Matthew Boyd and Nick Wilson

Rapid Developments in Artificial Intelligence

how might the New Zealand government respond?

> 'Faced with the possibility of an intelligence explosion, how can we maximize the chances of a desirable outcome?' — Chalmers, 2010

Introduction

Advances in artificial intelligence (AI) have opened opportunities in a range of human endeavours (NSTC Committee on Technology, 2016). In response to the speed of these developments there has been a burst of analysis and dialogue in New Zealand. The New Zealand Institute of Directors commissioned a white paper (Chapman Tripp, 2016); the Ministry of Business, Innovation and Employment published *Building a Digital Nation* and the Strategic Science Investment Fund 2017–24 business plan (Ministry of Business, Innovation and Employment, 2017, 2016), and supports the new Artificial Intelligence Forum of New Zealand (www.aiforum.org.nz).

Matthew Boyd is the Director of Adapt Research Ltd, Wellington. **Nick Wilson** is a Professor of Public Health in the Department of Public Health, University of Otago, Wellington.

Intelligent systems are here and are likely to bring about a 'fourth industrial revolution' (Government Office for Science, 2015; Helbing et al., 2017; NSTC Committee on Technology, 2016). Systems with general intelligence, more capable than humans at most tasks, are more probable than not within 20–30 years (Muller and Bostrom, 2016). According to Nicholas Davies, head of the World Economic Forum society and innovation department, such systems will 'fundamentally change the way we create value and do business, and value ourselves as human beings' (New Zealand Herald, 2017).

AI is a global issue and presents great opportunities for benefit, but also great risk. Risks range from economic to social and psychological, to existential. We argue that these risks are insufficiently articulated in New Zealand government reports to date, and there is an obligation

for New Zealand government agencies to consider deep questions about the kind of society we wish to live in and our role in the emerging global transition to an AI world. We explain AI briefly, identify four key risks that the policy response needs to reflect, and conclude that New Zealand needs to support serious research into the risks of AI. We need informed debate, formal public engagement and emergent policy.

What is AI?

AI systems are digital systems that automate or replicate intelligent behaviour. AI systems use data to infer patterns and learn solutions to problems. Programmers provide AI systems with goals rather than strict methods. Present intelligent systems may augment the intelligence of a user, or provide domain-specific expertise (narrow AI). AI development also has the potential to produce artificial general-purpose intelligent systems (AGI) – even sentient systems according to some speculators.

One reason AI has such enormous potential is that the world has been stockpiling information. This information takes the form of our internet search histories, purchase histories, social media posts, blogs, media reports, GPS data, government databases, marketing databases, sensor databases; any form of data you can imagine that is stored digitally. The quantity of information available is doubling year on year (Helbing et al., 2017; IBM, 2016). Such vast data sets arise in part because of the 'surveillance capitalism' that pervades the globe (Zuboff, 2015). We have created a rich representation of social reality, filled with sequences of cause and effect, associations, beliefs, emotions, goals, hopes, dreams, memories and behaviour. These data are what AI systems learn from and the methods and rules that AI employs in learning and making inference are often opaque to human observers.

AI brings great economic and social possibility, with potential for deep insight and productivity gains, novel methods, delegation of decision making to automatic systems, task competence exceeding expert humans, and the possibility of artificial superintelligence capable of things we cannot yet conceive. If we soon design AI systems that are very much better than humans at designing AI systems then an explosion of intelligence may occur spectacularly thereafter. Rapid evolution of intelligence may occur as well through Darwinian selective processes (Association for Computing Machinery, 2017). But subsequent generations of AI may contain 'bugs' that lead to unforeseen consequences, which humans are unable to remedy.

The applications of AI are catalogued elsewhere (Chapman Tripp, 2016; Government Office for Science, 2015; IBM, 2016; NSTC Committee on Technology, 2016). It is very clear that AI will influence policy, but we need policy about AI. Policy around AI is going to have to be flexible to accommodate rapid change and will need to be based upon principles of upholding core societal values.

> A UK government report on AI states that 'at present there is generally agreement that there should be a human in the loop … the nature of their role is likely to evolve as the technology develops'…

In this article we intentionally avoid the kind of 'theological speculation' found in many radical assessments of AI (Chalmers, 2010; Muller, 2014), and focus on the social and personal risks of rapid, unfettered development and implementation of AI.

What are the plausible risks of AI?

Broad and non-specific risk statements are easy to ignore. Details matter and are necessary. Inspired by the New Zealand Institute of Directors report, we now elaborate on the potential for AI to radically transform our world. We cannot catalogue all the risks of AI here, so we have chosen four particularly challenging risks, not yet detailed in key New Zealand policy documents.

The risks of bias and injustice

Algorithmic bias leads AI systems to demonstrate unexpected behaviour on occasion. Microsoft's Twitter chatbot learned to be racist from human data sets (Devlin, 2017; Gibbs, 2016). AI is not immune to human prejudice (IBM, 2016). The public may be much less forgiving of a biased machine than a biased individual, especially in critical domains.

The New Zealand government makes use of an integrated data infrastructure (IDI) to help target social investment by using past data to predict likely actions or qualities of different groups. Certain objective facts about individuals make the probability of them needing social assistance very much higher (McLeod et al., 2015). Statistical profiling is also used in insurance, law enforcement and many other domains. The IDI is a governmental database, and subject to government ethics, but databases outside government control do not necessarily benefit from such protection. Even within government, ethical checks may not always occur in intelligence or law enforcement activities. This is especially a risk if humans don't fully understand the systems. Even ethical safeguards may not be able to overcome the bias of machines that learn from human data sets.

Sophisticated AI systems, analysing the stockpile of data representing the structure of human social reality, may infer great insight about the probabilities of risks to individuals, augmenting human social work, police work or health care and improving human decisions through prescience. Social media can already be used to predict some future events (Phillips et al., 2017).

A UK government report on AI states that 'at present there is generally agreement that there should be a human in the loop … the nature of their role is likely to evolve as the technology develops' (Government Office for Science, 2015, p.10). What happens when AI is demonstrably more reliable than human decision makers? Do we remove the human from the loop? No system of prediction is perfect and the inference rules of AI systems may be inaccessible to human understanding. Ought such systems be used in insurance, law enforcement, social investment, or in the interests of profit? What happens when human decisions and AI conflict? New Zealand's chief science advisor Professor Sir Peter Gluckman notes that:

> While prediction based on risk factors is a key objective … such predictive approaches will identify risk and resilience factors based on group characteristics and there are significant limits and dangers in extrapolating this to a specific individual. (Morton, 2017)

These situations (which could involve issues of health care or prejudice) raise significant questions about liability, control, fairness, privacy and society. Will we, and ought we to, accept more and more delegation of authority to machines?

The risks of AI dominance of media discourse

Human rationality is bounded. We are subject to biases and heuristics of thinking that control the information we believe and may confound our best intentions (Gigerenzer, 2008; Kahneman, 2011; Richerson and Boyd, 2004). This means that our psychology is exploitable.

Indeed, our psychology is already exploited by advertising, propaganda and rhetoric. Connectivity and digital platforms make it much easier to share and spread information. It can be relatively easy to manipulate the public's perception of reality, and technological manipulation of public opinion is a daily occurrence. The 'fake news' phenomenon illustrates this (Gu, Kropotov and Yarochkin, 2017; Woolley and Howard, 2017). In 2017, 45% of Twitter activity in Russia was estimated to be automated (Woolley and Howard, 2017). Furthermore, content is also becoming more individualised.

Private and public entities already use 'big nudging' to provide information that exploits the relationship between psychological biases and behaviour (Benartzi et al., 2017). Such 'mind hacking' works probabilistically on a population level. We also know that 'fake news' can drive real behaviour.

Psychological exploitation is possible on an unprecedented scale with the help of intelligent machines exploiting the structure and function of social media and vast data sets. For example, a Trend Micro report claims that it costs $200,000 to generate fake social media and authentic-seeming news that results in a real-life demonstration about an issue that doesn't exist (Gu, Kropotov and Yarochkin, 2017). AI systems could enact such hacks much more efficiently with fabricated truth 'that panders to its audience's ideologies … enough to compel people to join an imagined cause' (ibid., p.60).

Combine big nudging, fake content, a greater understanding of human psychology and its vulnerabilities, and the ability of AI to individualise content: this has the potential to undermine not only truth, but free will, autonomy and democracy (Helbing et al., 2017). If the AI systems of interest groups become proficient at exploiting patterns of cause and effect we aren't even aware of, reality may recede in a storm of artificial content while we remain oblivious to our own manipulation (Woolley and Howard, 2017).

It may be that attempts to control opinions are doomed to fail; however, what results from such attempts is unpredictable. No one knows where 'persuasive computing' and 'computational propaganda' may lead us.

> Even if humans remain in control of the intelligent systems we design, … AI technologies threaten to make us vulnerable, alienated and, paradoxically, 'automated masters' of our creations.

The risks of economic chaos and the transformation of work

AI has the potential to massively disrupt our core economic systems. Many reports detail mass unemployment due to automation. New jobs created may not be jobs that New Zealand's labour market is equipped to capitalise upon. A 2017 OECD report cited New Zealand's low productivity and weakness in mathematics as barriers (OECD, 2017). The twin forces of job loss and profit gain create widening inequality. This is a major policy concern given the relationship between socio-economic conditions and health (Marmot and Allen, 2014).

Whether society values something, or someone, is contingent on the norms, beliefs and needs of the time (Sandel, 2010). There is a risk that present systems will become unfair with the arrival of AI. A new and just distribution of resources is needed as many workers begin to suffer through no fault of their own. We risk having large numbers of economically valueless citizens and a minority of technologically literate people acquiring unprecedented wealth and influence.

No one really knows where this will end up. But the endgame of an intelligence explosion might be a post-scarcity economy, where economic growth rates increase dramatically (Bostrom, Dafoe and Flynn, 2016), supply outstrips demand and the value of money collapses (Starkey, 2017). One possible solution to this issue is a universal basic income funded by taxing robots that supplant human workers (James, 2017; Nauman, 2017). This is the preferred position of

global tech leaders such as Bill Gates and Elon Musk.

Even if humans remain in control of the intelligent systems we design (as opposed to being influenced and swayed by them), AI technologies threaten to make us vulnerable, alienated and, paradoxically, 'automated masters' of our creations. We risk falling into a state where we lack know-how, and are dependent on algorithmic processes that control our lives and undertake the meaningful work we once did. This 'tragedy of the master' (Coeckelbergh, 2015) has profound implications for power, knowledge and experience. Increasing dependence on AI could ultimately lead to loss of meaning as human work is replaced by robots (Nauman, 2017) and we voluntarily submit, letting algorithms rule our lives (White, 2015).

Security and existential risks

The US Intelligence Community outlines the risks of AI in a 2017 report (Coats, 2017). These include the vulnerability of AI systems to cyber attack, and advances in foreign weapon and intelligence systems (in particular, autonomous weapon systems). Autonomous weapon systems could be made extremely difficult to 'turn off' to evade enemy interference, but this could make them inherently dangerous.

Some of the risks of AI seem to be genuinely existential (Bostrom, 2014; Chalmers, 2010; Danaher, 2015). These are particularly concerning given that AI research and development might be faster than expected and catch policymakers off guard as we face systems we do not understand. Existential risk from AI could be possible in one of several scenarios: first, if AI is programmed to do something devastating; second, if AI chooses a destructive or perverse method to pursue benevolent goals (Bostrom, 2014). In either case, very competent AI would pursue goals that are misaligned to those of humans.

Alternatively, AI could pose an existential threat by doing something accidental or unexpected (think firing nuclear weapons without a human-like grasp of the consequences, or devising a potent biological pathogen without knowing it will infect humans). AI systems don't have to be robotic to pose a physical existential threat to humans; there is a lot that could be controlled and interfered with through an internet connection. These critical systems include power grids, food supply chains and quarantine systems. They could potentially include future geoengineering systems that could, if interfered with, cause ecological havoc.

Pervading the four risks outlined here (and other risks identified elsewhere) are a set of moral and ethical themes, which beg for debate and policy. These themes centre on the locus of power and control (at levels of society and human–AI interaction). There are themes of privacy, freedom and autonomy, liability, regulation and safety, and curtailing malicious intent.

> If [humans] question the advice they receive, however, they may be thought reckless, more so if events show their decision to be poor … departments will need to be transparent about the role played by artificial intelligence in their decisions. (Government Office for Science, 2015, p.10)

Some of the solutions proposed in the UK report are local (such as certification for AI engineers) and so may not address global risks in a connected digital world. The UK government also has a Data Science Ethics Framework (UK Cabinet Office, 2016) which goes some way to guiding public sector data use, but we need regulation around the private sector too. A Royal Society report identifies 'social issues', 'implications for data use' and 'security and control' as issues (Royal Society, 2017); but it contains little actual detail of what these risks entail.

> AI is a problem space where ideologically diverse parties must come together over ethical issues, and New Zealand has a history as a flexible legislator and innovator in the space of social protection

International response to the risks of artificial intelligence

The US government calls for a whole-of-government response to AI, and outlines 23 recommendations to ensure that the long-term consequences of AI are beneficial (NSTC Committee on Technology, 2016). Identified risks, alongside regulatory ones, include inequality, employment disruption, challenges in trying to understand and predict the behaviour of AI systems, and the safety of AI 'when exposed to the full complexity of the human environment' (ibid., p.2).

The US recommends mandatory ethical training of AI practitioners, policy consistent with international humanitarian law, monitoring of milestones in AI development, bilateral talks with foreign governments, and ongoing public engagement.

A UK government report identifies similar risks, and in our opinion a critical risk pertaining to advice from AI:

The Canadian government has initiated a $125 million Pan Canadian AI Strategy aimed at making Canada a world leader in AI and attracting top talent (Canadian Institute for Advanced Research, 2017). Google, Amazon, Facebook, IBM and Microsoft have created the Partnership on Artificial Intelligence to Benefit People and Society to conduct research, including on ethics and fairness (Hern, 2016). IBM identifies issues of safety, control and trust, and that a fact-based dialogue is needed to inform progressive social and economic policy (IBM, 2016). Elon Musk is funding open AI to advance digital intelligence in a way that is most likely to benefit humanity as a whole. The Future of Life Institute has published an open letter regarding the safe development of AI and a list of research

priorities to ensure that AI is beneficial (Russell, Dewey and Tegmark, 2016).

A collective of European academics has recently published an opinion piece in *Scientific American* offering warnings about some of the most insidious risks of AI (Helbing et al., 2017). Their views are critical warnings about change in the nature of society and human reality. The European Union (EU) has taken official steps towards implementing civil law rules on robotics and requirements to register advanced robots (European Parliament, 2017). Also, the EU's new General Data Protection Regulation effectively creates a 'right to explanation', whereby a user can ask for an explanation of an algorithmic decision that was made about them. This will drive global standards for anyone who wants to deploy their AI products in the EU (Goodman and Flaxman, 2016).

It is somewhat surprising that policy documents produced by governments pay little attention to the outputs of organisations such as the Centre for Public Impact, the AI Initiative of the Future Society at Harvard Kennedy School, the Future of Humanity Institute at Oxford University, and others. Many of these academics concur that we need some form of global governance board (Bostrom, Dafoe and Flynn, 2016). Given the risks, the transition to machine superintelligence requires a set of 'policy desiderata', these authors argue. Policymakers must pay attention to:

- efficiency (providing technological opportunity, mitigating AI risk and ensuring global stabilisation, e.g. through the use of a single AI governance body);
- allocation (all people are exposed to the risks of AI so there must be resource shuffling to recognise risk externalities and the need for justice, given that we are behind a veil of ignorance regarding a post-AI world);
- population (we must consider how to treat AI systems and what kinds of new entities to bring into existence); and
- context transformation (responsibility and wisdom are needed in a radically unfamiliar environment).

By 'governance' the authors refer not only to the actions of states but also to transnational governance.

The New Zealand Law Foundation ... is funding a $400,000 University of Otago project which aims to explore the possible implications of AI innovations for law and public policy in New Zealand.

What has been the response to AI risks in New Zealand?

We argue that in comparison with the response of some nations, New Zealand lacks a governmental response. We also argue that the global response yet lacks the coordination required to deal with truly global risks.

The New Zealand Institute of Directors, noting the lack of dialogue about AI in New Zealand policy, published a horizon scan of AI in New Zealand (Chapman Tripp, 2016). This report surveys the opportunities and risks and identifies potential inequality, unemployment, and legal and regulatory needs. However, there is little discussion of the threat to freedoms and autonomy, to social power, of the risk of autonomous weapons, or the many other potential risks of AI.

The Institute of Directors' report poses two critical questions: what ethical challenges does widespread use of AI raise?, and what controls and limitations should be placed on AI technology? The report calls for a high-level working group to research these issues, and for a whole-of-government and whole-of-nation approach. We agree with these calls to action. These are necessary – but not sufficient – responses to the risks posed by AI. The New Zealand government needs to see AI as a wider issue than merely an instrumental tool for increasing GDP, and needs to be transparent in communicating the changes AI poses for society.

The Ministry of Business, Innovation and Employment writes that we ought to 'accelerate the safe adoption of AI technologies' (Ministry of Business, Innovation and Employment, 2017, p.7). It favours collaboration between the government, Callaghan Innovation and industry, and supports the AI Forum – yet nascent – to undertake research into AI to identify opportunities and mitigate risks. The AI Forum has an agenda for open discussion around policy and awareness of AI, the economy, and capability and skills needs. It also aims to balance the conversation by providing evidenced arguments against AI doomsayers. The forum's first research project is a stocktake of issues around the economy, society, education and government. This includes New Zealand's readiness for AI, the direct and indirect impacts, skills needed and government opportunities. The AI Forum looks set to provide important information for the government to consider in policymaking. More of this sort of activity focused on the New Zealand context is needed – the sooner the better.

These sentiments were reinforced in the 2017 Royal Society of New Zealand's regional lecture series, where Professor Alistair Knott spoke of employment issues, machine bias, transparency, accountability and ethics. He argues that interdisciplinary structures are required, which would include AI researchers, AI companies, economists, lawyers, social scientists and ethicists (AI Forum New Zealand, 2017b). The New Zealand Law Foundation, an independent charitable trust, is funding a $400,000 University of Otago project which aims to explore the possible implications of AI innovations for law and public policy in New Zealand. The study is a collaboration between the

Faculty of Law and the departments of Philosophy and Computer Science.

However, the New Zealand government remains strikingly upbeat about AI and articulates few risks. Coupled with Ministry of Business, Innovation and Employment's strategy that we ought to promote New Zealand as a test bed for emerging technologies, this is concerning. We don't yet know whether a fully informed New Zealand public would concur with this position.

More is needed and more global coordination

We need to create national and global norms surrounding AI. We need to ensure that current regulation around data access, use, privacy and consent are robust at international level. These are truly global issues, and regulating within borders will not prevent abuse across borders. AI is an opportunity for us to revisit inequality and justice on a global scale.

Key existing policy recommendations include, but are not limited to, the following:

- monitor AI development milestones to aid prediction;
- devise mechanisms to ensure fairness of benefit distribution;
- support informational self-determination and popular participation;
- improve transparency and remove 'information pollution';
- improve collaboration at national and global levels;
- promote responsible behaviour through digital literacy and digital ethics. (Bostrom et al., 2016; Helbing et al., 2017; NSTC Committee on Technology, 2016)

Given the risks, policy development is critical so that we don't throw away advances in democracy and human rights by succumbing to insidiously anti-democratic risks like persuasive computing.

Workers and other ordinary citizens must be engaged or decisions will be made for them by people who care more about personal interests. We may need to move toward more collective notions of responsibility, and this needs to have the means and scope to include non-human actions (White, 2015). A global response is needed (Lee, 2017).

What New Zealand might do

The arrival of AI is a collective choice problem at a national and global level. It is not simply a matter of ensuring that New Zealand stays 'ahead', as the Institute of Directors white paper argues. The issue of AI bears much in common with climate change. The New Zealand government and private and public organisations need to focus on the relationship between AI and core values. More ventures like the AI Forum are needed, along with extensive public engagement.

If we agree that '[i]t is totally unacceptable … to use these technologies to incapacitate the citizen' (Helbing et al., 2017, p.15), then we need to negotiate a new social contract, with a policy framework which sees citizens as partners and protects the right of people to clean information, to allow them to lead the truthfully informed, self-determined lives critical to a functioning democracy.

> … for what world of affordances do we want to be held accountable?

Key questions that require a local answer include:

- Do our present legal tools provide suitable options for dealing with the issues posed by AI?
- Ought the world to permit autonomous weapon systems?
- Ought we to permit individually targeted persuasion systems that threaten to undermine a truthfully informed public?
- What are the limits of nudging in the public interest, and the permissibility of nudging for private interests?
- Is a universal basic income one of the solutions to the potential for dramatic inequality?
- Are there emerging tools which might offer solutions to some of the threats posed by AI?

Given the above, we suggest that New Zealand policymakers ought to pursue the following five actions:

Research the risks and impact of AI: Government should fund research and reports on AI that include the ethical/philosophical/social and psychological issues; fund engagement with the public and a societal discourse; and ensure mechanisms so that the findings of studies such as Otago University's can inform policy. 'Funding could come from the government's $410m Strategic Science Investment Fund' (Ministry of of Business, Innovation and Employment, 2016).

Inform and engage the public: The government has a responsibility to digest the outputs of research initiatives such as Google Brain, OpenAI, the Machine Intelligence Research Institute and the Future of Humanity Institute, and a range of academic publications, and translate these so that the New Zealand public remains informed. Embedding a programme of digital ethics within the New Zealand educational curriculum is another option.

Produce clear recommendations: Existing policy needs to be analysed, international policy co-opted as appropriate, and new policy around the legitimate and low-risk use of AI developed. Policy needs to cover risks of: bias and injustice, dominance of media discourse, autonomy, economic chaos and security.

Take a global lead: The government support the formation of a single global body on AI similar to the Intergovernmental Panel of Climate Change, or an 'AI Club' as suggested by Nordhaus for addressing climate change (Nordhaus, 2015). It should advocate for social justice and equity on the global stage; and advocate for ethics around AI and protection of rights, privacy and safety, so we are not forced to follow the lead of others.

Maintain a vision for New Zealand: We must maintain a vision of New Zealand as a society of equality, empowerment and autonomy, with rights to truthful information, where we are protected from weapons of mass destruction. These are non-partisan issues.

We must prepare for a qualitatively different kind of society and move on from present thinking. The critical question is, 'for what world of affordances do we want to be held accountable?' (White, 2015). Issues of foreseeability and negligence must be central to these discussions. AI is a problem space where ideologically diverse parties must come together over ethical issues, and New Zealand has a history as a flexible legislator and innovator in the space of social protection (for example, the Accident Compensation Corporation).

Conclusion

The risks of AI range from comparatively minor issues of privacy and liability, through major societal and economic issues, to issues of existential risk. In general, the lack of detail on risk in government reports proffers a false sense of security and of the absence of fundamental risks to society. This appears to be especially the case in the limited New Zealand policy material on AI produced so far. One important reason that this is concerning is the fact that governments are not immune from causing accidental, or indeed intended, harm. Many of the examples we have presented focus on threats from the private sector, but governments can be just as capable of AI-driven 'Big Brother' social control as private entities.

None of the responses to the risks of AI we have seen fully addresses the problem of profound social change (relating to autonomy, vulnerability, disconnection from decision processes and the ethics of manipulation), let alone existential issues.

New Zealand punches above its weight on global issues and has been a world leader on women's suffrage, nuclear policy and addressing colonial injustice. New Zealand can encourage and work with other countries to move in the right direction, but we need to decide collectively what that direction looks like. This article is not the place for reaching normative conclusions, but these questions need New Zealand answers. With policy decisions being heavily dependent on values and high uncertainty, New Zealand can act as a global honest broker for forging international policy solutions.

References

Association for Computing Machinery (2017) 'Researchers are using Darwin's theories to evolve AI, so only the strongest algorithms survive', press release, https://cacm.acm.org/news/214830-researchers-are-using-darwins-theories-to-evolve-ai-so-only-the-strongest-algorithms-survive/fulltext

AI Forum New Zealand (2017) 'How should society prepare for advances in artificial intelligence?', https://aiforum.org.nz/news/2017/8/16/how-should-society-prepare-for-advances-in-artificial-intelligence-ai, retrieved 21 August

Benartzi, S., J. Beshears, K. Milkman, C. Sunstein, R. Thaler, M. Shankar, W. Tucker-Ray, W. Congdon and S. Galing (2017) 'Should governments invest more in nudging?', *Psychological Science*, 28 (8), pp.1041-55, epub 5 June, doi: 10.1177/0956797617702501

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

Bostrom, N., A. Dafoe and C. Flynn (2016) *Policy Desiderata in the Development of Machine Superintelligence*, Oxford: Future of Humanity Institute, Oxford Martin School, Oxford University

Canadian Institute for Advanced Research (2017) 'Canada funds $125 million Pan-Canadian artificial intelligence strategy', press release, retrieved from http://www.newswire.ca/news-releases/canada-funds-125-million-pan-canadian-artificial-intelligence-strategy-616876434.html

Chalmers, D. (2010) 'The singularity: a philosophical analysis', *Journal of Consciousness Studies*, 17 (9–10), pp.7-65

Chapman Tripp (2016) *Determining our Future: artificial intelligence opportunities and challenges for New Zealand: a call to action*, Auckland: Institute of Directors

Coats, D. (2017) *Worldwide Threat Assessment of the US Intelligence Community*, Washington: Office of the Director of National Intelligence

Coeckelbergh, M. (2015) 'The Tragedy of the Master: automation, vulnerability, and distance', *Ethics and Information Technology*, 17, pp.219-29

Danaher, J. (2015) 'Why AI doomsayers are like sceptical theists and why it matters', *Minds and Machines*, 25 (3), pp.231-46, doi: 10.1007/s11023-015-9365-y

Devlin, H. (2017) 'AI programs exhibit racial and gender biases, research reveals', Guardian, 13 April, https://http://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals

European Parliament (2017) 'Robots and artificial intelligence: MEPs call for EU-wide liability rules', press release, http://www.europarl.europa.eu/news/en/press-room/20170210IPR61808/robots-and-artificial-intelligence-meps-call-for-eu-wide-liability-rules

Gibbs, S. (2016) 'Microsoft's racist chatbot returns with drug-smoking Twitter meltdown', *Guardian*, 30 March, https://http://www.theguardian.com/technology/2016/mar/30/microsoft-racist-sexist-chatbot-twitter-drugs

Gigerenzer, G. (2008) *Rationality for Mortals: how people cope with uncertainty*, Oxford: Oxford University Press

Goodman, B. and S. Flaxman (2016) 'European Union regulations on algorithmic decision-making and a "right to explanation"', retrieved from arXiv:1606.08813

Government Office for Science (2015) *Artificial Intelligence: opportunities and implications for the future of decision making*, London: Government Office for Science

Gu, L., V. Kropotov and F. Yarochkin (2017) 'The fake news machine: how propagandists abuse the Internet and manipulate the public', Trend Micro, 13 June

Helbing, D., B. Frey, G. Gigerenzer, E. Hafen, M. Hagner, Y. Hofstetter, J. van den Hoven, R. Zicari and A. Zwitter (2017) 'Will democracy survive big data and artificial intelligence?', *Scientific American*, 25 February

Hern, A. (2016) '"Partnership on AI" formed by Google, Facebook, Amazon, IBM and Microsoft', *Guardian*, 28 September, https://http://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms

IBM (2016) *Preparing for the future of artificial intelligence: IBM response to the White House Office of Science and Technology Policy's request for information*, IBM

James, M. (2017) 'Here's how Bill Gates' plan to tax robots could actually happen', *Business Insider,* 20 March

Kahneman, D. (2011) *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux

Lee, K. (2017) 'The real threat of artificial intelligence', *New York Times*, 24 June, https://http://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html

Marmot, M. and J. Allen (2014) 'Social determinants of health equity', *American Journal of Public Health*, 104 (supplement 4), pp.S517-9

Ministry of Business, Innovation and Employment (2016) *Strategic Science Investment Fund: investment plan 2017–2024*, Wellington: Ministry of Business Innovation and Employment

Ministry of Business, Innovation and Employment (2017) *Building a Digital Nation*, Wellington: Ministry of Business Innovation and Employment

McLeod, K., R. Templeton, C. Ball, S. Tumen, S. Crichton and S. Dixon (2015) *Using Integrated Administrative Data to Identify Youth Who Are at Risk of Poor Outcomes as Adults*, Wellington: Treasury

Morton, J. (2017) 'Top scientist discusses big data, social policy', *New Zealand Herald*, 19 June, http://www.nzherald.co.nz/nz/news/article.cfm?c_id=1andobjectid=11878987

Muller, V. (2014) 'Risks of general artificial intelligence', *Journal of Experimental and Theoretical Artificial Intelligence*, 26 (3), pp.297-301

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

Nauman, Z. (2017) 'AI will make life meaningless, Elon Musk warns', *New York Post*, 17 February, http://nypost.com/2017/02/17/elon-musk-thinks-artificial-intelligence-will-destroy-the-meaning-of-life/

New Zealand Herald (2017) 'How to survive the jobs apocalypse', *New Zealand Herald*, 26 May, http://www.nzherald.co.nz/business/news/article.cfm?c_id=3andobjectid=11863989

Nordhaus, W. (2015) 'Climate clubs: overcoming free-riding in international climate policy', *American Economic Review*, 105 (4), pp.1339-70

NSTC Committee on Technology (2016) *Preparing for the Future of Artificial Intelligence*, Washington, DC: US National Science and Technology Council

OECD (2017) 2017 OECD *Economic Survey of New Zealand*, Paris: OECD

Phillips, L., C. Dowling, K. Shaffer, N. Hodas and S. Volkova (2017) 'Using social media to predict the future: a systematic literature review', retrieved from arXiv:1706.06134

Richerson, P. and R. Boyd (2004) *Not by Genes Alone: how culture transformed human evolution*, Chicago: University of Chicago Press

Royal Society (2017) *Machine Learning: the power and promise of computers that learn by example*, London: Royal Society

Russell, S., D. Dewey and M. Tegmark (2016) 'Research priorities for robust and beneficial artificial intelligence', retrieved from arXiv:1602.03506

Sandel, M. (2010) *Justice: what's the right thing to do?*, London: Penguin

Starkey, D. (2017) 'Elon Musk: automation will force universal basic income', Geek.com, 29 May, https://http://www.geek.com/tech-science-3/elon-musk-automation-will-force-universal-basic-income-1701217/

UK Cabinet Office (2016) *Data Science Ethical Framework*, London: Cabinet Office

White, J. (ed.) (2015) *Rethinking Machine Ethics in the Age of Ubiquitous Technology*, Hershey, PA: IGI Global

Woolley, S. and P. Howard (2017) *Computational Propaganda Worldwide: executive summary*, Oxford: Oxford Internet Institute, University of Oxford

Zuboff, S. (2015) 'Big other: surveillance capitalism and the prospects of an information civilization', 30, pp.75-89