives in government agencies to delegate decision-making powers to such a system.[13]

These technical subjects must also be paired with analysis of the social, human and institutional settings that sit around the deployment of such technologies in particular contexts. This could extend the scope of our literature search to areas like management, leadership, economics, human resources, sociology, psychology, legal and political systems, and public or private governance.

At this point, any of these technical or sociological subjects is the focus of an enormous body of literature, some stretching back for the best part of a century. While it is not possible to conduct a literature review of all this information, we think it is an important insight to realise that ADM systems, as socio-technical arrangements, are comprised of all of these things.

What are algorithms?

> Algorithms are sets of instructions. Although they can be non-digital, the most common use is to describe the instructions that a computer follows to process data and solve a problem.

An algorithm refers to a set of instructions given in defined steps. This is put succinctly by Frissen et al:[14]

> *To put it simply, an algorithm is a recipe for solving a problem step by step. It has been around for a long time ... It is said that a computer cannot be used without algorithms, but an algorithm can be applied without a computer*

The description from Mittelstadt et al supports this, and notes that the breadth of the term "algorithm" makes it difficult to draw coherent principles about how algorithms should or should not be used:[15]

> *Discussion of a concept as complex as 'algorithm' inevitably encounters problems of abstraction or 'talking past each other' due to a failure to specify a level of abstraction (LoA) for discussion, and thus limit the relevant set of observables (Floridi, 2008). A mature 'ethics of algorithms' does not yet exist, in part because 'algorithm' as a concept describes a prohibitively broad range of software and information systems.*

While we frequently discuss "an algorithm" as if it operates in isolation, an ADM system is likely to be comprised of a complex arrangement of algorithms which also include, for example, the software used to implement an ADM system and to process or capture the data that feeds

[12] Stats NZ and DIA "Algorithm Assessment Report" (October 2018) <https://data.govt.nz/assets/Uploads/Algorithm-Assessment-Report-Oct-2018.pdf>.

[13] See reference list.

[14] Prof. dr. Valerie Frissen, dr Marlies van Eck, Thijs Drouen "Research Report on Supervising governmental use of algorithms" (2 January 2020) Hooghiemstra & Partners <https://hooghiemstra-en-partners.nl/wp-content/uploads/2020/01/Hooghiemstra-Partners-rapport-Supervising-Governmental-Use-of-Algos.pdf>

[15] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter and Luciano Floridi, "The ethics of algorithms: Mapping the debate" (2016) Big Data & Society July-December 1-21.

it.

Stats NZ initially adopted a distinction between operational algorithms, policy and research algorithms, and business rules. This framing was helpful for illustrating the ways that the use of algorithms in decision-making could be analysed in terms of social and policy systems, as well as technical systems.[16] The definitions were as follows:

> *Operational algorithms: These impact significantly on individuals or groups. These analytical processes interpret or evaluate information (often using large or complex data sets) that result in, or materially inform, decisions that impact significantly on individuals or groups. They may use personal information about the individuals or groups concerned, but do not need to do so exclusively.*
>
> *Algorithms used for policy development and research: These include analytical tools used to analyse large and varied data sets to identify patterns and trends, to support policy development, forecast costs, and to model potential interventions. For this review the key distinction between these and operational algorithms is that they have no direct or significant impact on individuals or groups. They may inform policy development but have no significant or direct impact on service delivery. These algorithms are not the focus of this review, but agencies were asked to describe these in general terms.*
>
> *Business rules: These are simple algorithms created by people that use rules to constrain or define a business activity. They make determinations about individuals or groups, without a significant element of discretion. This review asked agencies to provide an illustration of their use of these types of algorithms but did not seek an exhaustive list of such processes.*

The emphasis on business rules reflects the fact that non-digital algorithms are frequently followed by humans in organisational settings. Researchers at the University of Otago noted that the distinction between "operational" and "policy and research" algorithms was based on the suggestion that policy and research algorithms had lesser effect, because they didn't directly determine the rights of New Zealanders.[17] Subsequently, this framing has been abandoned in New Zealand's algorithm charter in favour of a risk matrix that orients attention toward the likelihood of an impact occurring, and the scale of that impact in terms of breadth and severity.

Commonly, policy instruments attempt to deal with the broad scope of "ADM systems" by limiting definitions that focus primarily on decisions with "significant impact", or similarly flexible language, even in instruments like the GDPR.[18] In other cases, this issue has led to the adoption of "risk matrices", that compare the magnitude of the impact of a system with the likelihood of that impact occurring, and imposing graduated restrictions accordingly. This allows systems with minimal "direct impact" to go largely without scrutiny. The Algorithm Charter and the Canadian Directive on ADM are key examples of this and the Canadian Directive has been noted with approval in publications by the World Economic Forum.[19]

[16] Stats NZ and DIA "Algorithm Assessment Report" (October 2018) <https://data.govt.nz/assets/Uploads/Algorithm-Assessment-Report-Oct-2018.pdf>.

[17] Colin Gavaghan, Alistair Knott, James Maclaurin, John Zerilli, Joy Liddicoat "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, Wellington, 2019).

[18] General Data Protection Regulation 2016/679, L119, 4 May 2016, p 1-88, 2, 4 and 22.

[19] World Economic Forum "AI Procurement in a Box: AI Government Procurement Guidelines" (June 2020).

What is artificial intelligence?

> Artificial intelligence refers to the use of computers to perform tasks that would normally require human intelligence.

One of the more frank and practical definitions for AI is offered by Hagendorff in *"The Ethics of AI Ethics: An Evaluation of Guidelines:* "AI" is just a collective term for a wide range of technologies or an abstract large-scale phenomenon."[20]

This makes discussions about norms for AI, including frameworks and ethical principles, largely superficial. Rarely if ever do they identify which of the variety of technology groups they apply to, and how specifically they should be applied:[21]

> *"Ultimately, it is a major problem to deduce concrete technological implementations from the very abstract ethical values and principles. What does it mean to implement justice or transparency in AI-systems? What does a "human-centered" AI look like? How can human oversight be obtained? The list of questions could easily be continued."*

In a wide analysis of various AI principles or ethical guidelines, Hagendorff continues:

> *"The ethics guidelines examined refer exclusively to the term "AI". They never or very seldom use more specific terminology. However, "AI" is just a collective term for a wide range of technologies or an abstract large-scale phenomenon. The fact that not a single prominent ethical guideline goes into greater technical detail shows how deep the gap is between concrete contexts of research, development, and application on the one side, and ethical thinking on the other. Ethicists must partly be capable of grasping technical details with their intellectual framework. That means reflecting on the ways data are generated, recorded, curated, processed, disseminated, shared, and used (Bruin and Floridi 2017), on the ways of designing algorithms and code, respectively (Kitchin 2017; Kitchin and Dodge 2011), or on the ways training data sets are selected (Gebru et al. 2018)."*

In practice, AI systems perform narrow, functional tasks. Recently, advances have been made in machine learning techniques (notably, in the use of neural networks) as a result of computer sciences research, access to massive datasets, and access to computational processing power greater than ever before. This has led to some algorithms performing very well at tasks that were previously thought to be very difficult for machines, such as image recognition, speech recognition, and other headline-generating applications. It has also led to a massive surge in the marketing and promotional appeal of "AI".

Automated decision-making systems can use rule-based data processing, or machine learning, or both, however using some kinds of machine learning approaches (particularly neural networks) can have implications for the transparency and human auditability of those computational processes, either because of their design, or because of their sheer complexity. [22] Simple rule based systems can have equally significant impacts for individuals, depending on context.

[20] Hagendorff, T. "The Ethics of AI Ethics: An Evaluation of Guidelines" (2020) Minds and Machines 30, 99-120.

[21] Ibid.

[22] Ministry of Health "Emerging health technology: Introductory Guidance for safely developing & using algorithms in healthcare" (January 2019, Ministry of Health).

What do we mean by “decision” and “decision-making”?

> We advise avoiding philosophical discussions about whether machines are capable of making decisions or not, as compared to the way humans make decisions.

The law has a history of stipulating process requirements around “decision-making”, and the means of challenging “decisions”, particularly in an administrative law or governmental context.[23] As a matter of common language, we also frequently refer to algorithms and computer systems as “deciding” to do things or not. This engages philosophical discussion about whether computers can “decide” to do anything.

We do not think it necessary to engage extensively with this discourse in the present investigation. For instance, whether a logic gate “decides” to open or close based on a series of inputs (if, and, or, only if, etc) does not change the output - that the gate either opens or closes. The current investigation is not equipped, or particularly interested, in going beyond this. During our investigation, one interviewee with expertise in modelling and technical analysis of predictive risk models was given the following proposition: “Algorithms process data. Humans make decisions.” Her answer was to reject the dichotomy created as being irrelevant from the perspective of the person subject to the automated decision-making system. Instead, it is important to consider the complex social, organisational and technical processes that make up that system from the perspective of someone subject to it, whose questions will be: what are you doing to me; why are you doing it; and when are you doing it. She continued to say that, when assessing the risks and opportunities of automated decision-making systems, it would be a mistake to act as if automated processes, whether human or technical, do not already play a significant role in ordering and organising human society.

We do not think there is any merit to investigating this definitional and philosophical distinction any further in the present context, particularly given our adoption of the proposition that ADM systems are socio-technical in nature, and primarily involve a delegation of a decision-making exercise which fundamentally rests with human agents. We are confident that this approach is sound, especially given the kind of feedback collected during Toi Āria’s participatory research and the comments made at the Te Kotahi wananga.

How are humans part of ADM processes?

> ADM processes generally involve many different humans at many different stages. These include designers, operators, end-users and more.

Aspects of the cooperation of humans and computers within ADM processes is neatly captured by Dave Heatley:[24]

> *Traditional computer programming – by far the most common variety – relies on a human to pre-specify what the computer is to do with every combination of data it might encounter. That pre-specification is the “algorithm”, which in many cases equates to a computer “program”. But how does one know if an algorithm is doing what it intends to do? The answer is testing. This involves creating or obtaining data, feeding it into the program and checking that it produces the desired result. The selection of test data and*

[23] See by way of generic example Philip Joseph *Constitutional & Administrative Law in New Zealand* (Thomson Reuters, 2014, online edition) ISBN 9780864728432 at chapter 22 et seq.

[24] Dave Heatley “Biased algorithms – a good or bad thing?” New Zealand Productivity Commission FutureWorkNZ Blog (2 October 2019) <https://www.productivity.govt.nz/futureworknzblog/biased-algorithms-a-good-or-bad-thing/> accessed 20 Nov 2020.

the quality of checking clearly matter – and both selection and checking rely on humans.

Our analysis leads us to conclude that any given ADM process will involve multiple humans throughout the many different stages of the ADM process. These are all participants working in conjunction with automation to deliver the required workload. As such, to properly understand the role of trust in ADM systems, we have advised the Council to examine "trust" in relation to each of these kinds of participants, and not solely trust by the people subject to automated decisions.

In any given ADM process, the people involved are likely to include several (or perhaps all) of the following:

- Automation designers and developers;
- Engineers and roboticists;
- Data collectors and interpreters;
- Operators;
- Trainers and teachers;
- Spokespeople and communicators;
- Senior decision makers at Ministerial and Executive levels;
- Middle managers in organisational hierarchies;
- People subject to decisions;
- People not subject to decisions, but otherwise interested in the welfare of people subject to decisions.

Some of these people will work within the ADM process and others sit outside of it. Each will have a trust relationship of some degree with some part, or the whole of, the ADM system. It is important that they have the right levels of trust towards the system, as well as towards the other people involved in the system, so that they can do their job effectively.

One point we heard strongly in our interviews was that senior decision-makers, especially those who face public scrutiny, cannot be excluded from any investigation of trust and ADM. Some people in our interviews treated the concept of "social licence" with extreme skepticism, to the point of suggesting it should simply be understood as a measure of whether senior decision-makers will suffer reputation damage from the public if they are found to be using ADM systems. One consequence of that potential reputation damage was said to be that senior decision-makers also need to have trust in the people building and operating ADM systems, otherwise they will unjustifiably worry about backlash from the use of those systems, even where that backlash is entirely without merit.

Other people may have no influence on the operation of the ADM process but are still subject to the consequences of it. Their trust towards the ADM system and its output is important to encourage appropriate engagement with the system. A loss of trust in any part or person within an ADM system may result in adverse outcomes.

This means that building, maintaining, and assessing trust in ADM systems is extremely complex. In essence, users must trust the people at each level of the ADM system.

All automation involves human decision-making

> All automation is influenced by human choices and decisions. No aspect of automation is completely detached from human agency.

In reality, there are no ADM systems which are not tethered to human agency in some way. We draw attention to this fact to make it clear that human agency is integral to all elements of an ADM process. As a result, there will always be some degree of relationship between different humans through ADM, even when this relationship is mediated by automation.

Consider this from the European Commission's White Paper on artificial intelligence:[25]

> *While AI-based products can act autonomously by perceiving their environment and without following a pre-determined set of instructions, their behaviour is largely defined and constrained by its developers. Humans determine and programme the goals which an AI-system should optimise for.*

To convey this point we propose the following:

> All automation is designed and built by human designers. It is the product of decisions, such as what the automation should optimize for, how it should function, and what should be done with its outputs. When this automation is used as part of an ADM process, the people who must live with the consequences of that process have, in fact, been affected by *human decision-making.*

At a minimum, ADM systems are affected by the decisions of the people who designed the automation. Usually, there are many more people whose choices these systems have been affected by: the people who operate the automation, the organisation that procured the automated system, and so on. A subsequent conclusion could be that every ADM system must have an identifiable responsible person, since no ADM system exists without having been procured, built, operated, or enforced by humans. Moreover, this seems to have been the approach adopted by legislative drafters and government policy makers in the legislation we identified referring to "automated electronic systems".[26]

ADM systems are multi-layered and complex

> Any given ADM process is likely to be a complex system. Usually, such systems involve varying interactions of data collection, data inputs, data processing, data outputs, human interpretation of data outputs, and human decisions. Each of these layers of a system can raise trust issues.

The diagram below demonstrates the way that we see the different "layers" of an automated decision-making system. Each of these layers can raise trust issues from the perspective of different humans involved in the creation, maintenance and operation of an ADM system.

The reference in the diagram to "existence, impact, constitution, and purpose" on the left hand side relates to Upturn & Omidyar's report on non-technical means of enabling public auditing

[25] White Paper of the European Commission (2020) "On Artificial Intelligence – A European approach to excellence and trust", p 16.

[26] See reference list.

of ADM systems, specifically:[27]

- **existence:** knowledge and transparency about whether the ADM system exists and is being used;
- **impact:** what its impact is socially, when measured empirically by comparing a given state of affairs before the system was implemented, and again after it has been allowed to operate for sufficient time;
- **constitution:** what procedures frame the system's use (ie, a legal constitution that limits the system or attributes accountability to humans or organisations using it); and
- **purpose:** how its purpose is defined, and consequently when it can be said to be successful or not, and how such measures of success are framed.

A theoretical model for ADM systems

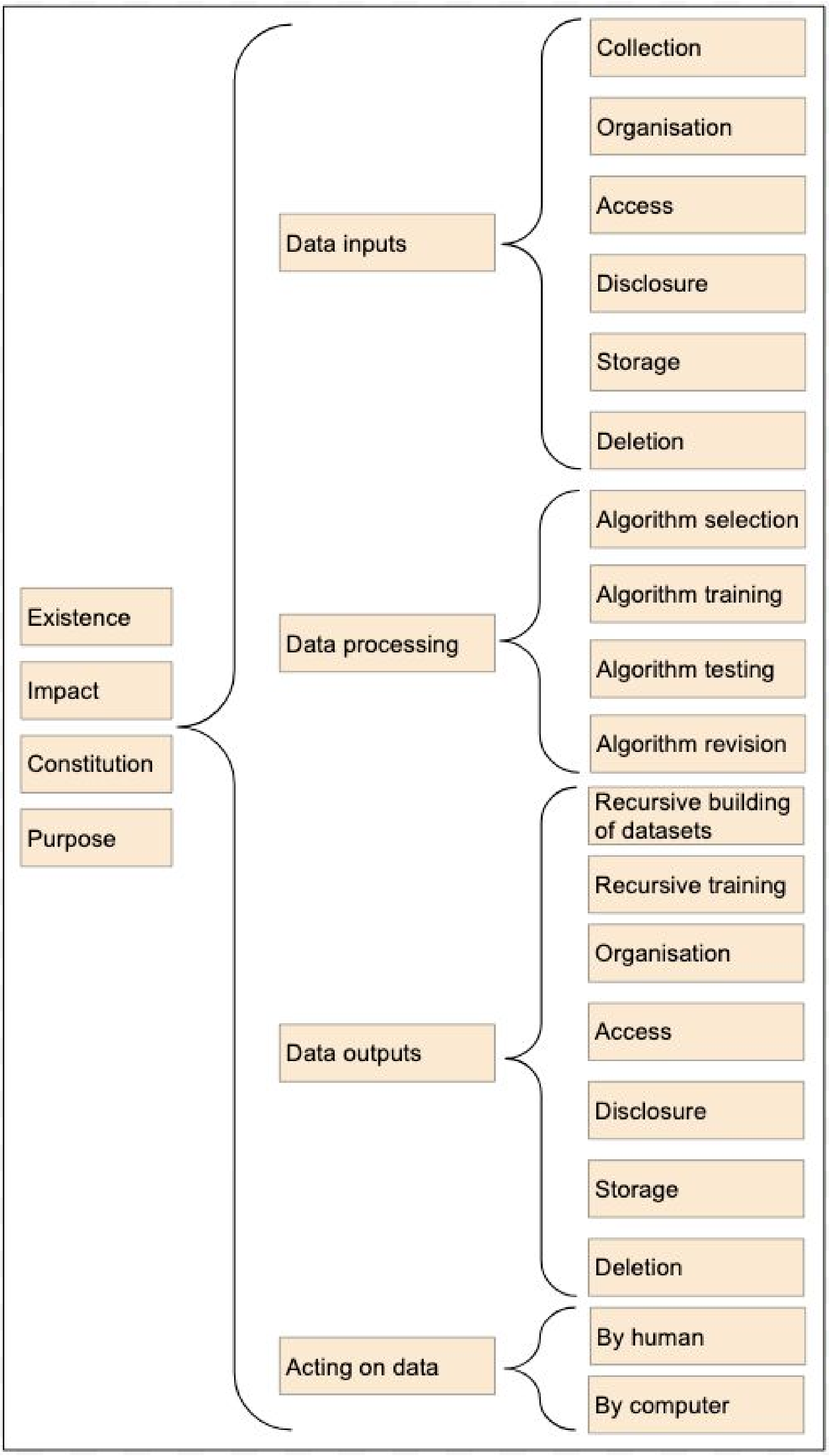
Based on our summary from the literature and from interviews about how ADM systems should be understood, there are potential trust points at all levels of an ADM system.

Firstly, the decision to implement an ADM system is made at some point, regardless of the eventual shape of that system. At that point, there are a range of ways to set the parameters of a system that will enable it to be audited even without technical assessment.[28] Trust can be affected by: knowing whether or not the system exists; by measuring the impact of the system on the world; by assessing the policies and procedures that frame the system; and by articulating what the purpose of the system is intended to be. Transparency around these non-technical parameters can be important for building trust in a range of human groups as outlined above.

In terms of the technical parameters of a system, it is broadly composed of four parts: data inputs; data processing, data outputs; and actions taken in reliance on those data outputs. These are set out in the middle column of the diagram. Within each of these four technical component parts of a system, there can be issues with reliability, accuracy, oversight and effectiveness, which are all important components of trust. Different groups can have different responsibilities in relation to each of those components. Importantly, it is highly likely that the person or group developing the ADM system does not have exclusive oversight and responsibility for all these different bits: many aspects of a system may be pulled from other agencies, acquired in commercial markets, or implemented beyond the designer's control.

[27] Aaron Rieke, Miranda Bogen, David G Robinson "Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods" (February 2018) <https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods>.

[28] Ibid.


Existence
Impact
Constitution
Purpose
Data inputs
Collection
Organisation
Access
Disclosure
Storage
Deletion
Data processing
Algorithm selection
Algorithm training
Algorithm testing
Algorithm revision
Data outputs
Recursive building of datasets
Recursive training
Organisation
Access
Disclosure
Storage
Deletion
Acting on data
By human
By computer

On the far right of the diagram are the smallest component bits of each of the four stages in a system for explanatory purposes. Each of these can be deployed or performed in ways that can undermine trust. Here are a few illustrative examples.

- **Automated decision-making requires data inputs.**

  Trust can be undermined by non-consensual collection of data, as understood through a typical privacy lens and reflected in the principles for safe and effective use of data and analytics.[29] Trust may have been undermined by non-consensual data collection even before it is used in an ADM system, and perhaps even beyond the control of the designer of that system. We heard through the wider research stream work by Toi Āria and te Kotahi that an emphasis on "deficit data" that only measures what is wrong with people can also undermine trust in an ADM system.

  Another feature of preparing data inputs for ADM systems is the way that collection may be biased because of biases held by human collectors who created those datasets. Equally, pre-existing datasets may be cleaned or organised in ways that affect their reliability, undermining the subsequent effectiveness of the ADM system.

  Data inputs from a system may be drawn from an incredibly broad range of sources, some of which are within the control of the system designer and operator, and some of which are not.

  When it comes to AI systems, a crucial point to consider is the way that data outputs from an ADM system may be fed back into data used to train the system itself: this can lead to compounding errors that drastically undermine the reliability of the system and the ability of auditors to test its operation. This illustrates the complexity of analysing trust at any one point in an ADM system in isolation from the wider system.

- **Automated decision-making systems process data that has been input into the system**.

  Humans are responsible for selecting how input data will be processed, including for example the operator's tolerance for false positives or false negatives in a system, in light of its purpose and the context for its intended use.

  Data issues may affect the way the algorithm develops over time, again in response to the way the parameters of a system have been set. Where algorithms change as a result of automated processes (ie, through machine learning), this is where significant errors can be introduced, affecting the reliability of data processing. Equally though, a simple rule based system like an email filter can produce errors by reliably processing data in an unintended way and diverting a message to spam instead of your inbox.

  Trust in data processing can be affected by issues such as whether data processing methods are transparent or a "black box", their complexity, and whether they can be

[29] Stats NZ, Privacy Commissioner "Principles for the safe and effective use of data and analytics" (May 2018).

disclosed or not in relation to commercial or proprietary considerations.

- **Automated decision-making systems produce a data output.**

  Sometimes these data outputs are fed back into the system in some way as data inputs, or may affect the algorithm responsible for processing data inputs. Many of the issues arising from data inputs also arise from data outputs, including traditional privacy concerns like who has access to those data points and whether they are created or collected for one purpose then used for another, perhaps without consent. These issues can affect the reliability of a system, public perception of it, or trust in the system held by operators, auditors or subjects.

- **Finally, automated decision-making systems involve acting on the data output produced in some way.**

  Initially, some think that a definitional line can be drawn that distinguishes between ADM systems and other systems based on whether the system acts on data outputs entirely without human input. The GDPR defines an ADM system as a making "a decision based solely on automated processing".[30] Our advice is that this line is unlikely to be tenable as a matter of practice, law and policy. That is why we have adopted a greater focus on how to assess the reliability and trustworthiness of ADM systems as a socio-technical human/machine partnership, rather than a completely autonomous system.

  We note that trust may be influenced by the extent to which ADM systems act on data outputs automatically. Generally, this is thought to be done by computer systems, although we note that humans may also act automatically on a data output without following a separate decision-making process or exercising judgement.

The difficulty in analysing "trust in automated decision-making" as a single subject in the abstract is that the factors that influence trust are so diverse, and ADM systems are so complex and diverse themselves, that the ways for trust to be bolstered or undermined in each part of that system are essentially incalculable. There is a kind of exponential effect to analysing trust and reliability of ADM systems, which is amplified even further when considering the different contexts in which ADM systems can be deployed, and the different perspectives from which trust can be analysed by different stakeholders in a system (ie operators, auditors, subjects).

For this reason, the Council's decision to test specific scenarios with identified groups of people will add significant value to an otherwise theoretical discussion. This is also why it is essential that the Council's recommendations focus on ways to capture specific case studies where ADM has been used successfully or unsuccessfully, to enable others to learn directly from the specific design features adopted in those cases.

[30] General Data Protection Regulation 2016/679, L119, 4 May 2016, p 1-88, 2, 4 and 22.

Each layer of an ADM system can be automated to various degrees

> Automation may occur at many different points in an ADM process, and usually will occur at multiple points.

Following the diagram above, which maps the various layers in an ADM system, we note that each of the steps in that system might be automated. For example, data collection can be automated or manual, for example by scraping web pages or entering information from a human into a database. Equally, data outputs may be disclosed or acted upon by a human receiving that data point, or by a system acting on it automatically.

UNDERSTANDING “TRUST” IN ADM

Why trust is relevant to ADM

> The benefits of automation are relatively straightforward. However, in many cases these benefits are limited or lost where humans are unwilling to trust the automation, or where they trust it too much, or too little.

Korber et al put this issue succinctly when they say that, “operators tend to use automation that they trust while rejecting automation that they do not.”[31] We believe that the same can be said for society at large, with some interesting exceptions.

The importance of appropriate levels of trust in ADM systems is encapsulated Dzindolet et al, summarising previous research:[32]

> *A recent and dramatic increase in the use of automation has not yielded comparable improvements in performance. Researchers have found human operators often underutilize (disuse) and overly rely on (misuse) automated aids (Parasuraman and Riley, 1997).*

As a result, the benefits may be fully derived where the humans involved in an ADM process have “the right levels of trust” – rarely absolute trust, or a complete absence of trust. Rather, the right level of trust will be an *appropriate level* based on the capabilities of the automation, its limitations, the context in which the ADM process is operating, and the other people relevant to the ADM process.

We note that in some contexts people may have little option but to use or rely on ADM systems, even where they do not trust them. Essentially, they are willing to be vulnerable to them even though they do not have positive expectations, or they are extremely uncomfortable with their vulnerability. Their lack of trust may have no influence on the operation of the ADM system, to the detriment of the system in various ways. For example, people may deliberately enter incorrect data into the system, or may fail to participate in feedback loops between operators and low-end users which might help improve the system.

At the same time, we note that people frequently do not trust an ADM system, and yet they regularly and willingly engage with it. In these instances, risk versus reward analysis takes place, and a certain quid pro quo to the use of digital services. It also draws attention to the fact that ADM systems are context specific.

Social licence

Social licence is a concept that, in this context, refers to public acceptance of the use of data-driven technologies. It was used widely in a range of public reports by Stats NZ[33] and the

[31] Korber, M., Baseler, E., Bengler, K. Introduction matters: Manipulating trust in automation and reliance in automated driving. Applied Ergonomics 66 (2018) 18-31.

[32] Dzindolet, M.T., Peterson S.A., Pomranky R.A., Pierce, L.G., Beck, H.P. (2003) The Role of Trust in Automation Reliance, Int. J. Human-Computer Studies 58 697-718.

[33] Stats NZ "A social licence approach to trust" (August 2018): “
Stats NZ’s social licence is defined as the permission it has to make decisions about management and use of the public’s data without sanction."

Data Futures partnership[34]. It is also discussed in guidance published by the Ministry of Health, but mainly linked to a project's "legitimacy, credibility and trust in the eyes of the public or key stakeholders".[35] It is an ambiguous concept and it lacks value when engaging in empirical work with people in the community, particularly Māori.[36] Our advice is that, while social licence can be a useful shorthand for a complex topic, it is vital to move past that shorthand as soon as possible.

Understanding trust in the ADM context

> Trust in ADM has two complementary elements. (1) A social element: referring to the kind of trust between the people that relate to an automated system. (2) A technical element: the kind of trust that humans have towards the technical components of an automated system itself.

At the social level, trust is a psychological state where a person accepts being vulnerable to another because they have positive expectations of how that person's actions will affect them. This is not a choice or a behaviour, but an underlying state of mind. For a range of reasons, different people have different "trust propensity", which relates to the way that people may have pre-existing dispositions to be trusting or distrustful where they have no previous relationship or experience with a system or a person.[37]

As regards human-to-automation trust, the relationship is influenced by a number of factors. In short, these concern the human's perception of the automation (its reliability, predictability, accuracy, etc), which is in turn influenced by a combination of human experience with the system (positive and negative), and information provided to the human about the system (its capacity for error, its specific strengths and limitations).

The human-to-human element of trust

> Humans' trust in ADM processes is significantly influenced by their perception of the other humans involved in that process, and by extension the institutions those humans represent.

We are particularly influenced by the definition presented here by Rousseau et al:[38]

> *"Trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another."*

Greater depth and texture is provided by Hoffman et al in a corroborating definition:[39]

[34] Data Futures Partnership "A Path to Social Licence Guidelines for Trusted Data Use (Summary document)" (August 2017, NZ Government). "What 'social licence' is: These guidelines encourage data use practices that will build acceptance. This acceptance is referred to as social licence."
[35] Ministry of Health "Emerging health technology: Introductory Guidance for safely developing & using algorithms in healthcare" (January 2019, Ministry of Health)
[36] Tūhono Trust "Sharing information for wellbeing: Māori engagement on social license report 2017" (Data Futures Partnership, June 2017).
[37] Jason A. Colquitt, Brent A. Scott, and Jeffery A. LePine "Trust, Trustworthiness, and Trust Propensity: A Meta-Analytic Test of Their Unique Relationships With Risk Taking and Job Performance" (2007) 92(4) Journal of Applied Psychology 909-927.
[38] Rousseau et al (1998) at 394-395.
[39] Robert R. Hoffman, Matthew Johnson, and Jeffrey M. Bradshaw "Trust in Automation" Human-Centered Computing, IEEE Computer Society, p 84.

> *"Interpersonal trust has been defined as a trustor's willingness to be vulnerable to a trustee's actions based on the expectation that the trustee will perform a particular action important to the trustor. Research has shown that interpersonal trust depends on several factors, including perceived competence, benevolence (or malevolence), understandability, and directability — the degree to which the trustor can rapidly assert control or influence when something goes wrong. Any one of these factors, or dimensions, could be more or less important in a given situation.*

This definition has in turn been informed by several other propositions, particularly that trust exists in situations of risk and interdependence. By way of summary, Rousseau et al conclude that:[40]

> *… scholars do appear to agree fundamentally on the meaning of trust. Trust, as the willingness to be vulnerable under conditions of risk and interdependence, is a psychological state that researchers in various disciplines interpret in terms of "perceived probabilities" … "confidence", and "positive expectations" … Trust is not a behavior (e.g.,, cooperation), or a choice (e.g., taking a risk), but an underlying psychological condition that can cause or result from such actions. … Finally, because risk and independence are necessary conditions for trust, variations in these factors over the course of a relationship between parties can alter both the level and, potentially, the form that trust takes.*

However, these definitions still leave certain elements unsettled. This is best described by Miller and Perkins:[41]

> *"The Mayer et al (1995) definition is the most widely accepted definition of trust, "A willingness to be vulnerable to another party when that party cannot be controlled or monitored." … However, the definition still begs questions. Vulnerable to what extent? Vulnerable to what outcome? How willing? What are the ramifications of being vulnerable? Does the context matter? Monitored or controlled to what extent?"*

These questions are extremely important to the current investigation, because they highlight the importance of context in assessing the use of ADM systems. The answers to these questions will vary greatly based on the specifications of the system, what is at stake, the context in which it is implemented, and the characteristics of people who must suffer the consequences of the ADM system.

The human-to-automation element of trust

Human trust in ADM is influenced by someone's perception of the automated system. If systems are perceived to be of good quality, humans will engage with them more.

It is intuitive that perception of the quality of a computer system influences human propensity to trust that system, and rely on it for the appropriate range of activities. Per Korber et al:[42]

> *We define trust in automation as "the attitude of a user to be willing to be vulnerable to the actions of an automation based on the expectation that it will perform a particular action important to the user, irrespective of the ability to monitor or to intervene.*

We also note that Korber et al include variations in individual as an important factor

[40] Rousseau et al (1998) at 394-395.

[41] Miller, E.J., Perkins, L. "Development of Metrics for Trust in Automation", Proceedings of the 15th International Command and Control Research and Technology Symposium(ICCRTS '10), Santa Monica, CA, June 22-24, 2010

[42] Korber, M., Baseler, E., Bengler, K. Introduction matters: Manipulating trust in automation and reliance in automated driving. Applied Ergonomics 66 (2018) 18-31

influencing human-to-automation trust dynamics:[43]

> *This definition implies that trust is a multidimensional construct that is based on relevant characteristics of the automated system (e.g., reliability, predictability) and the trustor himself (e.g., propensity to trust).”*

Technical factors influence trust in ADM

> Human trust in automated systems is influenced by its technical components. These are the factors which comprise the perceived quality of the automated system. As expected, systems which are perceived as being better quality tend to be more trusted by the humans that interact with them.

The literature proposes a wide variety of components or factors that influence human trust in automation. Often these overlap, and sometimes appear to be synonymous. We collect a variety of these below, accompanied by our contextualised understanding of what they mean. We are optimistic that they can be used to investigate perceptions of ADM within New Zealand, and influence the design of ADM systems that are ready for adoption.

Dr Janet E. Miller and LeeAnn Perkins use the following metrics as a framework for measuring levels of human trust in automated systems:[44]

- “Competence” – the system’s ability to do the prescribed task
- “Predictability” – the extent to which the operator can predict the system’s behaviour
- “Dependability” – the system’s ability to perform in all anticipated situations
- “Consistency” – low variability in the system’s behaviour
- “Confidence – the extent to which the operator can act on the system’s outputs

Araujo et al investigated population trust in ADM by reference to participant perception of the following factors, graded on a scale:

- “Fairness” – perceived to be not unfair, or not used unfairly
- “Usefulness” – perceived as making a positive contribution to society
- “Riskiness” – perceived as potentially socially harmful or undesirable

Communication and information influences trust in ADM

> Trust in automated systems is influenced by more than just the technical components of that system. It is also influenced by the communication and information provided about that system.

Irrespective of the quality of an automated system, a person’s level of trust will be influenced by the information they are given about it. In short, it is possible to have an excellent automated system, but without appropriate communication, humans may trust the system too little. Alternatively, there is a risk that humans may trust a system too much when they are

[43] Ibid.

[44] Miller, E.J., Perkins, L. "Development of Metrics for Trust in Automation", Proceedings of the 15th International Command and Control Research and Technology Symposium(ICCRTS ’10), Santa Monica, CA, June 22-24, 2010.

given information about it that makes them over-confident in its capabilities.

We are influenced here by the work of Korber et al in their investigation of the importance that "introduction" plays in manipulating levels of trust in automation:[45]

> *Prior information influences trust … The degree of competence and reliability of an automated system that is described in an introduction and is experienced in an introductory drive is positively correlated to the subjective trust in automation.*

As such, contextual information and communication about an automated system plays an important part in ensuring that people have the right levels of trust. What is appropriate will vary depending on the context in which the system is used and its specific strengths and limitations.

Context influences trust in ADM

> A person's tolerance or opposition towards the use of automation is likely to be affected by contextual factors, e.g. whether it is a high stakes or low stakes scenario. At the same time, ADM in different contexts should be subject to different considerations e.g. greater or lesser scrutiny and control.

This is corroborated by the following from Miller and Perkins:[46]

> *…trust in the context of corporate financial dealings would be quite different in detail from trust with respect to internet chat rooms. Also, various contexts can entail differing levels of attributes such as vulnerability, risk, and reward all of which affect levels of trust. Levels of automation should also be included in the description of context and domain of interest.*

These insights are a valuable reminder against one-size-fits-all analysis of ADM, whether in considering its potential benefits, potential risks, or the way in which it operates. Distinctions between low impact and high impact scenarios are core parts of assessing ADM systems under the algorithm charter and the Canadian directive on ADM. Organisations frequently ask exactly which ADM systems they use are subject to generalised rules and governance frameworks which might be absurdly interventionist in some contexts, and dangerously ambiguous in others. For example, they might use ADM in their HR departments, but also in entirely different contexts. Are both of these subject to the same governance rules, even when the context and risk is very different? In practice. the "risk matrix" approach discussed above in relation to the Algorithm Charter and the Canadian Directive aim to enable this kind of contextual judgement about what level of audit and oversight mechanisms are required..

Other factors influencing trust in ADM

People trust ADM less in the abstract, more in specific scenarios

> Many people have an aversion to ADM in the abstract. However, when given examples of ADM in specific scenarios, they tend to be more positive about it.

Based on trust, different people (or the same people in response to different scenarios)

[45] Araujo, T., Helberger, N., Kruikemeier, S. et al. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Soc (2020).

[46] Miller, E.J., Perkins, L. "Development of Metrics for Trust in Automation", Proceedings of the 15th International Command and Control Research and Technology Symposium(ICCRTS '10), Santa Monica, CA, June 22-24, 2010.

experience what the literature describes as "algorithmic appreciation" or "algorithmic aversion".[47] However, the literature suggests that people are more averse to ADM in the abstract, and more appreciative of it when presented with its use in specific scenarios.

We suggest this may be one reason why there is growing popularity around a generally critical public narrative around ADM. Increasingly common messages about the risks and limitations of ADM seem to have been widely influential. These may be skewing public perception of what is, in reality, a highly broad and ubiquitous group of technologies. We suspect this is also influenced by conceptual imprecision, e.g. conflating ADM with concepts like "artificial general intelligence".

Based on the literature, our opinion is that while an appropriate degree of criticism and skepticism about ADM is valuable, it is important that this does not lapse into wholesale opposition without nuance. Again, the best outcome is to develop the "right levels of trust" based on specific scenarios and context in which a system is deployed, the data it relies upon, and the outcomes it is producing. When presented with such specificity, the literature suggests people tend to have much greater support for the benefits that ADM use can provide, and the way it operates.

Personal characteristics influence levels of trust in ADM

> Levels of trust in ADM may vary based on personal characteristics, including age, gender, and levels of education.

The literature suggests that people may have different opinions about ADM, and different levels of enthusiasm towards it, based on their personal characteristics. The main support for this idea comes from a study by Araujo et al, in a study conducted with Dutch participants.[48]

It is possible that these differences may be relatively consistent across societies, although this is speculative. We are tentative about the strength of this finding, as well as its applicability to different populations. There is opportunity to conduct similar work in the New Zealand context to develop an accurate local picture.

In the study by Araujo et al, 958 Dutch participants responded to a survey which measured their perception of ADM in terms of perceived fairness, usefulness, and risk. Without repeating the contents of their research unnecessarily, the study found variation across the sample group in the following ways:

- Age – older participants generally perceived ADM as less useful to society, and tended to prefer manual (human) alternatives where possible.
- Education – greater levels of education correlated with greater enthusiasm towards ADM. Conversely, people with greater levels of education in computer science or mathematics perceived ADM as less fair and riskier.
- Gender – men and women broadly shared the same perception of ADM, though women perceived it to be significantly less useful.
- Self-efficacy and privacy – people that considered themselves more capable of controlling their personal data online were more appreciative of ADM and its

[47] Araujo et al (2020).

[48] Araujo, T., Helberger, N., Kruikemeier, S. et al. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Soc (2020).

usefulness, and more willing to engage with ADM systems.

- Beliefs about equality – individuals who described themselves as believing that equality is important to society were more positive about the usefulness of ADM, while being more negative about its risks.

These provide interesting insights into possible variations in how ADM is perceived across populations. However, we warn against relying too heavily on the findings of a single study. We also consider it likely that these same trends may not be reflected in a different country with a different socio-cultural profile. It would be useful to develop this kind of granular understanding within the New Zealand context, by surveying representative samples according to a robust methodology.

Holding ADM to higher standards can result in reduced trust propensity

> Sometimes people interacting with automation systems demand higher standards from those systems compared to human equivalents. This can result in automation being disfavoured, even where it outperforms human equivalents in the same task or role.

ADM systems are tools to help deliver a particular product or service, in the same way that human labour is a tool to deliver products or services. However, the literature discusses a prevailing propensity to hold ADM systems to higher standards than human alternatives.

Consider the following from Araujo et al:

> *For example, compared to human decision-makers, ADMs or recommendation systems can be seen as inscrutable, which might impact the user's willingness to accept the system, or its recommendations (Yeomans et al. 2019). People seem also to be far less forgiving towards ADM than to humans: Recognizing that an algorithm makes a mistake—even when its overall performance is better than of a human—was seen to be sufficient to make people choose the human decision-maker, thus leading to algorithmic aversion.*

We supplement this with the following analysis from the Ministry of Health:[49]

> *No algorithm is perfect, but the question is whether it is better than the next option. For example – is automating a process with the potential of an network outage or scheduled downtime maintenance, better than the current process which relies on a fax machine not 'running out of paper' therefore risking catastrophic data loss? In order to understand this, understanding the context and impact of the algorithm is key.*

Sometimes there is a misconception that ADM lacks bias compared to humans

> People have tended to wrongly believe that automated systems are without biases. However, increased public and academic scrutiny around this issue has made the potential for biases in automated systems more commonly known.

This phenomenon is commonly described as the "machine heuristic". It is a rule of thumb about machine superiority to human counterparts. Most often this is due to a perception that machines cannot and do not possess biases, whereas humans do. It also manifests in perceptions that machines are more secure and trustworthy than humans.[50]

[49] Ministry of Health "Emerging health technology: Introductory Guidance for safely developing & using algorithms in healthcare" (January 2019, Ministry of Health).

[50] Sundar, SS., Kim, J. "Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information", Conference on Human Factors in Computing Systems (May 2019) pages 1-9.

Consider the following findings from various analyses:[51]

> *Earlier studies show a general tendency to consider an "expert system to be more objective and rational than a human adviser" (Dijkstra et al. 1998, p. 160). This tendency, often based on the assumption that statistical methods out perform human judgement (Dawes et al. 1989), gives rise to the idea of algorithmic appreciation, showing in several instances that people prefer judgement or recommendations by algorithms when compared to human recommendations (Logg et al. 2018). This tendency may be partially attributed to the notion of the machine heuristic (Sundar 2008), which suggests that the less a user anthropomorphizes an interface, the more she or he will consider its decisions and selections to be objective and free of (ideological) biases.*

We note that faith in the machine heuristic has been substantially eroded as algorithmic biases have become the epicenter of fields of study (across law, sociology, governance, and critical studies) into the risks of increased use of data-driven technologies across society. We venture that it is now quite commonly accepted in media reporting that machines not only can be, but frequently *are* inherently biased according to the rules by which they are programmed, design choices made during their development, and other factors such as the data used to develop machine learning models.

Limitations of "trust" as an organising concept

> These are inherent limitations that result from relying on trust as the core concept to guide analysis of ADM, particularly at the theoretical level.

While the literature offers some useful definitions for trust which may be widely applicable, actual trust relationships are real and require nuanced analysis. Moreover, one of the prominent concerns around ADM systems is that they are deployed in situations where people are compelled or obliged to interact with them, despite lack of trust. Alternatively, situations frequently exist where ADM systems are applied to people unwillingly or unknowingly, thus rendering their trust in that system irrelevant.

It is important to note that a core concept of "trust" may not accurately represent the full variety of socio-cultural phenomena relevant to the way people interact with and within ADM processes. In fact, reliance on an English language-based conception of "trust" unavoidably compels the dialogue along one cultural interpretation, possibly at the expense of others. We also note that individual rights based approaches to privacy and data rights are being challenged by the use of algorithmic classification of groups of individuals, which might lead to harmful effects when an individual is classified as being part of this group, but are not dealt with through individualistic rights-based frameworks.[52]

We noted that the Social Wellbeing Agency favoured the use of concepts and language from te Ao Māori when aiming to summarise their findings about trusted data use.[53]

We draw attention to Te Mana Raraunga and its 16 data principles reflective of Māori values.[54] These may be equally useful conceptual vehicles for understanding human interaction with ADM, and are likely to be more appropriate in the discussion of Māori interaction with such

[51] Ibid.

[52] Brent Mittelstadt "From Individual to Group Privacy in Big Data Analytic" Philos. Technol. (2017) 30:475–494.

[53] For example, the SIA adopts five principles from te reo Māori in relation to data protection and use: he tāngata, manaakitanga, mana whakahaere, kaitiakitanga, mahitahitanga. Social Investment Agency "Data protection and use policy" (December 2019, New Zealand Government).

[54] Te Mana Raraunga Principles of Māori Data Sovereignty (October 2018).

processes.[55]

> *For Māori, trust in government is a key issue, and very much influenced by what has happened in the past. Māori are looking for a true partnership with the Crown and government, based on Treaty of Waitangi obligations and principles. We heard that a collective approach to wellbeing is needed, which focuses on all aspects of a person and their whānau, and works towards positive goals and aspirations, not simply addressing 'problems'. You also emphasised that Māori want to own their own data, and their own measures of wellbeing.*
>
> ...
>
> *For Pacific peoples, strengths-based and holistic approaches to wellbeing are important, which reflect Pacific worldviews. Pacific peoples are also looking for better relationships with government and to be empowered to develop solutions for their communities. Pacific peoples want to have a say about how their data is used, and by whom."*

This idea came through strongly in the wānanga held to inform the literature review on Māori data sovereignty, conducted as part of the Council's research.

Risks of automated decision-making

Speed of operation, risk of harm, and scale of harm

The reason why ADM has attracted increased attention from a trust perspective is apparent from the risk matrices adopted by the Algorithm Charter and the Canadian Directive on Automated Decision-Making. These require agencies to assess the safeguards required for ADM systems according to the anticipated magnitude of harm the system could cause, and risk that such harms could result.

The nature of the harms that can be caused by ADM systems are now well known. We say they are also widely accepted as a result of public advocacy around the ethics of AI systems. One of the most significant recent examples of an ADM system that has led to calls for a royal commission of inquiry and repayments and compensation in the billions of dollars is the robodebt "fiasco"[56] in Australia. "Robodebt" compared datasets about income from two different government agencies (inland revenue and a benefits payment agency), assessed the financial difference between these two data points, then raised a debt and issued a demand for payment to vulnerable people. These debts were a result of miscalculations and relying on erroneous datasets out of context and have led to significant impacts for individuals and for the Australian government. There were multiple legal challenges declaring the resulting decisions invalid, however these were never appealed or escalated in a way that meant they were dealt with effectively.[57]

[55] Social Investment Agency "What you told us: Findings of the 'Your voice, your data, your say' engagement on social wellbeing and the protection and use of data" (November 2018, New Zealand Government).

[56] Peter Whiteford "Robodebt was a policy fiasco with a human cost we have yet to fully appreciate" (16 November 2020) <https://theconversation.com/robodebt-was-a-policy-fiasco-with-a-human-cost-we-have-yet-to-fully-appreciate-150169>.

[57] Anna Huggins "We need human oversight of machine decisions to stop robo-debt drama" (2 July 2019, The Conversation).

General potential for harm

Harmful use of ADM can occur in three broad ways: disuse, misuse, and abuse.

This thinking is the product of Parasuraman and Riley, discussed below by Miller and Perkins:[58]

> *Parasuraman and Riley (1997) discuss such types of technology usage issues as misuse, abuse, and disuse. They define use as the voluntary employment of an automation technology and discuss the factors that influence the decision to use, misuse, disuse or abuse a specific technology.*

This framework proceeds as follows:

- Disuse – the discontinuation or underutilization of technology
- Misuse – an overreliance on a specific technology in the wrong context
- Abuse – an inappropriate application of technology by designers and managers

It is immediately easy to see how levels of trust are associated with potential harms. On the one hand, lack of trust can result in missed opportunities for societal benefit. This must not be ignored – for instance, where there is a social cost if we choose not to use an ADM system (e.g. a potentially life-saving cancer detecting system) because it has risks related to privacy or discrimination. On the other hand, excessive levels of trust can also result in ADM being applied uncritically, and with levels of certainty that are not reflected in the technology's capabilities. For example, ADM might be used to help make decisions about a person, based on algorithmic predictions of their future which can never be entirely reliable.

The potential for abuse is likely what most people think of when they imagine the risks of ADM. The range of abuses that might occur are as various as the ADM applications themselves and the areas in which they operate. A frequent concern is that ADM can be used unfairly to further racial inequality, particularly in areas of health and justice.

Based on our understanding of the literature, we think it best to consider risks of ADM in relation to the context that the process is being used. Some ADM is deployed in obviously high-stakes areas, e.g. algorithms for sentencing or parole applications. Others are deployed in much less immediately risky situations, e.g. the algorithm that decides which song to recommend to you next. Of course, all of these ADM processes will have real-world impacts that affect somebody and even companies like Spotify that simply recommend music are engaging in discussions about algorithmic ethics. It would also be very rare that ADM systems don't rely on the use of personal data, engaging privacy issues.

ADM systems rely on data, which creates risks and benefits

Because of the importance of data to ADM systems, ordinary data governance and ethics principles should apply.

The literature covering the area of data ethics and governance is immense. We can do no more than note the high level points from this literature by way of illustration throughout this report. There are risks in using data collected for one purpose for a new purpose. Further, there can be bias in datasets because of the way they are collected, or because of what is

[58] Miller, E.J., Perkins, L. "Development of Metrics for Trust in Automation", Proceedings of the 15th International Command and Control Research and Technology Symposium(ICCRTS '10), Santa Monica, CA, June 22-24, 2010.

collected and what is not collected. All datasets have some level of bias and inaccuracy. They are at best representations of the world, not the world itself. Some communities are excluded from datasets because of historical prejudice or being excluded from access to services. Algorithmic models are also trained from datasets that have particular biases in them. People from marginalised communities may be over or under-represented in those datasets. Finally, one significant risk is that use of an ADM system can skew the data held by an organization, and undermine that organization's ability to assess the impact and efficacy of its ADM system.

Harms may be difficult to detect

> It may be hard to spot where and how ADM is causing harm, especially where the technology is not completely transparent or able to be explained

ADM systems tend to operate at scale and in a distributed way. While their effects may be felt at a personal level, some of these personal effects can be invisible when looking across the data as a whole. Moreover, harm may not be obvious or intentional, which makes the task of spotting it and how it is occurring particularly pernicious. For instance, ADM systems might not explicitly gather data on a person's race, but can infer such facts from social media data - e.g. their preference in movies or music. This information might become a disproportionate factor in a decision-making process in ways that are difficult to detect at the outset or in retrospect.

> Because ADM is scalable, it may cause harm at the societal level while causing negligible harm to any given individual, failing to trigger audit mechanisms.

The characteristics that make ADM a valuable tool are the same characteristics that make its risks particularly concerning. That is, ADM can operate at speed and scale, with limited oversight, and without relief. This can mean that, by the time harm is identified and action taken, the impacts of that harm are significant.[59] Methods of testing ADM systems should be incorporated into system design at the outset where they pose any risk of high impact.

Groups subject to historical discrimination may be particularly at risk

> Some of the areas where ADM is being applied relate to people who are very poorly positioned to object to the use of ADM, or who are already vulnerable

A good example of this is Canada's use of predictive ADM processes in decisions around refugee induction and immigration. For more information, see the report of the University of Toronto School of Law, "Bots at the Gate: A Human Rights Analysis of ADM in Canada's Immigration and Refugee System":[60]

> *Vulnerable and under-resourced communities such as non-citizens often have access to less robust human rights protections and fewer resources with which to defend those rights. Adopting these technologies in an irresponsible manner may only serve to exacerbate these disparities.*

The research raised the possibility of human rights abuses arising where areas like immigration are used as a laboratory for computational experimentation:

[59] The Robodebt saga is an example of this.

[60] University of Toronto, "Bots at the Gate: A Human Rights Analysis of ADM in Canada's Immigration and Refugee System" <https://it3.utoronto.ca/wp-content/uploads/2018/10/20180926-IHRP-Automated-Systems-Report-Web.pdf>.

> *The use of algorithmic and automated technologies to replace or augment administrative decision-making in this context threatens to create a laboratory for high-risk experiments within an already highly discretionary system.*

Māori data sovereignty and colonisation

> Some ADM processes overlap Māori life, communities, and values. Because of this, it is of paramount importance to consider these overlaps, and engage with Māori around the management of risk. This is particularly relevant where ADM is used by the Crown given its obligations under te Tiriti o Waitangi.

When ADM is used by agencies of the Crown, trust in ADM cannot be separated from the history of colonization in New Zealand. Equally, private sector organisations have also played a role in colonization historically.[61] In some cases, Māori are excluded from official datasets, such as where ethnicity is under-reported or misreported in the health system. Equally, Māori are subject to over-surveillance, for example by police and social welfare agencies. Further, Māori are subject to data collection that focuses on deficits rather than strengths, leading to a picture of Māori people in ways that do not otherwise adequately represent the strengths of individuals and Māori as a group.

We identify at least three broad areas of overlap that create a risk of harm to Māori:

1) Because ADM uses and processes data, and particularly personal data, it engages important questions around Māori data sovereignty.[62]

2) Because some ADM applications are used in areas of governance, it must be considered in relation to tino rangatiratanga, or Māori sovereignty.

3) Because ADM is used in applications across important areas that affect Māori wellbeing, like health and justice, it must be considered in light of the Crown’s obligations under the Treaty of Waitangi.

These are all extensive areas for debate, discourse, and investigation, and as such are beyond the scope of this document. At minimum, we recommend a dedicated programme of engagement around these issues with Māori communities, and to utilise the extensive knowledge and expertise of appropriate researchers. This should be grounded by a focus on the high-quality work that has already been done, much of which is relevant to the Council’s investigation. The Te Mana Raraunga website includes a list of qualified experts with experience in Te Ao Māori.

We note that during the course of the research stream, the Council elected to seek a dedicated piece of work from Te Kotahi and we strongly encourage readers to read that work in detail as an essential component of the Council's overall workstream.

[61] Donna Cormack, Tahu Kukutai, Chris Cormack "Not one byte more: from data colonialism to data sovereignty / Kia Kaua Tētahi paita anō: Mai i Ngā raraunga whakatōpū ki te mana motuhake o ngā raraunga" (2020, Bridget Williams Books) Shouting zeroes and ones (ed Andrew Chen).

[62] See, for example, the following: Te Mana Raraunga Principles of Māori Data Sovereignty (October 2018); Data Futures Partnership "A Path to Social Licence Guidelines for Trusted Data Use (Summary document)" (August 2017, NZ Government); Caleb Moses "The Integrated Data Infrastructure / Te hanganga o ngā raraunga kōmitimiti me te whakaaetanga ā-iwi" (2020, forthcoming, Bridget Williams Books) Shouting zeroes and ones (ed Andrew Chen); Karaitiana Taiuru (2020) “Treaty of Waitangi/Te Tiriti and Māori Ethics Guidelines for: AI, Algorithms, Data and IOT.

ILLUSTRATIVE EXAMPLES OF ADM WITHIN NZ

Based on our definition, there are a vast number of ADM systems deployed and developed within New Zealand. These canvas most industries, and exist within both the private and public sphere. Their implications are nuanced and context specific. There are also ADM systems operating within New Zealand that are, substantially, extraterritorial and beyond the jurisdiction of the New Zealand legal system.[63] That is, they are under the control of organisations and corporations which are based elsewhere, in different legal and social settings.

Common examples of ADM New Zealanders use and encounter

> People interact with ADM processes every day. Some of these are not only commonplace, but essential systems for regulating, protecting, and facilitating social life. Many people are likely to be surprised by how often ADM impacts upon them.

Here, we give some examples of ADM systems that might be encountered commonly, occasionally or more rarely. We pause to note that we assess these frequencies based on our own experience. One point to reflect upon is the way that various communities may engage more or less frequently with ADM systems generally, or particular kinds of ADM systems, depending on their sociopolitical position. This is one reason why diversity of experience and training is important when it comes to analysing the operation and trustworthiness of ADM systems.

The sorts of ADM processes people might engage with most days include things like:

- Email filters
- Web filters
- Traffic lights
- Conventional computer technologies
- Facial recognition and social media camera applications
- News aggregators and social media news feeds
- Sorting algorithms for streaming music and videos
- Internet search engines
- Navigation applications

The sorts of ADM processes people might engage with occasionally include things like:

- Online pricing systems
- Autopiloting systems
- Chatbots
- Blood tests

[63] We note the impact of the new reforms around extraterritorial effect in the Privacy Act 2020 without delving deeper into these here.

- ROC*ROI

Other kinds of ADM processes which people engage with more rarely include things like:

- Applications for government services and entitlements
- Policing algorithms
- Automated systems relating to justice and the Courts
- Robotics systems, e.g. in industrial settings

Some of these examples are derived from and discussed in the algorithm assessment report by Stats NZ,[64] as well as the University of Otago report on the use of algorithms in the New Zealand public sector.[65]

Aiming for "increased adoption" of ADM might be better understood as "increased automation" of ADM systems

> Because ADM systems are already widely adopted, any focus on increasing the adoption of ADM might be better understood as attempts to increase the degree of automation in existing decision-making processes, or the use of automation in more complex tasks.

The more complicated decisions become, the more they can rely on data, which can have the effect of increasing the risks of poor data usage or poor algorithm selection. Equally, reliance on better use of complex data can enhance the benefits of automation too, depending on context.

NZ law explicitly and implicitly authorises public use of automated electronic systems to make decisions

> We found examples of statutes that explicitly authorise the use of "automated electronic systems" to make decisions. Equally, ADM is just a tool like any other that can be used by the public service without requiring lawful authorisation.

ADM is one kind of policy and operational tool like any other, even if it has features that give it particular risks and benefits. Accordingly, government agencies are empowered to use it like any other tool by their status as corporate bodies (ie, Crown Entities), so long as they comply with other regulation.[66] Because we include the social and institutional settings within these government agencies as part of an ADM system, any regulation that limits the way these institutions can act generally may also touch upon the ADM system: one example is the way that the Ombudsman or the Official Information Act 1982 can influence the amount of information available about an ADM system.

Literature in New Zealand and elsewhere concludes that ADM is frequently already covered by existing legal systems, although some have concluded that data protection law is

[64] Stats NZ and DIA "Algorithm Assessment Report" (October 2018) <https://data.govt.nz/assets/Uploads/Algorithm-Assessment-Report-Oct-2018.pdf>.

[65] Colin Gavaghan, Alistair Knott, James Maclaurin, John Zerilli, Joy Liddicoat "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, Wellington, 2019).

[66] Consider for example the Crown Entities Act 2004, the Ombudsmen Act 1975, the Privacy Act 2020 and the Official Information Act 1982.

insufficient to give protection against "unreasonable inferences."[67]

Some statutes in New Zealand explicitly authorise the use of "automated electronic systems" to make decisions, exercise powers, and fulfil functions under relevant enactments through a process of delegation.[68] There is a common drafting pattern which suggests to us that there have already been de facto policy decisions made by Parliament and various executive agencies about the kinds of safeguards expected of the use of automated electronic systems to make decisions. The kinds of safeguards include that there is recourse to a human review without undue delay[69] and that the relevant official responsible for delegation to the system is "satisfied that the system has the capacity to make the decision ... with reasonable reliability". [70] There will be an extensive body of information held in government as policy, operational material and data to show how agencies have continued to satisfy themselves of the reasonable reliability of an automated electronic system, including what if anything they have done to generate public trust and confidence in use of the system. We suggest this existing material is the best starting point for understanding how trust and ADM are interacting in New Zealand.

Aside from making decisions using heavily automated means, agencies are also essentially unable to function or to implement complex legislation without the use of software tools. These can also be classified as automated decision-making systems. There is an emerging literature around the use of new governance approaches that lead to "digital-ready legislation" or "better rules", that make statutes easier to convert reliably into algorithmic form, and thereby enhance their incorporation into ADM systems.

Case Study to investigate further: Customs New Zealand and "SmartGate"

> Customs New Zealand has been using an automated electronic system to make decisions about passenger processing and immigration since 2009. The system illustrates the complex layers in an automated decision-making system where trust may be engaged.

Customs uses "e-Gates" or "SmartGates", which incorporate:

- A chip in a user's passport, which is encrypted, and holds a static image of the passport holder. These static images can be recorded by users at home before being uploaded and submitted to the passport office using an automated decision-making system that is also subject to human review.
- A physical "gate", which is surrounded by signage which limits who can use the gate. The gate opens and closes to control human movement through it.
- A scanner, which detects and reads the information in an e-passport when the passport is physically held to the scanner by the user.

[67] Sandra Wachter & Brent Mittelstadt "A right to reasonable inferences: re-thinking data protection law in the age of big data and AI" (2019) 2 Columbia Business Law Review.
[68] See for example the Customs and Excise Act 2018, ss 274A-274E, 296-297 and "Arrangement for Use of an Automated Electronic System" (9 December 2010) New Zealand Gazette No 2010-go9328.
[69] See submission by the Privacy Commissioner about the role of "undue delay" in relation to ADM systems: Office of the Privacy Commissioner "Submission to the Government Administration Committee on the Border (Customs, Excise and Tariff) Processing Bill 2009" (30 October 2019).
[70] See s 296(2)(b), Customs and Excise Act 2018 and similar drafting in other enactments.

- A camera system which takes multiple frames from multiple angles to create a composite of the person standing in the gate.
- A biometric system that detects features of a person's face using measurements like distance between the eyes and the shape of the ears, then uses that information to compare the static image of the person with the composite image from the gate.
- A database within Immigration NZ that detects any flags or alerts against the passport holder, and can trigger a "rejection" by the gate.
- An audit system using an external contractor that vets the operation of the biometric systems in the gates.
- A system of review and testing that records data about the Gate's operations.
- A "confidence interval" system that is set by Customs and errs toward false positives and false negatives depending on the context and requirements of Customs and Customs' overall confidence in the system.
- A diplomatic and political system which widens or narrows the scope of people who can use the e-Gate system depending on diplomatic and other considerations.
- A legal authorisation system in the Customs and Excise Act.
- Declarations in the New Zealand Gazette about the Chief Executive's satisfaction of the system's "reasonable reliability".
- Oversight requirements around consultation with the Privacy Commissioner.
- Procurement and other processes around the creation and use of the system and acquisition of the hardware that comprises the gates.
- A kind of architectural exclusion, which renders the gates too narrow to be used by a person using a wheelchair unless it is an airport wheelchair that is narrow enough to fit into the gate.
- A human manual review system, whereby people rejected by the gate or people unwilling to use the gate, or unable to use the gate, can be referred for manual processing.

During the period smart gates have been in operation, passenger volumes at Auckland Airport have massively increased, but the overall resourcing required to process them have decreased. Customs performs "customer satisfaction" testing, but otherwise does not conduct trust and social licence work with the New Zealand public on its use of the SmartGate system.

Case Study: imprisonment for interfering with an automated electronic system

> Under the Food Act 2014, an individual can be imprisoned for up to 3 months for interfering with an automated electronic decision-making system.

The Food Act 2014 governs how people trade in food, including matters of food safety and accountability for unsafe food.

Sections 374 and 375 of that Act empower the Chief Executive to use an automated electronic system to make decisions, exercise a power, carry out a function or duty, or other actions. The

Chief Executive can only make such an arrangement if satisfied that the "system has the capacity to do the action with reasonable reliability" and there is a process of review by a human for affected people without undue delay. Section 375 describes the effect of the use of an ADM system, and sub (4) states that "if the system operates in such a way as to render the action done or partly done by the system clearly wrong, the action may be done" by a nominated human.

Notably for the Council, section 239 of the Act creates a criminal offence for anyone who intentionally obstructs or hinders an automated electronic system under the Act. It is also a criminal offence to knowingly damage or impair an automated electronic system. The offence is punishable by a fine of $250,000 for a body corporate, or for individuals to a fine of $50,000 and imprisonment of up to 3 months. We think this is notable because, far from being something new to government or with limited legal authority, in this case a person can be imprisoned for interfering with how an ADM system operates.

Potential Case Study: Election 2020 referendum electoral packs

In the most recent general election in New Zealand, multiple complaints to the Electoral Commission alleged that information pamphlets relating to one of the referenda had been wrongly included within the envelopes containing standard electoral forms and posted to electors. According to reporting by Radio New Zealand, the electoral commission said that all 24 complaints were not justified because their systems were automated and, in effect, it was impossible for these systems to have erred or been manipulated in the manner alleged.[71] RNZ reported the situation as follows [emphasis added]:

> *Chief electoral officer Alicia Wright said they had investigated how it happened but had ruled out anything being added through the packaging machines.*
>
> *"Every scenario for the brochure to be inserted into the pack or swapped out with other material has been looked at and eliminated."*
>
> *She said when lots of election material was delivered, people thought the items had arrived together. The voting packs are compiled at NZ Post sites in Auckland and Christchurch* ***using a fully automated process****.*
>
> *The commission has confirmed that* ***NZ Post had run various tests and determined that no additional material could have entered the process without detection by its machines****.*
>
> *"The machines are finely calibrated and reject any overweight items. These tests, along with our other enquiries, have* ***led us to conclude that only material meant to be in the packs, went into the packs,"*** *Wright said.*

This is an opportunity for an interesting case study into a specific application of an ADM system in a unique context, where a complete level of trust has been established in that process, to better understand how the electoral commission could be placed to exhibit such confidence in the face of 24 separate complaints.

[71] Radio New Zealand "Electoral Commission: Anti-euthanasia flyers not included in EasyVote packs" (9 October 2020) <https://www.rnz.co.nz/news/national/427991/electoral-commission-anti-euthanasia-flyers-not-included-in-easyvote-packs>.

SUPPLEMENTARY INSIGHTS

Apart from the key insights covered above, we register the following thoughts to the Council about how to proceed with its work based on our time spent with the literature.

1	ADM systems are socio-technical in nature, not exclusively technological. It is difficult to extract “decision-making” from other layers of an automated decision-making system. This makes ADM a difficult subject to investigate from a digital technologies perspective.	Human decisions influence ADM systems at multiple levels, including: data selection; algorithm selection and training; setting levels of confidence when it comes to generating scores and making predictions; interpreting data outputs; incorporating machine outputs into decision-making; and finally the extent of human delegation to the automated system.
2	Some kinds of decision may be less amenable to automation, because they are poorly represented by digital data.	At base, ADM systems process data. They are automated through computational processing of that data. That means, for any ADM system, the decision involved must be reducible to data at some level. Data inputs are required, and data outputs must be produced in order to implement or support a decision. Some decisions that we make about the world around us cannot be adequately or fairly reduced to digital data. They will not be amenable to ADM systems. Some decisions, and some data processing, should not be delegated to a computer to the same extent as others: specifically, the decision should not be automated to the same degree. That is the case where the impacts of a decision may have disproportionately severe effects, or where minor errors may compound rapidly at computational scales to create significant problems.
3	Don’t get hung up on “kinds of algorithms” or whether a machine is capable of “making decisions”.	In many ways, focusing on algorithms is a secondary consideration to focusing on data. We advise the Council not to get caught up on distinguishing between different kinds of algorithms, unless the structure of an algorithm affects its transparency and auditability. Many algorithms with various features and effects will be used together in relation to a wide range of datasets. Instead, only focus on the kind of algorithm involved to the extent that it affects the ethical and epistemic issues caused by using

		that algorithm, such as its transparency and auditability. For example, neural networks can be opaque and complex machine learning weighting algorithms can be very time-consuming to assess. There are a range of ways to facilitate public scrutiny over ADM systems that do not rely on close technical analysis of the computational aspects of the system.[72]
4	Profiling, risk prediction, and forecasting are hazardous, but hazards can be managed	The existence of risk, and even bias, is not fatal to an ADM system's viability. Rather, it simply means that such things must be consciously identified and taken into account in risk mitigation strategies.
5	Focusing on safety and efficacy can lead to trust. Few people will trust an unreliable system.	Trust is inextricably linked with whether a system achieves its intended purpose or not. If it is not effective, or it is unsafe (as in unpredictable or unreliable), then we can hardly suggest it should be trusted. This link between trust and trustworthiness makes it difficult to focus only trust without focusing more generally on other elements of system design and the operational and policy processes around those systems. Sometimes starting with trust is the wrong thing to do: instead, trust will follow from safety, efficacy and accountability.
6	New Zealand needs to foster confidence in the specific benefits and limitations of ADM systems among policy and technical communities.	We heard during interviews that frequently key decision-makers would avoid the use of ADM systems even where that avoidance wasn't justified: ie, there were low levels of trust, even though the systems might have been trustworthy. This was suggested to be a result of the fact that these communities lacked the skills to confidently assess the risks and limitations of the system involved. It can be difficult to scrutinise ADM systems at a technical level. A certain degree of expertise is required. Further, analyses can be long and technical in a way that is inscrutable to many.

[72] Aaron Rieke, Miranda Bogen, David G Robinson "Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods" (February 2018) <https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods>

		That is not unique to ADM. There are other areas where senior leaders or policy makers are required to deal with scientific or technical material outside their training. One area worth focusing on in future is how to up-skill people dealing with ADM systems, to both avoid overconfidence and underconfidence, and instead focus on contextual merit.
7	There are enough sets of high-level principles: there is a demand for more detailed, case-specific guidance.	Expert stakeholders, corroborated by the literature, point out that there are an extensive array of high-level principles on AI and ethics. The greater the generality of these, the more widely they can be applied, but in all likelihood the less effect they can have. Organisations repeatedly ask to understand how principles and frameworks apply to them, in their specific operational context, and with their specific conditions. We suspect the reason that this level of detailed guidance is less attractive to organisations conducting research in this space is because it is less scalable and more resource intensive than general, abstracted statements of principle. We think it is important to shift towards more detailed guidelines, away from principles. This should include specific case studies, because trust in ADM is affected massively by who is doing it, what is being done, the context, the potential impacts. Resist being forced into abstract discussions. Insist on specifics.
8	Foster expert communities in New Zealand who can share examples of good practice	It would be enormously helpful to contribute to building networks of technical expertise across New Zealand. These networks could exchange experience and information, just as other professional bodies do (e.g. in medicine, law, and engineering). Exchanging ideas about best practice and use cases is one of the more effective and implementable mechanisms available. At the same time, this will need to be done in a way that respects commercial sensitivity and the perceived hazards of being

		transparent in government. The way that we deal with this in other areas of society is to rely on the capability and integrity of bodies of experts or other mechanisms for verifying the reliability of those communities' reporting and advice. We recommend that the Council examine how New Zealand might form bodies of excellence around the use of automated decision-making systems with an intentional view to sharing case studies that describe how those systems were developed, implemented, used and audited. This will enable new entrants and others to gain close contextual insight into how such systems are used and what good practice looks like.
9	Senior decision-makers also need to trust ADM and that their staff are competent to work with ADM	We have heard compelling cases that change must be led from the top, by champions at the executive (private sector) and Ministerial and executive (public sector) levels. At the same time, these individuals must have the technical expertise to ask the right questions. Frequently, senior decision-makers might be policy analysts, lawyers, or doctors, who might be averse to the risk of digital projects primarily because they lack the skills to confidently assess the magnitude of the impacts and likelihood of those risks occurring. If senior decision-makers are up-skilled in digital technologies and how ADM systems work, they can ask better questions of the people responsible for working with the systems under their supervision.
10	New Zealand must engage in a discussion about Māori data sovereignty	New Zealand scholars are internationally influential on the issue of how data and data-driven technologies can operate to perpetuate colonisation. Both public and private entities have played a role in colonisation, and private entities ought to still consider how their technologies and datasets could have disproportionate impacts on different groups that are unjustified or cause harm. Māori scholars should be supported to exercise tino rangatiratanga around Māori data, and to

		consider the ways that mātauranga Māori can or cannot co-exist with, complement, or supplant automated decision-making processes.
11	Don’t just rely on a human in the loop “Human in the loop” is one tool, for specific systems in specific contexts, but it is not a silver bullet solution	Sometimes, the sheer volume of decisions being made by an automated system would overwhelm any manual review system. Sometimes the nature of what the algorithm is doing is beyond the capabilities of a human to review, process and scrutinise. In general, it is cognitively hard to second-guess the high-tech system, or to be alert to errors when they are very rare. Empirical research shows that some systems can reduce the impact of human bias in carefully tailored situations, but not in all. In any case, analysis must be specific enough to consider which “human”, in what fashion “in”, and within what “loop”.
12	Trust must be considered from multiple perspectives	In practice, the Council's focus has been on levels of trust held by people subject to automated decision-making systems. This is only one area that could have been investigated. In an ADM system, trust can be held from various perspectives, including: the operator; the people overseeing the operator and accountable for operator’s actions; the decision-maker whose judgement is being automated; and finally the person subject to “decision” or processing. Whenever trust is being analysed, we suggest always asking “trust by who in what, in what situation”. Trust will vary heavily depending on specifics.
13	Automated decision-making is a development, not a revolution	Humans rely on computers, data, or the product of algorithmic processing for nearly every decision they make today, in a business, government or professional context. When considering ADM, focus on the extent to which humans are removed from decision-making processes. This is seldom if ever a binary matter, instead existing on a spectrum.

14	Trust and ADM cross into the extensive literature on AI ethics, data governance, related topics	AI research goes through periods of “winter” and “summer” as enthusiasm and technical capability wax and wane. Currently, we are in an AI summer period, although some are forecasting a coming “AI winter” again. Because of the hype around AI, there has been extensive discussion about the ethics, legality and implications of AI. That has led to a proliferation of “principles”. Fortunately, the literature on principles and ethics has also reached the stage of meta-analysis. Mittelstadt et al condense all relevant principles to these: 1. Epistemic concerns 2. Normative (ethical, legal) concerns 3. Transparency They call for future work on the ethics of algorithms to adhere to their framework in order to facilitate better progress. In New Zealand, we have a range of principle-based guidance, all of which are referred to in the Algorithm Charter. There is no need for another set of principles unless these are to be nested within clear guidance, with a conscious appreciation of how they will be applied and measured. They should be developed with technologists and data scientists who are intended to use them.
15	ADM has to be scrutinised like any other policy or operational tool	It can be difficult to scrutinise a technical system. There are also areas where government agencies legitimately have discretion as to how legal or policy objectives are achieved. However, just because a computational system is being used, that does not excuse the government from obligations of transparency and accountability for how that system works. If an agency cannot explain how a system works, it should not be using it. The level of detail in the explanation varies: so, for example, it might be acceptable to use inscrutable systems so long as there are other means for verifying the system’s safety and efficacy.

16	Trust is dynamic and contextual, as are the requirements of ADM systems to be safe and effective	Trust is not a static state of affairs. It will wax and wane over time depending on context, including factors completely beyond the control of direct parties to a relationship. Equally, what makes a good ADM system relies heavily on contextual matters like what it is meant to achieve. We suggest close attention to context and detail when it comes to analysing ADM systems.

Areas that merit further investigation

In future research in New Zealand relating to trust and ADM, we recommend focusing on the following areas:

- The empirical testing of ADM systems, including methods, best practice, and case studies demonstrating how an actual system has been tested.
- Empirical testing of trust in systems by operators to understand how operators of ADM systems do or do not feel confident relying on them. We found some examples of situations where law required public service chief executives to satisfy themselves of the “reasonable reliability” of a system: how did they satisfy themselves of that reliability?
- What does good practice look like when it comes to including affected people in the building, operation and auditing of ADM systems? There are strong suggestions this is required, but how have organisations gone about doing so and what can we learn from them?
- The Council’s research has tended to focus on use of ADM systems in the public service. In future, we recommend a closer focus on how private groups are building and using ADM systems.
- While there are frequent calls for transparency around ADM systems, are there any good examples of how transparency has been achieved, in New Zealand or elsewhere? How can such methods be adopted?
- A strong recommendation heard in our interviews was that mid-level managers and procurement staff in large organisations will do what is required to comply with the expectations and demands of senior decision-makers and management: in the public service, that includes Ministers of the Crown. How might we ensure that better advice or training is available to senior decision-makers, or that experts feel greater confidence to advise them on the strengths and limitations of ADM systems in particular contexts?

REFERENCE LIST

Blogs, news articles and other web materials

- Algorithm tips: <Algorithmtips.org>.
- Drive Tesla Canada "Tesla Model 3 owner walks away uninjured after crashing into overturned truck on highway" (1 June 2020) <https://driveteslacanada.ca/model-3/tesla-model-3-owner-walks-away-uninjured-after-crashing-into-overturned-truck-on-highway/>.
- Jacob Goldstein "How The Electronic Spreadsheet Revolutionized Business" NPR All Things Considered <https://www.npr.org/2015/02/27/389585340/how-the-electronic-spreadsheet-revolutionized-business>.
- Julia Carrie Wong "Greyball: how Uber used secret software to dodge the law" (4 March 2017) The Guardian.
- Marlies Van Eck "Automated Administrative Decisions and the Law" (11 May 2018) <https://automatedadministrativedecisionsandthelaw.wordpress.com/>.
- Marlies Van Eck "Study on Supervising the use of algorithms by the government." (8 January 2020) <https://automatedadministrativedecisionsandthelaw.wordpress.com/2020/01/08/study-on-supervising-the-use-of-algorithms-by-the-government/>.
- New Zealand Council for Civil Liberties "Algorithm Charter – Panel Discussion" <https://nzccl.org.nz/event/algorithm-charter-panel-discussion>.
- Nick Hopkins "Revealed: Facebook's internal rulebook on sex, terrorism and violence" 21 May 2017, The Guardian.
- The Data Transfer Project "Data transfer project overview and fundamentals" (White paper, 20 July 2018) <https://datatransferproject.dev/dtp-overview.pdf>.
- Transparency International New Zealand "Algorithm Charter puts focus on people" <https://www.transparency.org.nz/algorithm-charter-puts-focus-people-2/>.

Research-based materials including conference publications, research reports and journal articles

- A Zaheer, B McEvily, V Perrone "Does trust matter? Exploring the effects of interorganisational and interpersonal trust on performance" (1998) Organisation Science 9(2) at 141.
- Aaron Rieke, Miranda Bogen, David G Robinson "Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods" (February 2018) <https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods>
- Alexa Hagerty, Igor Rubinov "Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence" (18 July 2019)

<https://arxiv.org/pdf/1907.07892.pdf>.

- Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK. ACM, New York, NY, USA, 12 pages. https://doi.org/ 10.1145/3290605.3300271

- Anna Huggins "We need human oversight of machine decisions to stop robo-debt drama" (2 July 2019, The Conversation).

- Anna Pendergrast, Kelly Pendergrast "Digital Inclusion / Te Whakaōrite i e Urunga Tuihono" (2020, Bridget Williams Books) Shouting zeroes and ones (ed Andrew Chen).

- Araujo, T., Helberger, N., Kruikemeier, S. et al. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Soc (2020)

- Bart Nooteboom "Trust and innovation" in Handbook of Advances in Trust Research (Reinhard Bachmann and Akbar Zaheer, eds) Edward Elgar 2013.

- Bickmore TW, Utami D, Matsuyama R, Paasche-Orlow MK. Improving Access to Online Health Information With Conversational Agents: A Randomized Controlled Experiment. J Med Internet Res. 2016;18(1):e1. Published 2016 Jan 4. doi:10.2196/jmir.5239

- Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter and Luciano Floridi, "The ethics of algorithms: Mapping the debate" (2016) Big Data & Society July-December 1-21.

- Brent Mittelstadt "From Individual to Group Privacy in Big Data Analytic" Philos. Technol. (2017) 30:475–494

- Briony Blackmore "Predictive Risk Models In Criminal Justice Te Whakatauira Tūraru Matapae Me Te Ture Taihara I Aotearoa" (2020, Bridget Williams Books) Shouting zeroes and ones (ed Andrew Chen).

- Caleb Moses "The Integrated Data Infrastructure / Te hanganga o ngā raraunga kōmitimiti me te whakaaetanga ā-iwi" (2020, Bridget Williams Books) Shouting zeroes and ones (ed Andrew Chen).

- Colin Gavaghan, Alistair Knott, James Maclaurin, John Zerilli, Joy Liddicoat "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, Wellington, 2019).

- D. Harrison McKnight and Norman L. Chervany "Trust and Distrust Definitions: One Bite at a Time" in R. Falcone, M. Singh, and Y.-H. Tan (Eds.): Trust in Cyber-societies, LNAI 2246, pp. 27–54, 2001.

- Donna Cormack, Tahu Kukutai, Chris Cormack "Not one byte more: from data colonialism to data sovereignty / Kia Kaua Tētahi paita anō: Mai i Ngā raraunga whakatōpū ki te mana motuhake o ngā raraunga" (2020, Bridget Williams Books)

Shouting zeroes and ones (ed Andrew Chen).

- Dzindolet, M.T.., Peterson S.A., Pomranky R.A., Pierce, L.G., Beck, H.P. (2003) The Role of Trust in Automation Reliance, Int. J. Human-Computer Studies 58 697-718
- ECCE '06: Proceedings of the 13th European conference on Cognitive ergonomics: trust and control in complex socio-technical systems (Association for Computing Machinery, New York NY, United States)
- Ella Brownlie ""Encoding Inequality: The Case for Greater Regulation of Artificial Intelligence and Automated Decision Making in New Zealand"" (Victoria University of Wellington Legal Research Paper No. 8/2020).
- Emile Kolthoff, Michael Macaulay, Frank Anechiarico "Introduction: integrity systems for safeguarding ethics and integrity of governance" International Review of Administrative Sciences 79(4) 593-596.
- Eva C. Kasper-Fuehrer, Neal M. Ashkanasy "Communicating trustworthiness and building trust in interorganizational virtual organizations" (2001) 27 Journal of Management 235-254.
- Horizon Research "Using and sharing information for wellbeing: completed for the Tuhono Trust" (June 2017, Horizon Research).
- Jake Goldenfein, 'Algorithmic Transparency and Decision-Making Accountability: Thoughts for buying machine learning algorithms' in Office of the Victorian Information Commissioner (ed), Closer to the Machine: Technical, Social, and Legal aspects of AI (2019)
- James Downe, Richard Cowell, Alex Chen, Karen Morgan "The determinants of public trust in English local government: how important is the ethical behaviour of elected councillors?" (2013) 79(4) International Review of Administrative Sciences 597-617
- James Larus and Chris Hankin "Regulating Automated Decision Making" Communications of the ACM, August 2018, vol 61 no. 8 DOI:10.1145/3231715
- James Larus et al "When Computers Decide: European recommendations on Machine-Learned Automated Decision Making" (2018) Informatics Europe & EUACM.
- Jason A. Colquitt, Brent A. Scott, and Jeffery A. LePine "Trust, Trustworthiness, and Trust Propensity: A Meta-Analytic Test of Their Unique Relationships With Risk Taking and Job Performance" (2007) 92(4) Journal of Applied Psychology 909-927
- Karl Lofgren, Michael Macaulay, evan Berman, Geoff Plimmer "Expectations, Trust, and 'No Surprises': Perceptions of autonomy in New Zealand crown entities" 2018 Australian Journal of PUblic Administration 77(4) 672-684.
- Korber, M., Baseler, E., Bengler, K. Introduction matters: Manipulating trust in automation and reliance in automated driving. Applied Ergonomics 66 (2018) 18-31
- Lee, J.D., See, K.A., "Trust in Automation: Designing for Appropriate Reliance", University of Iowa, Iowa City, Iowa

- M.L. Cummings (MIT) “Automation Bias in Intelligent Time Critical Decision Support Systems” (paper presented to American Institute of Aeronautics and Astronautics 1st Intelligent Systems Technical Conference 20-22 September 2004) <https://arc.aiaa.org/doi/abs/10.2514/6.2004-6313> at 1.

- Mahtab Ghazizadeh, John D. Lee, Linda Ng Boyle "Extending the Technology Acceptance Model to assess automation"" Cogn Tech Work (2012) 14:39–49 DOI 10.1007/s10111-011-0194-3

- Mayer, Roger C;Davis, James H;Schoorman, F David "An integrative model of organizational trust" The Academy of Management Review; Jul 1995; 20, 3; ProQuest pg. 709

- MHAF Lokin "Wendbaar wetgeven" PhD Thesis, Vrije universiteit Amsterdam, 31 October 2018.

- Michael Macaulay "Plenary 2 to ANZSOG: Transparency, trust and public value" (31 August 2015) <https://www.youtube.com/watch?v=1TbDf23-80k> accessed 10 June 2020.

- Miller, E.J., Perkins, L. "Development of Metrics for Trust in Automation", Proceedings of the 15th International Command and Control Research and Technology Symposium(ICCRTS ’10), Santa Monica, CA, June 22-24, 2010

- OECD (2020), Personal Data Use in Financial Services and the Role of Financial Education: A Consumer-Centric Analysis www.oecd.org/daf/fin/financial-education/Personal-Data-Use-in-Financial-Services-and-the-Role-of-Financial-Education.pdf.

- Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. Hum Factors. 2010;52(3):381-410. doi:10.1177/0018720810376055

- Peter Peng Li "Inter-cultural trust and trust-building: the contexts and strategies of adaptive learning in acculturation" in Handbook of Advances in Trust Research (Reinhard Bachmann and Akbar Zaheer, eds) Edward Elgar 2013.

- Philip Bromiley and Jared Harris "Trust, transaction cost economics, and mechanisms" in Handbook of Trust Research (ed Reinhard Bachmann, Akbar Zaheer) Edward Elgar, 2006.

- Prof. dr. Valerie Frissen, dr Marlies van Eck, Thijs Drouen LLM "Research Report on Supervising governmental use of algorithms" (2 January 2020) Hooghiemstra & Partners <https://hooghiemstra-en-partners.nl/wp-content/uploads/2020/01/Hooghiemstra-Partners-rapport-Supervising-Governmental-Use-of-Algos.pdf>

- R Bachmann, A C Inkpen "Understanding Institutional-based Trust Building Processes in Inter-organizational Relationships" (2011) 32(2) Organization Studies 281.

- Reinhard Bachmann "Trust, Power and Control in Trans-Organizational Relations" (2001) 22(2) Organization Studies 337-35.

- Reinhard Bachmann and Akbar Zaheer "Handbook of Advances in Trust Research" Elgar Online.
- Reinhard Bachmann and Akbar Zaheer "Introduction" in Handbook of Advances in Trust Research (Reinhard Bachmann and Akbar Zaheer, eds) Edward Elgar 2013.
- Rob Kitchin (2017) Thinking critically about and researching algorithms, Information, Communication & Society, 20:1, 14-29, DOI: 10.1080/1369118X.2016.1154087
- Robert R. Hoffman, Matthew Johnson, and Jeffrey M. Bradshaw "Trust in Automation" Human-Centered Computing, IEEE Computer Society, p 84.
- Rousseau, D.M., S.B. Sitkin, R.S. Burt and C. Camerer (1998), ‘Not so different after all: a cross-discipline view of trust’, Academy of Management Review, 23(3), 292–404.
- Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz "Guidelines for Human-AI Interaction" CHI 2019, May 4–9, 2019, Glasgow, Scotland, UK.
- Sandra Wachter "Data protection in the age of Big Data" Nature Electronics Volume 2, pages 6-7 (2019).
- Sandra Wachter & Brent Mittelstadt "A right to reasonable inferences: re-thinking data protection law in the age of big data and AI" (2019) 2 Columbia Business Law Review.
- Schartum, D. W. (2016). Law and algorithms in the public domain. Etikk I Praksis - Nordic Journal of Applied Ethics, 10(1), 15-26. https://doi.org/10.5324/eip.v10i1.1973
- Scott Mishler , Jing Chen , Edin Sabic , Bin Hu , Ninghui Li , Robert W. Proctor "Description-Experience Gap: The Role of Feedback and Description in Human Trust in Automation" (Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting) at 313.
- Transparency in Social Media: Tools, Methods and Algorithms for Mediating Online Interactions (Sorin Adam Matei, Martha G. Russell, Elisa Bertino eds) Springer International Publishing Switzerland, 2015.
- Tūhono Trust "Sharing information for wellbeing: Māori engagement on social license report 2017" (Data Futures Partnership, June 2017)
- UToronto Law, The Citizen Lab "Bots at the Gate: A Human Rights Analysis of ADM in Canada's Immigration and Refugee System" (2018) <https://it3.utoronto.ca/wp-content/uploads/2018/10/20180926-IHRP-Automated-Systems-Report-Web.pdf>.
- Weibel A, Six F "Trust and control: the role of intrinsic motivation" in Handbook of Advances in Trust Research (Elgar Online).

Reports and statements on ADM by public agencies

- Council of Europe Ad Hoc Committee On Artificial Intelligence "Draft table of contents of the feasibility study" (2 July 2020) <https://rm.coe.int/cahai-2020-18-table-des-matieres-table-of-contents/16809ee914>.

- Data Futures Partnership "Our Data, Our Way: What New Zealand people expect from guidelines for data use and sharing" February/March 2017.
- Dave Heatley “Biased algorithms – a good or bad thing?” New Zealand Productivity Commission FutureWorkNZ Blog (2 October 2019) <https://www.productivity.govt.nz/futureworknzblog/biased-algorithms-a-good-or-bad-thing/> accessed 20 Nov 2020.
- New Zealand Customs Service "Border (Customs, Excise and Tariff Processing) Bill - Departmental Report" (4 November 2009).
- New Zealand Customs Service "Initial briefing – Border (Customs, Excise and Tariff Processing) Bill" (20 October 2009).
- Office of the Australian Information Commissioner "Summary of the OAIC's assessment of Department of Immigration and Border Protection's handling of personal information using SmartGate systems" (24 October 2019)
- Social Investment Agency "Data exchange" (4 May 2018, New Zealand Government, SIA-2018-0352).
- Social Investment Agency "From listening to learning" (December 2018, New Zealand Government).
- Social Investment Agency "Place-Based Initiatives" (May 2018, NZ Government, SIA-2017-0359).
- Social Investment Agency "Social Investment Analytical Layer" (May 2018, NZ Government, SIA-2017-0360).
- Social Investment Agency "What you told us: Findings of the ‘Your voice, your data, your say’ engagement on social wellbeing and the protection and use of data" (November 2018, New Zealand Government).
- Social Investment Agency "What you told us: quick guide" (undated, New Zealand Government).
- Stats NZ ""Partnering with Māori on real-world issues: Improving data access and building capability"" (October 2018, NZ Government).
- Stats NZ "A social licence approach to trust" (August 2018).
- Stats NZ "Data knowledge Centre" (27 July 2017) accessed 29 June 2020.
- Stats NZ "Data leadership and capability: leading New Zealand's data capability" (November 2018, New Zealand Government).
- Stats NZ "Data Stewardship Framework" (March 2018, NZ Government).
- Stats NZ "Empowering agencies to use data more effectively" (March 2018, New Zealand Government).

- Stats NZ "Experimental initiatives" (accessed 29 June 2020).

- Stats NZ "Legislative review: Flexible, future-focused data and statistics legislation" (March 2018, NZ Government).

- Stats NZ "NZ Government data system map" (April 2018, New Zealand Government).

- Stats NZ "Partnering with Māori on real-world issues: Improving data access and building capability" (April 2018, NZ Government).

- Stats NZ and DIA "Algorithm Assessment Report" (October 2018) <https://data.govt.nz/assets/Uploads/Algorithm-Assessment-Report-Oct-2018.pdf>.

- The Agency for Digitisation "Digital-ready legislation" (accessed 23 June 2020) <https://en.digst.dk/policy-and-strategy/digital-ready-legislation/>.

- Waitangi Tribunal "Tu Mai te Rangi! Report on the Crown and Disproportionate Reoffending Rates" (2017) Wai 2540.

Submissions on policy and law related to automated decision-making

- Auckland Airport "Submission by Auckland Airport to Customs, Excise and Tariff Processing Bill" (29 October 2009) <www.parliament.nz>..

- Charles Chauvel "Comments by Chairperson of regulations Review Committee on the Border (Customs, Excise and Tariff) Processing Bill." (29 October 2009) <https://www.parliament.nz/resource/en-NZ/49SCGA_ADV_00DBHOH_BILL9612_1_A17262/42744ddd128db174e144ebb357c67f446304d5df>.

- Independent Monitoring Mechanism of the Convention on the Rights of Persons with Disabilities "Making Disability Rights Real Whakatūturu Ngā Tika Hauātanga 2014-2019 " (June 2020) ISBN: 978-0-473-52499-9 (Print) / 978-0-473-52500-2 (PDF)

- Internet New Zealand "Submission on Draft Algorithm Charter" <https://internetnz.nz/sites/default/files/submissions/InternetNZ_submission_Algorithm_Charter.pdf>.

- New Zealand Council for Civil Liberties "NZCCL Submission on Draft Algorithm Charter" <www.nzccl.org.nz>

- Office of the Privacy Commissioner "Submission to the Government Administration Committee on the Border (Customs, Excise and Tariff) Processing Bill 2009" (30 October 2019).

- Stats NZ "Submissions summary: draft algorithm charter" <https://data.govt.nz/use-data/analyse-data/government-algorithm-transparency-and-accountability/submissions-summary-draft-algorithm-charter/>.

- Transparency International New Zealand "TINZ Submission on Draft Algorithm Charter" <https://www.transparency.org.nz/wp-content/uploads/2020/01/Algorithm-Charter-Submission-to-Stats-NZ-by-TINZ.pdf>.

Legislation, regulation and legal and policy instruments

- "Arrangement for Use of an Automated Electronic System" (9 December 2010) New Zealand Gazette No 2010-go9328.
- Bills Digest "Legal Assistance Amendment Bill 2011" (Supplementary Order Paper No 250) 11 June 2013, Parliamentary Library of New Zealand.
- Border (Customs, Excise, and Tariff) Processing Bill
- Customs (Arriving Passenger and Crew Declarations) Amendment Rules 2019
- Customs and Excise Act 1996, ss 274A-274E.
- Customs and Excise Act 2018, section 296-297.
- Food Act 2012, ss 239, 374, 375.
- General Data Protection Regulation 2016/679, L119, 4 May 2016, p 1-88, 2, 4 and 22.
- Hansard (24 November 2009) 659 NZPD 8220.
- Legal Assistance Amendment Bill 2011 (Supplementary Order Paper No 250), Digest no 2051.
- Notices pursuant to Customs and Excise Act, 9 December 2010 to 31 October 2019 New Zealand Gazette <https://gazette.govt.nz/home/NoticeSearch?act=Arrangement+for+use+of+an+automated+electronic+system&soloRedirect=false>.
- State Sector Act 1988, section 86.
- Statistics Act 1975.
- Statistics New Zealand Algorithm Charter <https://data.govt.nz/assets/Uploads/Draft-Algorithm-Charter-for-consultation.pdf>.
- Stats NZ, Privacy Commissioner "Principles for the safe and effective use of data and analytics" (May 2018)
- Te Mana Raraunga Principles of Māori Data Sovereignty (October 2018).
- The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems (2018)
- Wine Act 2003, s 118A "Arrangement for System".

Selected operational guidelines, guidance and advice

- Canada Treasury Board "Directive on Automated Decision-Making" <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592&section=html>.
- Commonwealth Ombudsman (Australia) "Automated decision-making better practice guide" (2019, Ombudsman.gov.au)

- Council of Europe "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment" (17 July 2020) <https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>.

- Data Futures Partnership "A Path to Social Licence Guidelines for Trusted Data Use (Summary document)" (August 2017, NZ Government).

- Data Futures Partnership "A Path to Social Licence: Guidelines for Trusted Data Use" (August 2017, NZ Government).

- Government of Canada "Algorithmic Impact Assessment (AIA) questionnaire" <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>.

- Independent High-level Expert Group On Artificial Intelligence Set Up By The European Commission "Ethics Guidelines for Trustworthy AI" (8 April 2019). European Commission.

- Ministry of Health "Emerging health technology: Introductory Guidance for safely developing & using algorithms in healthcare" (January 2019, Ministry of Health)

- OECD "OECD Principles on Artificial Intelligence" (22 May 2019, OECD).

- OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449

- Social Investment Agency "Commissioning and Partnering" (May 2018, NZ Government SIZ-2017-0358).

- Social Investment Agency "Data protection and use policy" (December 2019, New Zealand Government).

- World Economic Forum "AI Procurement in a Box: AI Government Procurement Guidelines" (June 2020).