Navigating Transformative AI: myopia, existential risk, and regulation

I. Introduction

Emerging technology is an exploration of a vast and novel landscape. The path ahead is perilous, and we must proceed with caution. On the journey of technological progress, we find ourselves at a "high mountain pass … the only route onward a narrow path along a cliffside, a crumbling ledge on the brink of a precipice … if we fall, everything is lost … it is the greatest risk to which we have ever been exposed."[1] Our rapidly advancing technology has the potential to careen out of control unless our wisdom can match its pace. One such technology is artificial intelligence (AI). How can regulation ensure that we navigate this landscape safely without falling from the precipice?

In this essay, I propose three regulatory strategies to address challenges surrounding the potential for AI to destroy humanity (existential risk). Firstly, we need to understand the nature of existential risk and accurately define the type of AI which might cause it. This is akin to drawing a map of the landscape ahead. Secondly, we can broaden the law's myopic view of these issues by shifting the purpose of regulation toward long-term precaution. This is like charting a broad route for the whole journey, focusing on the big picture and end goal. Finally, we can prevent AI developers from taking dangerous risks by putting in place specific structures and regulations, providing directions to those leading the journey. With these steps, humanity can wisely traverse the landscape ahead.

Among the many regulatory challenges of safe AI development, I focus on 'general control.'[2] This is the goal of ensuring that AI systems do what humans want them to do, also known as 'value alignment.'[3] This essay will only address the broad challenge of 'accident risks' from creating an unaligned AI that causes unintentional harm as opposed to 'misuse risks' or

---

[1] Toby Ord *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury Publishing, Great Britain, 2020) at 31.
[2] Matthew U. Scherer "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies" (2016) 29 Harvard Journal of Law & Technology 354 at 359.
[3] Stuart Russell *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Books, United States of America, 2020) at 137-138.

'structural risks.'[4] I will not discuss mitigating lesser risks such as employment or privacy, nor will I mention fairly distributing the benefits of AI advancements. While these issues are important, they pale in comparison to existential risk.

II. Mapping the Landscape

A. Existential Risk: Overarching Challenge

What is existential risk and why does it demand attention? Nick Bostrom defines an 'existential catastrophe' as an event that "causes human extinction or permanently and drastically curtails humanity's potential."[5] This would not only be a disaster for all 7.8 million people currently alive, but also for the trillions of people in the future who would no longer come into existence. However undeniably bad this is, we first need to know the chance of it happening. This is the 'risk' part of the equation. Is it high enough to deserve our thought, effort, and regulation? In his recent book, Toby Ord estimates a *one in six* chance of an existential catastrophe *this century*.[6] In other words, if humanity continues on its current course, it is leaving its fate to a dice roll.

Ord estimates the probability of an existential catastrophe caused specifically by AI at one in ten, the vast majority of the overall risk.[7] The most obvious way this could happen is through 'misalignment', where an objective function of a powerful AI system creates unintended and harmful outcomes.[8] Nick Bostrom imagines an example where an immensely powerful AI is given the objective 'make paperclips' and promptly turns all matter in the universe into paperclips, including humans.[9] A more realistic example is instructing an AI to cure cancer, only to have it give people cancer in order to conduct experiments.[10] Therefore, preventing the development of misaligned AI systems that pose an existential risk is crucial for safeguarding humanity's future.

B. Defining the Regulatory Target: Overarching Strategy

[4] Remco Zwetsloot and Alan Dafoe "Thinking About Risks From AI: Accidents, Misuse and Structure" (11 February 2019) Lawfare < https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure>.
[5] Nick Bostrom "Existential Risk Prevention as a Global Priority" (2013) 4 Global Policy 15 at 15-16.
[6] Above n 1 at 167.
[7] Above n 1 at 167.
[8] Above n 3 at 137-138.
[9] Nick Bostrom "Ethical Issues in Advanced Artificial Intelligence" (2003) 2 Cognitive Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence 12 at 16.
[10] Above n 3 at 138.

The first step in creating regulation that reduces existential risk from misaligned AI is properly defining the regulatory target. In other words, the law must focus on the type of AI that could cause an existential catastrophe. This is no small task. The definition of *any* kind of AI is notoriously contested. The most well-known definition is from Stuart Russel: "artificial intelligence is often used to describe machines or computers that mimic the cognitive functions that humans associate with the human mind, such as learning and problem-solving."[11] Similarly, the AI Forum New Zealand defines AI as "artificial digital technologies that enable machines to reproduce or surpass the abilities that would require intelligence if humans were to perform them."[12] It is also important to note there is a difference between 'narrow AI' which can only perform specific tasks and 'broad AI' (or artificial general intelligence) which has the wide-ranging and transferable cognitive abilities of a human.[13] While it is more likely that a broad/general AI causes an existential catastrophe, it is also conceivable that an extremely powerful narrow system could do the same. One of the most alarming possibilities is the advent of a 'superintelligence', an AI "that greatly exceeds the cognitive performances of humans in virtually all domains of interest."[14] One way in which general intelligence could become superintelligent is through 'recursive self-improvement', the ability of an AI system to make itself iteratively more intelligent.[15] Irving John Good describes a situation where "an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind."[16] Other less likely but still plausible scenarios are 'whole brain emulation' where human minds become digitally simulated,[17] or 'multi-agent systems' where a collection of AIs work together.[18]

It is clear from these myriad possibilities that the 'form' of an AI system is very hard to predict when regulating with existential risk in mind. Therefore, I propose that regulation instead targets the potential 'impact' of AI systems. A focus on impact rather than form allows regulation to be more future-proof against unanticipated forms of advanced AI. It achieves

[11] Stuart Russell and Peter Norvig *Artificial Intelligence: A Modern Approach* (4th ed, Pearson, New Jersey, 2021) at 1.
[12] AI Forum NZ "Introducing Aotearoa's Proposed AI Cornerstones" (29 April 2021) AI Forum New Zealand <https://aiforum.org.nz/2021/04/29/introducing-aotearoas-proposed-ai-cornerstones/>.
[13] Above n 3 at 46-47.
[14] Nick Bostrom *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, Oxford, 2014) at 53.
[15] Irving John Good "Speculations Concerning the First Ultraintelligent Machine" (1966) 6 Advances in Computers 31 at 33.
[16] Above n 15 at 33.
[17] Above n 14 at 56.
[18] K. E. Drexler "Reframing Superintelligence: Comprehensive AI Services as General Intelligence" (2019) Technical Report #2019-1, Future of Humanity Institute, University of Oxford at 53.

many benefits of 'technological neutrality' without the downsides.[19] This framework is also in line with Jonas Scheutt's idea of a 'risk-based' definition[20] and the European Commission's recent categorisation of AI into different risk levels.[21] In my view, the best application of this framework is making the regulatory target *'transformative artificial intelligence' (TAI)*. A broad definition of TAI is "AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution."[22] Experts give a ~20% chance that TAI will be developed by 2036, a ~50% chance by 2060, and a ~70% chance by 2100.[23] This regulatory target defined in terms of impact captures *any* form of AI that might pose an existential risk. Therefore, it is not an under-inclusive definition, which is a common problem.[24] On the other hand, the framing is possibly over-inclusive. However, given the magnitude of the risk at hand, over-inclusiveness is tolerable – perhaps even desirable – as a means of ensuring safety.

III. Charting the Course

A. Techno-legal Myopia: Broad Regulatory Challenge

Every journey into the unknown requires looking forward and anticipating the path ahead. Regulators are woefully unprepared for the journey of TAI development. Winston Churchill famously stated that "technology gets halfway around the world before moral philosophy can put its pants on." The same can be said for the relationship between law and technology: "the hare of science and technology lurches ahead … the tortoise of the law ambles slowly behind."[25] Nowhere is this dynamic more striking than in the field of existential risk. Governments spend more on ice cream each year than ensuring that emerging technologies do not destroy humanity.[26] Regarding AI in particular, "it could be said that public policy on artificial general intelligence does not exist."[27]. The European Commission's newly proposed

[19] Lyria Moses "Recurring Dilemmas: The Law's Race to Keep Up With Technological Change" (2007) 21 U. Ill. JL Tech. & Pol'y at 57.
[20] Jonas Schuett "A Legal Definition of AI" (2019) available at SSRN 3453632 at 7.
[21] EU Commission *EU Commission Proposals for Artificial Intelligence* (21 April 2021) European Commission <https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682>.
[22] Holden Karnofsky "Some Background on Our Views Regarding Advanced Artificial Intelligence" (6 May 2016) Open Philanthropy <https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence#Sec1>.
[23] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang and Owain Evans "When Will AI Exceed Human Performance? Evidence from AI Experts" (2018) 62 Journal of Artificial Intelligence Research 729 at 730.
[24] Above n 20 at 6.
[25] John H. Pearson "Regulation in the face of technological advance: Who makes these calls anyway?" (1999) 13 Notre Dame JL Ethics & Pub. Pol'y 1 at 1.
[26] Above n 1 at 32.
[27] Tom Everitt, Gary Lea and Marcus Hutter "AGI Safety Literature Review" (2018) International Joint Conference on Artificial Intelligence available at arXiv: 1805.01109 at 19.

rules on artificial intelligence do not mention 'existential risk' at any point.[28] The Charter of the AI Forum New Zealand does not even contain the word 'safety'.[29] An appropriate diagnosis for this challenge is 'techno-legal myopia.' Regulators rush toward the landscape of TAI barely looking past their own feet, blind to the precipice that looms ahead.

B. Longtermism and Precaution: Broad Regulatory Strategies

To combat the challenge of techno-legal myopia, the law should widen its gaze to future possibilities. The philosophy of 'longtermism' posits that:

> "From a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years."[30]

Current regulations are disconnected from posterity and this must change. The law need only accept a very weak version of longtermism to warrant contemplating an existential catastrophe caused by TAI within the next 100 years. One way to achieve a more longtermist mindset is to formally acknowledge that existential risk is real and poses a real threat to humanity, especially those who are not yet born.

How might this look in practice? The European Commission's new rules outline three levels of risk from AI and state that an 'unacceptable risk' is:

> "The placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm."[31]

While it is possible to interpret this section as capturing existential risk, further clarity may be beneficial. 'Physical harm' could include 'death on a global scale' and 'a person' could be defined as 'any human being who currently exists or may exist in the future.' Perhaps this rephrasing requires that a fourth level of risk be added to the existing framework.

Additionally, regulatory precaution should take an expected value approach to future possibilities. There is significant uncertainty regarding the development of TAI and the advent

[28] Above n 21.
[29] AI Forum New Zealand "Charter Document" (October 2018) AI Forum NZ <https://aiforum.org.nz/wp-content/uploads/2018/10/AIFNZ-Charter-approved-24-October-2018.pdf>.
[30] Nick Beckstead "On the Overwhelming Importance of Shaping the Far Future" (Doctor of Philosophy Dissertation) Graduate School New Brunswick 2013.
[31] Above n 21.

of an existential catastrophe. Expected utility theory provides a framework for making decisions under uncertainty. Put simply, an actor should compare the positive and negative impact of an action with the probability of it happening.[32] While putting concrete numbers on this calculation is outside the scope of this essay, it appears that any action which even minutely increases the chance of an existential catastrophe will have a net negative expected value regardless of its upside. Unfortunately, this action is precisely the one the law is taking now; allowing for the attempted development of powerful AI systems without strict regulatory safety mechanisms in place. The very same actions were taken when testing nuclear weapons despite speculation that the nuclear chain reaction could *ignite the entire atmosphere*.[33] Such an approach to balancing risk and progress is analogous to throwing oneself blindly off a cliff expecting there to be a soft landing just below. Just because there happened to be a soft landing once, does not make the jump wise. And if we make this jump again, we may not be so lucky. The upside of our current permissive regulatory tilt is that we may see the benefits of TAI slightly sooner. However, when compared to the disastrous consequences of an existential catastrophe, expected value dictates that regulatory tilt should favour precaution over progress. There is significant disagreement about what the precautionary principle actually entails or whether a weak or strong version is appropriate. The idea of expected value is also helpful for resolving this issue; different levels of precaution are required for different technologies that have different worst-case scenarios. In this context, the worst-case scenario is perhaps the worst thing imaginable, which suggests a very strong version of the precautionary principle is appropriate. An example is Bostrom's 'Maxipok Rule': "maximise the probability of an 'OK outcome', where an OK outcome is any outcome that avoids existential catastrophe."[34]

How might this look in practice? An example of a relatively weak precautionary principle regarding environmental protection is principle 15 of the Rio Declaration:

> "Where there are threats of serious or irreversible damage, lack of full scientific certainty shall be not used as a reason for postponing cost-effective measures to prevent environmental degradation."[35]

An appropriately strong precautionary principle regarding TAI and existential risk that incorporates the Maxipok Rule might look like:

---

[32] R.A. Briggs "Expected Utility Theory" (15 August 2019) Standford Encyclopedia of Philosophy <https://plato.stanford.edu/entries/rationality-normative-utility/#DefExpUti>.
[33] Above n 1 at 91.
[34] Above n 5 at 19.
[35] *Rio Declaration on Environment and Development* A/CONF.151/26 (14 June 1992) at 3.

> "Where there is even a minute threat (>.000001% i.e. one a million) of an existential catastrophe resulting from the attempted creation of transformative artificial intelligence, full scientific certainty shall be not used as a reason for postponing restrictive safety measures being placed upon developers."

Such a principle could also be added to the European Commission's risk framework.

IV. Guiding the Journey

A. Legal Structures: Specific Regulatory Strategy

Current regulatory structures cannot keep pace with AI development and this is dangerous. It is imperative to proceed wisely, thinking carefully about each step forward. Matthew Scherer proposes regulation of AI in general through an Act (the Artificial Intelligence Development Act or AIDA) and a subordinate Agency.[36] The Act would lay out the overall purpose of AI regulation, establish the Agency, and delegate powers to it.[37] Made up of AI experts, the Agency would certify the safety of AI systems before developers could deploy them.[38] The agency would also be able to update legal rules and definitions with the ratification of legislators.[39] This structure addresses the need for strict oversight and flexibility. Unlike legislators, the members of the agency would be able to dedicate their time to ensuring safe AI development. Furthermore, their ability to update the Act would allow the law to stay connected to rapidly advancing technology. Perhaps most importantly, delegating substantive AI policy to an Agency run by experts addresses the issue of legislators not being equipped to deal with technical problems.[40]Adapting this framework to TAI would require a more specific Act (perhaps called the TAIDA) and an Agency of experts on all forms of powerful intelligence systems and existential risk. When it comes to the *enforcement* of regulations, however, this framework is inadequate. Scherer suggests that uncertified agencies be subject to limited tort liability and that any uncertified AI systems sold be subject to joint and several liability.[41] While this may be a good incentive structure, it is still possible that developers would release an uncertified system anyway. When it comes to existential risk, this possibility is unacceptable. Therefore, under the hypothetical TAIDA, certification of potentially transformative AI systems should be mandatory and any developer failing to comply should

[36] Above n 2 at 394.
[37] Above n 2 at 394.
[38] Above n 2 at 395.
[39] Above n 2 at 395.
[40] Above n 2 at 395.
[41] Above n 2 at 393.

face a high penalty. A potentially suitable penalty could be the crime of 'omnicide' (the intentional destruction of humanity).[42] This penalty would only make sense if the crime was *threatening* existential catastrophe rather than *causing* it, because if one actually occurred, the law would become irrelevant. Additionally, as the average developer would never release an AI system with the *intention* of creating an existential catastrophe, knowledge is a more appropriate mens rea for this offence. In fact, strict liability may even more suitable given that many developers are unaware of existential threats posed by their technologies.

B. Goals, Principles, and Rules: Specific Regulatory Strategies

The first regulatory step in directing AI developers as they move forward is defining the purpose of the hypothetical TAIDA. Stuart Russell makes the argument that while the general conception among AI developers is 'the more intelligent the better', it ought to be 'the more beneficial the better.'[43] Scherer supports this paradigm shift by arguing that regulation should:

> "ensure that AI is safe, secure, susceptible to human control, and aligned with human interests, both by deterring the creation of AI that lack those features and by encouraging the development of beneficial AI that includes those features."[44]

Simultaneously deterring unsafe AI and encouraging aligned AI is an example of Bostrom's principle of 'differential technological development': the idea that society should "retard the development of dangerous and harmful technologies … and accelerate the development of beneficial technologies".[45]

Combining these ideas, an appropriate purpose is as follows:

- Goal: TAI that benefits humanity
- Strategy: differential technological development
- Tactic: preventing developers from creating TAI with a 0.000001% causing an existential catastrophe by only permitting the development and deployment of AI that is:
  - Safe
  - Secure
  - Susceptible to human control
  - Aligned with human interests

[42] Phil Torres "International Criminal Law and the Future of Humanity: A Theory of the Crime of Omnicide (2019) available at SSRN 3777140.
[43] Above n 3 at 172.
[44] Above n 2 at 394.
[45] Nick Bostrom "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards" (2002) 9 Journal of Evolution and Technology 1 at 23.

The second step is creating specific regulations that will achieve the aforementioned tactic. There are four broad levels of market regulation: prescriptive, management-based, performance-based, and meta.[46]

1. Prescriptive regulation

It may be desirable to follow certain principles when creating AI systems. Three such principles are proposed by Russell:[47]

1. The only objective of AI should be to fulfil human preferences (the altruism principle).
2. AI should be initially uncertain about these preferences (the humbleness principle).
3. The ultimate source of information about human preferences should be human behaviour (the learning principle).

Stricter rules might require developers to train AI algorithms with particular methods to achieve alignment. Some possible methods are: inverse reinforcement learning, iterated amplification, and debate games.[48] The issue is that demanding specific training methods is so draconian and inflexible that it may completely halt AI development in its tracks. Reasonable regulation must ensure safety but cannot be so rigid as to prevent moving forward at all or denying the ability to adapt to a changing landscape. Therefore, it may be more appropriate to simply impose general principles. However, there may still be issues of inflexibility. For example, what if it is impossible to write uncertainty into an algorithm? A more mild form of regulation compared to principles or specific training methods could be simply prescribing broad goals to developers. For example, Open AI's charter[49] commits to 'long term safety.' Similarly, the Machine Intelligence Research Institute's research agenda[50] outlines their mission of ensuring that 'the creation of smarter-than human-intelligence has a positive impact.' Regulation could require developers to include goals such as these in their publicly available mission statements, thus creating accountability and an incentive to focus on safety. It is not yet clear what type of prescriptive regulations are most appropriate. But it is crucial to strike a balance between strictness and flexibility.

[46] Christopher Carrigan and Cary Coglianese "The politics of regulation: From new institutionalism to new governance" (2011) 14 Annual Review of Political Science 107 at 109.
[47] Above n 3 at 171-184.
[48] Above n 27 at 10-17.
[49] Open AI "Open AI Charter" (9 April 2018) Open AI <https://openai.com/charter/>.
[50] Nate Soares and Benya Fallenstein "Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda" (2017) Machine Intelligence Research Institute <https://intelligence.org/files/TechnicalAgenda.pdf>.

2. Management-based regulation

The regulatory Agency must remain updated on the progress of AI development and have means of certifying safety. For the first issue, an informative model is Helsinki's open AI 'registers'.[51] These registers outline:

> "what, where, and how AI applications are being used … which datasets were used for training purposes; how algorithms were assessed for potential bias or risks; and how humans use the AI services."[52]

A similar register could be created for any company attempting to develop powerful AI systems. Developers would be required to log their AI projects on this register well before they are complete and define their desired end product, making clear how close it might be to TAI. Having all AI organisations on this register would help the Agency know which ones to prioritise in their investigations. The means of certifying safety could come in the form of a 'risk audit.' Companies would have to disclose the following information:[53]

1. The complete source code.
2. A description of all hardware/software environments in which the AI has been tested.
3. How the AI performed in the testing environments.
4. Any other information pertinent to the safety of the AI.

Next, the Agency could verify that any prescribed regulations regarding the creation of the systems had been followed. Then, they could test the system as they deemed fit.

3. Performance-based regulation

Testing could be conducted through a performance-based system, giving algorithms tasks to perform in a completely simulated environment and verifying that the algorithm's actions are aligned with human preferences. This approach has two advantages. Firstly, it does not require a deep understanding of the algorithm's internal workings, only an observation of its actions, thus bypassing the 'black box problem.'[54] Secondly, if the algorithm was not aligned, it would not be about to do any damage in the real world. Such a testing system may require very advanced virtual realities. For example, simulating all of the world's resources and people, then giving the AI system a command such as 'make paperclips' and seeing if Bostrom's imagined catastrophe came to fruition. After completing this risk audit, the agency would only provide

[51] Luciano Floridi "Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki" (2020) 33 Philosophy and Technology 541 at 541.
[52] Above n 51 at 541.
[53] Above n 2 at 397.
[54] Yavar Bathee "Black Box and the Failure of Intent and Causation" (2017) 31 Harv. JL & Tech 889 at 889.

certification if they concluded that the AI system had a less than 0.000001% chance of causing an existential catastrophe. Only then would further development and deployment to the public be permitted.

4. Meta regulation

An appropriate focus for this type of regulation is the balance between strictness and flexibility. The law must ensure the safe development of AI while remaining malleable to new situations. One way to achieve this is through 'sunset clauses,' which would make regulations imposed by the TAIDA and Agency expire unless they were reaffirmed or changed. This would encourage regular re-examine of regulation and allow for any necessary changes. With sunset clauses in place, the law could maximise safety by imposing stricter prescriptive regulations and means of testing AI systems with the knowledge that – if they were flawed – changes would be made in the future.

V. Conclusion

We must carefully navigate the landscape of emerging technology. While regulators of AI face numerous challenges, strategies are available. By understanding the nature of existential risk and making TAI a regulatory target, we can create an accurate map. By responding to techno-legal myopia with longtermism and precaution, we can chart a course into the unknown. By creating a dedicated Act and Agency along with specific rules and principles, we can guide those leading the journey. If we fall from the precipice, we fail not only ourselves but trillions to come. But if we proceed with wisdom, posterity will be eternally grateful for their existence.

VI. Bibliography

A. Books

Nick Bostrom *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, Oxford, 2014).

Stuart Russell and Peter Norvig *Artificial Intelligence: A Modern Approach* (4th ed, Pearson, New Jersey, 2021).

Stuart Russell *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Books, United States of America, 2020).

Toby Ord *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury Publishing, Great Britain, 2020).

B. Journal Articles

Christopher Carrigan and Cary Coglianese "The politics of regulation: From new institutionalism to new governance" (2011) 14 Annual Review of Political Science 107.

Irving John Good "Speculations Concerning the First Ultraintelligent Machine" (1966) 6 Advances in Computers 31.

John H. Pearson "Regulation in the face of technological advance: Who makes these calls anyway?" (1999) 13 Notre Dame JL Ethics & Pub. Pol'y 1.

Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang and Owain Evans "When Will AI Exceed Human Performance? Evidence from AI Experts" (2018) 62 Journal of Artificial Intelligence Research 729 at 730.

Luciano Floridi "Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki" (2020) 33 Philosophy and Technology 541.

Matthew U. Scherer “Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies” (2016) 29 Harvard Journal of Law & Technology 354.

Nick Bostrom “Ethical Issues in Advanced Artificial Intelligence” (2003) 2 Cognitive Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence 12.

Nick Bostrom “Existential Risk Prevention as a Global Priority” (2013) 4 Global Policy 15.

Nick Bostrom “Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards” (2002) 9 Journal of Evolution and Technology 1.

Yavar Bathee “Black Box and the Failure of Intent and Causation” (2017) 31 Harv. JL & Tech 889.

C. Dissertations

Nick Beckstead “On the Overwhelming Importance of Shaping the Far Future” (Doctor of Philosophy Dissertation) Graduate School New Brunswick 2013.

D. Internet Resources

AI Forum New Zealand “Charter Document” (October 2018) AI Forum NZ <https://aiforum.org.nz/wp-content/uploads/2018/10/AIFNZ-Charter-approved-24-October-2018.pdf>.

AI Forum NZ “Introducing Aotearoa’s Proposed AI Cornerstones” (29 April 2021) AI Forum New Zealand <https://aiforum.org.nz/2021/04/29/introducing-aotearoas-proposed-ai-cornerstones/>.

Holden Karnofsky “Some Background on Our Views Regarding Advanced Artificial Intelligence” (6 May 2016) Open Philanthropy <https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence#Sec1>.

EU Commission *EU Commission Proposals for Artificial Intelligence* (21 April 2021) European Commission <https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682>.

Open AI "Open AI Charter" (9 April 2018) Open AI <https://openai.com/charter/>.

Nate Soares and Benya Fallenstein "Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda" (2017) Machine Intelligence Research Institute <https://intelligence.org/files/TechnicalAgenda.pdf>.

R.A. Briggs "Expected Utility Theory" (15 August 2019) Standford Encyclopedia of Philosophy <https://plato.stanford.edu/entries/rationality-normative-utility/#DefExpUti>.

Remco Zwetsloot and Alan Dafoe "Thinking About Risks From AI: Accidents, Misuse and Structure" (11 February 2019) Lawfare < https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure>.

E. Other Sources

Jonas Schuett "A Legal Definition of AI" (2019) available at SSRN 3453632.

K. E. Drexler "Reframing Superintelligence: Comprehensive AI Services as General Intelligence" (2019) Technical Report #2019-1, Future of Humanity Institute, University of Oxford.

Lyria Moses "Recurring Dilemmas: The Law's Race to Keep Up With Technological Change" (2007) 21 U. Ill. JL Tech. & Pol'y.

Phil Torres "International Criminal Law and the Future of Humanity: A Theory of the Crime of Omnicide (2019) available at SSRN 3777140.

Rio Declaration on Environment and Development A/CONF.151/26 (14 June 1992).

Tom Everitt, Gary Lea and Marcus Hutter "AGI Safety Literature Review" (2018) International Joint Conference on Artificial Intelligence available at arXiv: 1805.01109.