Submission to the Open Government Partnership

Matters to be included in the second National Action plan - 2018-2020 .

# Reducing the hazards created by digitisation and artificial intelligence to government accountability and citizen freedom.

Dr Julia Thornton

This submission is being made by me as a private citizen. By way of introduction, my PhD completed not many years ago is based in the sociology of technology. I retain a strong interest in various branches of this discipline. In addition to this I am a member of organisations which research the workings of parliaments and others which promote government accountability and integrity and take citizen rights seriously.

Before I embark on the substance of this submission I would like to remark on how difficult the OGP Australia website is to navigate and how hard it is to find information on making submissions. The fact that the registration page does not show that there is an alternative means of submitting comments (by email) is a case in point. Good governance processes start with good communication and effective listening. A website is the digital version of this.

To structure this submission, I will begin by outlining the parameters of the problems with both algorithms and data and the two combined, and then set out proposed checks and balances

**SECTION 1. Why do algorithms, especially Artificial Intelligence algorithms pose a threat to accountability and transparency?**

There has been rapid adoption of automated advice and decision-making in finance, welfare, judicial and policing decisions, immigration and other public service areas.

These automate public service processes make

judgements about citizen entitlements and many other citizen interactions with government processes.

Moreover use of sophisticated artificial intelligence to undertake surveillance and identity tasks for government and by private corporations in addition to their use in psychological profiling and political propaganda as evidenced by the activities of Facebook and Cambridge Analytica and other similar profiling software suggest that we are facing something of a watershed in attempting to regulate and control for democratic freedoms like accountability, transparency, personal privacy, rights and justice.

These innovations have the capacity to pose huge threats to transparency, accountability and citizen freedom, some of which are new versions of old problems, especially those associated with corruption and political manipulation of citizen behaviour, and some of which pose problems not encountered before.

Some of this is a consequence of scale - more data is available and more ways to collect and analyse it are being developed. More ways of sharing data and its outcomes are also scaling up.

But algorithms themselves pose special problems for openness and transparency which are new to regulators and others who would wish to see society kept open and democratic.

**1. Algorithms are not value free.**

As an opening statement, it is obvious that algorithms are not value free in the same way that any government or public service decision is not value free. They encode the same normative underpinnings to a desire to fix problems as humans do. And as humans do they should be subject to the same ethical interrogation as to the intent and effect of their operation. This is not a philosophically difficult step, although it may prove practically difficult, as does any form of transparency requirement that considers values.

Numbers are inherently not value-free as attempting to transform qualities into quantities will show. However, algorithms form a special case of difficulty in determining what these inherent values might be. As algorithmic processes might be used for aiding corruption or criminality but also for making more ordinary value judgements, which could lead to social or political bias, political manipulation or abrogation of fairness and rights, discovering and subjecting these values to scrutiny must be part of any transparency process.

The difficulty of determining the social values locked into algorithms is double problem. It is a problem on the human side - recognising one's own values is as difficult as maintaining constant awareness of the air you breath - and it is a problem on the machine side. Values are set at the planning stage of a platform or program - and are apparent in determining the goals of the algorithm; they are embedded in the logic of the process as it is executed; and they can be seen in the consequences - unintended and otherwise - of the outcomes of its operation.

Biases are, inter alia, built in by the classification method used for producing data sets, non representative data collection, lack of match between the context for which algorithms or AI is constructed and the one for which it is used, the lack of demographic and economic diversity of its designers, [[1]](#footnote-1) and the way language is used in programs and platforms. [[2]](#footnote-2)

How to understand these values, and whether they are indeed understandable when they are written into an algorithm are matters to be addressed in this and later sections of this submission.

Values in predictive modelling.

The intent of the program designer is clearly an issue for anti corruption action. Intending to destructively hack or defraud is a clear cut case of antisocial values, although even here, "white hat" hackers and digital whistleblowers can be acting for the greater good.

Things get more complicated when trying to determine the ethical base of predictive modelling.

Governments are most enthusiastic in adopting automated processes for everything from answering human services queries to common Enterprise Resource Planning (ERP) services throughout the public service. [[3]](#footnote-3) Commitment to further digitisation of government processes is inherent in the OGP National Action Plan Commitment 2.3 "Digitization of Government Services". [[4]](#footnote-4) Many of these, especially those dedicated to automating human service requests and processes, will use predictive models.

Predictive models allow for forecasting and forward decision-making, and are a hallmark of more sophisticated software.

However as Cathy O'Neill points out [[5]](#footnote-5), the algorithms which drive predictive models of human services risks are also inimical to openness transparency and justice on three deeper levels than merely finding the values in the intent of the writers of the code.

A. Opaque model rules

The data subjects may be aware that data being collected from them will be used for algorithmic modelling, but the model itself is usually opaque to the subject. As she points out, "Opaque and invisible models are the rule and clear ones the exception." [[6]](#footnote-6)

Most companies treat their predictive models as intellectual property, and go out of their way to keep the secret. Similarly, the modelling algorithms that are used for public service decisions are rarely build by public servants. Most come as proprietary software. Again the opacity of the models workings are deliberate.

While Google and Facebook have recent y been pushed to make data collection more controllable by data subjects, no such transparency applies to their predictive behavioural models.

Furthermore opacity is increased by sheer socio-technical complexity.

Simon and Rieder explain.

"... even if the details of a specific algorithm were made accessible and the necessary technical expertise to investigate could be mustered, chances are that a 'smoking gun', i.e. evidence of 'hardcoded' bias or discrimination, could not be readily found. This is because algorithmic systems do not function as standalone boxes, but as networked socio-technical assemblages that include a multitude of human and non-human actors, with people debating models, setting target goals, cleaning training data, adjusting parameters, and choosing the specific context of application. Hence, algorithmic accountability is also difficult to achieve because algorithmic systems are fundamentally complex". [[7]](#footnote-7)

Here the opacity is not deliberate but caused by the nature of the technology and the technological system. It is equally hard to render transparent, but cannot be controlled by exhortation, regulation or culture change.

B. Detrimental to subjects interests

O'Neill asks the question, "Does the model work against the subject's interests? In short, is it unfair? Does it damage or destroy people's lives?" [[8]](#footnote-8) There are more and more examples of algorithmic models being used to anticipate risk. These range from insurance risk and financial risk, not just to private companies but to governments, but human and social risk. They are increasingly being used to evaluate human services risks like criminal recidivism, and anticipatory likelihood of criminality including social and racial profiling, and the likelihood of fine non repayment and welfare fraud. [[9]](#footnote-9)

This automation has been used far more widely in the USA, but there is no reason to suspect that it will not increase in Australia. Predictive privacy harms have been well documented. [[10]](#footnote-10). [[11]](#footnote-11)

Those who give up the most data to governments- those who use the welfare system or who are part of the criminal justice system - are also the most at risk of discrimination, incorrect data holdings and over policing.

If these models were human, this discrimination would be illegal. The same standards should be applied to algorithms used in profiling risk prediction systems.

All of these models include the aforementioned assumptions about the value which should be accorded certain social characteristics and therefore effectively criminalise people before a crime has been committed. They do not, by and large attempt to find the causes of non compliance, and direct attention away from systemic causes of criminality, fraud and risk. A focus on the individual as risky then leads to a failure to address accountability and integrity measures at the systemic level.

Integrity measures which apply to predictive risk modelling entail both attention to human rights and to systemic integrity.

C. Scalable to a social norm

The third ethical dimension identified by O'Neill is whether the models have the capacity to "scale" or grow. [[12]](#footnote-12)

The underlying issue is the turning of an opinion previously the province of one or two people into a norm that is used as a more or less universal operating rule. Before algorithms, if a bank manager evaluated a customer as a borrowing risk, that opinion would be confined to that person. However these days, as O'Neill points out,

"*If a bank's (algorithmic) model of a high risk borrower ,for example, applies to you, the world will treat you as just that - a deadbeat, - even if you are horribly misunderstood. And when that model scales as the credit model has, it affects your whole life - whether you can get an apartment, a job or a car ..."* [[13]](#footnote-13)

Here the problem is the transformation of an individual judgement ("this person is a borrowing risk") into a social judgement ("these people are a borrowing risk"). As the model scales, so the capacity to negotiate a judgement with the judging apparatus reduces. It is hard to appeal against the judgement of a machine and even harder if the judgement applies to a class of people rather than one.

D. Impunity from privacy legislation.

There is a fourth ethical consequence of predictive behaviour modelling not mentioned by O'Neill that also renders them inimical to openness transparency and justice.

This is the fact that because organisations are using predictive behaviour modelling they do not directly use data that they have sourced themselves or have purchased from a third party. Instead they use the modelling algorithms that have been built on some other data by software companies. These algorithms can build 'personas', imagined customers (or citizens) who conform to character typologies invented through classification of data on real people. These models are then applied to real people to exploit some behaviour like buying or voting, with the consequence that they have a real personal outcome for the person so targeted but without use of their own data and so without regulation.

Crawford and Schultz use the example of Target (USA), the clothing and homewares company, as an example of how these models evade personal data legislation.

"*For example, in the New York Times article about Target predicting pregnancy, Target had never collected data showing that any particular female customer was pregnant — a fact that most people would almost assuredly consider to be very personal and intimate information.*

*Instead, Target predicted this information. Furthermore, the prediction was just as personally sensitive as if it had been collected or shared inappropriately.*

*Target also used the predictive privacy information in a similarly personally sensitive manner by exploiting it for marketing purposes.*

*Nevertheless, because it did not collect the information from any first or third party, Target had no obligation under current privacy regimes to give notice to, or gather consent from its customers in the same way that direct collection protocols require*." [[14]](#footnote-14)

All of the caveats about values which apply to predictive models used by government for service delivery also apply to predictive models used by political parties for election processes. They also apply to the algorithms used for the ever increasing surveillance of citizens through everything from speed cameras to public transport ticketing and the compilation of large data sets through government collected demographic and welfare information.

Values in algorithms - Implications for integrity, transparency and accountability

Algorithms are tools for encoding any tasks we put our minds to creating, good or bad. To use the language of sociology, they reify values. They have unintended consequences like creating barriers to accountability and transparency, and reinforcing discrimination against classes of people.

Values cannot be removed from algorithms any more than they can be removed from society, but it is important for the equality implicit in democracy that we are aware of the and we do our best to promote diversity and self awareness amongst the creators of code, and to check if government and public service use of risk management programs abrogates rights and entrenches inequality and bias.

We should also be cautious of using algorithmic decision-making to penalise individuals while ignoring the systemic issues that have created the problem.

**2. Algorithms tend to be trusted as value neutral.**

While algorithms and the attendant predictive models they use are not value free, the fundamental problem is that governments, the public service, corporations and even private citizens tend, by default, to treat them as if they are.

This misplaced trust means that their opacity is tolerated as a commercial or government entitlement, their underlying assumptions are unexamined and their application is not subject to any significant level of probity.

The tendency for people to trust algorithmic decisions over human ones once sufficient time has passed for the novelty of a shift to a machine to wear off, has huge attractions for risk shifting by governments.

Two phenomena have the effect of making machine decisions and processes seem more value neutral than human ones. Firstly turning qualities into quantities (or values into numbers) has a lulling effect on alertness to bias. Numbers are seen to be intrinsically superior to value or qualitative statements in terms of appearing to imply no politics.

Secondly, as machines up until fairly recently have been dumb, (but not value free - a gun, for instance, has definite values built into it, and the idea of the personal computer embodies the ideal of working alone), so it has been easier to consider that they act impartially. If you are hit by a car in an accident it is easy to blame the driver. If you are hit by a falling car with no driver its harder to blame the car; the car will hit anyone, regardless of status or politics under the same circumstances, while the driver of the car is presumed some agency.

The third attraction of risk shifting to algorithms is not their apparent neutrality but their lack of contextual judgement and the non negotiability of algorithmic decisions. Phone calls that were previously interactions with public servants who knew about their systems, have now become conversations with chatbots that have crystallised common answers into routines that are variable but lead to the same conclusions. Such answers cluster around the needs of a population mean and frequently fail to recognise specific disabilities or special needs that are experienced by a small percentage of the population, thus discriminating against minorities.

Their non negotiability is often compounded by the removal of a human fall back contact, so complex problems cannot be discussed with a human being with decision-making discretion.

These factors were briefly mentioned as effects of scaling earlier in this submission, but they have efficiency dividends in that they can circumvent any individualisation of government or public service decision-making and speed up processing.

However as we have seen, values and judgements are integral to the development of algorithms and can be used to the detriment of the people they are purported to serve. Governments must guard against the temptation to risk shift to programs as it could be argued occurred with "robo-debt". [[15]](#footnote-15)

The same argument that machines were responsible, this time represented as a software fault was initially tried by the Commonwealth Bank with regard to the defect in its automatic teller machines that allowed money laundering. [[16]](#footnote-16)

This kind of deflection of human responsibility to machines should not be able to occur.

Greg Medcraft, in a speech to the FINSIA regulator Panel invoked three principles of accountability and responsibility.

*"There are three principles to keep in mind to make sure there is appropriate accountability for technology platforms and algorithms:*

*1. Responsible person: for any algorithmic system, there needs to be a person who is responsible for its design – and its outcomes.*

*2. Capable of explanation: automated decisions must be able to be meaningfully explained to customers, to the regulator, and to any other interested stakeholders.*

*3. Redress: if and when algorithms make mistakes, whether because of data errors in their inputs, or because of issues with their design, there need to be avenues for redress.*

*We cannot use technology platforms or algorithms to simply shift risk to the consumer or other areas of society."* [[17]](#footnote-17)

Unfortunately for reasons described below, these principles are almost impossible to execute in practice.

But before embarking on the reasons we cannot apply Greg Medcraft's principles, (see heading 3 below," Decisions made by Artificial intelligence cannot have their workings retrospectively understood") I want to explore more fully the pressure to trust in algorithmic systems.

A. Blockchain - a special case of 'trust me, I'm an algorithm'.

It is quite likely that current crypto currencies will not survive, they are certainly volatile forms of value, [[18]](#footnote-18) and blockchain in its current forms is both extraordinarily energy inefficient and takes up to 10 minute to make a transaction, which is long by digital standards. [[19]](#footnote-19)

However, one aspect of their design, 'trustless ' transactions looks set to persist. One does not need to know the workings of cryptocurrency or how it rests on blockchain to understand the attraction of "trustless transaction systems". [[20]](#footnote-20) [[21]](#footnote-21)

There is boundless faith in the attraction of trustless transaction systems to eliminate all sorts of corruption by eliminating the fiduciary 'middlemen' of financial and contract exchanges, in favour of machine verification of exchanges between peers, as currently represented by blockchain transactions. [[22]](#footnote-22) [[23]](#footnote-23)

Such systems are called 'trustless' because no-one needs to trust a person, the system itself creates the trust through the manner of its transaction. Blockchains simultaneously enable the transaction transfer and ensure **sender authenticity** and **currency validity**.

Without engaging in too much technical description, however, in addition to the problems created by the generic values encoding problems of all algorithms, blockchains pose their own special ethical dilemmas. They irreversibly record transactions, even mistakes, which are not retrospectively correctable.

In addition, public 'permissionless' blockchains like those used for Bitcoin and Etherium are open to everyone and need no-one's permission to join which makes them look ideal for audit by machine and general use as a way to control corruption and errors. On the other hand, private 'permissioned' blockchains such as those being considered by banks (to prevent the existential threat to banking posed by the existence of open public trustless systems) are dependent on the permission of the owner to enter, and only show the transactions that pertain to the particular user. [[24]](#footnote-24) They are effectively confidential systems, closed to inspection.

At a minimum, irreversible transactions (in public blockchains) and permissions and the confidentiality built into private blockchains pose a difficulty for investigation and audit by ordinary anti corruption systems.

Further regulatory barriers are posed by decentralisation and confidentiality. Blockchains can also perform a co-ordinating function between transactions.

*"It ... allows for the execution and interconnection of a variety of smart contracts that interact with one another in a decentralized and distributed manner. Multiple smart contracts can be bound together to form decentralized organizations that operate according to specific rules and procedures defined by smart contracts and code."* [[25]](#footnote-25)

Decentralized blockchain-based organizations could be more difficult for governments to control and regulate, and "the pseudo-anonymous nature of blockchain technology presents significant regulatory challenges. Its widespread use could potentially undermine the ability of law enforcement to uncover and clamp down on illegal activity". [[26]](#footnote-26)

The temptation to governments and political parties to move their own transactions to such private trustless systems once they have been refined, may also prove detrimental to openness and transparency.

The temptations of implicit trust - Implications for integrity, transparency and accountability.

Our tendency to treat algorithms as value-neutral combined with their reification of particular norms may result in granting a new rigidity and obduracy to problems of government transparency, party honesty, amplification of the power effects and interests of small unrepresentative groups and abrogation of human rights and democratic processes.

**3. Decisions made by artificial intelligence cannot have their workings retrospectively understood.**

The reason that decisions made by artificial intelligence cannot have their workings retrospectively understood lies in the nature of algorithms and especially in the nature of artificial intelligence. Ordinary algorithms are programmed to a specific end by someone who writes code to produce an effect. These kind of algorithms are explicable, even if not value free as explained above.

They can be investigated by asking about the goal, the algorithmic process and the output, and articulating the decision and value logic of all three in an audit process.

However, the growth of instances where artificial intelligence is applied to government decision-making or control of resources, finances, or citizen behaviour is steadily increasing. But as artificial intelligence becomes increasingly sophisticated, so the processes by which AI algorithms reach decisions or give advice become intrinsically inscrutable. [[27]](#footnote-27)

This does not simply arise through the complexity of the number of algorithms. Artificial intelligence introduces a qualitative change which means that while the goals may be interrogated, the processes and perhaps also the outputs cannot.

As transparency of decision-making in government requires that those answerable have some way to explain the logic of their decisions, this is an issue that must be addressed. Already the genie is out of the bottle and government and public service decisions that rely on artificial Intelligence for automation are effectively inexplicable.

The ethical question is not just whether those responsible can explain their AI assisted decisions, it is how to allocate responsibility when the process is a black box and the outcome may be both unpredictable and unexpected.

A. Fundamental inscrutability.

According to Selbst and Barocas, "The requirement that those affected by decisions be able to make sense of them is a central tenet of due process and administrative law" [[28]](#footnote-28). But the problem with advanced artificial intelligence is that it is not just those who are affected who are eliminated from making sense of AI, even the engineers who created them cannot elucidate their processes.

The cause of this is well explained by Knight [[29]](#footnote-29);

"*The workings of any machine-learning technology are inherently more opaque, even to computer scientists, than a hand-coded system. This is not to say that all future AI techniques will be equally unknowable. But by its nature, deep learning is a particularly dark black box.*

*You can’t just look inside a deep neural network to see how it works. A network’s reasoning is embedded in the behaviour of thousands of simulated neurons, arranged into dozens or even hundreds of intricately interconnected layers. The neurons in the first layer each receive an input, like the intensity of a pixel in an image, and then perform a calculation before outputting a new signal. These outputs are fed, in a complex web, to the neurons in the next layer, and so on, until an overall output is produced. Plus, there is a process known as back-propagation that tweaks the calculations of individual neurons in a way that lets the network learn to produce a desired output.*

*The many layers in a deep network enable it to recognize things at different levels of abstraction."*

Probably the most influential example of this type of program is Alpha Zero, which instead of using "handcrafted knowledge and domain-specific augmentations" used "deep neural networks and a tabula rasa reinforcement learning algorithm" [[30]](#footnote-30) and was able to learn how to play chess from the rules of chess, and not from coded instances of how to play. [[31]](#footnote-31)

The problem this poses is that we "need to be able to explain their decisions — including building confidence about how they will behave in the real-world, detecting model bias and for scientific curiosity ". [[32]](#footnote-32)

Some possible solutions include:

* Extending legal rights of data/algorithmic subjects to an explanation.

Wachter et al highlight the promise of the European *General Data Protection Regulation (GDPR) [[33]](#footnote-33)* in granting openness to AI decisions.

*" Since approval of the EU General Data Protection Regulation (GDPR) in 2016, it has been widely and repeatedly claimed that the GDPR will legally mandate a ‘right to explanation’ of all decisions made by automated or artificially intelligent algorithmic systems. This right to explanation is viewed as an ideal mechanism to enhance the accountability and transparency of automated decision-making." [[34]](#footnote-34)*

Wachter et al, nevertheless conclude that " *However, there are several reasons to doubt both the legal existence and the feasibility of such a right." [[35]](#footnote-35)*

One major stumbling block is that information afforded to data subjects must be meaningful, in order that they understand what is being done to their data, their rights and any privacy implications, which information from "black box" processes patently isn't.

Furthermore the legal framework of the GDPR falls short. They argue that "the GDPR does not, in its current form, implement a right to explanation, but rather what we term a limited ‘right to be informed' ". [[36]](#footnote-36)

* Using "audit study" to understand algorithmic decision making

Sandvig et al advance the proposition that "*Luckily, a method exists for researchers to look inside these complicated, algorithmically driven computer decision systems: the “audit study”. This method, which serves as the most respected social scientific method for the detection of racial discrimination in employment and housing, uses fictitious correspondence. "* [[37]](#footnote-37) They test outcomes by preparing two or more test documents which are equal in every respect except for the variable in question which may be causing for instance, discrimination. The difference in responses reveals the bias. ie they are testing an algorithmic outcome as a substitute for understanding the logic by which it was derived.

However, this method only effectively applies to fixed algorithms where machines are not constantly modifying the way they calculate by programming themselves. However it could work on artificially intelligent programs to reveal where the object of the program was purposeful discrimination.

The major difficulty is the size of the task. It is difficult to imagine a group of human auditors keeping up with auditing the exponential proliferation of AI based on neural networks inside and outside government where they impact on the public interest.

* Ensuring algorithmic results are available for audit.

A similar suggestion relies on increasing the transparency of proprietary algorithms - especially social media algorithms to ensure they are at least accessible to audit tests.

Ghonim and Rashbass propose, " We believe that all platforms using algorithms to distribute content must develop a standardized public interest API (a standard interface for sharing and accessing data) that provides a detailed overview of the information distributed on their networks, while respecting concerns for user privacy, trade secrets and intellectual property." [[38]](#footnote-38)

Essentially they suggest that it is the inputs and results of social media (or other) algorithms which should be opened up. Technically this can be done by requiring the addition to such algorithms of an application called an "open application programming interface". This would allow for the building of third-party algorithms to monitor and report on the inputs and effects of the original AI program.

How and what forms of audit and reporting could take place are a matter of both imagination and technical skill. The authors suggest as key public interest matters, identifying influencers - presumably both human and bots - and identifying what has been censored. However other results of algorithms in use could equally be targeted where there is a suspicion that public interest or ethical outcomes are in danger.

An "open results" approach could ultimately be combined with an "ethical design" approach so that both the inputs and outputs of algorithms are scrutinised in the absence of human capacity to follow the inscrutable logic of the algorithm itself.

This solution overcomes the speed and capacity problem of the last solution, since the audit is machine based, but still does not open up the actual internal logic for scrutiny. We have no idea why the program reached the conclusions it did.

* Interrogating the machine about its processes

It has been suggested that as AI gets better, it may be possible to interrogate the decision machine as an intelligent reporter on its own internal rules and states, in order to get it to explain how it arrived at a decision. In this case, the machine itself is being treated as an equivalent responsible agent. [[39]](#footnote-39)

In the DARPA case, the ethical rationale is trust in the machine in war. "*Explainable AI—especially explainable machine learning—will be essential if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners*". [[40]](#footnote-40)

However*, "just as many aspects of human behavior are impossible to explain in detail, perhaps it won’t be possible for AI to explain everything it does. Even if somebody can give you a reasonable-sounding explanation [for his or her actions], it probably is incomplete, and the same could very well be true for AI, ... It might just be part of the nature of intelligence that only part of it is exposed to rational explanation. Some of it is just instinctual, or subconscious, or inscrutable.” [[41]](#footnote-41)*

Faulty as they are, machine self interrogation offers the best hope to date of constructing something like an explanation for AI processes and actions. The chance of meaningful information is increased when a combination of interpretability techniques is used simultaneously. [[42]](#footnote-42)

B. Additional opacity

As well as fundamental inscrutability through "Black boxing", additional opacity is added in much the same way as that described for data models above. The data set on which the AI algorithm is trained is not always open and transparent and the weighting of the processes to which it is subject, is also frequently commercial in confidence. [[43]](#footnote-43)

C. AI builds AI

Another layer of difficulty in making sense of algorithms or being responsible for their outcome may occur as AI builds AI.[[44]](#footnote-44)

This enables human designers to have even less understanding of the processes, including normative assumptions and values, or the decision making logic of the processes built into their products than current AI designers do. They include those with no machine learning (ML) expertise. [[45]](#footnote-45)

In addition, Google is making their machine learning platform openly available in the hope that this will encourage other developers to use AI to build AI. As the New York Times points out, Google believes "that if more people and companies are working on artificial intelligence, it will propel their own research". [[46]](#footnote-46)

Demonstrating that artificial intelligence is already taking over both routine and high level creative skills that take years to learn, the consequences of this development will be a proliferation of artificial intelligence about which there is little understanding and less oversight, used for all sorts of processes, including those of government and the public service.

D. Things own things?

A whole new dimension of accountability and explicability is introduced when, as may occur in the near future, things own themselves or other things. The idea is a somewhat journalistic title currently in use for an exploration of the limits of autonomous machines. At the moment the concept is under consideration for autonomous vehicles, [[47]](#footnote-47) but once developed there is clearly potential application elsewhere.

The idea was first suggested by Mike Hearn at the 2013 Turing Festival in Edinburgh as a way to make vehicles so cheap they could undercut the huge corporations that now represent a digital cartel. By needing to generate no more profit than was needed to run itself, it could own a bank account, pay tax, send itself in for maintenance, calculate the most profitable way to spend its day and turn itself off if opportunities were not there, all driven by autonomous AI as is being developed for self driving cars. The cars would communicate with each other and people through a public blockchain-like trading application which like the bitcoin blockchain, would not be privately owned.

Belk reflects on the consequences, *".... self - owning objects would exhibit a primary agency that initiates and motivates actions that affect other objects, people, and environments. They would own themselves in a strong agentic sense not only behaviorally, but also legally and morally." [[48]](#footnote-48)*

Do we then argue that because algorithms are strongly agentic, like humans, they are also subject to the same requirements for moral constraint, social responsibility and self explanation?

Self owning objects would presumably own their own AI, and they may also own the means to develop it further (see AI builds AI above). The 'smart contracts' enabled by blockchain may facilitate private secure trustless co-ordination between machines. [[49]](#footnote-49) There is a minefield of transparency regulation and ethical systems management to be negotiated here.

In explicability and opacity in algorithms - Implications for integrity, transparency and accountability

At the moment, neither legal tools nor technological tools appear fully equipped to deal with the problems of inexplicability and opacity in AI algorithms. The fact that these are both increasing in use and developing in inscrutable expertise exponentially, only makes the problem worse. It may be that this issue is the bellether for regulation of the whole new-technology economy, in that the speed of government institutions and the speed of regulation, including the development of the technological means of regulation will never again match the speed of technological change. For trust in government and the capability of the state to maintain control of the society and its phenomena that it is supposed to legislate for, this could spell a social, political and democratic disaster.

While there may be a principle in law that people can be held responsible even for things they do not understand, this responsibility stands to look increasingly puny.

**4. Data privacy**

A. Genres of data.

Data privacy is the starting point for most discussion on "Open Government". [[50]](#footnote-50) However not even this, despite widespread awareness of its problematic outcomes, can be guaranteed, even with the best promises of de-identification of data. What level of privacy data should be subject to is significantly determined by the source and purpose of the data.

Government data on itself.

Some government data does not warrant privacy protection. Indeed it is the very data which should form the basis of government openness and transparency. In a speech to the Victorian Branch of the Fabian Society in August 2018, Chris Culnane asserts that with one exception, there is no case for privacy of government data on itself.

"*In the case of data related to the government itself, there is clearly no justifiable expectation that the everyday actions, decisions, and operations of the government should be private. With the exception of some areas of national security, government should be transparent and accountable to the population, and as such, cannot have an expectation to privacy in the course of its operation*. " [[51]](#footnote-51)

I would suggest there are in fact three exceptions - matters of national security, preliminary investigation and prosecution of alleged breaches of the law, and cabinet in confidence discussions.

Political Party data

While governments may be inappropriately unwilling to expose their own data the opposite is true when it comes to data held on citizens, particularly by political parties. They are quite content to use other peoples data as if it were their property, but with negligible privacy and security protections for the data subjects. . There are hardly any legal limitations on the use of personal data for political purposes in Australia and no obligation to disclose to voters what information is held on them by politicians, political parties, subcontractors or party volunteers. Political parties, and campaign organisations are exempt from the Privacy Act [[52]](#footnote-52) and also exempt from the Spam Act, and the Do Not Call Register.

Both David Vaile, Chair of the Australian Privacy Foundation, and Timothy Pilgrim, the Australian Privacy Commissioner disagree with this level of immunity from legal restriction on using voters personal data in secret, but both major political parties have no plans to remedy the situation. [[53]](#footnote-53) [[54]](#footnote-54)

This leaves voters unaware of what information is held on them and open to undue political manipulation and discrimination, with no way to correct mistakes and misapprehensions, while leaving them vulnerable to hacks of either direct political party information or of the third party proprietary systems which are used for analysing it.

Public service data

At the most basic level of privacy we expect the public service to keep our records private and safe. Data sharing however is set to become extensive, driven by deliberate government policy, including the OGP goal, 2.3 - 'Digitisation of Government services' [[55]](#footnote-55) and the Australian Government 'Digital Transformation Agency'. [[56]](#footnote-56)

While it is roughly possible to divide data sharing issues between internal government agencies and departments, and external to government agencies, external businesses and service providers, practically it is difficult to separate the functions and privacy requirements for data. It is easier in fact to define 'public service data' as 'all data held by the government that may be used by government and non government personnel and agencies'.

At a primary level we expect good encryption and security housekeeping within government departments. [[57]](#footnote-57) This should include:

1. Protecting privacy
2. Protecting data
3. Restricting access [[58]](#footnote-58)

Each of these require a different form of thinking about privacy and encryption. These restrictions should apply within the public service equally to data shared outside it, guarding against overuse of public service data for government surveillance purposes.

While it is impossible to know as an outsider, how effectively these encryption and security measures are being administered - theft and corruption being by nature, secretive activities - at least there is some citizen protection in the Notifiable Data Breaches scheme. This scheme "mandates that Australian Government agencies and the various organisations with obligations to secure personal information under the Privacy Act 1988 (Cth) (Privacy Act) notify individuals affected by data breaches that are likely to result in serious harm." [[59]](#footnote-59) However it has elements of shutting the barn door after the data-hacker horse has breached it. What is needed are proactive processes to tackle privacy issues and data release at both the human and machine level.

General data distribution.

Dr Jake Goldenfein from Swinburne University, interviewed on ABC RN "Future Tense", summed up the changes to the socio technical domain needing data protection.

"*Data protection was originally developed in the technological environment of centralised government filing systems, the computerisation of these large filing systems. The technological environment of contemporary society is vastly different, it's vastly more decentralised.*

*The quantities of data that we are talking about are incomparably greater, and the way that data is processed is very, very different. So whereas it used to be a lot of human processing or quite basic automatic or algorithmic processing, we are now talking about very, very sophisticated data mining mechanisms, artificial intelligence machine learning mechanisms that are processing data, as you said, in real time all the time, and we are not really dealing with centralised government databases anymore, we are dealing with situations where an individual through their use of their smart phone is providing data to numerous entities very, very quickly. We are talking about a situation where physical objects are increasingly connected to the internet through internet of things technologies. The amount of sensors which are picking up or collecting data and providing that data to be processed is increasing exponentially*. "[[60]](#footnote-60)

What this means is that corporations and governments are now knitted more closely together by mutual dependence on the same digital products for acquiring and managing data, their data and data collection is interwoven, and breaches of privacy in one sector can easily disrupt the other.

Breaches by accident or hacking

Breaches by de-identification.

Privacy breached without your data

Third party data; selling on data sets and buying in personal information.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

This submission is unfinished. I have written it over the four days of Easter, and am already relying on the generosity of the OGP NAP team to allow it to be accepted.

There is much more to be said about the four final headings to do with privacy matters; government relations with commercial entities; protection from political surveillance and psychological manipulation though both working on beliefs and addictions as powered by AI and psychometrics, and the overall effects of big data, social media, the internet of things, and the power of commercial internet based companies on governments, democracy and society.

I am aware that much of this submission does not fit under the themes identified for the NAP, but it is consistent with the overarching OGP principles. Given the speed of development and omnipresence of digital technologies in government and in all our lives, its lack of fit with predetermined themes does not make the issues raised any less important for governments to grasp.

Should there be any extension of time for submission, I should be very pleased to know. An update will be forwarded as soon as possible.

Dr Julia Thornton

03/04/2018

1. **Funnell, A. (2017, December 12). AI, algorithmic decision making, ethics and the under-representation of women in tech** [MP3]. *Radio National*. ABC RN. Retrieved from <http://www.abc.net.au/radionational/programs/futuretense/ai,-algorithmic-decision-making,-ethics-and-the-under-represent/9250270> [↑](#footnote-ref-1)
2. **Selbst, A., & Barocas, S. (2017). Regulating Inscrutable Systems**. Presented at the We Robot 2017, Yale Law School. pp 13 - 28 [↑](#footnote-ref-2)
3. **Department of Finance. (2017, September 26). Release of Request for Proposal to establish an Enterprise Resource Planning (ERP) Software as a Service (SaaS) Panel** [Text]. Retrieved March 31, 2018, from <https://www.finance.gov.au/blog/2017/September/26/release-of-request-for-proposal-to-establish-an-enterprise-resource-planning-%28erp%29-software-as-a-service-%28saas%29-panel/> [↑](#footnote-ref-3)
4. **Open Government Partnership Australia. (2018). 2.3 Digitization of Government Services**. Retrieved March 31, 2018, from <https://www.opengovpartnership.org/current-commitments/23-digitization-of-government-services> [↑](#footnote-ref-4)
5. **O’Neil, C. (2016*). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown/Archetype.** [↑](#footnote-ref-5)
6. Ibid p 28 [↑](#footnote-ref-6)
7. **Simon, J., & Rieder, G. (2015). Big Data, Corporate Governance, and the Limits of Algorithmic Accountability**. *Proceedings of ISIS Summit Vienna 2015—The Information Society at the Crossroads*. Retrieved from <https://www.academia.edu/22924819/Big_Data_Corporate_Governance_and_the_Limits_of_Algorithmic_Accountability> [↑](#footnote-ref-7)
8. **O'Neill** Op cit p 29 [↑](#footnote-ref-8)
9. **Ferguson, A. G. (2017). *The rise of big data policing: surveillance, race, and the future of law enforcement***. New York: New York University Press. [↑](#footnote-ref-9)
10. E.g. O'Neill pp 141 - 178 [↑](#footnote-ref-10)
11. **Wilcock, S. (2016). Policing Welfare: Risk, Gender and Criminality**. *International Journal for Crime, Justice and Social Democracy*, *5*(1), 113–130. <https://doi.org/10.5204/ijcjsd.v5i1.296> [↑](#footnote-ref-11)
12. O'Neill Op Cit. [↑](#footnote-ref-12)
13. Ibid p 30 [↑](#footnote-ref-13)
14. **Crawford, K., & Schultz, J. (2014). Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms**. *Boston College Law Review*, 55, 93.p 98 [↑](#footnote-ref-14)
15. **McIlroy, T. (2017, September 13). “Absolutely shocking”: 20,000 sent robo-debt notices found to owe less or nothing**. Retrieved March 30, 2018, from <http://www.canberratimes.com.au/national/public-service/20000-people-sent-centrelink-robodebt-notices-found-to-owe-less-or-nothing-20170912-gyg8mm.html> [↑](#footnote-ref-15)
16. **Parry, T. M. S. Y., & Ockenden, W. (2017, August 7). How CBA’s smart ATMs and a coding error caused a massive mistake** [Text]. Retrieved March 30, 2018, from <http://www.abc.net.au/news/2017-08-07/commonwealth-bank-how-smart-atms-and-coding-error-caused-mistake/8781066> [↑](#footnote-ref-16)
17. **Medcraft, G. (2017). *Digital disruption and data:  An ASIC update. A speech by Greg Medcraft, Chairman,  Australian Securities and Investments Commission.*** (p. 4). Sydney , Australia: FINSIA Regulator Panel. Retrieved from <http://download.asic.gov.au/media/4451299/greg-medcraft-finsia-published-8-september-2017.pdf> [↑](#footnote-ref-17)
18. **Mohamed** **Damak. (2018, February 19). The Future of Banking: Cryptocurrencies Will Need Some Rules to Change the Game**. Retrieved April 1, 2018, from <http://www.spglobal.com> [↑](#footnote-ref-18)
19. **Malmo, C. (2015, June 30). Bitcoin Is Unsustainable**. Retrieved April 1, 2018, from <https://motherboard.vice.com/en_us/article/ae3p7e/bitcoin-is-unsustainable> [↑](#footnote-ref-19)
20. **The Economist. (2015, October 31). The Trust Machine**. *The Economist*. Retrieved from <http://www.economist.com/node/21677198/print> [↑](#footnote-ref-20)
21. **Kasireddy, P. (2018, February 3). ELI5: What do we mean by “blockchains are trustless”?** Retrieved April 1, 2018, from <https://medium.com/@preethikasireddy/eli5-what-do-we-mean-by-blockchains-are-trustless-aa420635d5f6> [↑](#footnote-ref-21)
22. **The Economist. (2017, July 15). Disrupting the trust business**. *The Economist*. Retrieved from <https://www.economist.com/news/world-if/21724906-trust-business-little-noticed-huge-startups-deploying-blockchain-technology-threaten> [↑](#footnote-ref-22)
23. **Swan, M. (2015). *Blockchain: Blueprint for a New Economy***. O’Reilly Media, Inc. [↑](#footnote-ref-23)
24. **Turner, E., Zakrzewski, T., Stout, C., & Kingston, J. (2018, January 31). Cracking the Code of Privacy in the Blockchain.** Retrieved April 1, 2018, from <http://www.spglobal.com> [↑](#footnote-ref-24)
25. **Wright, A., & De Filippi, P. (2015). *Decentralized Blockchain Technology and the Rise of Lex Cryptographia*** (SSRN Scholarly Paper No. ID 2580664). Rochester, NY: Social Science Research Network. Retrieved from <https://papers.ssrn.com/abstract=2580664> p 15 [↑](#footnote-ref-25)
26. **Wright, A., & De Filippi, P. (2015)** p 21 [↑](#footnote-ref-26)
27. **Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate.** *Big Data & Society*, 3, 1–21 [↑](#footnote-ref-27)
28. **Selbst, A., & Barocas, S. (2017). *Regulating Inscrutable Systems*. Presented at the We Robot 2017, Yale Law School.** http://www.werobot2017.com/wp-content/uploads/2017/03/Selbst-and-Barocas-Regulating-Inscrutable-Systems-1.pdf [↑](#footnote-ref-28)
29. **Knight, W. (2017, April 11). The Dark Secret at the Heart of AI***.* MIT Technology Review. Retrieved from https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [↑](#footnote-ref-29)
30. **Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., … Hassabis, D. (2017). *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm*. ArXiv:1712.01815** [Cs]. Retrieved from http://arxiv.org/abs/1712.01815 [↑](#footnote-ref-30)
31. See also **Gibbs, S. (2017, December 7). AlphaZero AI beats champion chess program after teaching itself in four hours. *The Guardian***. Retrieved from http://www.theguardian.com/technology/2017/dec/07/alphazero-google-deepmind-ai-beats-champion-program-teaching-itself-to-play-four-hours [↑](#footnote-ref-31)
32. **Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability**. *Distill*, *3*(3), e10. <https://doi.org/10.23915/distill.00010> [↑](#footnote-ref-32)
33. **European General Data Protection Regulation. (2017). Home Page of EU GDPR.** Retrieved November 7, 2017, from <http://eugdpr.org/eugdpr.org.html> [↑](#footnote-ref-33)
34. **Wachter, S., Mittelstadt, B., & Floridi, L. (2016). *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*** (SSRN Scholarly Paper No. ID 2903469). Rochester, NY: Social Science Research Network. Retrieved from <https://papers.ssrn.com/abstract=2903469> [↑](#footnote-ref-34)
35. Ibid [↑](#footnote-ref-35)
36. Ibid p 4 [↑](#footnote-ref-36)
37. **Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). An Algorithm Audit**. In S. P. Gangadharan, V. Eubanks, & S. Barocas (Eds.), *Data and Discrimination: Collected Essays*. Open Technology Institute/New America. Retrieved from <https://na-production.s3.amazonaws.com/documents/data-and-discrimination.pdf> [↑](#footnote-ref-37)
38. **Ghonim, W., & Rashbass, J. (2017, October 31). Opinion | It’s time to end the secrecy and opacity of social media**. *Washington Post*. Retrieved from <https://www.washingtonpost.com/news/democracy-post/wp/2017/10/31/its-time-to-end-the-secrecy-and-opacity-of-social-media/> [↑](#footnote-ref-38)
39. **Gunning, D. (2016, August 11). Explainable Artificial Intelligence**. Retrieved April 2, 2018, from <https://www.darpa.mil/program/explainable-artificial-intelligence> [↑](#footnote-ref-39)
40. Ibid. [↑](#footnote-ref-40)
41. **Knight, W. (2017, April 11).** Op Cit**.** [↑](#footnote-ref-41)
42. **Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018**). Op Cit [↑](#footnote-ref-42)
43. **Genevieve Bell** talking on **Van Extel, C. (2018, March 27). Techtopia** [MP3]. *Big Ideas*. WOMAD 11 March 2018: ABC RN. Retrieved from <http://www.abc.net.au/radionational/programs/bigideas/techtopia/9592580> [↑](#footnote-ref-43)
44. **Metz, C. (2017, November 5). Building A.I. That Can Build A.I**. *The New York Times*. Retrieved from <https://www.nytimes.com/2017/11/05/technology/machine-learning-artificial-intelligence-ai.html> [↑](#footnote-ref-44)
45. **Li, F. F., & Li, J. (2018, January 17). Cloud AutoML: Making AI accessible to every business**. Retrieved March 30, 2018, from <https://www.blog.google/topics/google-cloud/cloud-automl-making-ai-accessible-every-business/> [↑](#footnote-ref-45)
46. **Metz, C. (2017, November 5). Building A.I. That Can Build A.I**. *The New York Times*. Retrieved from <https://www.nytimes.com/2017/11/05/technology/machine-learning-artificial-intelligence-ai.html> [↑](#footnote-ref-46)
47. **Kelion, L. (2015, February 16). *Could driverless cars own themselves?*** BBC News. Retrieved from <http://www.bbc.com/news/technology-30998361> [↑](#footnote-ref-47)
48. **Belk, R. (2017). *Things Owning Things***. In J. F. Sherry & E. M. Fischer (Eds.), Contemporary Consumer Culture Theory. Taylor & Francis [↑](#footnote-ref-48)
49. **Wright, A., & De Filippi, P. (2015)** Op Cit pp 14-15 [↑](#footnote-ref-49)
50. **Open Government Partnership - Australia. (2018, March). 2.2 - Build and maintain public trust to address concerns about data sharing and release**. Retrieved March 31, 2018, from <https://ogpau.pmc.gov.au/commitment/22-build-and-maintain-public-trust-address-concerns-about-data-sharing-and-release> [↑](#footnote-ref-50)
51. **Culnane, C. (2017, August). *De-identification – what is it good for…*. Unpublished speech for Fabians Victoria Event “Digitisation of government data: democratic benefits and possible costs”** 30/08/2017, Melbourne. Retrieved from [OpenGovData.pdf](file:///C:\Users\Julia\Dropbox\Documents\PROJECTS%20-%20COMMUNITY\ACCOUNTABILITY%20ROUNDTABLE\Open%20Government%20Partnership\2018\OpenGovData.pdf) (Video at https://vimeo.com/231940675) [↑](#footnote-ref-51)
52. **Australian Law Reform Commission. (2010). *Exemption for registered political parties, political acts and practices*** (Australian Privacy Law and Practice No. ALRC Report 108). Australian Law Reform Commission. Retrieved from <https://www.alrc.gov.au/publications/41.%20Political%20Exemption/exemption-registered-political-parties-political-acts-and-pract> [↑](#footnote-ref-52)
53. **Munro, K. (2018, March 21). Australia’s major parties defend privacy exemption over Cambridge Analytica.** Retrieved March 31, 2018, from <http://www.theguardian.com/australia-news/2018/mar/22/australias-political-parties-defend-privacy-exemption-in-wake-of-cambridge-analytica> [↑](#footnote-ref-53)
54. **Vaile, D. (2018, March 22). Australia should strengthen its privacy laws and remove exemptions for politicians.** Retrieved March 31, 2018, from <http://theconversation.com/australia-should-strengthen-its-privacy-laws-and-remove-exemptions-for-politicians-93717> [↑](#footnote-ref-54)
55. **Open Government Partnership Australia. (2018). 2.3 Digitization of Government Services.** Op Cit [↑](#footnote-ref-55)
56. **Australian Government. (2018, April). Digital Transformation Agency**. Retrieved April 2, 2018, from <https://www.dta.gov.au/> [↑](#footnote-ref-56)
57. **MacDermid, T. (2017, November 2). Can government really keep data safe across agencies?** *The Mandarin*. Retrieved from <https://www.themandarin.com.au/85560-can-government-really-keep-data-safe-across-agencies/> [↑](#footnote-ref-57)
58. From a talk given by Dr Vanessa Teague of University of Melbourne at the Office of the Victorian Information Commissioner, on 7th March 2018. [↑](#footnote-ref-58)
59. Office **of the Australian Information Commissioner (OAIC). (2018, February 19). Mandatory data breach notification comes into force this Thursday**. Retrieved April 2, 2018, from [/media-and-speeches/media-releases/mandatory-data-breach-notification-comes-into-force-this-thursday](file:///C:\media-and-speeches\media-releases\mandatory-data-breach-notification-comes-into-force-this-thursday) [↑](#footnote-ref-59)
60. **Funnell, A. (2017, October 4). Perspectives on Privacy**. *Radio National*. ABC RN. Retrieved from <http://www.abc.net.au/radionational/programs/futuretense/perspectives-on-privacy/9012328> [↑](#footnote-ref-60)