Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Legal and Ethical Issues

This input paper was prepared by Herbert Smith Freehills

Suggested Citation

Herbert Smith Freehills (2018). Legal and Ethical Issues. Input paper for the Horizon Scanning Project "The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing" on behalf of the Australian Council of Learned Academies, www.acola.org.

The views and opinions expressed in this report are those of the author and do not necessarily reflect the opinions of ACOLA.

Briefing Note

Artificial Intelligence - legal and ethical issues

1 The brief

This paper outlines the key legal and ethical issues arising from artificial intelligence (‘**AI**’).

AI is part of a cluster of technologies, including data analytics, blockchain and robotics, which will transform our society. This transformation will create great opportunities, the breadth of which we are already beginning to glimpse. However, these technologies have such power, and their workings are so complex and impenetrable, that they also pose great risks.

Society and governments need to consider these risks thoughtfully to ensure they are sensibly mitigated and that the opportunities are prudently taken.

> “*Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct.”* — Nick Bostrom, Superintelligence: Paths, Dangers, Strategies, 2014.

In this paper, we consider the legal and ethical issues arising from AI in two broad categories:

- questions of responsibility and ownership that arise from what AI produces; and
- issues arising out of the increasingly powerful and pervasive nature of AI, and how it will interact with society and individuals.

We conclude with some thoughts on how to respond to these issues.

2 What do we mean by ‘Artificial Intelligence’?

Before that, however, a preliminary issue: what is AI?

The term has various meanings; from technical to cultural, and all points in-between. For the purposes of this paper, we use this definition:

> AI is a system which uses massive computational power and big data to learn and independently come to conclusions.

Technically, AI is built from what are called ‘machine learning’ processes; these are processes that come out of the fields of computer science and statistics.[1] You may have heard of deep learning, neural networks and genetic algorithms – these are some of the more common machine learning tools.[2]

It may seem trite, but it is worth pointing out that AI is not actual intelligence in the way humans understand it. Although initial attempts to create AI were premised on the idea of ‘cracking the intelligence code’ and turning ‘intelligence’ into a computer program, that was found to be too complex.

Instead, prompted by recent growth in computing power and data availability, AI simulates intelligence by processing vast amounts of data and learning from it.[3]

[1] Jacob Biamonte et al, ‘Quantum Machine Learning’ (2017) 549(7671) *Nature* 195.

[2] Select Committee on Artificial Intelligence, House of Lords, *AI in the UK, Ready Willing and Able* (2017-19) 14.

[3] Dominique Hogan-Doran, 'Computer says "no": automation, algorithms and artificial intelligence in Government decision making' (2017) 13 *The Judicial Review* 1, 23.

This leads to a critical characteristic of AI: there is no intelligent process underlying the outcomes it produces. If a conclusion AI reaches is correct, it is correct because the data says it is correct, not because the AI has intelligently reasoned that it is so.[4] This is the root of what is called the 'explainability' issue – AI decisions may not be able to be explained in ways that we are used to.

AI is already prevalent in society:

- AI powered driverless cars are on our roads and creating a host of issues.[5] Many jurisdictions are grappling with regulating their use. A key question is: if a driverless car is involved in an incident, who is responsible?
- Facial recognition technologies using AI now make it possible to identify and locate individuals with a high degree of accuracy.[6] This creates enormous privacy issues.
- AI technology is now being used to make data-based decisions in a variety of fields – insurance vetting, loan applications, even sentencing decisions in the legal system. How does this square with society's traditional desire to have important decisions be transparent, explainable and reviewable?

Finally, on the subject of 'what is AI', it is worth noting that inevitably, in time, we will need to consider whether advanced manifestations of AI so approximate intelligence that they can be said to have achieved sentience, thereby arguably entitling them to legal rights and perhaps even obligations.[7] AI this advanced – not in existence yet, but inevitable – has been termed 'artificial general intelligence'.[8] We have not dealt with this issue in this paper, as it is some way down the track – though perhaps not as far as many realise.

"Artificial intelligence will reach human levels by around 2029. Follow that out further to, say, 2045, we will have multiplied the intelligence, the human biological machine intelligence of our civilization a billion-fold." — Ray Kurzweil, Director of Engineering at Google

3 The 'output' issues

One category of legal and ethical issues arising from AI covers the vexed questions of responsibility and ownership that arise from what AI produces – its 'output'.

In particular:

- When an AI makes a decision, it may not be transparent, explainable and reviewable in the way that decisions made by a human are. How do we respond to this?
- In addition, when AI makes a decision, who is responsible for that decision?
- And conversely, when AI creates property, who owns it?

[4] Dominique Hogan-Doran, 'Computer says "no": automation, algorithms and artificial intelligence in Government decision making' (2017) 13 *The Judicial Review* 1, 23.

[5] Brian Clegg, 'Turing's Taxi (autonomous artificial intelligence cab driver)' (2017) 60(8) *Communications of the ACM* 104.

[6] Tech Emergence, *Facial Recognition Applications – Security, Retail and Beyond* (25 June 2018) <https://www.techemergence.com/facial-recognition-applications/>.

[7] Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology* (NY Penguin Group 2005) 98.

[8] Ben Goertzel, 'Artificial General Intelligence: Concept, State of the Art, and Future Prospects' (2014) 5(1) *Journal of Artificial General Intelligence* 1, 1.

3.1 The ‘explainability’ issue

Until recently, the ‘explainability’ of computer system outputs was generally not an issue. Computers were programmed to run in accordance with a set of rules. If necessary, the basis on which decisions were made could be explained.

However, decisions made by more advanced AI may not be explainable, because the decisions are being made by reviewing vast sets of data, and not on the basis of clearly outlined reasoning. That means if the decision is wrong – and it may be wrong if the data is flawed – then society has no way of reviewing the decision.

Examples of poor data – or poor AI design – leading to wrong AI decisions are already happening:[9]

- an image recognition system characterising an African American couple as gorillas;
- a translation engine associating the role of engineers with being male; and
- a policing tool disproportionally targeting minority communities.

In short, AI does not always get it right: it is not intelligent, it merely processes a lot of data. If that data is wrong, or incomplete, or biased, then the decision it makes – the output – may be too. Indeed, much of the existing data that AI is presently using comes from humans, and so inevitably that data bears the imprint of inherent human biases.

Traditionally, on the basis of principles of fairness, society has valued processes that allow important decisions to be reviewed, for example:

- many administrative decisions can be appealed, such as an application for a building permit;
- court decisions, including sentencing, can obviously be appealed; and
- many organisations have their own formal or informal review processes for their customers and employees; for example, applying for a loan or a promotion.

For reviews of this nature to be effective, the reasoning behind the decision must be understood. This is the premise for legal rules relating to the transparency of decision making.

This is the so-called ‘black box’ problem with AI.[10] Increasingly, as data sets get bigger, and processes more complex, it will simply not be possible to explain the reasoning behind an AI’s decision. In such a world, the ability to review decisions is compromised.

So:

- When important decisions are increasingly being made on the basis of large data sets, how do we ensure that the data is accurate and how do we ensure that the public has faith in that? Does there need to be data-quality regulation?
- In an environment where decisions will increasingly be made which are not capable of being explained or reviewed in traditional ways, how do we respond? Does there need to be regulation requiring AI-based decisions to be explainable?

3.2 The ‘responsibility’ issue

One characteristic of AI powered systems is that they can make decisions independently of humans. As the dissociation between the controller of the system and the decisions the system makes becomes more pronounced, it will become increasingly difficult to allocate responsibility for those decisions.

[9] Ryan Calo, ‘Artificial Intelligence Policy: A Primer and Roadmap’ (2017) *The University of California Davis School of Law Review* 399, 411.

[10] Dominique Hogan-Doran, 'Computer says "no": automation, algorithms and artificial intelligence in Government decision making' (2017) 13 *The Judicial Review* 1, 32.

This means that when the decision has consequences that give rise to issues of responsibility – most notably, questions of legal liability – our traditional legal concepts, which require there to be someone 'at fault', do not work.

At their present stage of evolution, most AI systems would be considered to simply be 'tools', in the sense that they are controlled by humans. This aligns with traditional legal principles: if there is any liability, it is attributable to the controller.[11]

However, as AI use develops and the idea of a 'controller' becomes more irrelevant, this analysis will become more difficult. For example, who is responsible when an AI-powered driverless car causes an accident? The vehicle manufacturer? The software programmer? The owner of the car? There is a point at which our current laws and regulations will be unable to effectively deal with issues raised by AI.

So:

- In a world where more and more decisions will be made by AI systems independently of humans, who is responsible when something goes wrong? Does there need to be regulation attributing responsibility for AI-based decisions?

3.3 The 'ownership' issue

"The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform"
– Ada Lovelace's notes on the Analytical Engine, 1843.

Having considered issues concerning the obligations triggered by AI outputs, we now turn to the converse: when an AI creates something, in a world where there is no direct connection between what it is told to do and what it creates, who owns the creation?[12]

An AI system can produce a variety of tangible and intangible 'things' that can be characterised as property. At present, this mainly comprises intellectual property, such as copyright and confidential information. However, as AI is increasingly used in combination with robotics and automation, AI will increasingly create tangible property as well. For example, artificial intelligence software has already been used to produce original paintings.[13] Who should own the copyright in artwork when there has been no human input at all?

Legally, property can only be owned by a legal entity.[14] As with liability for decisions, ownership of property arising from AI will likely be attributed to the legal entity that 'controls' it. However, this analysis starts to break down as AI systems start to act independently.

So:

- In a world where property is arising from AI systems independently of humans, who is entitled to the ownership rights – and obligations – that attach to that property? Does there need to be regulation allocating ownership rights for AI created property?

[11] See for example *Convention on the Use of Electronic Communications in International Contracts (UN)* article 12, which states that a person (whether natural or legal entity) on whose behalf a computer was programmed should ultimately be responsible for any message generated by the machine.

[12] David Poole, Alan Mackworth, *Artificial Intelligence: Foundations of Computational Agents* (Cambridge University Press, 2010) (Web version) http://artint.info/html/ArtInt_47.html accessed 9 August 2018.

[13] David Pogue, 'Is Art Created by AI Really Art?' *Scientific American* (1 February 2018) https://www.scientificamerican.com/article/is-art-created-by-ai-really-art/.

[14] *Copyright Act 1968* (Cth) ss 32 and 35.

4 The ‘freedoms’ issues

Another category of legal and ethical issues arising from AI relates to the concerns about how increasingly powerful and pervasive AI’s interaction with society and individuals will be, and the effect it may have on our traditional freedoms.

In particular:

- AI, in combination with massive data collection, gives rise to significant privacy issues.
- Cybersecurity can exacerbate AI risks, and AI can exacerbate cybersecurity risks.
- The societal disruption that AI will contribute to puts pressure on our democratic institutions.
- AI has been spoken of as creating an existential threat.

4.1 The ‘privacy’ issue

AI systems process large amounts of data. The pressure to obtain more and more data will test society’s commitment to the privacy of its citizens and the sovereignty of personal information. Presently, we have laws that require consent before personal information is obtained and used[15]: how will these laws fare in an AI enabled world?

In addition, as AI processes become more complex, it will become harder to determine what information about people is being obtained, and what is being done with it.[16]

There is an increasing, and disconcerting, range of activities that can be undertaken with people’s personal information, beyond the simple act of disclosing it, including:

Profiling: this is the use of data about someone to make generalisations about them. AI is very good at this. Such profiles can be used for marketing, which is now extremely prevalent, especially in digital platforms. Facebook offers the ability to target consumers based on their predicted consumer preferences.[17] However, AI profiles can also be used for a host of other purposes, such as making decisions about you: in insurance, loan applications, job applications, healthcare and so on.

Location: personal information about individuals can also be used to identify and locate them. Facial recognition technology plays a part here, as discussed, but the issue is broader than that. Several years ago, the US military gave thousands of soldiers wearable devices to provide data on their exercise activity. Surprisingly, the company that collected the data posted the information online. Unsurprisingly, people used that data to locate the soldiers, and thereby several U.S. military compounds.[18]

So:

- How do we respond to the fact that people’s personal information can be used for a host of activities which they may not have consented to, especially when, with sufficiently advanced AI systems, it may not be possible to know what information is held and what is done with it, and it may not be possible to opt out?

[15] *Privacy Act 1988* (Cth) Schedule 1 Part 2 Principle 3.

[16] Dominique Hogan-Doran, 'Computer says "no": automation, algorithms and artificial intelligence in Government decision making' (2017) 13 *The Judicial Review* 1, 27.

[17] The Intercept, ‘Facebook uses artificial intelligence to predict your future actions for advertisers’ (13 April 2018) https://theintercept.com/2018/04/13/facebook-advertising-data-artificial-intelligence-ai/.

[18] Alex Hern, ‘Fitness tracking app Strava gives away location of secret US army bases’ *The Guardian* (29 January 2018) https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases.

4.2 The 'cybersecurity' issue

Cybersecurity is a significant issue in the digital environment, even absent AI. However, AI adds a few nuances to general cybersecurity risk.

The first is the significant potential for AI to be used maliciously to power more effective and damaging cybersecurity attacks. For example, 'spear phishing' is a type of cybersecurity attack involving an email that is specifically tailored to an individual or organisation, often using AI.[19] The specificity is what gives this type of attack its power, and that is achieved through AI.

The second is that AI systems themselves are susceptible to cybersecurity attacks. This is true of all IT systems; however, as AI becomes more integral to the making of significant decisions, this becomes a greater danger. For example, an AI-powered driverless car can be fooled by subtle alterations of road signs.[20] It is also possible to develop AIs that fool other AI systems into making incorrect classifications or decisions.[21]

Further, due to the 'explainability issue' detailed in 3.1 above, cybersecurity breaches have even greater potential for damage in situations where organisations using AI need to troubleshoot failures in their security systems. The lack of explanation given by 'black box' AI systems will mean finding the root cause of a problem will be extremely difficult. This could make the system susceptible to further attacks and heighten vulnerability.

So:

- Should we regulate to mitigate the risk of AI being used to breach our systems?
- Should we regulate to protect our critical AI systems from cybersecurity attacks?

4.3 The 'institutions' issue

Whist AI can deliver benefits to society, it can also create societal risks, because of its ability to disrupt existing norms.

By way of example, AI-powered technologies have, to some degree, been involved in the displacement of workers from jobs, distorting financial markets, curating our newsfeeds and creating quasi-monopolies. How do our current institutions (such as ASIC, the ACCC and APRA) respond to this? Do they have the power? Do they have the resources?

An illustration of this is the trend towards AI systems taking over roles which traditionally have required specific qualifications, certification or training; for example, legal advice or healthcare. Should AI systems be allowed to perform these kinds of roles if they achieve a level of 'competence'?

The monopoly risk derives from the potential for a small number of operators in a market to have the resources to adopt AI systems on an immense scale, thereby pushing out smaller players and reducing competition.

So:

- Given the inevitable disruption that AI will bring to society, how do we ensure that our institutions, or at least the democratic principles that underlie them, are protected? What additional powers and responsibilities do these institutions need?

[19] Juan Martinez, 'Spear-Phishing Attacks: What You Need to Know' *PCMagazine* (12 June 2017) https://au.pcmag.com/feature/48348/spear-phishing-attacks-what-you-need-to-know.

[20] Jonathan M. Gitlin, 'Hacking street signs with stickers could confuse self-driving cars', *arstechnica (*9 February 2017) https://arstechnica.com/cars/2017/09/hacking-street-signs-with-stickers-could-confuse-self-driving-cars/.

[21] Select Committee on Artificial Intelligence, House of Lords, *AI in the UK, Ready Willing and Able* (2017-19) 97-8.

4.4 The ‘existential threat’ issue

The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.” – Eliezer Yudkowsky, Artificial Intelligence as a Positive and Negative Factor in Global Risk, 2006.

Whilst there is great difference of opinion on how significant the risk is, most commentators agree that, to some degree, uncontrolled AI could present a threat to our existence.[22] AI systems may prove to be impossible to control and may, through accident or malice, develop an agenda that does not align with ours.

In this context, so called ‘lethal autonomous weapons’ are in use, and fatalities have occurred.[23]

In many ways, the risk imposed by AI in this regard is analogous to that posed by nuclear power.

So:

- How can we mitigate the risk of ‘runaway’ AI, including the risks posed by lethal autonomous weapons. Does AI itself need to be regulated?

5 What considerations are relevant to any government response?

As mentioned several times in this paper, society and government need to give a lot of thought to how to deal with these issues – the opportunities are great, but so are the risks.

It is likely that regulation will provide some of the answers.

Equally important will be education, thought leadership and guidance, and government can play a key role here as well. Apart from anything else, the best regulation is more likely to emerge out of an educated community and an informed discussion.

In that regard, we can see benefits in considering:

(a) the establishment of a government supported AI institute, tasked with further research of legal, ethical and other issues arising from AI;

(b) out of that initiative, the government facilitating the development of an overarching set of values and principles to guide our response to AI issues;

(c) on the basis of those values and principles, overseeing the creation of guidelines and frameworks for the development of regulations that can be provided to relevant departments, sectors and industries; and

(d) where appropriate, encouraging industry specific regulations which are tailored to the specific issues that AI applications are creating in that industry.

Finally, many of these legal and ethical issues will only be truly effective if a global approach is taken. Governments will need to engage globally on this issue.

“Our approach to existential risks cannot be one of trial-and-error. There is no opportunity to learn from errors. The reactive approach — see what happens, limit damages, and learn from experience — is unworkable. Rather, we must take a proactive approach.” – Nick Bostrom, Superintelligence: Paths, Dangers, Strategies, 2014.

[22] See Denise Garcia, ‘Lethal Artificial Intelligence and Change: The Future of International Peace and Security’ (2018) 20 *International Studies Review* 334, 335; Noel Sharkey, ‘Saying ‘No!’ to Lethal Autonomous Targeting’ (2010) 9(4) *Journal of Military Ethics* 369.

[23] Denise Garcia, ‘Lethal Artificial Intelligence and Change: The Future of International Peace and Security’ (2018) 20 *International Studies Review* 334, 335.

Tony Joyner
Technology, Media, Telecommunications Sector
Lead Partner
Herbert Smith Freehills

+61 8 9211 7582
+61 409 787 971
tony.joyner@hsf.com

27 August 2018

Note: Oli Tod, Annabel Beech, Ebony Garlick, Jack Joyner and Casey Lickfold assisted in preparing this paper