

Recommendations for Artificial Intelligence policy in Australia

Effective Altruism ANZ

Contact: [REDACTED]

Executive Summary

This document, responding to the white paper released by the Australian Human Rights Commission, outlines some key societal considerations and recommendations for the regulation of AI in Australia. The white paper thoroughly covered many of the nearer term issues in AI, including data privacy, accountability, bias in decision making from AI and the safety/security of systems that are currently being used or are relatively close to deployment like self-driving cars. It also touched upon some longer term considerations like large-scale unemployment.

This document focuses on the link between near-term and long-term issues, and further elaborates on some of the other most important longer-term issues relevant to AI regulation. We first include an introduction to the literature linking near term and long term issues. We then discuss some of the long-term issues experts around the world consider the most important and why appropriate policy is still absolutely necessary despite these issues more speculative nature. We then provide some recommendations on what Australia's policy response should be, addressing the questions posed in the white paper, with a particular focus on recommendations for managing the issues discussed in this document.

This document has been compiled by Effective Altruism ANZ, an organisation that aims to support Australians and New Zealanders to improve the world as effectively as possible. The development of AI technologies is expected to have sweeping societal implications in both the near- and long-term future, and poses various risks. Given this, we see effective AI policy as crucial for safeguarding the future wellbeing of Australians.

Contents

The link between nearer term and longer term issues	3
Key longer-term considerations	4
Malicious uses of AI	4
Automation of social engineering attacks	4
Prioritising targets for cyber attacks using machine learning	4
Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)	5
The main reasons for policy makers to care about AGI and SGI	5
Summary	5
The prospect of superintelligence	6
When is AGI or ASI likely to arrive, and what would it mean for us?	7
What can we do now?	7
What useful role is there for government, given the long time horizon?	8
Low likelihoods still need risk management	9
Recommendations	9
Have access to informed experts	10
External experts	10
Internal experts	10
Collaborate with international stakeholders	10
Responses to consultation questions	10
What should be the main goals of government regulation in the area of artificial intelligence?	10
Considering how artificial intelligence is currently regulated and influenced in Australia:	11
What existing bodies play an important role in this area?	11
What are the gaps in the current regulatory system?	11
Would there be significant economic and/or social value for Australia in establishing a Responsible Innovation Organisation?	11
Under what circumstances would a Responsible Innovation Organisation add value to your organisation directly?	11
How should the business case for a Responsible Innovation Organisation be measured?	12
If Australia had a Responsible Innovation Organisation:	12
What should be its overarching vision and core aims?	12
What powers and functions should it have?	12
How should it be structured?	12
What internal and external expertise should it have at its disposal?	12
How should it interact with other bodies with similar responsibilities?	13

How should its activities be resourced? Would it be jointly funded by government and industry? How would its independence be secured?	13
How should it be evaluated and monitored?	13

Concluding remarks	13
---------------------------	-----------

The link between nearer term and longer term issues

Nearer term issues are already being seen today or are likely to happen in the coming few years. They include data privacy, accountability, bias in decision making from AI and the safety/security of systems that are currently being used or systems that are close to deployment (eg. self-driving cars).

Longer term issues are much less certain but may still be worthy of concern due their large societal implications. They include wide-scale loss of jobs from AI being able to replace people in entire industries and other problems that arise from developing AI that can solve a similar range of problems to a human.

This section outlines why near term and longer term issues surrounding AI should be considered together. [This article](#), published in *Nature* earlier this year, articulates the current disconnect seen between near and long term issues and why this is mistaken:

“These two sets of issues are often seen as entirely disconnected. Researchers [and policy-makers] working on near-term issues see longer-term issues as a distraction from real and pressing challenges, or as too distant, uncertain or speculative to allow for productive work now. On the other hand, those focused on longer-term challenges argue that their potential impact dwarfs that of present-day systems, and that these issues therefore deserve a proportionate share of research attention.

We believe that this perception of disconnect is a mistake. There are in reality many connections between near- and long-term issues, and researchers [and policy makers] focused on one have good reasons to take seriously work done on the other. Those focused on the long term should look to the near term because research directions, policies and collaborations developed on a range of issues now could significantly affect long-term outcomes. At the same time, those focused on the near term could benefit from considering work on

long-term forecasting and contingency planning, which takes seriously the disruptive potential of this powerful new technology.’

Knowledge about near term issues is useful for making predictions about long term issues and vice versa. Having internal experts that have an accurate view of the current state of technology and the rate of progress is useful both for near term and longer term policy considerations. Being in touch with technical experts allows one to prepare both for immediate issues and to be informed about longer term issues happening on the horizon. Collaborations with national and international stakeholders helps ensure effective action to ensure technology does more good than harm to society both in the near and long term.

Key longer-term considerations

Malicious uses of AI

[This report](#) summarises how AI could be used by malicious actors. It was drafted by authors from 14 institutions, including specialist research institutes at the University of Oxford, University of Cambridge, Yale and Stanford.

It outlines some plausible scenarios in the coming years. A couple of examples have been included below.

Automation of social engineering attacks

“Victims’ online information is used to automatically generate custom malicious websites/emails/links they would be likely to click on, sent from addresses that impersonate their real contacts, using a writing style that mimics those contacts. As AI develops further, convincing chatbots may elicit human trust by engaging people in longer dialogues, and perhaps eventually masquerade visually as another person in a video chat.”

Prioritising targets for cyber attacks using machine learning

“Large datasets are used to identify victims more efficiently, e.g. by estimating personal wealth and willingness to pay based on online behavior.”

We strongly recommend reading [the whole report](#).

Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)

As computers have become more advanced, they have begun to solve certain problems as well as, and frequently, better, than people.

If we are able to develop a technology that can learn to solve new problems it encounters like people can, and if it can do so as well as, or better than, people can, a whole range of social issues arise from this.

Many experts in the field believe there is the potential for intelligence to be created at a human level or above, and along with this, there are many risks that need to be managed.

The main reasons for policy makers to care about AGI and SGI

We believe that the following excerpt, written by the founder of the Cambridge University Institute: “Centre for the Study of Existential Risk”, explains what the risk of a human level or higher artificial intelligence is plausible, concerning and relevant to policy makers. Only the components of the report that we endorse and think are pertinent to Australian policy have been quoted below. The full report can be found [here](#).

Summary

“There are good reasons to take seriously the possibility that artificial intelligence (AI) will eventually outstrip human intelligence, perhaps greatly so. This may happen in the lifetimes of our children and grandchildren. Its impacts, for better or worse, are likely to be immense.

Our best prospect of ensuring that this development is beneficial is to tackle it cooperatively, and as early as reasonable foresight allows.

[...] Government can play an important role in fostering the academic, technological and policy-level coordination this process is likely to require, both nationally and internationally.

In particular, by establishing an appropriate standing body, the Government can help to ensure that we do not fall into the trap of ignoring this important long-term issue, when short-term issues always seem more pressing and more tractable.”

The prospect of superintelligence

“ I J Good was a Cambridge-trained mathematician, who worked with Alan Turing at Bletchley Park, and at Manchester after the War. In their free time, Good and Turing often talked about the future of machine intelligence. Both were convinced that machines would one day be smarter than us. In the 1960s, when Good emerged from a decade at GCHQ, he began to write about the topic.

In his first paper Good tries to estimate the economic value of an ultra-intelligent machine. Looking for a benchmark for productive brainpower, he settles impishly on John Maynard Keynes. He notes that Keynes' value to the economy had been estimated at 100 thousand million pounds, and suggests that the machine might be good for a million times that – a mega-Keynes, as he puts it.

But there's a catch. "The sign is uncertain" – in other words, it is not clear whether this huge impact would be negative or positive: "The machines will create social problems, but they might also be able to solve them, in addition to those that have been created by microbes and men." Most of all, Good insists that these questions need serious thought: "These remarks might appear fanciful to some readers, but to me they seem real and urgent, and worthy of emphasis outside science fiction."

In one sense, the prospect that concerned Good remains the same, fifty years later. It boils down to four key points:

- 1) We have no strong reason to think that high-level intelligence is less possible in non-biological hardware than it is in our skulls.
- 2) We have no reason to suppose that "human level" marks an interesting limit, in non-biological systems, freed of constraints (e.g., of size, energy consumption, access to memory, and slow biochemical processing speeds) that apply in our own case.
- 3) So we should take seriously the possibility that AI will reach places inaccessible to us – kinds and levels of intelligence that are difficult for us to understand or map, and likely to involve capabilities far beyond our own, in many of the tasks to which we apply our own intelligence.
- 4) Finally, there's a prospect that AI systems will themselves contribute to improvements in AI technology, at some point. As Good saw clearly, there's then a potential for an exponential rate of development.

The big change since 1965 is that the incentives that will lead us in this direction are now much more obvious. We don't know how long the path to

high-level machine intelligence is, but we can be certain that its individual steps will be of huge commercial value, and immensely important in other ways – for security purposes, for example. So we can be sure that these pressures will take us in that direction, by default. AI is already worth trillions of dollars – perhaps not yet a mega-Keynes, but well on the way.

At present, however, AI is very good at (some) narrowly-defined tasks, but lacks the generality of human intelligence. The term artificial general intelligence (AGI) is used to characterise a (hypothetical) machine that could perform any intellectual task that a human being can, including tasks not tied to specific set of goals. The term artificial superintelligence (ASI) refers to an AGI that greatly exceeds human capacities, in these respects.”

When is AGI or ASI likely to arrive, and what would it mean for us?

“A time-line for the development of AGI is difficult to predict, in part because it may depend on an unknown number of future conceptual advances. A recent survey of AI researchers reported that most regarded AGI as more likely than not, well within this century. It does not seem alarmist to say that while it is not on our doorsteps, it may be only “decades away” (as a leading AI researcher puts it recently, intending to dispell the popular impression that it is just around the corner).

Concerning the impact of AGI or ASI, we know little more than Good. It does not seem controversial that its impact is likely to be very big indeed. The world-leading AI researcher Professor Stuart Russell (UC Berkeley) is convinced that – for better or worse – it would be “the biggest event in human history.” But the sign is still uncertain, as Good put it. The potential benefits are immense, not least in the light of AGI’s potential to solve many other problems. But there’s also a risk. As Turing himself put it, “It seems probable that once the machine thinking method has started, it would not take long to outstrip our feeble powers. ... At some stage therefore we should have to expect the machines to take control.””

What can we do now?

“These issues are going to be with us for a long time, and are likely to become more pressing as AI develops. In the short term, the obvious strategy is to attempt to foster the level of interest, expertise, and cooperation that the task is likely to require in the future. In effect, we should be trying to steer some of

the best of human intelligence to the job of making the best of artificial intelligence. Most of all, in my view, we should avoid the mistake of putting off the issue to another decade or generation, on the grounds that it seems too hard or too much like science fiction, or because other issues in the same area simply seem more pressing.

There are encouraging recent signs of rapidly growing interest in these issues, for example in an open letter now signed by many AI professionals and others, following an international meeting in Puerto Rico in January 2015. In particular, there is a growing sense of the desirability of cooperation between technology, policy, and academic partners. Some of this cooperation will necessarily be pre-competitive sharing, for commercial and other reasons – but all the more reason to engineer the kind of trust and cooperation that make such sharing possible.”

What useful role is there for government, given the long time horizon?

“The likely time-scale of these developments, and their dependence on ongoing research and progress in the field, makes a decisive intervention at one point in time impractical. More than in most cases, we are bound to be scanning a moving horizon. Nevertheless, there is a clear role for government that is likely to be beneficial, no matter how the field develops. It can foster, promote, and add its voice to a cooperative effort, both nationally and internationally, to monitor developments in the field, to flag opportunities and challenges as they arise, and generally to try to ensure the community of technologists, academics and policy-makers is as well prepared as possible to deal with both.”

While technology is clearly quite far from human level intelligence today, technology moves quickly while information takes time to disseminate and policy takes time to be enacted. While this is a speculative risk, we believe that it is the government’s responsibility to engage in appropriate risk management as part of innovation regulation, as discussed in the next section.

Having policy experts ready to advise when this technology goes from the realm of science fiction to science is extremely important. Having experts that are aware of where the current state of technology is, both allows enactment of appropriate longer term policy but also will help with creating forward thinking policy that deals appropriately with near term issues.

Australia can become a thought leader in holistic thinking about issues in innovation by becoming exceptionally well-informed without much additional cost. Australia engaging in international collaborations with other governments and institutes is an extremely helpful step towards ensuring that, if AGI/ASI happens, it goes exceptionally well instead of exceptionally poorly.

Low likelihoods still need risk management

The risks discussed above are highly uncertain and it is quite possible that they do not eventuate. But the same is also true when doing risk management of any kind.

When building a bridge, there is a very healthy amount of regulation on what risk is acceptable and how to minimize risk. A bridge that had a 1 in 100 chance of collapsing in the coming decades would be unlikely to pass safety regulations. Technology that could potentially harm all of humanity but only has a 1 in 10,000 chance of being invented in the coming decades should also have an appropriate amount of regulation.

Recommendations

We thoroughly endorse the creation of a regulatory body that is targeted at and has expertise in innovative new technologies. Its goals should be to tackle both near term and long term considerations of technological advances.

By having a body that tackles both the near term and long term issues together, it can better solve both sets of concerns. We suspect that recommendations for near term issues will be well covered by other submissions so our recommendations mainly concern policy/regulation that also regard the more speculative and longer term issues outlined above.

As these issues are highly uncertain at this point, the best concrete steps to take that are relevant to long term but also near term issues are:

- 1) Be exceptionally well-informed
- 2) Be exceptionally collaborative nationally and internationally

Have access to informed experts

In order to be exceptionally well informed, the RIO should have access to internal and external experts.

External experts

The leading external experts on the longer term risks from AI and other similar technologies include:

- Future of Humanity Institute, University of Oxford
- The Centre for the Study of Existential Risk, University of Cambridge
- The Future of Life Institute (founders included Skype co-founder Jaan Tallinn and board members include Elon Musk and, before his passing, Stephen Hawking)
- The Open Philanthropy Project (a non-profit foundation funded by Dustin Moskovitz, co-founder of Facebook)

Internal experts

Have internal experts that remain in contact with external experts, that are informed about the current rate of progress in AI technology and are up-to-date on the latest research relevant to AI development.

Collaborate with international stakeholders

It is essential for the best policy response to the risks outlined in this document that there is an international collaborative response. Australia should aim to collaborate wherever possible with other countries, with leading researchers in the field, companies producing these technologies and other relevant stakeholders. This is important for tackling both near and long term issues. An example opportunity to collaborate with international stakeholders is [The Partnership on AI](#) (partners include Google, Amazon and the University of Oxford's The Future of Humanity Institute). Other opportunities to collaborate should be looked for and taken advantage of as they arise.

Responses to consultation questions

We have included our explicit responses to the consultation questions below.

What should be the main goals of government regulation in the area of artificial intelligence?

The goal of regulation in Australia should be to ensure AI and similar technologies benefit society more than they harm society. It should aim to:

- protect society from near term threats such as data privacy and decision making by AI in systems currently being deployed; and
- protect society from longer term risks that are less likely but may do more harm in the future if appropriate policies are not enacted early.

Considering how artificial intelligence is currently regulated and influenced in Australia:

a. What existing bodies play an important role in this area?

No existing bodies in Australia play an important role in regulating the longer term implications of innovation.

b. What are the gaps in the current regulatory system?

Many of the key near-term gaps were outlined in the white paper and we also imagine that they will be thoroughly addressed by other submissions.

The issues outlined in this document are currently no-one's responsibility within government. Without a division responsible for being aware of where the technology currently is, where it is going and what the societal implications of it are, no appropriate policy could or will be enacted to minimize the risks of such technologies.

Would there be significant economic and/or social value for Australia in establishing a Responsible Innovation Organisation?

As has been argued for above, there would be significant social value for Australia in establishing a Responsible Innovation Organisation. Without one with an appropriate agenda being created, Australia and the international community put the general public at significant and unnecessary risk.

Under what circumstances would a Responsible Innovation Organisation add value to your organisation directly?

EA ANZ exists to promote high impact ways for Australians and New Zealanders to help others and society at large.

If the Responsible Innovation Organisation stayed on top of leading research on best policy response to both near term and longer term issues in relation to innovation, it could help achieve our broader vision of Australia becoming a key player in making society better today and for following generations.

How should the business case for a Responsible Innovation Organisation be measured?

The RIO's business case should be measured based on its ability to ensure regulation is more aligned with societal interests. If it is well-informed on the longer term issues there is a strong case for it being able to fulfil this function in regards to finding an appropriate policy response to longer term issues.

If Australia had a Responsible Innovation Organisation:

c. What should be its overarching vision and core aims?

A regulatory body should also aim to:

- Ensure policy is enacted that helps to protect society from near term threats such as data privacy and decision making by AI in systems currently being deployed
- Ensure that the regulation is in line with expert opinions on appropriate policy to minimize risk in response to longer term, larger and more speculative risks like AGI/ASI (and the threats that arise as technology approaches this such as mass unemployment)
- Encourage Australia to collaborate with the international community when it comes to the societal implications of current and emerging technologies

d. What powers and functions should it have?

Powers and functions should be those that allow the above recommendations to be fulfilled. What is best here depends on the structure of the body and is not our area of expertise so we have left this section brief.

e. How should it be structured?

It should be a government body that can achieve the functions described above. The best structure to achieve the functions described is not our area of expertise so we have left this section brief.

f. What internal and external expertise should it have at its disposal?

See the above section on having access to informed experts.

g. How should it interact with other bodies with similar responsibilities?

There are no Australian bodies with similar responsibilities in regards to longer term issues. It should collaborate with international bodies with similar responsibilities as much as possible.

h. How should its activities be resourced? Would it be jointly funded by government and industry? How would its independence be secured?

It should be government funded if possible to ensure independence, but better to be (partially) industry funded than not funded at all.

It could perhaps obtain philanthropic funding from stakeholders interested in ensuring that the longer term trajectory of AI is as safe as possible (if this regulative body was interested in having longer term considerations on its agenda, our organisation could help put you in touch with potential funders).

i. How should it be evaluated and monitored?

To check whether the RIO appropriately addresses longer term issues, it should seek feedback from leading experts in these areas (such as those listed in the 'informed experts' section above).

Concluding remarks

We are excited to see Australia take seriously the societal implications of innovation.

We look forward to seeing a regulatory body that influences Australian policy in ways that ensure AI and similar technologies do more good than harm to society, both in the near and long term, and both nationally and internationally.