Matt Boyd and Nick Wilson

Catastrophic Risk from Rapid Developments in Artificial Intelligence

what is yet to be addressed and how might New Zealand policymakers respond?

Abstract

This article describes important possible scenarios in which rapid advances in artificial intelligence (AI) pose multiple risks, including to democracy and for inter-state conflict. In parallel with other countries, New Zealand needs policies to monitor, anticipate and mitigate global catastrophic and existential risks from advanced new technologies. A dedicated policy capacity could translate emerging research and policy options into the New Zealand context. It could also identify how New Zealand could best contribute to global solutions. It is desirable that the potential benefits of AI are realised, while the risks are also mitigated to the greatest extent possible.

Keywords artificial intelligence, catastrophe, governance, international cooperation, risk analysis, risk mitigation

Matt Boyd researches health, technology and catastrophic risk, has a PhD in philosophy, and is the owner of Adapt Research Ltd. **Nick Wilson** is a research professor of public health at the University of Otago, Wellington.

The authors declare no competing interests. Matt Boyd received funding from the AI Forum of New Zealand in 2019 to contribute research work and writing of AI Forum publications.

Artificial intelligence (AI) is not one technology but many and includes machine learning applications and a number of types of advanced algorithms. The development and deployment of these technologies promises to advance economies, wellbeing and sustainability (AI Forum New Zealand, 2019). However, AI is both a general purpose technology, and a dual use technology of concern. This means that AI has a diverse set of uses both beneficial and harmful. This technology is now widely distributed in a world full of complex interacting threats. In the longer term, AI could plausibly even pose an existential threat to humanity.

In a previous issue of *Policy Quarterly* we outlined the emerging risks posed by AI and presented broad options for a New Zealand policy response (Boyd and Wilson, 2017). In the two years since that publication a lot has changed. Many of the developments are summarised in a major report by the Australian Council of Learned Academies (ACOLA), which notes that AI has global impact and that an international response is needed (Walsh et al., 2019). Cédric O, the secretary of state for the digital sector of France, highlighted this shared international concern when he said, 'An international platform will be necessary in order to ensure a sustainable development of artificial intelligence and serve humanity as a whole' (Marrs, 2019).

In what follows we resurvey the emerging AI landscape from a New Zealand perspective, identify potential catastrophic risks from AI, and argue for policies and action to anticipate and mitigate these risks in order that AI might predominantly benefit New Zealand society.

The AI Forum identifies risks due to the domination of AI technology by a handful of advanced corporations and the potential for an economy enabled by AI that results in wider inequality.

Recent advances in AI and the policy response

Our previous article outlined four AI risk domains: bias and injustice; economic chaos and the transformation of work; AI dominance of media discourse; and security and existential risks. Public sector work (internationally and in New Zealand) has focused on addressing some of these issues over the last two years.

For example, the risk of algorithmic injustice due to biased data, explicit or implicit algorithmic rules, and even unjustifiably neutral algorithms (Susskind, 2018) has entered mainstream thought. The AI Forum of New Zealand's report *Towards Our Intelligent Future* (AI Forum New Zealand, 2019) discusses these; an Algorithm Assessment Report (Statistics New Zealand, 2018) assesses government algorithm use; and there is a forthcoming Digital Government Strategy currently (as of late 2019) at consultation stage. Statistics New Zealand has also released a draft *Algorithm Charter* for consultation (Statistics New Zealand, 2019). Related work on the safe and appropriate use of data and the importance of transparency, explainable algorithms and the right to review are helpful early steps.

Employment prospects and economic stability in an automated world are the focus of work by the New Zealand Productivity Commission (Productivity Commission, 2019) and the Prime Minister's Business Advisory Council (Business Advisory Council, 2019). There is a Future of Work Tripartite Forum, and also a New Zealand Digital Skills Forum. Ensuring growth in a world enabled by AI is being taken seriously and policy approaches are proposed.

With respect to AI and media discourse, the threat that recommendation algorithms are serving up harmful content has reached global awareness through such initiatives as the New Zealand-initiated 'Christchurch Call' (Ministry of Foreign Affairs and Trade, 2019).

The risks of physical harm, use of technology as a weapon, and risks of accidental technological catastrophe are probably growing. A number of international researchers see the catastrophic risks of AI as the most likely near-term threat to humanity (Turchin and Denkeberger, 2018b). This is especially so when use of AI might enhance the threats posed by nuclear weapons and advanced biotechnology. While many measures of human wellbeing, such as life expectancy, infant mortality, murder rates and tolerance, have all been trending for the better (Pinker, 2011), the risk that we cause great harm to ourselves with advanced technology is probably growing (Bostrom, 2019).

The AI discussion takes centre stage

There has been an explosion of AI-related publications, including reports with a New Zealand focus. In addition to those mentioned above, the Royal Society of New Zealand's report *The Age of Artificial Intelligence in Aotearoa* (Royal Society of New Zealand, 2019) accompanied the major report by ACOLA.

Most national AI strategies focus on research, talent and industrial strategies (Dutton, 2018). Finland and the Netherlands have started to systematically educate their populace on AI (Delcker, 2019; University of Amsterdam, 2019). The United States and China also have a substantial focus on developing AI (Future of Life Institute, 2019; Select Committee on Artificial Intelligence, 2019).

Although many of these publications mention well-known and near-term risks (such as job displacement, bias and injustice), other reports have an explicitly upbeat tone. Economic analyses by PricewaterhouseCoopers (PwC, 2018) and McKinsey (McKinsey Global Institute, 2018) trumpet the boon AI will bring to profit and productivity.

Towards Our Intelligent Future is the most comprehensive report on AI in New Zealand to date. The AI Forum identifies risks due to the domination of AI technology by a handful of advanced corporations and the potential for an economy enabled by AI that results in wider inequality. Issues of algorithmic bias, injustice, transparency, fairness, autonomy, privacy, inclusiveness and safety are all touched on. So too are the risks of citizen manipulation, cyber-attack and totalitarian practices. Some of these risks may have an impact on only a small proportion of the population, but others could be the seeds of greater problems. The AI Forum provides a policy map, but in what follows we move beyond these day-to-day policy needs and focus on the larger-scale risks of AI.

The growing AI risk

AI could cause large-scale harm if it is programmed to do something devastating, or if it develops a destructive method to achieve its goals (Bostrom, 2014). Deployment of AI could also create structural risks that might lead to, or exacerbate, other threats (Zwetsloot and DaFoe, 2019). Structural threats will often require collective action to counter.

Risks from AI have been outlined in a number of recent papers (Brundage et al., 2018; The Workshop, 2019; Turchin and Denkeberger, 2018a; Yampolskiy and Spellchecker, 2016). The *Malicious Uses of AI* report catalogues these as threats to digital security, physical security and political security (Brundage et al., 2018). The probability and seriousness of catastrophic AI failures will likely increase with time (Yampolskiy and Spellchecker, 2016).

Recent technical developments underscore this growing risk. For example, Google DeepMind has made very rapid progress in mastering strategic games (AlphaStar Team, 2019). The significance of this is that DeepMind's AI applications are now exhibiting strategic capability that was until recently considered an engineering challenge. DeepMind also developed AlphaFold to predict the three-dimensional structure of biological proteins from their primary amino acid sequence. This is a very difficult problem in biology, and AlphaFold won the annual protein-folding prediction contest on its first attempt (Evans et al., 2018).

Open AI has developed an application for generating text content, which was deemed 'too dangerous to make public' and so only a partial version as open source has been released (Radford et al., 2019; Whittaker, 2019). Deepfake technology can now produce realistic video content depicting events that never occurred with convincing resemblance to actual subjects. This technology is now easily accessible and widely deployed (Barnes and Barraclough, 2019): for example, the video of Mark Zuckerberg describing how he was influenced by fictional villainous entity Spectre (O'Neill, 2019). Deepfake technology was also used to impersonate the voice of a CEO for economic gain (Stupp, 2019).

In sum, the ability of AI to produce synthetic text and multimedia, generate insights in domains such as biotechnology, and engage in strategic activity is rapidly progressing.

Even this present AI technology gives reason to be concerned. Allan DaFoe of the Centre for the Governance of AI at the Future of Humanity Institute has argued that even if we stopped scientific improvement in AI now, there are extreme systemic risks, including: mass labour displacement, unemployment and inequality; of AI as a key strategic industry, with monopolistic frontrunners; that surveillance could empower suppressive regimes and robotic repression could circumvent any human reluctance to fire upon protestors; of AI undermining global strategic stability by allowing for a successful pre-emptive nuclear strike (for example, if satellite image analysis and ocean sensors could reliably reveal the location of the nuclear-capable submarines necessary for a retaliatory strike) (DaFoe, 2018).

> AI is emerging as a threat to the functioning of democracy, which may result in a broken system that produces erratic, non-representative outcomes.

Global catastrophic risks

Global catastrophic risks are those which would bring crippling damage to human wellbeing on a global scale (Bostrom and Cirkovic, 2008). Such events may currently be of low probability, but they are potentially high impact and warrant attention because even a small decrease in the probability of their occurring has large pay-offs. Some of the risks from AI fall within this category. In addition, several scenarios show that AI could pose an existential threat to the survival of humanity or to the continuation of a flourishing technological civilisation. Such threats are identified in the work of Nick Bostrom (Bostrom, 2014), and have been catalogued (Turchin and Denkeberger, 2018a). We now examine six persisting risks associated with AI for which there does not yet appear to be an adequate policy response.

The risk that democratic processes erode

Access to clean information and tracking actual states of affairs in the real world underpins all well-functioning democratic processes. AI is emerging as a threat to the functioning of democracy, which may result in a broken system that produces erratic, non-representative outcomes. In the wake of the Cambridge Analytica scandal, the concept of 'weaponised advertising' emerged, where big data helps to target individuals and sway opinions. Non-legitimate powers are using AI and campaigns of disinformation in strategic efforts to bypass the democratic process (Mazarr et al., 2019; Polyakova, 2018). The ACOLA report on AI in Australia and New Zealand notes that platforms have been hijacked and websites, social media accounts and links created and inserted in connection with the Brexit referendum and the 2016 United States presidential election (Walsh et al., 2019).

There is some evidence that it is difficult to convince people to change their minds. However, clouding the argument with misinformation can inhibit political discourse, thereby advancing strategic ends (Bridle, 2018). Synthetic text and deepfake media are likely to increase the cloudiness. Algorithmic content recommendations amplify these effects. Also, modelling studies demonstrate that these algorithms can prevent populations converging on agreed beliefs (Sirbu et al., 2019). In this increasingly chaotic information environment, it is almost impossible to distinguish actors, motives, fake news, paranoid fiction and state propaganda (Bridle, 2018).

A New Zealand perspective on digital threats to democracy has been taken by one group, The Workshop, which has identified three core problems: platform monopolies, opaqueness of the algorithms, and business models that reward amplification of engagement without regard for wellbeing (The Workshop, 2019). The *Perception Inception* report on synthetic media (Barnes and Barraclough, 2019) examines these threats but recommends that no new law is needed in New Zealand at present. However, even if laws were to be changed, this won't prevent illegal activities. Better content moderation might be a step in the right direction, but the scale of the problem demands a wider and more effective response (The Workshop, 2019).

New Zealand society will need to decide to what degree we accept machines inferring psychological information about us in order to manipulate our beliefs (Burr and Cristianini, 2019), and whether content targeting interferes with the human right to free belief formation (UN Special Rapporteur, 2018). We may need to act to avoid a future political sphere where intelligent algorithmic code enables those with vested interests to exert power through ubiquitous surveillance and the control of perception (Susskind, 2018).

New forms of attack do not just steal information, but aim to shut down public infrastructure, cause physical damage and erase data.

The risk of totalitarianism

It is one thing for democracy to degrade and cease functioning as intended; it is another for surveillance and control systems to incrementally push a functioning democracy towards totalitarianism.

The increasing business use of intelligent surveillance systems, such as facial recognition, coupled with state surveillance approaches, including the array of Chinese social credit scoring systems (Kobie, 2019), along with rogue apps harvesting massive caches of user data means there may be very little that remains private in the 'age of surveillance capitalism' (Zuboff, 2019). Especially concerning is the ability of AI systems to detect emotions, identified by the World Economic Forum as a tool by which 'oppressive governments could … exert control or whip up angry divisions' (World Economic Forum, 2019, p.73). These trends will potentially change human behaviour on a mass scale.

Privacy can be waived by an informed individual, but privacy has benefits for society beyond those of the individual. For example, when privacy is absent, dissenting thought is suppressed. This has implications for the ability of individuals to form activist groups and hold employers, or public institutions, to account (for example, through whistle-blowing).

Erosion of civil liberties through automated censorship or manipulation could slowly emerge as the new normal. Monopolistic corporates could squeeze morality into their products, such as an in-home digital device such as Alexa that could 'snitch' to parents (or authorities) about adolescents' drug use, for example.

Digital code can force us to act a certain way and transgressions can be instantly logged and punished. Taken to the limit, increasing surveillance could become a societal panopticon where everyone is surveilled all the time, without the ability to watch the watchers. Imbalances in power such as this are easily entrenched and, without vigilance, human societies could sleepwalk into AI-facilitated totalitarianism.

Differential adoption of such technologies by powerful regimes and corporations could lead to profound disruptions in the world order, which New Zealand policymakers should be concerned about. This is particularly so given this country's extreme dependence on the rest of the world for trade, tourism and the exchange of new ideas and technologies.

The risk of violence and conflict

Fully autonomous vehicles and drones raise the possibility of a wide range of near-future lethal autonomous weapon systems (LAWS). Russia has opened a 'technopolis' hub to pursue advanced military technologies, including AI (Bendett, 2019); the 2018 US Department of Defense strategy calls for AI to be pushed across the military (Department of Defense, 2018); and China has a very ambitious AI development plan with a focus on civil–military fusion (Johnson, 2019).

Among the catastrophic risks from AI identified in one review is the 'wrong command sent to a robot army' (Turchin and Denkenberger, 2018a). But a wrong command is not necessary. A significant threat is not that lethal systems obey commands, but that they run amok. In 2010 financial algorithms caused a flash crash of the US stock market, wiping 9% off the Dow Jones in 30 minutes (Bridle, 2018). A 'flash crash' event involving autonomous weapons, such as a massive AI-coordinated swarm of drones or 'slaughter bots', could be catastrophic. The European Parliament has called for a ban on LAWS (European Parliament, 2018), but at the September 2019 meeting on the Convention on Certain Conventional Weapons, Russia and the US continued to resist the requirement for 'human control' of weapon systems.

AI could also be harnessed for a digital attack by states or by terrorists. The World Economic Forum sees the threat of AI-enabled cyber-attack as a major concern (World Economic Forum, 2019). Cyber-attacks posing a catastrophic threat include ransomware attacks on cloud-computing providers, attacks on electricity suppliers, and attacks directed at weapons systems. New forms of terrorism could attempt to disrupt automated global markets by manipulating algorithmic processes (Bridle, 2018).

New forms of attack do not just steal information, but aim to shut down public infrastructure, cause physical damage and erase data (Greenberg, 2019). The stakes are high when attacks have successfully penetrated nuclear power stations that are not connected to the internet. Often, witting or unwitting humans are used as attack vectors. A worrying risk would be cyber-attacks escalating to real-world attacks (Das, 2019). New Zealand will need to ensure that robust cyber security is a priority moving forward.

The risk of AI in combination with other threats

As a general purpose technology, like the steam engine, electricity or the internet, AI has the potential to enhance other catastrophic risks. For example, AI is integral to many advances in genetic technologies and other biotechnologies. Building on AlphaFold, AI could be used in ways that increase the risk of a biotechnological catastrophe, such as an accidental or intentional extreme pandemic (from genetically-engineered biological agents), or catastrophic disruption to critical ecosystems or food supplies. One survey estimates a 2% probability of human extinction from engineered pandemics by the year 2100 (Sandberg and Bostrom, 2008).

AI is also a concern for nuclear weapons and fissile materials safety. A 2019 report on the impact of AI on nuclear threats and strategic stability finds that AI could amplify risks (Boulanin, 2019). Key vulnerabilities include brittle nuclear systems, the threat of AI-driven cyber-attack, and misperceptions about the activities and intent of rival states with respect to AI and nuclear capability.

Quantum computing in conjunction with AI could plausibly pose new risks. Google has recently published a paper claiming that its quantum computers can outperform standard computers on a particular problem (Arute et al., 2019). Known as 'quantum supremacy', this new advance could be the first step towards vulnerabilities in encryption and other high-stakes systems.

Bostrom has advanced a 'vulnerable world hypothesis', which contemplates technological discoveries that could threaten humanity (Bostrom, 2019). It is possible that five scientists tinkering in a lab for a year, with the aid of machine learning and a digital–biological converter (Boles et al., 2017), could accidentally or intentionally bring about Bostrom's 'moderately easy bio doom', thereby proving his vulnerable world hypothesis correct (Bostrom, 2019). Potential strategies have been advanced to mitigate some of these combination threats, and New Zealand policymakers should become familiar with them.

The risks of artificial general intelligence

Artificial general intelligence (AGI) that possesses human-level capability, or superintelligence that vastly outperforms humans in all tasks, are as yet only theoretical. However, if successfully developed, AGI would pose additional risks, in part because it could be used by one organisation or state to achieve an unassailable strategic advantage. But AGI could also be problematic if its goals are poorly specified or not aligned with those of human wellbeing. This is a serious risk in part because of the technical difficulties in designing risk-free systems (Amodei et al., 2016; Bostrom, 2014). Concerns about AGI are not particularly pressing now, but estimates for the arrival of human-level intelligence, which might then exhibit an explosion of very rapid self-improvement, sit at 50% by 2040 (Muller and Bostrom, 2016). Policymakers therefore need to at least agree on a timetable for when we start to think about this issue, and what signals would trigger earlier action.

Unknown risks

Rapid technological progress, and the associated interactions between technology, society and the environment, could lead to dynamic, difficult-to-predict threats – 'technological wildcards' (Ó hÉigeartaigh, 2017). We know that the history of the world is partly driven by 'black swans', rare and hard-to-predict events that change everything (Taleb, 2007). Without further analysis we don't know which risks from AI will become most salient, and we don't know if AI is the most salient risk (it could well be) in the near term or at later stages of AI maturity. The large number of failure modes described above, and elsewhere, suggests that we haven't yet contemplated all of them (Turchin and Denkeberger, 2018a). In light of these unknowns, an agile, broad and adaptable policy response to the future of AI appears warranted.

Potential New Zealand policy response options

AI is a diverse technology touching every branch of government and society. We therefore reason that AI risk mitigation should be seen as one component of a general catastrophic risk mitigation strategy. The scale of global catastrophic risks, their uncertain probability, and the long time frame across which risks may emerge mean that individual governments and groups of like-minded ones contributing to global governance is really the only place from which to mount an effective response.

> Increased multilateralism may be a productive response to a range of threats and New Zealand diplomats should join and advance New Zealand and other initiatives.

The Cambridge Centre for the Study of Existential Risk (CSER) has published a list of policy options for governments to consider with regard to global catastrophic risks (CSER, 2019). This identifies five barriers to effectively dealing with catastrophic risks such as those posed by AI. These are:

- lack of incentives for long-termism in national policy;
- lack of government agility to respond to new perspectives on risk;
- insufficient risk management culture in government;
- lack of technical expertise; and
- failure of imagination.

We believe that these problems all apply in New Zealand and see five key areas – described below – where New Zealand policy could help mitigate the catastrophic risks from AI. A key obstacle, as Tom Barraclough, co-author of the *Perception Inception* report, notes, is that 'It's not clear who is responsible in government for anticipating these issues' (Kenny and Livingston, 2019).

It is true that there are many immediately pressing demands on the public sector, but there is enormous value in the future which may be put at risk if we do not take time for big-picture thinking (Boyd and Wilson, 2018). For this reason, an imperative first step is to designate responsibility for investigating and advocating on these issues. Some mechanism for distilling the information and advising various key decision makers is necessary to ensure comprehensive coverage and avoid redundant analysis (or omission). Furthermore, risks need to be evaluated as a portfolio so that prioritisation can occur. We should focus on the *most* important, not the *merely* important, risks.

> Given the specialist nature of and rapidly growing international literature on AI policy and safety, a dedicated unit is required to evaluate risks and possible responses.

Advocate for international cooperation

It is clear that the threats described above could have global impact and that a global response is needed. There have been some steps in this direction. In July 2018 the United Nations secretary general appointed a High-level Panel on Digital Cooperation to support 'cooperative and interdisciplinary approaches to ensure a safe inclusive digital future for all taking into account relevant human rights norms' (Walsh et al., 2019). With respect to AI, the Canadian and French governments have instigated an International Panel on Artificial Intelligence. This body is modelled on the Intergovernmental Panel on Climate Change and aims to bring together policy experts and researchers in AI, the humanities and social sciences to ensure that AI development is grounded in human rights (Marrs, 2019). At meetings associated with the G7 summit in 2019, New Zealand expressed interest in joining this panel, and this would seem highly desirable. There also exists an International Grand Committee on Disinformation and 'Fake News': representatives from 12 nations, including Australia, Finland and the United Kingdom, met in Ireland in November 2019 (Houses of the Oireachtas Communications Unit, 2019). Increased multilateralism may be a productive response to a range of threats and New Zealand diplomats should join and advance New Zealand and other initiatives.

In a number of precedents New Zealand has taken a key role in global coordination around threat, such as the anti-nuclear stance. The country's recent lead role in the Christchurch Call shows that the Ministry of Foreign Affairs and Trade is well placed to progress such initiatives.

There are currently insufficient coordinating and monitoring mechanisms to prevent an AI catastrophe should it arise. A general ability to stabilise a world vulnerable to technological risk might require greater capacities for preventive policing and global governance (Bostrom, 2019). Ironically, some degree of intrusive surveillance (for example, in certain risk domains around AI such as military applications and biotechnology) might be required to effectively monitor risks and eliminate serious threats. As well as advocating for international cooperation, New Zealand could take action beyond the Christchurch Call and set a standard for other nations to follow.

The need to structure and resource public institutions so that solutions are possible

Any diplomatic response by New Zealand must be well informed. It is important to be clear who is responsible for this advice and to ensure they are well resourced, or quality advice will not be forthcoming. Given the specialist nature of (and rapidly growing international literature on) AI policy and safety, a dedicated unit is required to evaluate risks and possible responses. This necessarily ongoing process (for AI and other technologies) should be institutionalised through the creation of enabling structures in the New Zealand public sector. 'Small government' thinking is inappropriate when a nation faces major threats and needs to support getting global governance working.

Whatever form this specialist unit takes, it could logically reside in the Department of the Prime Minister and Cabinet, or the Ministry for Foreign Affairs and Trade, or even the Ministry of Business, Innovation and Employment. The unit will need some level of independence, and a broad mandate for long-term thinking across a range of AI risks (and opportunities). This would allow the catastrophic risks from AI to be studied and managed with input from NGOs and academic institutions not constrained by legacy and near-horizon political thinking.

Given the rapidly advancing risks in the field, this unit must be formed now, and be flexible and agile. Its success will depend on the quality of thinking it harbours. This means that a competent multidisciplinary team must be involved, including exceptional AI technical experts and engineers, experts from other dual-use technology disciplines, such as biotechnology, and historians, social scientists and ethicists with knowledge of social and political transformations underpinned by technology. The personnel recruited will be key to determining the success of this task.

The future value of free and flourishing New Zealand lives justifies at least a modest investment in protection (Boyd and Wilson, 2018), and even more so if New Zealand is particularly well placed as an island nation to survive some existential threats (Boyd and Wilson, 2019). In the context of catastrophic risks, it seems reasonable to apportion perhaps 0.01% of GDP (approximately $25 million in the first year) to analysing the threats of AI and other potential catastrophic risks to New Zealand. This kind of approach provides something of an insurance policy against future risks. After the initial scoping, future investment needs, and options, will be clearer.

Understand the risks and take a risk management approach

A number of global research institutes have examined the risks posed by AI: these include the Future of Humanity Institute, the Machine Intelligence Research Institute, Open AI and CSER, among others. However, research publications are not effective unless used to inform policy. The initial investment outlined above must ensure that the outputs of these global institutions are digested to inform local policy; the investment could also fund secondments of government staff (and from New Zealand-based NGOs and universities) to centres such as the Future of Humanity Institute, Open AI and CSER in order to bolster local capability to think productively in this space.

When managing risks, the priority of action depends on the likelihood that the risk will transpire, the magnitude of the resulting harm, and the ability to mitigate the risk. Even if the probability is low, if the potential harm is very great, and a solution possible, then some resources should rationally be allocated (even if this is a collective action problem requiring coordination of multiple countries). Initial work could focus on deducing the probabilities and magnitudes of the risks we have outlined. Monitoring needs to occur so that these values can be updated regularly over time. A scenario-based approach along with signal monitoring could determine which scenarios are eventuating. Table-top simulation exercises would help identify legislative and policy gaps requiring closer examination.

Implement mitigation strategies

Government inaction and the hope that the 'ethics policies' of the developers of technology will mitigate risk is not sufficient (Nemitz, 2018). Humans may now have the power to rapidly destabilise and destroy institutions and assets that we have built incrementally over centuries. Sustaining life and civilisation are inherently valuable projects and therefore essential. When any issue is 'essential', the principle of essentialism dictates that this issue must be dealt with by law (ibid.). We already have a set of agreed law that must be adhered to globally: the International Declaration of Human Rights, and other associated principles. Preserving human rights should be the benchmark for all risk mitigation pertaining to AI.

Beyond ensuring that human rights obligations are met, mitigation strategies should ensure that we can benefit from AI without suffering the harms. Strategies might include regulating certain technologies, certifying developers, banning some practices (such as impersonating humans), monitoring developments in AI, performing safety research, and other activities specific to particular threats. AI itself might form part of the solution to some of the risks posed by AI. This could be the case for the risk of cyber-attack, or the dissemination of dangerous information.

Overall, a focus on risk and reliability is important. 'Concrete problems in AI safety' are already known (Amodei et al., 2016) and guidelines for responsible AI systems are being produced, such as the Alan Turing Institute's 'responsible design of AI systems in the public sector' (Leslie, 2019). The aim should be to accelerate this safety research.

Consider the longer term

Although there is a need to move quickly to address AI risks, we note that AI risks are situated within a suite of emerging global catastrophic risks and there is also need for this work to feed into an aggregating mechanism that can prioritise risk response across a range of threats. We therefore argue that the New Zealand government should invest in futures analysis and horizon scanning, to increase its capability for foresight and shed light on the possible consequences of the choices humans make today. We agree with sociologist Elise Boulding that modern society is suffering from 'temporal exhaustion': because we are 'mentally out of breath all the time from dealing with the present, there is no energy left for imagining the future' (Boulding, 1978). Government has a deep responsibility to future generations and for long-term incentives that transcend individual interests. Even if some risks are far distant, our experience with climate change is telling. It takes a long time to mount a coordinated national and international response. It is helpful to remember that the Kyoto Protocol was signed in 1997, yet the threat posed by climate change is still very far from being solved.

> Government has a deep responsibility to future generations and for long-term incentives that transcend individual interests.

This call for futures thinking has come from a number of quarters. NZ Tech has called for a 'Ministry for the Future' (Muller, Carter and Matthews, 2017). The Environmental Defence Society suggests a Futures Commission (Environmental Defence Society, 2019). Boston, Bagnall and Barry have argued at length for improved foresight and oversight to make government more accountable to Parliament for the quality of its long-term decision making. While noting that there is no obvious single best approach, these authors list a number of specific options for reform (Boston, Bagnall and Barry, 2019). Some futures ideas are not new in New Zealand or overseas. Indeed, the Muldoon government disestablished a Commission for the Future in the early 1980s. Sweden created a Ministry for the Future in 2014 (Ma-Dupont, 2016).

AI is a heterogeneous package of technologies that pose risks unequal in probability and scale, and which do not stand alone. But we also need to place AI in a coordinated portfolio of interdependent global catastrophic risks. Institutionalised systematic assessment and response to all major catastrophic risks will be needed in New Zealand to ensure a thriving future.

Conclusion

New Zealand has a number of important assets, including people, culture and the environment, to protect over the very longest time horizons. This is particularly true when one takes a perspective of guardianship consistent with a te ao Māori world view. Such a perspective mandates that we understand the possible catastrophic risks of AI, which include threats to democracy, the risk of totalitarianism, threats to physical and digital safety, and as yet unknown risks. These risks need to be evaluated within the set of related global catastrophic risks. One way to achieve this is to advocate for a coordinated international response and to designate responsibility for evaluating the risks from AI and planning for their mitigation within the New Zealand public sector. Mitigation of catastrophic risks should be a critical component of public policy, and is an undertaking that only governments are positioned to perform, in conjunction with other like-minded governments.

References

AI Forum New Zealand (2019) *Towards Our Intelligent Future: an AI roadmap for New Zealand*, Auckland: AI Forum New Zealand

AlphaStar Team (2019) 'AlphaStar: mastering the real-time strategy game StarCraft II', *DeepMind*, 24 January, https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mane (2016) 'Concrete problems in AI safety', arXiv:1606.06565 [cs.AI]

Arute, F., K. Arya, R. Babbush, D. Bacon, J. Bardin, R. Barends et al. (2019) 'Quantum supremacy using a programmable superconducting processor', *Nature*, 574, pp.505–10

Barnes, C. and T. Barraclough (2019) *Perception Inception: preparing for deepfakes and the synthetic media of tomorrow*, Auckland: Brainbox Institute

Bendett, S. (2019) 'The rise of Russia's hi-tech military', American Foreign Policy Council, https://www.afpc.org/publications/articles/the-rise-of-russias-hi-tech-military

Boles, K., K. Kannan, J. Gill, M. Felderman, H. Gouvis, B. Hubby, K.I. Kamrudm, J.C. Ventor and D.G. Gibson (2017) 'Digital-to-biological converter for on-demand production of biologics', *Nature Biotechnology*, 35, pp.672–75

Boston, J. (2017) *Safeguarding the Future: governing in an uncertain world*, Wellington: Bridget Williams Books

Boston, J., D. Bagnall and A. Barry (2019) *Foresight, Insight and Oversight: enhancing long-term governance through better parliamentary scrutiny*, Wellington: Institute for Governance and Policy Studies

Bostrom, N. (2014) *Superintelligence: paths, dangers, strategies,* Oxford: Oxford University Press

Bostrom, N. (2019) 'The vulnerable world hypothesis', *Global Policy*, doi.org/10.1111/1758-5899.12718

Bostrom, N. and M. Cirkovic (eds) (2008) *Global Catastrophic Risks*, Oxford: Oxford University Press

Boulanin, V. (ed.) (2019) *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, Stockholm: Stockholm International Peace Research Institute

Boulding, E (1978) 'The dynamics of imaging futures', *World Future Society Bulletin*, 12 (5), pp.1–12

Boyd, M. and N. Wilson (2017) 'Rapid developments in artificial intelligence: how might the New Zealand government respond?', *Policy Quarterly*, 13 (4), pp.36–43

Boyd, M. and N. Wilson (2018) 'Existential risks: New Zealand needs a method to agree on a value framework and how to quantify future lives at risk', *Policy Quarterly*, 14 (3), pp.58–65

Boyd, M. and N. Wilson (2019) 'The prioritization of island nations as refuges from extreme pandemics', *Risk Analysis*, doi: doi.org/10.1111/risa.13398

Bridle, J. (2018) *New Dark Age*, London: Verso

Brundage, M., S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel et al. (2018) *The Malicious Use of Artificial Intelligence: forecasting, prevention, and mitigation,* Future of Humanity Institute, Center for the Study of Existential Risk, Center for New American Society, Electronic Frontier Foundation and Open AI

Burr, C. and N. Cristianini (2019) 'Can machines read our minds?', *Minds and Machines*, 29 (3), pp.461–94

Business Advisory Council (2019) *A Future that Works: harnessing automation for a more productive and skilled New Zealand,* Wellington: Prime Minister's Business Advisory Council

CSER (2019) *Managing Global Catastrophic Risks: part 1: understand,* Cambridge: Centre for the Study of Existential Risk

DaFoe, A. (2018) 'Prof Allan Dafoe on trying to prepare the world for the possibility that AI will destabilise global politics', in R. Wiblin (ed.), *80,000 Hours*, podcast

Das, D. (2019) 'An Indian nuclear power plant suffered a cyberattack: here's what you need to know', *Washington Post*, 5 November, https://www.washingtonpost.com/politics/2019/11/04/an-indian-nuclear-power-plant-suffered-cyberattack-heres-what-you-need-know/

Delcker, J. (2019) 'Finland's grand AI experiment', https://www.politico.eu/article/finland-one-percent-ai-artificial-intelligence-courses-learning-training/

Department of Defense (2018) *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*, Washington, DC: Department of Defense

Dutton, T. (2018) *Building an AI World: report on national and regional AI strategies*, Toronto: CIFAR

Environmental Defence Society (2019) 'EDS releases second working paper in phase 2 of its resource management reform project', press release, 20 October, https://www.eds.org.nz/our-work/media/media-statements/media-statements-2019-2/media-release-nbsp-eds-releases-second-working/

European Parliament (2018) 'European Parliament resolution of 12 September 2018 on autonomous weapon systems', European Parliament

Evans, R., J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin et al. (2018) 'De novo structure prediction with deep-learning based scoring', paper presented at the 13th meeting of the Critical Assessment of Techniques for Protein Structure Prediction, Mexico, 1–4 December

Future of Life Institute (2019) 'AI Policy – China', https://futureoflife.org/ai-policy-china/?cn-reloaded=1

Greenberg, A. (2019) *Sandworm: a new era of cyberwar and the hunt for the Kremlin's most dangerous hacker*, New York: Doubleday

Houses of the Oireachtas Communications Unit (2019) 'International Grand Committee on Disinformation and "Fake News" Dublin, Ireland

– Wednesday 6th and Thursday 7th November 2019', press release, 24 October, https://www.oireachtas.ie/en/press-centre/press-releases/20191025-international-grand-committee-on-disinformation-and-fake-news-dublin-ireland-wednesday-6th-and-thursday-7th-november-2019/

Johnson, J. (2019) 'Artificial intelligence and future warfare: implications for international security', *Defense and Security Analysis*, 35 (2), pp.147–69

Kenny, K. and T. Livingston (2019) 'Can Kiwis tell fact from fake news in the leadup to the 2020 elections?', *Stuff*, 5 September, https://www.stuff.co.nz/national/politics/115518413/can-kiwis-tell-fact-from-fake-news-in-the-leadup-to-the-2020-elections

Kobie, N. (2019) 'The complicated truth about China's social credit system', *Wired*, 7 June, https://www.wired.co.uk/article/china-social-credit-system-explained

Leslie, D. (2019) *Understanding Artificial Intelligence Ethics and Safety: a guide for the responsible design and implementation of AI systems in the public sector*, London: Alan Turing Institute

Ma-Dupont, V. (2016) 'Sweden: a "Ministry of the Future" to think about tomorrow's public policy', *Responsive Public Management*, 82, pp.1–2

Marrs, C. (2019) 'French and Canadians to launch global panel on ethical AI', Global Government Forum, 22 May, https://www.globalgovernmentforum.com/french-and-canadians-to-launch-global-panel-on-ethical-ai/

Mazarr, M., R. Bauer, A. Casey, S. Heintz and L. Matthews (2019) *The Emerging Risk of Virtual Societal Warfare: social manipulation in changing information environment*, Santa Monica: RAND Corporation

Ministry of Foreign Affairs and Trade (2019) 'Christchurch Call', http://www.christchurchcall.com

McKinsey Global Institute (2018) *Notes from the AI Frontier: modeling the impact of AI on the world economy*, McKinsey Global Institute

Muller, G., J. Carter and P. Matthews (2017) *New Zealand's Digital Future: 2017 manifesto*, NZ Tech, IT Professionals New Zealand and InternetNZ

Muller, V. and N. Bostrom (2016) 'Future progress in artificial intelligence: a survey of expert opinion', in V. Muller and N. Bostrom (eds), *Fundamental Issues of Artificial Intelligence*, Berlin: Springer

Nemitz, P. (2018) 'Constitutional democracy and technology in the age of artificial intelligence', *Philosophical Transactions of the Royal Society A*, 376 (2133), https://doi.org/10.1098/rsta.2018.0089

Ó hÉigeartaigh, S. (2017) 'Technological wild cards: existential risk and a changing humanity', in *The Next Step: exponential life*, Madrid: BBVA

O'Neill, L. (2019) 'Doctored video of sinister Mark Zuckerberg puts Facebook to the test', *Guardian*, 12 June, https://www.theguardian.com/technology/2019/jun/11/deepfake-zuckerberg-instagram-facebook

Pinker, S. (2011) *The Better Angels of Our Nature: a history of violence and humanity*, London: Penguin

Polyakova, A. (2018) *Weapons of the Weak: Russia and AI-driven asymmetric warfare*, Washington, DC: Brookings Institution

PwC (2018) *The Macroeconomic Impact of Artificial Intelligence*, PricewaterhouseCoopers

Productivity Commission (2019) *New Zealand, Technology and Productivity: technological change and the future of work, draft report 1*, September, Wellington: New Zealand Productivity Commission

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever (2019) 'Language models are unsupervised multitask learners', https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Royal Society of New Zealand (2019) *The Age of Artificial Intelligence in Aotearoa*, Wellington: Royal Society Te Apārangi

Sandberg, A. and N. Bostrom (2008) *Global Catastrophic Risks Survey Technical Report #1 2008–1*, Oxford: Future of Humanity Institute

Select Committee on Artificial Intelligence (2019) *The National Artificial Intelligence Research and Development Strategic Plan: 2019 update,* Washington, DC: National Science and Technology Council

Sirbu, A., D. Pedreschi, F. Giannotti and J. Kertesz (2019) 'Algorithmic bias amplifies opinion fragmentation and polarization: a bounded confidence model', PLoS ONE, 14 (3), doi.org/10.1371/journal.pone.0213246

Statistics New Zealand (2018) *Algorithm Assessment Report*, Wellington: Statistics New Zealand

Statistics New Zealand (2019) *Algorithm Charter,* Wellington: Statistics New Zealand

Stupp, C. (2019) 'Fraudsters used AI to mimic CEO's voice in unusual cybercrime case', *Wall Street Journal*, 30 August, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Susskind, J. (2018) *Future Politics*, Oxford: Oxford University Press

Taleb, N. (2007) *The Black Swan*, Random House

The Workshop (2019) *Digital Threats to Democracy*, Wellington: The Workshop, https://www.theworkshop.org.nz/

Turchin, A. and D. Denkeberger (2018a) 'Classification of global catastrophic Rrisks connected with artificial intelligence', *AI and Society*, doi: 10.1007/s00146-018-0845-5

Turchin, A. and D. Denkeberger (2018b) 'Global catastrophic and existential risks communication scale', *Futures*, 102, pp.27–38

UN Special Rapporteur (2018) *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, New York: United Nations

University of Amsterdam (2019) 'National AI course launched', https://www.uva.nl/en/shared-content/faculteiten/en/faculteit-der-natuurwetenschappen-wiskunde-en-informatica/news/2019/01/national-ai-course-launched.html?1571862859492

Walsh, T., N. Levy, G. Bell, A. Elliott, J. MacLaurin, I. Mareels and F. Wood (2019) *The Effective and Ethical Development of Aritficial Intelligence: an opportunity to improve our wellbeing*, Melbourne: Australian Council of Learned Academics

Whittaker, Z. (2019) 'OpenAI built a text generator so good, it's considered too dangerous to release', https://techcrunch.com/2019/02/17/openai-text-generator-dangerous/

World Economic Forum (2019) *The Global Risks Report 2019*, Geneva: World Economic Forum

Yampolskiy, R. and M. Spellchecker (2016) 'Artificial intelligence safety and cybersecurity: a timeline of AI failures', *Arxiv*, arXiv:1610.07997

Zuboff, S. (2019) *The Age of Surveillance Capitalism*, London: Profile Books

Zwetsloot, R. and A. DaFoe (2019) 'Thinking about risks from AI: accidents, misuse and structure', *Lawfare*, https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure