The Good, the Bad, and the Ugly: A companion to the AI Forum NZ's Research Report '

6-May-2018
Matt Boyd – Adapt Research Ltd
matt@adaptresearchwriting.com

Last week the AI forum New Zealand released its report 'Artificial Intelligence: Shaping a future New Zealand'. In what follows I wish to commend the authors for an excellent piece of horizon scanning, which lays the foundation for a much-needed ongoing discussion about AI and New Zealand, because, like the Wild West there is much as yet unknown regarding AI. Microsoft was at pains to point this out in their 'The Future Computed' report published earlier this year. In what follows I comment on some of the content of the AI Forum NZ's report and also try to progress the discussion by highlighting areas that warrant further analysis. Like all futurism we can find the good the bad and the ugly within the report.

The Good

The report has done a thorough job of highlighting many of the opportunities and challenges that face us all in the coming years. It is a necessary and very readable roadmap for how we might approach the issue of AI and New Zealand society. The fact the report is so accessible will no doubt be a catalyst to meaningful debate.

> A 'flash crash' event involving autonomous weapons is not something we could simply trade out of a few minutes later.

It was good to see insightful comments from Barrie Sheers (Managing Director, Microsoft NZ) at the outset, which set the tone for what was at times was (necessarily) a whistle-stop tour of the web of issues AI poses. Barrie's comments were nuanced and important, noting that those who design these technologies are not necessarily those who ought to decide how we use them. This is a key concept, which I will expand on below.

The report is generally upbeat about the potential of AI and gives us many interesting case studies. However, the 'likely', 'many', benefits of AI certainly do not give us carte blanche to pursue (or approve) any and all applications. We need a measured (though somewhat urgent) approach. Similarly, there is omission of some of the key threats that AI poses. For example, AI is suggested as a solution to problem gambling (p. 76), yet AI can also be used to track and persuade problem gamblers online, luring them back to gambling sites. For every potential benefit there is a flip side. AI is a tool for augmenting human ingenuity, and we must constantly be aware of the ways it could augment nefarious intentions.

It was good to see the report highlight the threat of autonomous weapons and the fact that New Zealand still has no clear position on this. We need to campaign forcefully against such weapons as we did with the issue of nuclear

weapons. The reason for this is that in 2010 financial algorithms caused a $1 trillion dollar flash crash of the US stock market. Subsequent analysis has not satisfactorily revealed the reason for this anomaly. A ‘flash crash’ event involving autonomous weapons is not something we could simply trade out of a few minutes later.

The issue of risk and response lies at the heart of any thinking about the future of AI. One of the six key recommendation themes in the report centers on ‘Law, Ethics and Society’. There is a recommendation to institute an AI Ethics and Society Working Group. This is absolutely critical, and its terms of reference need to provide for a body that persists in its place for the foreseeable future. This working group needs to be tasked with establishing our values as a society, and these values need to shape the emergence of AI. Society as a whole needs to decide how we use AI tools and what constraints we place on development.

Ultimately, there probably ought to be a Committee for AI Monitoring, which distills evidence and research emerging locally and from around the world to quickly identify key milestones in AI development, and applications that pose a potential threat to the values of New Zealanders. This Committee probably ought to be independent of the Tech Industry, given Barrie Sheers comments above. Such a Committee would act as an ongoing AI fire alarm, a critical piece of infrastructure in the safe development of AI, as I discuss further below.

The Bad

Before I begin with the bad, I am at pains to emphasise that ‘Shaping a Future New Zealand’ is an excellent report, which covers a vast array of concepts and ideas, posing many important questions for debate. It is the quality of the report that draws me to respond and engage to further this important debate.

A key question this report poses is whether we will shape or be shaped by the emergence of AI. A key phrase that appears repeatedly in the document is ‘an ethical approach’. These two ideas together make me think that the order of material in the report is backwards in an important way. Re-reading Microsoft’s ‘The Future Computed’ report yesterday made me certain of this.

It may seem trivial, but in the AI Forum’s report, the section on ‘AI and the Economy’ precedes the section on ‘AI and Society’. This is to put the horse before the cart. Society gets to decide what we value economically, and also gets to decide what economic benefits we are willing to forgo in order to protect core values. We (society) get to shape the future, if we are willing and engaged. It is the societal and moral dimension of this issue that can determine what happens with AI and the economy. If we want to ‘shape’ rather than ‘be shaped’ then this is the message we need to be pushing. For this reason I think it is a mistake to give AI and the Economy precedence in the text.

A feature of the writing in this report is the abundance of definite constructions. These are constructions of the form ‘in the future X will be the case’. This is perhaps dangerous territory when we are predicting a dynamic,

exponential system. Looking to Microsoft's approach the phrase 'no crystal ball' stands out instead.

I'll digress briefly to explain why this point is so critical. Rapidly developing systems change dramatically in ways that it is not easy for our psychology to grasp. Say you have a jar containing a bacterium (let the bacterium represent technical advances in AI, or the degree to which AI permeates every aspect of our world, or the number of malicious uses of AI, or some such thing). If the bacteria doubles in number every minute, and fills the jar after an hour, then by the time the jar is a quarter full (you're really starting to notice it now, and perhaps are predicting what might happen in the future) you only have 2 minutes left to find another jar, and 2 minutes after that you'll need 3 more jars. In the classic Hanson-Yudowsky debate about the pace of AI advance, what I've just illustrated represents the 'AI-FOOM' (rapid intelligence explosion) position. This is a live possibility. The future could still look very different from any or all of our models and predictions.

Furthermore, a disproportionate portion of the AI and the Economy section focuses on the issue of mass unemployment. This is the 'robots will take our jobs' debate. The argument here is protracted, far more detailed than any other argument in the document, and the conclusion is very strong. I think this is a mistake. Straining through models and analyses of spurious accuracy to reach an unambiguous conclusion that 'AI will not lead to mass unemployment' appears to be predetermined. The length of the reasoning (certainly compared to all other sections) conveys the illusion of certainty.

But we're talking here about a tremendous number of unknowns. Including very many of Donald Rumsfeld's infamous 'unknown unknowns', the things we don't even know we don't know yet. The modeling projects 20 years through this indeterminate melee and it is hard to accept such a definite conclusion (I know as much from looking at what past labour market models have predicted and what actually transpired). Prediction is hard, especially about the future. This is why trader, risk analyst and statistician Nassim Taleb encourages us to anticipate *anything*. The history of the world is shaped by Black Swans. These are unpredictable events that we rationalize after the fact, but which change everything. The only response to such uncertainty is to build resilience.

I'm not saying that there will be mass unemployment, I'm saying that trying to prove one way or the other is a risky approach. What I am saying is that the conclusion is misplaced, as risk analysts we ought not burn bridges like this. Let's call a spade a spade. To me the argument in 'Shaping a Future New Zelaand' appears to be a rhetorical device put forward by those who don't want us to contemplate massive labour force disruption. If people are afraid for their jobs, they are less likely to authorize AI (and given the moral precedence of society over economy *authorize* is the correct term).

But to take this argument even further, what is the reason that we fear mass unemployment? It's not because of mass unemployment per se, it's because unemployment can deny people meaningful activity in their life, and it can also cause economic pain. However, mass unemployment is only one way to

cause these things. We should also be considering, for example, other ways that AI might deny us meaningful activity (with mass automation of decisions) or cause economic harm (through financial market collapse following an algorithmic mishap – financial error or financial terror) and so on. Mass unemployment is a side-show to the real discussion around value, meaning and risk that we need to be having.

If we don't plan for what to do in the event, then we have built a fragile system

By concluding that there is no risk, nothing to worry about, we risk being caught off-guard. A safer conclusion, and one that provides in fact much more security for everyone, is one that is reached without analysis. Maybe AI leads to mass unemployment, maybe it doesn't. The problem is that if we don't plan for what to do in the event, then we have built a fragile system (to use Taleb's term).

By accepting at least the possibility of mass unemployment, we can invest in resilience measures, pre-empt any crisis, and plan to cope. We put that plan into action if and when the triggering events transpire. What we need is an insurance policy, not to hide our head in the sand. What we need is a fire alarm. That would be the way to allay fears. That would be how to ensure the system is antifragile.

Society must precede Economy

Given the pace of AI innovation and surprising advances, we don't know how New Zealand will be affected by AI, but we can control what we are comfortable permitting. This is why Society must precede Economy.

In fact this has been a weakness of much contemporary political reasoning. Problems are tackled on an ad hoc basis, to determine how they might economically benefit us. What is lacking is a set of overarching values that we hold as a society and that we apply to each problem to determine how we must respond (whether or not it accords with our best economic interests). Max Harris tackles this issue in his recent book 'The New Zealand Project'.

So I return to the phrase, 'an ethical approach' which is the main theme of this report that needs unpacking. We need to decide as a society what our ethical approach is. We need a framework, which will determine whether each instance of AI is good, bad or ugly.

I'll turn to a concrete example. If I'm being critical (which I am in the interests of really pushing this debate deeper) there are some important omissions from the report.

Notably, very little mention is made of the advertising and communications industry. This is surprising given recent developments with fake news, the Cambridge Analytica saga and the associated Facebook debacle. All of which are merely the tip of the iceberg of an industry that has already taken advantage of the fact that the public is generally ill-informed about the uses

and possibilities of AI. Marketing is turning into manipulation. Attempts are being made to manipulate citizens to behave in ways that exploit them.

With Cambridge Analytica, Marketing is turning into manipulation

It's debatable to what degree these techniques have succeeded to date, but remember that bacteria has only been growing in the jar for 58 minutes so far, so the tools are rudimentary (to stick with our analogy, the tools employed by Cambridge Analytica were only one quarter effective, in 4 minutes we face tools with eight times that effect! – look at AlphaGo Zero and think about how the relatively rule-based human social network game might be learned, and what the intentions might be of those who control that technology)

The point is that we are facing a situation where we humans, who possess a psychology riddled with unconscious heuristics and biases, and are simply not rational, no matter how much we rationalize our behavior, are faced with AI systems that on the one hand are dreadfully incompetent compared to ourselves, and yet on the other hand have immense knowledge of us and our tendencies. This latter feature means there is potential for a power imbalance in these interactions and we are the victims. This is the fundamental premise of the industry of nudging. Which when deployed with less than altruistic goals we can plainly call manipulation.

The AI Forum report contains very little on manipulation and disinformation by AI, or the potential horror scenarios of AI impersonating (convincingly) powerful human individuals. We are going to need to solve the problem of trust and authenticity very quickly, and more importantly, to start to condemn attempts to impersonate and mislead.

We need more discussion about the degree to which we ought to require AI systems with which we interact to disclose their goals to us. Is this system's goal to make me buy a product? To stop me changing banks? To make me vote for Trump? To maximize the amount I spend online gambling? Perhaps we need regulation that makes AI developers ensure that AIs must declare that they are AIs.

The reason for this is because humans have evolved a very effective suite of defenses against being swindled by humans, but we are unfamiliar with the emerging techniques of AI. Unlike when I deal with a human, I'm unfamiliar with the knowledge and techniques of my potential manipulator. Private interests are going to flock to manipulation tools that allow them to further their interests.

There is one line in the report addressing this issue of manipulation by AI, but it is an important line. The Institute of Electrical and Electronics Engineers is in the process of drafting an engineering standard about ethical nudging. This certainly gets to the heart of this issue, but it remains to be seen what that standard is, what kinds of systems it covers, and who will adopt it. We could have done with such a standard before Cambridge Analytica, but we still need ways to make businesses adhere to it. New Zealand needs to be having values-based discussions about this kind of thing, and we need to be

monitoring overseas developments so that we have a say, and do not get dragged along by someone else's standards.

The Ugly

The report does a good job of laying out the strategies other nations are employing to maximize the probability of good AI outcomes. These case studies certainly make New Zealand look late to the party. However, there is no discussion of what is ultimately needed, which is a global body. We need an internationally coordinated strategy of risk management. This will be essential if nations do not want to be at the receiving end of AI use that they do not condone themselves. This is a coordination problem. We need to approach this from a values and rights perspective, and New Zealand has some successful history of lobbying the globe on issues like this.

The report highlights some potential threats to society, such as bias, transparency, and accountability issues. However, there are many further risks such as those that exploit surveillance capitalism, or threaten autonomy. Given that there are potential looming threats from AI, to individuals open to exploitation, to democratic elections from attempts at societal manipulation, to personal safety from autonomous agents, and so on, what we need is more than just a working group. It is very apparent that we need an AI fire alarm.

Even if we manage to approach AI development 'in an ethical way' (there's that phrase again) and ensure that no one should design AI that seeks to exploit, manipulate, harm or create chaos, we will need to be able to spot such malicious, and quite probably unexpected acts before they cause damage. Furthermore, many private entities are more concerned with whether their behavior is legal rather than ethical. The difference is substantial. This is why we need a Committee for Monitoring AI. I'll explain.

Fire is a useful technology with many societal and economic benefits, but it can go wrong. Humans have been worrying about these side-effects of technology since the first cooking fire got out of control and burned down the forest.

Humans have been worrying about these side-effects of technology since the first cooking fire got out of control and burned down the forest.

Eliezer Yudowsky has written a powerful piece about warning systems and their relevance to AI. Basically he notes that fire alarms don't tell you when there is a fire (this is because most times they ring there is no fire). But conversely the sight of smoke doesn't make you leap into action. This is especially true if you are a bystander in a crowd (perhaps it's just someone burning the toast? Surely someone else will act, and so on). What fire alarms do is they give you permission to act. If the alarm sounds, it's OK to leave the building. It's OK to get the extinguisher. You're not going to look silly. The proposed AI Ethics and Society working group, and my suggested Committee for Monitoring AI ought to act as fire alarms.

Perhaps a system of risk levels is needed that account for the scale of the particular AI risk, its imminence, and the potential impact; a colour-coded system to issue warnings. Importantly, this needs to work at a global not just local level due to the threat from outside and the lack of national boundaries for many AI applications. Our global interactions around AI need to extend beyond learning from foreign organisations and sharing gizmos.

Overall, we need to shift the focus around AI innovation from one of rapid development to market, to one concerned with risk and reliability. AI as a technology has more in common with anaesthesia or aviation than with sports shoes or efficient light bulbs. Like aviation, we need to ensure high-reliability AI infrastructure when AI is at the helm of logistics and food supply, power grids, self-driving cars and so on. We need redundancy, and I'm not confident this will be implemented especially given the single point of failure systems we still have commanding our telecommunications network in New Zealand. A human factors, safety systems engineering approach is needed, and this will require large changes to computer science and innovation training.

Conclusions

The AI Forum New Zealand is to be commended for a detailed yet accessible report on the state of play of AI in New Zealand. These are exciting times. Overall the urgency with which this report insists we must act is absolutely correct.

The Recommendations section begins, 'Overall, the AI Forum's aim is for New Zealand to foster an environment where AI delivers inclusive benefits for the entire country'. This must be the case. We just need to work hard to make it happen. The best way to ensure inclusive benefits is to settle on a value framework, which will enable us to unpack the elusive 'ethical approach'. By running each possibility through our values we can decide quite simply whether to encourage or condemn the application.

Like any tool, AI can be used for good or for bad, and no doubt most applications will simply be ugly. The report claims that some of the important potential harms, for example criminal manipulation of society, are as yet 'unquantified'. Well, it is not only criminals that seek to manipulate society, and to be honest, I'm not one for waiting around until harmful activity is quantified.

We need to decide what is OK and what is not, anticipating what might be coming. As the report indicates, this will require ethical and legal thinking, but also sociological, philosophical, and psychological. I would argue that a substantial portion of the Government's Strategic Science Investment Fund be dedicated to facilitating these critical allied discussions and outputs.

Most of all we need to design for democracy and build an antifragile New Zealand. As a Society we must indeed work to shape the future. What values are we willing to fight for and what are we willing to sell-out?