Demonstrably doing AI accountability by design: problems, practical processes and tools

Peter Leonard[1]

Collection, use, linkage and sharing of data sets for advanced data analytics, AL/ML applications and automated decision-making is creating novel opportunities and challenges for businesses, government agencies and other organisations.

Well-developed frameworks for assessing impact upon data privacy of identifiable individuals, and for ethics review of research projects, do not bring within scope a number of key emerging concerns.

In particular, many organisations are not yet using frameworks, tools and methodologies to embed responsible governance of decisions as to whether and how to use data in circumstances:

- that are not research projects subject to mandatory ethics review, and
- that fall outside the ambit of privacy impact assessment (because there is no relevant use or disclosure of personally identifying information about individuals ether identified or identifiable, from the data itself or through combination of that data with other available data);
- where the proposed uses of the data to create outputs, or the outcomes to be effected from use of those outputs, will not erode digital trust of affected individuals. Digital trust is at significant risk where outputs and outcomes are not transparent, explainable, explained, socially beneficial, and not significantly adverse to the affected individual.

A key issue is how to facilitate, while appropriately controlling, use of individual level data for data analytics and AI/ML applications to create useful outputs that can then be used to effect socially beneficial outcomes, without undermining trust of affected individuals.

That statement of the key issue should be entirely uncontroversial. What is often not recognised, however, is that addressing this statement requires transparency and accountability by design.

Over the last decade, data professionals have (or should have) become adept at processes and tools to implement the principle of protection of personal information about individuals

[1] Peter Leonard is a data, content and technology business consultant and lawyer advising data-driven business and government agencies. Peter is principal of Data Synergies and a Professor of Practice at UNSW Business School (IT Systems and Management, and Business and Taxation Law). Peter chairs the IoTAA's Data Access, Use and Privacy work stream, the Law Society of New South Wales' Privacy and Data Committee and the Australian Computer Society's Artificial Intelligence and Ethics Technical Committee. He serves on a number of corporate and advisory boards, including of the NSW Data Analytics Centre.

by design and by default. Privacy and information security protective governance frameworks and processes should now be relatively well understood.

By contrast, the principle of transparency and accountability by design is new. We haven't yet developed standards, or industry accepted best practice, processes and tools for building transparency and accountability by design.

Yet, and seemingly suddenly, the question of how to give effect to this principle has become both time critical and reputationally sensitive.

Why is this so?

Firstly, we haven't properly addressed the tectonic shift that was, and is, the smartphone. More about that later.

Secondly, we've been myopic in our focus upon protection of personal information and appropriate application of existing data privacy laws, rather than considering how to address the growing deficit in digital trust of citizens.

We've been focussed on important issues, but not all of the important issues.

Third, we have spent far too much time developing statements of ethical principles for artificial intelligence.

Statement of principles do not cause or enable organisations to do things differently.

The right incentives and sanctions cause organisations to do things differently.

The right methodologies, tools and processes enable organisations to do things differently.

AI is only one of the things that need to be done well.

Things - what that we do, when, how, in what way - that need to be done well, need to be done much better in relation to many things.

AI is not some 'thing' that is fundamentally different. AI just another application of advanced data analytics - albeit that sometimes, in some manifestations, AI presents further tricky issues, in particular by taking humans out-of-the-loop, or by being a black box with no explainability. More often, AI presents similar practical issues as to transparency and accountability by design issues as the issues that we should already be grappling with in relation to here and now digital applications and technologies, including:

- smartphones,
- personal wellness devices and other health IoT applications,
- smart homes, smart infrastructure and other IoT applications where the affected individual may have no idea that their activities are being observed or sensed;
- the MyHealth Record system, and
- many health data linkage and sharing initiatives.

A real danger in identifying and addressing AI risks is that we think that AI is fundamentally different to, or smarter than, other data driven automation, causing us to then fall victim to the common viruses, AI aversion, or its evil twin, AI deference.

The best way to avoid these afflictions is to be really clear about what AI risks we are addressing: specifically, uses and applications of data relating to humans in ways that affect those humans, or other humans, or our environment.

In other words, addressing how health data is captured and used is critical to ensure digital trust of citizens in relation to a wide range of health technology applications.

Addressing AI by focusing upon governance of data inputs, outputs and outcomes is not, of course, to assert that data governance is the only issue that AI raises. As so powerfully stated by our former Prime Minister Julia Gillard in relation to Australia's gender bias issue, "It doesn't explain everything, it doesn't explain nothing, it explains some things".

Data issues can't be understood without considering digital trust.

The growing deficit in digital trust of citizens

Digital trust of device users, of consumers and of citizens, is being eroded in myriad ways. Causes of erosion of digital trust include:

- the Facebook - Cambridge Analytica imbroglio, which exacerbated concerns to uncontrolled and disturbing secondary uses of data, in turn fuelled by (for fuelling) 'fake news' and other public controversies as to data misuse and abuse;
- growing concerns as to pervasive and unexplained surveillance;
- many custodians adopting a reflexive and default 'trust us, we know best' approach. Sometimes this reflexive responsibility is reliably implemented in practice, but 'we know best' often leads to hubris - over-confidence and under-explanation;
- 'Robodebt' and like controversies as to automated, algorithmically driven actions by both government agencies and other organisations. When automated decisions-making is perceived by a significant number of media commentators and other influencers and citizens as not sufficiently reliable, this concern undermines trust more generally as to data driven decision making, also fuelling (or fuelled by) existential fears as to unaccountable artificial intelligence, killer drones, rampant robots and incipient singularity;
- loss of trust in non-financial governance and ethical oversight of major institutions generally, particularly driven in Australia by findings of recent Royal Commissions as to failures of ethical governance in major organisations that should know better, and other reported misbehaviour by major organisations;
- continuing alerts as to cybersecurity and cyber-risk, creating fears that data custodians cannot be relied upon to reliably control, curate and secure data about individuals, by design and by default.

Notice and consent-based data privacy models are under severe strain, and increasingly recognised as such by users. We are gamed by some service providers to not read terms of use and privacy policies, and we are invited to click-through 'consent' to terms that are unintelligible to many readers. Almost all users take the Faustian bargain of giving access to data about us in exchange for free or subsidised services, and click straight through while hoping and praying for the best, then maybe go back and look at the privacy settings if we see a sufficiently disturbing report of something going awry. There are no widely accepted smartphone applications for individuals to conveniently give (or withhold) granular, use case specific (and therefore much simpler to understand) consents to specific uses and disclosures of data about them.

Would digital trust of citizens be nurtured by creating mechanisms for granular control by individuals of how and when data relating to them and their activities is applied the benefit of others? Quite possibly: we haven't really tried, and therefore don't know. In the meantime, the currency created by broad form digital consents is over-used by some service providers, and has therefore becomes debased.

Unless we focus upon this digital trust deficit, we will not understand why accepted views of good policy making and of independent ethics review don't cut it anymore.

Trust is not built through reassurance of trustworthiness.

Trust is no longer assured by good intentions, good status or blemish free history, or conduct of a privacy impact assessment.

Our most trusted Australian institution, the Australian Bureau of Statistics, discovered this new truism with so-called 'Census fail'.

The Federal Department of Health discovered a deficit in citizen trust with its policy switch from opt-in to opt-out for MyHealth Record.

The Department of Health rediscovered it again in relation to assurances that ADHA would use administrative practices to address concerns about default settings on access to MyHealth record data by third parties, and that therefore legislative protections were not required.

Facebook discovered that global billboarding across the globe about how Facebook took trust seriously did not stem fallout from Cambridge Analytica.

And so on, and on.

Many citizens now express views that it is no longer good enough for an organisation to make public statements:

- that the organisation takes user concerns or consumer concerns seriously (and for the organisation to then address concerns in some unexplained way behind closed doors),

- that the organisation adheres to principles of social responsibility (without also describing how the organisation ensures consistent and reliable implementation of these principles, and why the organisation should be trusted to do what it says),
- that there are distinguished, responsible and ethical individuals on the organisation's board, on an ethics advisory board, or otherwise working with them and overseeing what they do (where those individuals do not appear to be being deeply embedded in the culture of the organisation, or appear to be using unstructured or unsystematic reasoning as to what is good or bad that may not have traction within the organisation),
- that they use digital technologies and data for the benefit of society (without also explaining how any inherent detriments are managed so that this benefit does not come at significant cost to some individuals within society, or to environment detriment).

Digital trust requires transparency and accountability by design and default, in relation to many uses and applications of health-related data.

Transparency and accountability by design and default is particularly required in relation to uses of individual level health related data in transactor key (deidentification) based data analytics to create outputs and to effect outcomes that are targeted, differentiated and potentially discriminatory.

Good outputs properly curated and managed should not lead to bad outcomes.

Outcomes that are targeted to particular segments of individuals, including highly granular segments, need not be discriminatory, but they can be, and these discriminatory outcomes can be effected without knowing the identity of the affected individuals, and therefore entirely outside the existing scope of Australian data privacy laws.

Of course, this is not a new issue: the decades-old National Guidelines[2] address this very issue.

Many privacy advocates, and some privacy regulators, say: the solution is clear:

- properly inform affected individuals, and then ask for their consent, and,
- if the existing regulatory definition of research is too narrow, expand the ambit of projects that must be subject to mandatory ethics review. If HRECs works for research, and citizens accept independent research ethics review as an appropriate control mechanism, let us expand the range of project that are subject to ethics review, and extend their coverage to include commercial applications of individual level data for data analytics.

I don't agree.

[2] Guidelines under Section 95 of the Privacy Act 1988 as issued by the National Health and Medical Research Council.

Accepted Australian models for ethics review of research projects are cumbersome and slow. They are also unduly repetitive, partly because there is limited transparency, and therefore reuse, of past reasoning as to like projects, and partly because linkage data sets are often created on a project specific basis, also with no, or limited, reuse.

Entirely outside the ambit of not research projects subject to mandatory ethics review, collection and secondary uses of health-related data is exploding, largely through rapid expansion in the number of devices collecting health related data, increased standardisation of data formats and integration of data across multiple devices intermediated by the smartphone. More data is more readily available, on and across the data clouds, and more discoverable, at low cost. This explosion in devices, data points and cloud is occurring while at the same time there is a continuing decline in trust of affected individuals and other citizens as to how many organisations collect, curate and share digital data about individuals.

Inferences from this data can be used to differentiate between affected individuals, whether or not identified or identifiable, in whether a product or service is offered to them and as to the price and non-price terms at which a product or service is offered.

As algorithms are refined through experience, with more data and through machine learning, these inferences become more likely to be correct, and also less explainable.

Addressing the trust deficit

There are real and significant risks inherent in many applications of digital heath-related data, and not just advanced AI/ML applications.

Organisations need to know how to recognise, assess, mitigate and manage these risks, consistently and reliably, while not jumping at shadows.

Organisations need to ensure that applications of data drive decision making and AI/ML:

- comply with laws, including by being reliably safe and fit for purpose,
- do not undermine fundamental human rights. These rights include a right of no unfair discrimination and for individuals to be free to go about our private lives, in public and online, without being pervasively observed,
- reflect good design principles of data privacy by design and default, and of data security by design, accessibility (by individuals with special needs) by design, and environmental responsibility by design;
- reflect emerging standards and expectations as to fair, explainable and ethical, accountable and transparent uses of data.

Organisations need to weigh up benefits and detriments of applications of AI/ML, to ensure that benefits to broader society do not come at the cost of significantly adverse outcomes for a minority.

In considering risks of applications of AI/ML, organisations need to understand and apply the precautionary principle, being that when human activities may lead to unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that harm.

Making it real

Risks of AI/ML applications have been addressed by many statements of guidelines and principles for ethical AI, or good AI. Over the last year the inventory of guidelines and principles has grown to over 85. This inventory continues to rapidly grow.

Most statements elaborate four basic principles, often reduced to the acronyms FEAT or FATE: fairness (to affected individuals), ethics (including social equity and social beneficence), accountability and transparency (including by being explainable).

These guidelines and principles state in various ways what organisations should do or not do.

Only a few statements endeavour to elaborate on how organisations can reliably and verifiably do what is advocated to be done.

The less explored question is how to achieve good AI, not the now well traversed question of what is good AI. Accordingly, our focus should be to:

- to address the deficiency of practical operational guidance upon how organisations can reliably and verifiably assure good AI, through good AI/ML governance,
- to identify methodologies and tools that are already available and which we think deserve consideration for adoption by a wide range of organisations developing, deploying or implementing applications of AI/ML;
- to assist alignment of processes and tools for ensuring good or ethical AI with other risk and project management frameworks already used in many organisations, and
- to identify gaps in coverage of methodologies and tools that we believe need to be filled.

Good AI/ML governance requires organisations to apply a systematic framework to rigorously, reliably and verifiably, identify, evaluate, mitigate and manage risks arising from particular uses and applications of AI/ML, including through application of appropriate tools, methodologies, controls and safeguards.

Frameworks, methodologies, processes and tools for AI/ML review should be designed to be closely aligned with frameworks, methodologies and processes already implemented within organisations, to minimise duplication and cost and ensure that gaps do not arise through incorrect assumptions as to coverage.

Governance includes how AI initiatives are supervised, tracked and managed by boards and senior managers within organisations, and not good management of data flows associated with AI/ML applications.

Governance should enable AI/ML teams to assess what they do, or shouldn't do, and enable others to assess whether AI/ML teams are properly discharging this responsibility.

Essential elements of good governance include:

- oversight within an organisation,
- accountability of AI/ML teams within the organisations in which they work, and
- accountability of organisations developing, deploying and implementing of applications of AI/ML to persons outside the organisation that are affected by an organisation's activities, including adverse effects that arise through failure to assess, mitigate and manage residual risks of to humans, other lifeforms and the environment, of applications of AI/ML.

To address the deficit of trust of citizens in data driven processes, and ensure good governance of AI/ML, the right questions need to be posed and addressed by organisations developing or deploying AI/ML applications which may have significant effects outside an organisation. Answers to the right questions must then be fed back into AI/ML development or deployment decisions, at the right time. Feedback loops need to be established.

Good governance is not easy. Organisations need to ensure that they get quite a number of things right:

- building awareness within the organisation as to the need for good governance, as to what good governance looks like, and as to the consequences of failing to implement good governance,
- building capabilities within the organisation to consistently assure good governance,
- adoption of good principles and common purpose – an organisation must know what it is striving to achieve, and avoid,
- good culture, including relevant stakeholders within the organization caring about whether outcomes are good, and being empowered to achieve good outcomes and call out unacceptable behaviours and practices,
- revising incentives structures to reward good outcomes,
- revising lines of accountability to not encourage bad outcomes,
- good processes, so that good governance is reliably assured, and
- post-implementation checks and feedback loops, failsafes, and remediation mechanisms and procedures, so when things do go wrong, they are quickly identified and fixed.

In order to ensure that whatever approach to good AI is responsibly and reliably applied within relevant organisations, there needs to be processes that assure:

- a clear framing of the 'threshold' or 'gating' condition'- that is, when and how the threshold circumstances and level of risk are determined to be such as to require application of the relevant methodology;

- a good process for assessment of likelihood and magnitude of risks, for assessing appropriate mitigation strategies and steps, and for managing any residual risks that remain after those mitigation strategies and steps are implemented, including addressing contingencies such as when unforeseen circumstances arise or mitigation measures fail or are circumvented;
- a clear description of the 'handling conditions' (once within the gate) – that is, the various controls and safeguards, including mechanisms for oversight and accountability, that constitute an appropriate AI governance system;
- a clear framing of the 'ungating condition' – that is, when and how an output from an AI governance system is determined to be safe to effect a particular outcome in a particular context, and how the output is controlled and managed to mitigate risks that it is used to effect outcomes that are not determined to be safe.

Are we having the right debate?

There is significant policy focus upon the adequacy or otherwise of Australian data privacy laws. Some of this focus starts from the incorrect premise that notching up Australian privacy laws towards the misdescribed GDPR 'gold standard' will address what are, in reality, a much broader set of concerns.

A GDPR-like regulated requirement of unambiguous, express, fully informed consent of affected individuals as to collection and uses of data about them is a laudable ideal.

However, notching up consent requirements in many cases merely shift responsibility from service provider to user to make choices that users are not likely to make carefully or well – or worse, allows service providers to assert that it is not the provider's responsibility to work out what uses of data are far and responsible, and what are not.

This is not to disparage the importance of the global debate about appropriate scope and coverage of data privacy laws, and as to framing of further data privacy limitations upon secondary uses of data collected through use of digital technologies. But lets get real about what data privacy laws do, and can do.

How much can data privacy laws really do?

Discoverability of data relating to individuals may be created within a privacy protected data analytics environment. In many cases, substantial data value can be created and commercialised, without particular individuals becoming identifiable. Through pseudonymisation of identifiers and deployment of appropriate controls and safeguards protecting data analytics environments, many uses of user data need not be privacy invasive.

Of course, it is easier to link disparate data sets using personal identifiers than it is to deploy a properly isolated and safeguarded data analytics environment that uses only pseudonymised data linkage transactor keys. It is also easier to release outputs and insights

without taking reliable steps to ensure that the outputs cannot be used to re-identify affected individuals.

In short, data privacy governance is exacting. And frameworks, tools and methodologies for good data governance are immature and are therefore not well understood.

Good data handling on its own does not create good outputs. Executives of organisations often do not know how to evaluate the quality of their data scientist units and the reliability of data science outputs and insights. The term 'data science' carries, as the term 'management science' once did, an enticing ring of exactitude. However, algorithms however skilfully derived and applied may be based on poor data, or simply misapplied when used in particular contexts. Often poor data practices are implemented inadvertently, or as a result of cutting corners, rather than through bad intent.

Framing a discussion with affected individuals

As already noted, most user data is generated in circumstances where the relevant humans no longer understand or control the 'data exhaust' associated with their activities or transactions. Where users are unknowing creators of data exhaust, those users are particularly vulnerability to data uses that may be adverse to their interests. A simple example: I don't choose to be observed by my very smart rental car, but I am. When I drive it out of the parking slot, I don't reach for the vehicle manual to school up on the car's data analytics capabilities. Even when I am presented with terms explaining particular data uses, life is too short for me to read and evaluate the terms. Another example: that CCTV over there is capturing my image and applying facial recognition to work out who I am and whether someone, somewhere, is happy with what I am doing. And so on. Frequently data about me is used in ways for which I do not knowingly and reflectively give consent.

Sometimes I am asked, and I click 'I agree' without reading all the terms, Should recalcitrant consumers (such as me), who don't read all terms proffered to us, be punished for our unwillingness to engage with the torrent of privacy disclosures by organisations with whom we deal?

And even if I don't care at all about my data privacy, I might still want to join ranks with many millennials and demand to know who is doing what, and deriving how much value, using data about me. Many millennials do not care about privacy or transparency by right, but sense that value is being derived from data about them, that free services are great, but that no-cost may be less than fair value, and that they are not given enough information about what is going on to force a meaningful negotiation over fair allocation of data value.

Many businesses are reluctant to initiate a discussion as to what is fair to consumers, because they can't control that discussion, or they simply don't want to give away value.

Some early mover data platform businesses captured the data high ground and since then have engaged in tactical retreats, giving away certain data value if and when required to mitigate particular crises in digital trust of consumers.

Many other data driven organisations, such as some insurers and banks, are more willing to sacrifice short term data value in order to preserve longer term certainty and therefore sustainability for data value-adding investments. However, they are concerned that initiating a discussion with customers as to fair data exchange can lead to unpredictable and uncontrollable outcomes: explanations of many data applications and data value chains are devilishly tricky and can sound self-serving, or just plain creepy. Try explaining to sceptical citizens and consumer advocates how real time targeted advertising does not require any disclosure of a mobile user's identity to the advertiser or its media buyer, or explaining how audience segmentation value is allocated at points in the complex and multi-party advertising and publisher supply chain. Emerging expectations of consumers as to transparency and accountability in handling of data about them may lead to an imperative for a provider to restrict data flows within a multi-entity data ecosystem, while at the same time regulators seek to force opening up of supply-side data ecosystems to new data intermediaries.

Leaving aside the desire for greater transparency as to data uses to facilitate consideration of fair value of data, why should a consumer need to engage with a data collector as to whether a particular collection of data is a fair, proportionate and reasonable?

Regulators don't require consumers to take responsibility for determining whether a consumer product is fit for purpose and safe when used for the product's stated purpose, and unsuitable or unsafe when used for other purposes. Why should data driven services be any different?

A smartphone user may not want transparency and responsibility forced upon them, so that they then must make a sensible decision (or just click-through). Instead, a smartphone user may want accountability of the data controller, to ensure that the data controller responsibly and reliably does what is fair.

However, fairness is a notoriously normative concept, which is why competition law seeks exactitude of economic theory in evaluating effects on consumer welfare. Beneficence for the majority of consumers also results in less than 'fair' treatment of a few - at least, as those few perceive their treatment relative to treatment of the majority. It all turns on the particular context.

This isn't easy. But good data governance, and reliable and consistent application of tests as to what is fair, and responsible and socially beneficial, while also being reasonably equitable, will be key differentiators of organisations that build substantiable business models from those many other organisations that will fade away as they erode digital trust, causing damage along the way. Although implementation of good data governance and responsible AI is not easy, the choice should be.

Peter Leonard
16 October 2019