Enlightened Intelligence:
Towards an Ethical Framework for Strong AI

Maria Romanova Hynes

A Research Paper submitted to the University of Dublin,
in partial fulfilment of the requirements for the degree of
Master of Science Interactive Digital Media

2016

Declaration

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

I declare that the work described in this Research Paper is, except where otherwise stated, entirely my own work and has not been submitted as an exercise for a degree at this or any other university.

Signed: _____

Maria Romanova Hynes

13th May 2016

Permission to lend and/or copy

I agree that Trinity College Library may lend or copy this Research Paper upon request.

Signed: _____

Maria Romanova Hynes

13th May 2016

Acknowledgments

I would like to thank my supervisor, Dr. Mads Haahr, for his generous support and invaluable advice on how to improve this paper and my husband, Winter Romanov Hynes, for giving me all the love, help and encouragement in the world.

This paper is dedicated to my family and especially my mom, Alina Sukharnikova, and my *babushka*, Inna Romanova, without whom nothing of this would have been possible.

Summary

Considering Artificial Intelligence one of humanity's most ambitious scientific endeavours of our time, this paper argues that the fundamental problems of ethics should be at the heart of AI research. After answering the question of whether contemporary scientists believe that it is possible in principle to create an artificial mind and conducting a historic overview of the past and present advancements in the field, a new definition of intelligence that has an ethical dimension to it and an ethical framework within which further developments on AI might be undertaken will be proposed. This paper argues that human intelligence is inappropriate as the model of intelligence for AI as humans are notorious for destroying their social and natural worlds by engaging in wars and environmentally devastating activities. Instead this paper presents a model of intelligence, called Enlightened Intelligence, inspired by the Buddhist concept of *enlightenment*, that considers technology a part of the common environment and implies that the only form of truly rational and intelligent behaviour for thinking machines would be to overcome their immediate interests and to reach beyond themselves and out to others out of the desire to protect the environment they are themselves a part of. It will be finally suggested that an inter-disciplinary, cross-religious, cross-cultural study is needed to develop a complete operational ethical model that would meet the needs of AI.

Key-words: artificial intelligence, machine intelligence, AI, ethics, Buddhism, enlightenment, intelligence, rational behaviour, strong AI, history of AI.

Table of Contents

Introduction

"Conjectures are of great importance since they suggest useful lines of research."

Alan Turing

Google announces a self-driving minivan.[1] Google is developing a humanoid robot.[2] Robot surgeons perform successful operations on soft tissue all by themselves.[3] Artificial Intelligence creativity machine learns how to play Beethoven's "Ode to Joy" in diverse music genres.[4] Stephen Hawking and Elon Musk, among others, sign an open letter, titled *Research Priorities for Robust and Beneficial Artificial Intelligence*, calling for research on the prevention of potential hazards that the evolving technology of AI might bring to mankind.[5] Nick Bostrom calls for the necessity of ensuring that the super-intelligent computers of the future share human values before humans create the actual super-intelligent computers.[6]

There is no doubt that technology *is* getting smarter and it causes concern among leading scientists and philosophers. This paper seeks to outline a direction for future research in ethics in conjunction with the development of AI in response to this concern. Chapter I juxtaposes two opposing theoretical viewpoints with regard to the feasibility of creating Artificial Intelligence for the purpose of asking the fundamental question of whether it is wise at all to spend time devising ethical frameworks for AI. Chapter II examines the history of the field of Artificial Intelligence in order to show that there is a renewed hope of

[1] Popular Science. "Google Announces Self-Driving Minivan." Accessed May 12, 2016. http://www.popsci.com/googles-first-self-driving-minivan-is-coming.
[2] Popular Science. "Google's Human-Shaped Robot Takes First Walk Outside." Accessed May 12, 2016. http://www.popsci.com/google-sent-its-human-shaped-robot-outside.
[3] Science. "Video: Robot Surgeons Make a Big Advance." Accessed May 12, 2016. http://www.sciencemag.org/news/2016/05/video-robot-surgeons-make-big-advance.
Popular Science. "Autonomous Robot Performs Successful Surgery on Living Pig." Accessed May 12, 2016. http://www.popsci.com/new-robotic-surgery-tool-outperformed-human-surgeons.
[4] Open Culture. "Artificial Intelligence Creativity Machine Learns to Play Beethoven in the Style of The Beatles' 'Penny Lane'." Accessed May 12, 2016. http://www.openculture.com/2016/05/beethoven-in-the-style-of-the-beatles-penny-lane.html.
[5] Future of Life Institute. "An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence." Accessed May 12, 2016. http://futureoflife.org/ai-open-letter/.
[6] TED. "Nick Bostrom: What Happens When Our Computers Get Smarter Than We Are." Accessed May 12, 2016.https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are?language=en#t-973754.

endowing machines with sentience among AI specialists, inspired by recent scientific breakthroughs. Chapter III argues that the development of an ethical framework within which machine intelligence should be cultivated is absolutely necessary and must be put at the very heart of AI research. As part of this argument, I propose an ethical model, called Enlightened Intelligence and based on the Buddhist concept of enlightenment, and a new operational definition of intelligence, with a crucial ethical dimension to it. I therefore suggest that humans should be aiming at creating not duplicates of themselves but socially and environmentally responsible beings that are more ethical and more intelligent than their creators. The Enlightened Intelligence framework outlines what more ethical and more intelligent behaviour could imply. Finally, I argue that the finest models of ethical and moral conduct and intelligence found across cultures in diverse religious and philosophical systems should be investigated as part of a comprehensive inter-disciplinary, cross-religious, cross-cultural study conducted for the purpose of developing a complete operational ethical model that would meet the needs of AI.

I. Is Artificial Intelligence Possible in Principle?

"The search for the truth is in one way hard and in another way easy – for it is evident that no one of us can master it fully, nor miss it wholly. Each one of us adds a little to our knowledge of nature, and from all the facts assembled arises a certain grandeur."

Aristotle

Is the creation of an artificial mind a conceivable prospect? The aim of this chapter is to juxtapose two opposing points of view with regard to the feasibility of creating Artificial Intelligence for the purpose of asking the fundamental question of whether ethical frameworks for AI are needed at all. Following the procedure of *reduction ad absurdum*, it will be first proposed, in agreement with Ray Kurzweil's argument, that it is possible to create artificially intelligent agents, capable of thinking, learning and creating. Secondly, following Roger Penrose's argument, the possibility of creating such technologies will be disputed. By examining the limitations of both the algorithm and our current understanding of physics, it will be suggested that it is impossible to create a machine that would be truly reasoning, self-aware and capable of understanding. Finally, this view will be also challenged by showing that there is no evidence to prove that some new method of computation will not perchance succeed at endowing machines with sentience in the future.

How does one of the best known contemporary proponents of strong AI tackle this question? Ray Kurtzweil, computer scientist, inventor and a Director of Engineering at Google, bases his belief in the prospect of the creation of artificially intelligent computers on two laws which he outlines in his book, *The Age of Spiritual Machines*:

(1) "The Law of Time and Chaos," implying that the interval between salient events in *the history of the universe* increases (such as the evolution of gravity, the strong force and the electroweak force in less than a second after the Big Bang and the emergence of the first atoms some 300,000

years later) along with the increase in chaos/disorder according to the Second Law of Thermodynamics;[7]

(2) "The Law of Accelerating Returns," implying that the interval between salient events *in the history of civilisation* grows shorter as order increases.[8]

These two laws seem to contradict each other only at first sight. What Kurtzweil is stating is that "when there is a lot of chaos in a process, it takes more time for significant events to occur; conversely, as order increases, the time periods between salient events decrease."[9] Whereas the universe as a whole is dependent on the Second Law of Thermodynamics and hence is subject to Kurtzweil's first law, human civilisation manifests the growing order of the evolving technology itself and hence is subject to Kurtzweil's second law. The difference between the universe as a whole and human civilisation lies in the fact that whilst the former is a closed system, the latter is an open system that needs to draw its options from the 'diversity' of chaos manifested in the universe for its evolutionary process.[10]

In such a way, the Second Law of Thermodynamics, according to Kurzweil, does not contradict the prospect of creating artificially intelligent computers, moreover it only enhances it. As the universe grows more and more disorderly, life on planet Earth will grow more and more orderly, because "[evolution] builds on its own increasing order."[11] Consequently, as Moore's Law on Integrated Circuits states,[12] the capacities of computers will exponentially increase, and it will not take as long for them to develop into a sentient form of life as it has taken for humans, for "…human intelligence, a product of evolution, is far more intelligent than its creator"[13] and is capable of giving birth to another form of intelligence much quicker than evolution was capable of giving birth to human intelligence.

---

[7] Kurzweil, Ray, *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* (London: Texere, 2001), 29.
[8] Ibid., 30.
[9] Ibid., 29.
[10] Ibid., 31.
[11] Ibid., 32.
[12] Ibid., 22-25.
[13] Ibid., 47.

In such a way, Kurtzweil outlines the following 'recipe' for creating a mind: take the right set of formulas (namely, neural nets and evolutionary algorithms), add some built-in knowledge a machine can progressively build upon (assuming that the further knowledge-acquisition process is fully automated), and cleverly apply computation to the mix.[14] Thus strong AI takes the human mind as a model for artificial intelligence and believes that mental activity can be mimicked by some well-defined sequence of operations, because it is itself just a sequence of operations. Marvin Minsky, one of the fathers of the field of AI stated, for example, that humans are "actually machines of a kind whose brains are made up of many semiautonomous but unintelligent 'agents'."[15] Somewhat anticipating the neuroscientist Simon LeVay,[16] he writes in *The Turing Option*: "Minds are simply what brains do."[17] Minsky and Kurzweil believe that just by deciphering the workings of the human brain, people will eventually unveil and decode the intricate mechanism of the underlying architecture of the brain and will ultimately be able to recreate it in a machine, for there is no magic to it as far as science goes. Kurzweil's prediction in 1999 was that humans would build their first intelligent machines by 2020.[18] But is it not too far-fetched to propose that sentience could be inscribed into an algorithm?

Sir Roger Penrose, a mathematical physicist, philosopher of science and the author of the most powerful attack on AI yet written, *The Emperor's New Mind*, stands in drastic opposition to such proponents of strong AI as Kurzweil and would be of the latter opinion. For Penrose believes that our thinking process is not entirely algorithmic in nature, since it incorporates within itself intuition, instinct and insight – the mental qualities without which the outstanding mathematician, logician and philosopher Kurt Gödel would have been unable to prove his famous theorems of Incompleteness. So let us turn briefly to mathematics for the purpose of inquiring into the role of *insight* in the most rigid

[14] Kurzweil, *The Age of Spiritual Machines*, 101.
[15] Quoted in: The New York Times. "Marvin Minsky, Pioneer in Artificial Intelligence, Dies at 88." Accessed May 12, 2016. http://www.nytimes.com/2016/01/26/business/marvin-minsky-pioneer-in-artificial-intelligence-dies-at-88.html?_r=1.
[16] LeVay's quotation reads: "The mind is just the brain doing its job." Quoted in: Norden, Jeanette, *Understanding the Brain: Course Guidebook* (Chantilly: The Great Courses, 2007), 1.
[17] Minsky, Marvin, and Harrison, Harry, *The Turing Option* (London: Viking, 1992), 73.
[18] Kurzweil's last book published in 2012 is actually called *How to Create a Mind: the Secret of Human Thought Revealed* (London: Viking, 2012).

science of all and asking the question of whether there is something entirely non-algorithmic about the mind and the brain.

Gödel's story of discovery finds its roots in the beginning of the twentieth century when Bertrand Russell and Alfred N. Whitehead set about developing a highly formalized mathematical system of axioms and rules of procedure in order to prevent the paradoxical types of reasoning that led to Russell's own paradox.[19] The specific scheme they produced turned out to be rather limited, however the idea persisted: if a statement is true one must be able to prove it by way of following mathematical rules. David Hilbert extended the inclusion criteria for the scheme: he proposed that all correct mathematical types of reasoning should be incorporated. And then he asked the famous question that resulted in an overwhelming response from Hilbert's colleagues, namely: "*Is there some general mathematical procedure which could, in principle, solve all the problems of mathematics…one after the other?*"[20] [Emphasis added]

A number of mathematicians embarked on solving Hilbert's puzzle or battling the formalist approach to mathematics, among which the names of Turing, Church and Gödel are by far the most famous. While Turing formulated his concept of the Turing and Universal Turing machines (the latter being the modern computer) and showed that there are indeed classes of problems that cannot be solved algorithmically (for example, the non-stopping of the Turing machine action), Gödel arrived at the two Incompleteness theorems which struck a mortal blow to Russell's formalism, for he managed to show that a perfectly well-defined proposition that belonged to a formal mathematical system could be both true and formally unprovable.[21] How did he did he do that?

Complying with the formalists' logic and following the individual rules of procedure of a formal system, Gödel constructed a propositional function for some particular well-defined arithmetical statement $P_k(k)$ that asserted that there does not exist such an x for which x^{th} proof proves $P_k(k)$:

---

[19] Russell's Paradox states: A is the set of all sets which are not members of themselves. Is the very set A a member of itself?

[20] Penrose, Roger, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford: Oxford University Press, 1989), 34.

[21] For Gödel's theorems of Incompleteness see: Mind-Crafts. "Gödel's Incompleteness Theorems – A Brief Introduction." Accessed May 12, 2016. http://math.mind-crafts.com/godels_incompleteness_theorems.php.

$$\sim \exists x[\Pi_x \text{ proves } P_k(k)].^{22}$$

Assuming that one constructed the formal system well and laid out all the axioms and rules of procedure correctly, $P_k(k)$ can*not* be proven, for if there was a proof for it, it would mean that $P_k(k)$ would be false as an arithmetical proposition to begin with. Hence, one has to conclude that $P_k(k)$ must be a *true* statement even though it *cannot be proven* within the system[23] – an 'unfortunate' contradiction to what the formalists proposed!

Gödel's findings may give rise to feelings of existential uncertainty,[24] for they show that there are things that can never be proven, however, Penrose's understanding of the implications of the theorems is rather optimistic: yes, Gödel shows that some statements cannot be possibly proven by the very methods one trusts (rules); however, he also shows that the very way one can construct a statement might be true, even if one cannot prove its truthfulness by using the rules of the system.[25] What does it say about the human mind? The power of human understanding allows people to transcend the rules they were given before by virtue of pure insight, thus going beyond rigid algorithmic thinking which people feed into the computers they create.

The formalists' search was inspired by the desire to overcome ambiguity involved in the problem of deciding what counts as valid reasoning and what does not by way of pronouncing one final judgement upon it. However, Penrose argues, both Gödel and Turing showed that the one who holds the final judgement is not the enacted finite set of instructions (algorithm) but the human mind that somehow simply *knows* whether a mathematical statement is true or a particular Turing machine will stop or run indefinitely by "using insights into the meanings of operations."[26] What allows humans to know truth from falsity (or beauty from ugliness) is the act of *consciousness* that Penrose, being a dedicated Platonist, believes to be the ultimate "'seeing' of a necessary truth."[27] But what is it exactly that allows the brain to be more efficient than a Turing

[22] Penrose's notation in: Penrose, *The Emperor's New Mind*, 106.
[23] Penrose, *The Emperor's New Mind*, 108.
[24] Hofstadter, Douglas R., *Gödel, Escher, Bach: An Eternal Golden Braid* (New York: Vintage Books, 1979).
[25] YouTube. "Sir Roger Penrose - Consciousness and the Foundations of Physics." Accessed May 12, 2016. https://www.youtube.com/watch?v=eJjydSLEVlU.
[26] Penrose, *The Emperor's New Mind*, 116.
[27] Ibid., 445.

machine? Penrose believes that it is by harnessing quantum mechanical non-computable laws that the brain is capable of availing of such non-algorithmic process as, for example, deciding on the validity of an algorithm itself.

Quantum mechanics is by no means an easy science to grasp and I shall not attempt to dig deep into it. However, in order to fully appreciate Penrose's argument, it is necessary to discuss the role of quantum mechanical procedures in the brain that, according to him, might be implicated in giving rise to consciousness. The scientist's suggestion that there must be something quantum-mechanical to the human brain springs from the belief that humans must be part of the common world to which quantum laws apply along with the laws of classical physics. Unlike classical physics which deals with physical phenomena at an observable level, quantum mechanics, "a theory of uncertainty, indeterminism, and mystery,"[28] describes the behavior of molecules, atoms and subatomic particles.

One of quantum mechanics' fundamental differences from classical physics is that certain type of procedures (such as state-vector reduction), which Penrose denotes by **R**, introduce uncertainties and probabilities into the theory and hence they must be non-computable. Totally deterministic processes (like Schrödinger's evolution), which Penrose denotes by **U**, also play a part in quantum mechanics, but it is both **U** and **R** that are needed for the spectacular agreement with observation that quantum mechanics manifests. As one might expect, it is the non-deterministic, non-computable procedure **R** that is believed to play a role in conscious activities, however Penrose is uncertain about where exactly in the brain it might occur: "one clear place where action at the single quantum level can have importance for neural activity is the retina"[29] (for experiments have shown that a single photon 'landing' on the dark-adapted retina is enough to trigger a microscopic nerve signal);[30] his other hope is that

[28] Penrose, *The Emperor's New Mind*, 149.
[29] Ibid., 400.
[30] Ibid.

someday cells of single quantum sensitivity[31] will be discovered deep in the brain[32]…but for now both possibilities remain speculative.

Penrose goes even further in his argument, suggesting that our current understanding of physics is altogether not enough to describe the functioning of the human brain. According to his belief, even quantum theory is a "stop gap, inadequate in certain essentials for providing a complete picture of the world."[33] To illustrate this, he suggests that the deterministic equations of classical physics and the deterministic operation of **U** in quantum mechanics betray their inadequacies when it comes to the problem of time flow.[34] For "…whatever physics is operating, it must have an essentially time-asymmetrical ingredient,"[35] whereas the above-mentioned theories are time-symmetrical, meaning that for them the future determines the past in just the same way as the past determines the present, thus contradicting our experience and observation (a broken glass of water does not seem to be able to reassemble itself into its initial whole state). Ascertaining that the new theory, which Penrose calls *Correct Quantum Gravity* (CQG), has to be time-asymmetrical, he argues that some new procedure taking place at the quantum-classical borderline is to be found; a procedure that will contain a non-algorithmic, non-computable element implying that "the future would not be computable from the present, even though it might be determined by it."[36]

Thus having come full circle, Penrose arrives at his initial claim that consciousness cannot be evoked by a mere algorithm (albeit a very sophisticated one), for there is something non-computable about the human brain and the world we live in that yet has to be discovered. The quantum mechanical procedure **R** is an echo of that truth, as it must be a quantum gravity effect.[37] Going against the strong AI contention, he proposes that indeed the human brain is algorithmic, but only to a certain degree, namely when it comes to

[31] In particular, Penrose cites tiny structures in the neurons called microtubules. See: "Sir Roger Penrose - Consciousness and the Foundations of Physics."
[32] Penrose, *The Emperor's New Mind*, 400.
[33] Ibid., 226.
[34] Ibid., 306.
[35] Ibid., 304.
[36] Ibid., 431.
[37] Ibid., 366.

handling already mastered skills or learnt information (and it is the cerebellum, an ancient part of the human brain, that is in charge of this). The function of consciousness, however, is to form new judgements and to transcend the rules that were laid out beforehand.[38] And it was Nature itself that for some reason has evolved sentient beings like ourselves rather than mere automatons. "Consciousness is the phenomenon whereby the universe's very existence is made known,"[39] and it cannot be described by any algorithm. But does Penrose really rule out the possibility of creating artificially intelligent computers that can think, learn and create, by showing that the human mind avails of such non-algorithmic resources as insight; resources that are unavailable to modern-day computers?

As of 2016, the paradigm of computation remains inherently algorithmic in nature. Neither quantum computers, nor neural networks break free from the limitations of algorithmic procedure and thus do not signal a shift in paradigm which would possibly allow humans to create artificially intelligent machines; for as Penrose shows, consciousness and awareness go beyond following a well-defined set of instructions and imply understanding. However, the very possibility of developing a different method of computation, based on humans' better knowledge of what consciousness is, cannot be ruled out. Penrose himself admits that his argument against strong AI extends only to its current state of the art, rooted in the algorithmic paradigm:[40]

> If we ever do discover in detail what quality it is that allows a physical object to become conscious, then, conceivably, we might be able to construct such objects for ourselves – though they might not qualify as 'machines' in the sense of the word that we mean it now.[41]

Although acknowledging that Penrose's critique is indeed powerful, Kurzweil dismisses his two conjectures to human superiority over algorithms: first, he argues, problems of Gödelian type cannot be solved either by humans or computers, humans can only estimate them but computers make estimates too; secondly, if the human brain avails of quantum computing, then human neurons

[38] Penrose, *The Emperor's New Mind*, 411.
[39] Ibid., 227.
[40] "Sir Roger Penrose - Consciousness and the Foundations of Physics."
[41] Penrose, *The Emperor's New Mind*, 416.

exhibiting quantum action will be eventually replicable in a machine.[42] Whilst it is not yet known whether it is possible to create an artificial mind, people do acknowledge the visible growth in the capacity of their technology. The latter would be one of the very few points on which Penrose is actually in agreement with the supporters of strong AI,[43] however, he is still sceptical of their mere reliance on the number of transistors in a computer built to replicate the number of neurons in the human brain to simulate consciousness, for he writes:

> If we believe that it is simply the largeness of the neuron number that allows us to have conscious experiences, whilst present-day computers do not seem to, then we have to find some additional explanation of why the action of the *cerebellum* [in which about thirty thousand million neurons are found, $3*10^{10}$] appears to be completely *un*conscious, while consciousness can be associated with the *cerebrum*, which has only about twice as many neurons ($7*10^{10}$), at a much smaller density.[44] [Original emphasis.]

The argument between opponents and proponents of strong AI can go on for much longer without each running out of conjectures to make, still neither one nor the other can completely disprove the opposite view, because one can only disprove what one knows. As of now, science and philosophy do not have the definitive answer to the questions of where consciousness arises from or even of what it is, yet technology does evolve fast, and what was not possible for it yesterday is possible for it today, so the dream of creating artificially intelligent machines persists and even seems to become increasingly real as time goes by. One conclusion, however, can be drawn: if there is even the slightest possibility of succeeding at creating truly intelligent machines, humans must ensure in advance that they will not be a threat to humans, other animals and the environment.

[42] Kurzweil, *The Age of Spiritual Machines*, 111.
[43] Penrose, *The Emperor's New Mind*, 396.
[44] Ibid.

II. True Intelligence versus AI Applications

"Face to face with mind as artifact we're face to face with almost more themes in the human experience than we can count or comprehend. And there's the added zest that this idea may turn out to transcend the human experience altogether and lead us to the metahuman."

Pamela McCorduck

The concept of artificial intelligence implying the possibility of creating computational systems capable of thinking, learning and creating seems to be futuristic even to modern day science. Notwithstanding the fact that science has advanced dramatically since the first formulation of one of the main questions in the field of AI, i.e. "Can machines think?" in Turing's article "Computing Machinery and Intelligence,"[45] the idea of having machines that can act independently of their human programmers is still science fiction. Kurtzweil might have been correct in predicting growing sophistication of AI techniques and associated hardware design:[46] virtual personal assistants like Apple's Siri or Google's Voice Search, AI applications in computer games such as *Far Cry* and *Call of Duty*, Google and Tesla's self-driving smart cars and even Nest Labs' programmable, self-learning, sensor-driven security systems are becoming increasingly more familiar to our contemporaries. However, the mystery of what consciousness or sentience really are has not yet been unveiled and all of the aforementioned technologies work only within the bounds of their particular, limited programs that cannot be compared with the potential of human intelligence for multipurpose activities. They are AI *applications*, in other words, means to an end, and not independently reasoning, self-aware, *intelligent* actors in the world, or ends in themselves in Kantian terminology.[47]

The aim of this chapter is to distinguish between the centuries-old ambitions of breathing life into true, autonomous, artificially created, intelligent

[45] Turing, Alan M., "Computing Machinery and Intelligence," in *Artificial Intelligence: Critical Concepts Vol. II*, ed. Ronald Chrisley (London: Routledge, 2000), 19.
[46] See: Kurzweil, *The Age of Spiritual Machines*.
[47] Kant, Immanuel, *Grounding for the Metaphysics of Morals (Third Edition): with On a Supposed Right to Lie because of Philanthropic Concerns*, trans. James Wesley Ellington (Indianapolis: Hackett Publishing Company, 1993).

agents and the day-to-day reality of ever-emerging AI applications. *The narrow claim* of AI that only certain mental processes, involving the exercise of intelligence, are computationally realisable will be matched against *the wide claim* of AI which proclaims that all mental states are so realisable.[48] It will be discussed in particular how the narrow claim leads to the development of AI applications in order to outline the existing models of artificial intelligence. Finally, the chapter will close with a contemplation on how recent breakthroughs in the field have led to the revival of the wide claim and the renewal of the hope to create truly intelligent machines. But before proceeding to an overview of current AI technologies and ideologies, the history of human imagination with regard to the mechanical 'other' should be addressed in passing, for it is that history that has informed our present attitudes.

As far as cultural records trace, humans have always asked fundamental questions about their own nature. For a long time, those questions pertained to separating *Homo sapiens* as a special kind of species from the rest of the animal kingdom and attributing superior characteristics to them. René Descartes, well known for believing that non-human animals were mere *automata*, famously declared that if not for the immortal and immaterial 'soul,' humans would also be nothing but machines.[49] Having largely dispensed with extreme anthropocentrism since and having not found the mysterious substance called 'soul' in the human body, philosophy and science still cannot answer the question of whether we are but mere machines living out our biological functions, albeit of some impressive complexity. Furthermore, if there is nothing extraordinary about the workings of the brain or human mental phenomena in general, which have been deciphered at an unprecedented pace in the twentieth and twenty-first centuries, then maybe it is indeed possible to recreate consciousness? For obvious reasons, answers to these riddles will have a direct bearing on the future development of AI.

[48] Torrance, Steve, "Ethics, Mind and Artifice," in *Artificial Intelligence: Critical Concepts Vol. IV*, ed. Ronald Chrisley (London: Routledge, 2000), 285.
[49] Mazlish, Bruce, "The Man-Machine and Artificial Intelligence," in *Artificial Intelligence: Critical Concepts Vol. I*, ed. Ronald Chrisley (London: Routledge, 2000), 134.

But how old are these inquiries and how old is the dream of creating artificial non-organic life, the mechanical 'other' glancing at its human creator from the other side of the mirror? To a Western student, this discourse might seem inherently European. They learn as children about the fortunes and misfortunes of such literary characters as Pinocchio and Frankenstein's monster,[50] first encountering the clash between 'artifice' and 'natural creation.' Some might venture even further and familiarise themselves with Hans Christian Andersen's tale "The Nightingale"[51] in which the dichotomy between that which is born and that which is constructed is fleshed out in the figures of a human-made nightingale that is finite in its life and faulty and a real bird whose singing is spontaneous and eternal, as life conceives more life and the singing of a beautiful nightingale never dies.

In all these stories, the mechanical 'other' seems to embody both human fascination with the possibility of an artificial creation and also human fears of its potential of transcending the boundaries established by humans and endangering *Homo sapiens*' superiority over other forms of life. The earliest recorded parable of *automata* dates back to the third century BC and comes in fact from Ancient China that is known to have abound in early mechanical toys such as flying dragons, hydraulically-moving boats and puppet orchestras among other things.[52] In a passage from the *Lieh Tzu*, a certain artificer Yen Shih presents before King Mu of Chou his handiwork which the latter mistakes for a human:

> The king stared at the figure in astonishment. It walked with rapid strides, moving its head up and down, so that anyone would have taken it for a live human being. The artificer touched its chin, and it began singing, perfectly in tune. He touched its hand, and it began posturing, keeping perfect time. It went through any number of movements that fancy might happen to dictate.[53]

[50] Mazlish, "The Man-Machine and Artificial Intelligence," 140-44.
[51] Andersen, Hans Christian, "The Nightingale," in *The Annotated Hans Christian Andersen*, ed. Maria Tatar (New York: W. W. Norton & Company, 2007).
[52] Mazlish, "The Man-Machine and Artificial Intelligence," 135.
[53] Ibid.

So the fact is that the dream (or for some, like Nick Bostrom, nightmare)[54] of designing artificial life is as old as the water wheel. For it seems to be only natural for beings who are aware of themselves and their mortality to be questioning and challenging the limitations of their nature either by coming into contact with other living things brought into existence by nature or by contemplating the possibility of becoming "the great Author of Nature"[55] themselves.

Despite its deep roots in humanity's cultural imagination, it was not until the middle of the twentieth century that AI emerged in the West as a scientific field. S. Russell and Norvig state that the first work that is now generally recognised as AI was done by Warren McCulloch and Walter Pitts in 1943. It drew on the basic function of neurons in the human brain, a formal analysis of B. Russell and Whitehead's propositional logic as formulated in *Principia Mathematica* and Turing's theory of computation.[56] However, the full vision of AI was formulated in Turing's revolutionary paper "Computing Machinery and Intelligence" from 1950, for it not only argued that it is possible in principle to create machines that can think but it also defined a measurement for machine intelligence. From then onwards, the discipline grew quickly in a number of American universities such as Dartmouth College, the official birthplace of AI, MIT and Stanford among others.[57]

One document from the early days of AI is of particular importance, as it defined in non-ambiguous terms what the field's goals and aspirations were. This paper by J. McCarthy, M. L. Minsky, N. Rochester and C. E. Shannon, titled "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," was produced in 1955 for the purpose of securing funds for a "2 month, 10 man study of artificial intelligence."[58] The researchers' agenda was formulated thus:

---

[54] See: The New Yorker. "The Doomsday Invention: Will Artificial Intelligence Bring Us Utopia or Destruction?" Accessed May 12, 2016. http://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom.

[55] Mazlish, "The Man-Machine and Artificial Intelligence," 136.

[56] Russell, Stuart J., and Norvig, Peter, *Artificial Intelligence: A Modern Approach* (New Jersey: Pearson Education, 2010), 16.

[57] Ibid., 17.

[58] McCarthy, J., Minsky, M. I., Rochester, N., and Shannon, C. E., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," in *Artificial Intelligence: Critical Concepts Vol. II*, ed. Ronald Chrisley (London: Routledge, 2000), 44.

> The study is to proceed on the basis of the conjecture that *every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.* An attempt will be made to find how to make machines use language, form abstractions and concepts, solve the kinds of problems now reserved for humans, and improve themselves.[59] [Emphasis added]

Following from this statement, it can be suggested that the founders of AI supported what Steve Torrance called *the wide claim* of the discipline according to which "any and all aspects of [human] mentality can in principle be realized on computer systems of some arbitrary degree of complexity – or at least they can be explained in computational terms."[60] This observation is significant for it tells one something about the belief of the first AI researchers in the possibility of replicating human-like consciousness in a machine. In particular, their research was meant to contribute to the following aspects of the artificial intelligence problem: (1) improving programs in general; (2) finding out how to program a computer to use language; (3) doing more theoretical work into neuron nets; (4) providing a theory and criteria for efficiency of calculation; (5) exploring the problem of self-improvement; (6) describing machine methods for forming abstractions; (7) conjecturing how randomness and intuition can foster creative thinking in a machine.[61] In such a way, their research interests can be grouped under four rubrics which define what AI set to achieve from the dawn of its day: creating intelligent programs inspired by *brain models*, developing applications for *machine learning*, nurturing *creativity* in machine performance and studying the relation of *language* to intelligence.

What distinguished AI from other disciplines and prevented it from becoming a branch of mathematics or computer science is explained by AI's ambition to "[embrace] the idea of duplicating human faculties like creativity, self-empowerment and language use."[62] This is reflected in the "imitation game" as an operational test for intelligence, proposed by Turing and now known as the Turing Test, in the course of which one particular digital computer pretends to be a human while a human interrogator communicating simultaneously with the

[59] McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," 44.
[60] Torrance, "Ethics, Mind and Artifice," 287-288.
[61] McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," 44-45.
[62] Russell and Norvig, *Artificial Intelligence: A Modern Approach*, 18.

computer and another human being tries to decide which is computer and which is man.[63] Indeed, like in the tale about the artificer Yen Shih, whose 'robot' is made act like a human and who even advances to the ladies in the King's court thus trespassing the limits of what is permitted and presenting a sexual threat to the men,[64] the "imitation game" exemplifies the early definition of intelligence adopted by the field as being modelled after human intelligence: if a computer can be mistaken for a human in the course of blind communication then the computer can be considered intelligent.

There have been multiple attempts to formulate an adequate operational definition of intelligence that would be more fitting than that of Turing's, for notwithstanding its centuries-old legacy, the idea of modelling AI after human intelligence is problematic for a number of reasons. First, being "the only practical clue to the nature of intelligence which is readily available,"[65] human intelligence is nonetheless fraught with contradictions, since too often humans do not maximize their positive impact on the environment as rational agents are supposed to do according to S. Russell and Norvig's definition of rational behaviour for AI technologies, formulated thus: "For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has."[66] Secondly, analogous with artificial flight, the successful implementation of which was due to the Wright brothers' knowledge of aerodynamics and not imitation of a wing-flapping motion, it is necessary to understand intelligence as such rather than human intuitive knowledge of their own capacities.[67] Thirdly, because very little is known about the nature of human intelligence, a useful definition of it simply cannot be produced. Instead, what one might hope for is an account of

[63] Turing, "Computing Machinery and Intelligence."
[64] Mazlish, "The Man-Machine and Artificial Intelligence."
[65] Whitby, Blay, "The Turing Test: AI's Biggest Blind Alley?" in *Artificial Intelligence: Critical Concepts Vol. IV,* ed. Ronald Chrisley (London: Routledge, 2000), 197.
[66] Russel and Norvig, *Artificial Intelligence: A Modern Approach*, 37.
[67] Whitby, "The Turing Test: AI's Biggest Blind Alley?" 199.

intelligence "involving some sort of comparison with human beings"[68] that would make "no direct reference to either humans or machines."[69]

So what are some of the definitions of intelligence that would fit the latter criteria? Alexander D. Wissner-Gross proposes that "intelligence is a physical process that tries to maximize future freedom of action and avoid constraints in its own future."[70] He even proposes a physical formula that would describe this process:

$$F=T\nabla S\tau,$$

where F is a force "that acts so as to maximise future freedom of action…with some strength T [and] the diversity of possible accessible futures S up to some future time horizon τ."[71] However, ethical considerations are totally missing from this definition. At whose expense future freedom of action is maximised? How does intelligence act if constraints in its own future cannot be avoided? McCarthy and P. J. Hayes provide a less vague but longer definition that declared an entity intelligent if "it had an adequate model of the world…if it was clever enough to answer a variety of questions based on this model…if it could get additional information from the external world when required, and if it could perform such tasks in the external world as its goals demanded and its physical abilities permitted."[72] This definition, precise as it is, also lends itself to a number of questions. What is meant by an adequate model of the world and how does one decide what can be regarded as such? How can the goals of an artificially intelligent agent and its physical abilities be balanced? This illustrates multiple problems inherent in the construction of intelligent agents and explains why the first attempts to build convincingly intelligent AI in the early days of the field failed.

Some insight into the high hopes of AI in its infant stages can be gained from the "Tomorrow" television series covering the latest developments in computer research and artificial intelligence produced by CBS for MIT in 1961.

---

[68] Whitby, "The Turing Test: AI's Biggest Blind Alley?" 202.
[69] Ibid.
[70] TED. "A. Wissner-Gross: A New Equation for Intelligence." Accessed May 12, 2016. https://www.ted.com/talks/alex_wissner_gross_a_new_equation_for_intelligence.
[71] Ibid.
[72] McCorduck, Pamela, "Robotics and General Intelligence," in *Artificial Intelligence: Critical Concepts Vol. I*, ed. Ronald Chrisley (London: Routledge, 2000), 467.

In the program, the pioneers of the field confess their expectations of thinking machines: Oliver Selfridge, for example, states that even though he does not expect his daughter to marry a computer, he is nonetheless "convinced that machines can and will think in our lifetime."[73] Claude Shannon, the father of information theory, is also rather optimistic on the matter: "I…expect that within…ten or fifteen years something will emerge from the laboratory which is not too far from the robot of science fiction."[74] One of the first non-numerical applications of computers, the Electronic Brain, which was meant to translate Russian into English at the height of the Cold War, promised to replace human translators "within five years or so."[75] However, putting a dictionary into a computer's memory proved to be insufficient for the construction of intelligible sentences.

Being a general purpose machine, the computer was thought to be able "to do things which in humans required intelligence,"[76] with much programming behind the scenes, of course. What the AI pioneers had to duly discover was that the amount of programming required for producing any kind of remotely intelligent behaviour was astronomical indeed. A good example of the field's further exploration of intelligent behaviour in the 1960s and 1970s was Stanford Research Institute's robot Shakey that could move from room to room, avoid obstacles, react to unforeseen circumstances, like somebody hindering its progress through space, and display some of the ability associated with intelligence such as planning and learning.[77] Project manager Nils Nilsson provided an excellent account of what the researchers learnt from Shakey as well as of the degree of complexity involved in creating computer-machine-sensor systems:

> …when we said, Shakey, move forward three feet, the only thing we could be absolutely sure of is that he did not move exactly three feet. He probably would move three feet plus or minus epsilon according to some normal distribution, depending upon the errors in the calibration and slippage in the wheels; but maybe he moves one and a half feet and runs

[73] MITvideo. "'The Thinking Machine' (1961) – MIT Centennial Film." Accessed May 12, 2016. http://video.mit.edu/watch/the-thinking-machine-1961-mit-centennial-film-6712/.
[74] Ibid.
[75] Ibid.
[76] Ibid.
[77] McCorduck, "Robotics and General Intelligence," 470.

> into the wall, or maybe he doesn't move at all because the commands got garbled in transmission, or his batteries are low. So there's an interesting research area that we made some progress on – how to build robust systems, and what kinds of monitoring are needed and how the system has to check whether it accomplishes what it tries to accomplish. We developed ways of using the TV camera and sensory feedback to monitor and update Shakey's own model of the world.[78]

The funding of the Shakey project was terminated in 1972 because allegedly no immediate military application could be perceived.[79] One important lesson, nonetheless, that the first experiments in the field of AI taught was that intelligence *is* remarkably difficult to recreate. It is for this reason that Turing's prediction that by the end of the twentieth century computers would be playing the "imitation game" successfully[80] did not come true. Instead, by the end of the twentieth century, in 1997 to be precise, a certain digital computer, known by the name of Deep Blue, defeated world chess champion Garry Kasparov in a game of chess and marked the first true glory of intelligence for digital computers. Yet, it would be a fallacy to believe that Deep Blue's intelligence satisfies any of the above definitions of intelligence, as the computer certainly did not venture into the big world beyond chess. Deep Blue was an AI application that performed one task well, but one task only, whereas true intelligence requires a profound versatility and plasticity of the mind and thought processes.

It can be argued that Deep Blue is representative of *the narrow claim* of the field, as defined by Torrance, "according to which that portion of human mentality which involves the exercise of intelligence…can be reproduced in working computer programs with complete fidelity, so that when a computer is displaying a certain kind of behaviour it is – *to this extent* – exemplifying genuine mentality." In other words, Deep Blue was programmed to display only a limited number of mental phenomena and not 'the mind.' For example, what lay at the heart of the program was an *evaluation* function allowing it to assess each given chess position and decide whether the situation on the board was overall more favourable to White or Black. The evaluation function looked at four basic chess values: the worth of a particular chess figure, its position, King's safety and

[78] McCorduck, "Robotics and General Intelligence," 472.
[79] Ibid., 473.
[80] Turing, "Computing Machinery and Intelligence."

tempo. Additionally, Deep Blue employed a system called *selective* extensions that made the computer's search of possible moves more efficient by selectively choosing only promising paths to follow. Finally, one certain advantage that Deep Blue had over the human brain was the speed of computation which was several orders of magnitude faster: the program generated up to 200,000,000 positions per second when calculating the most optimal move.[81]

Deep Blue is only one example of AI applications performing several chosen well-defined tasks rather than trying to simulate human consciousness or to pass the Turing Test. AI technologies have been implemented in a vast number of diverse areas, such as banking and finance, communications and documentation, construction, energy, manufacturing, military, operations management, palaeontology, security auditing, software engineering and transportation to list only a few.[82] According to James Lighthill, an applied mathematician who conducted a general survey of the field including "specialised reports on the contribution of AI to practical aims"[83] in the 1970s, the main areas of AI research can be grouped under three distinct categories.[84] Since the AI research done in the twenty-five years of the discipline's existence prior to Lighthill's investigation proved that a high degree of generality of application was near to impossible to achieve, he suggests that AI specialists realised that their program designs would have to utilise a large quantity of detailed knowledge about the problem domain instead,[85] which arguably has led to a fine-tuned diversification of the field allowing for the production of specialised applications.

The first category that Lighthill lists is called "Advanced Automation." It has practical, technological goals, it makes use of the general-purpose digital computer's logical potentialities and it takes as its objective "to replace human beings by machines for specific purposes, which may be industrial or military on

[81] IBM Research. "How Deep Blue Works: Under the Hood of IBM's Chess-Playing Supercomputer." Accessed May 12, 2016. https://www.research.ibm.com/deepblue/meet/html/d.3.2.html.
[82] *Innovative Applications of Artificial Intelligence 2*, ed. Alain Rappaport and Reid Smith (Cambridge: MIT Press, 1991).
[83] Lighthill, James, "Artificial Intelligence: A General Survey," in *Artificial Intelligence: Critical Concepts Vol. III*, ed. Ronald Chrisley (London: Routledge, 2000), 497.
[84] Ibid.
[85] Ibid., 506.

the one hand, and mathematical or scientific on the other."[86] The three dominant lines of inquiry identified for this category are *pattern-recognition activities*, *data storage and retrieval* and *problem solving*. In particular, work is conducted in the following domains: character recognition; speech recognition and synthesis; machine translation; product design and assembly; container packing; cryptography; guided missiles; exploration and action in hostile environments for military, space and firefighting purposes; the automation of problems of logical deduction including theorem proving, inductive generalisation and analogy spotting; analysis of chemical structures; graph traversing; generating improved methods for industrial and economic planning and decision making; machine learning.

The second category has fundamental, biological aims and is called **"**Computer-based studies related to the Central Nervous System (CNS) in humans and animals." It has as its objective "theoretical investigations related to neurobiology and to psychology"[87] and is concerned with using computer-based models of neural nets to develop such special functions as, for example, visual pattern recognition and scene analysis, visual and auditory memory, associative recall, psycho-linguistics, classification, inductive generalisation and learning.

The third category classified by Lighthill is what he calls a "bridge activity," justified by the links that it creates between "Advanced Automation" and "Computer-based CNS research." It bears the name of "Building Robots" and, more closely than the other two categories, reflects the long-standing human dream of creating the mechanical 'other.' Lighthill defines a robot as "an automatic device that mimics a certain range of human functions"[88] and highlights the following objectives that robotics concerns itself with: coordination and movement; visual scene analysis; use of natural language; 'common-sense' problem solving.

So how exactly are AI applications used for practical aims today? A few answers spring to mind. NASA's Remote Agent program became the first on-board autonomous planning program ever, capable of scheduling operations for a

[86] Lighthill, "Artificial Intelligence: A General Survey," 499.
[87] Ibid., 501.
[88] Ibid., 503.

spacecraft.[89] Self-driving cars, alluded to before, use autonomous control and computer vision systems to navigate themselves along a lane. Medical diagnosis programs based on probabilistic analysis are being constantly improved for diagnosing illnesses.[90] Even more impressive is the use of robot assistants in microsurgery.[91] The application of robotics extends to space exploration also, with the newly introduced R5: Valkyrie, a humanoid robot designed to perform in extreme environments and announced by NASA in 2015.[92] The problems of language understanding are tackled with the help of language translation software with more success than in the past.[93] But there is much more in store than that.

In November 2015, Google released TensorFlow, an open-source platform for machine learning that was built over the last three years. Google's software library is used in more than fifty Google products and now is accessible to anyone in the world with a computer and internet connection. TensorFlow utilises the technology of *deep learning* availing of deep neural networks and is intended to be used for the improvement of one of the most lucrative AI applications, a true digital assistant in a smartphone capable of speech recognition, search, vision detection, etc., thus unleashing the whole potential of navigating the digital world. And according to Dave Gershgorn, within Google, this massive undertaking passes under the name of machine intelligence, for what the company hopes to achieve is true, not artificial intelligence, just in a machine.[94] (It is worth mentioning that Google is not alone in exploring the potential of deep learning. Microsoft's Peter Lee says there is also promising early research on its uses for industrial inspection and robot guidance, developing personal sensors for predicting medical problems and sensors for cities that could tell, for example, where traffic jams might occur.)[95]

---

[89] Russell and Norvig, *Artificial Intelligence: A Modern Approach*, 27.
[90] Ibid., 28.
[91] Ibid.
[92] SPACE.com. "Humanoid Robot R5: Valkyrie 'Dances' in NASA Music Video." Accessed May 12, 2016. http://www.space.com/31270-humanoid-robot-r5-valkyrie-dances-in-nasa-music-video.html.
[93] Russell and Norvig, *Artificial Intelligence: A Modern Approach*, 28.
[94] Popular Science. "How Google Aims to Dominate AI: The Search Giant is Making its AI Open Source so Anyone Can Use It." Accessed May 12, 2016. http://www.popsci.com/google-ai.
[95] MIT Technology Review. "Deep Learning." Accessed May 12, 2016. https://www.technologyreview.com/s/513696/deep-learning/.

Google's other initiative, described by Demis Hassabis as "the Apollo program of artificial intelligence,"[96] is arguably even more daring. Hassabis leads a team of computer scientists and neuroscientists at Google's DeepMind, the group behind the AlphaGo software that defeated the reigning world champion at Go in 2016. What makes Go a remarkable landmark for AI to achieve is the number of possibilities that the game offers for each move on the board. Unlike chess, that allows a player about 35 options per turn, Go accounts for 250, with more plausible Go positions than there are atoms in the universe.[97] In such a way, players have to rely on their intuition as well as calculation powers. How did the program do it? Like TensorFlow, AlphaGo avails of deep learning that attempts to mimic the activity of the neocortex.[98] But there is an added technique called *reinforcement learning* permitting the program to explore a new environment and adjust its behaviour to it, which results in the computer generating its own algorithm from the examples it has learnt. Thus DeepMind's software succeeded not only in the Go game world but also in those of Atari arcade games, such as "Space Invaders," likewise defeating human opponents.

Hassabis believes that the reinforcement learning approach can be used for a number of commercially more viable applications: DeepMind is currently cooperating with the U.K.'s National Health Service on a software for recognising signs of kidney malfunction as well as working with business divisions of Google aiming to improve recommendation systems (for products like YouTube or for advertising). There are also hopes of enhancing algorithms used in robotics that would permit robots to understand their environments better.[99] Hassabis' biggest ambition, however, is to create general artificial intelligence that, like a human, could learn how to solve a vast range of problems. It is for this reason that DeepMind has an internal ethics board of philosophers, lawyers and businesspeople whose job is to find keys to hard questions of a philosophical and ethical nature.

---

[96] MIT Technology Review. "How Google Plans to Solve Artificial Intelligence." Accessed May 12, 2016. https://www.technologyreview.com/s/601139/how-google-plans-to-solve-artificial-intelligence/.
[97] Ibid.
[98] "Deep Learning."
[99] "How Google Plans to Solve Artificial Intelligence."

Notwithstanding a certain amount of scepticism coming from people like Jean-Christophe Baillie who opposes the renewed high hopes of AI by reminding one that true intelligence requires not only sophisticated learning skills but also embodiment and the ability to communicate,[100] the last couple of years of research in the field has led to a renaissance of *the wide claim* stating that it is indeed possible in principle to recreate all the aspects of mentality in a computational system and achieve real machine intelligence. Scientists do acknowledge that the brain is much more complex than any of the existing neural networks but the amount of progress is nonetheless reassuring. The very idea of creating intelligence in a machine leads to fundamental ethical considerations that ultimately set the trajectory of research and ambitions regarding what humans desire to achieve. And as AI applications become increasingly more sophisticated and the centuries-old dream of creating artificial non-organic life persists, researchers realise that the problems of ethics lie at the very heart of the discipline together with computer science, mathematics, physics, biology and psychology.

[100] MIT Technology Review. "Five Lessons from AlphaGo's Historic Victory." Accessed May 12, 2016. https://www.technologyreview.com/s/601072/five-lessons-from-alphagos-historic-victory/.

III. An Enlightened Machine

> "One individual can conquer the entire world of objects, but he cannot conquer another person without destroying him as a person. The individual discovers himself through this resistance. If he does not want to destroy the other person, he must enter into communion with him. In resistance of the other person the person is born."
>
> Paul Tillich

The previous chapter concluded with a brief consideration of why AI research should concern itself with the fundamental problems of ethics. The aim of this chapter is to expound this claim and to propose an ethical framework, based on the Buddhist model of enlightenment and called Enlightened Intelligence, that might be considered as appropriate for AI research. Furthermore, it will be argued that a cross-disciplinary, inter-religious, inter-cultural study is necessary for developing a complete operational ethical model that would meet the needs of AI. Since neither the proponents of strong AI nor their opponents deny the possibility of creating artificially intelligent machines (although Penrose's argument, outlined in Chapter I, seems indeed to be persuasive) and there is a renewal of interest in creating truly intelligent computational systems among AI specialists in the present, I argue for the importance of developing an ethical system within which machine intelligence should be cultivated. For if there is even the slightest possibility of succeeding at this ambitious endeavour, humans should ensure in advance that the autonomous agents they create are truly rational in the sense proposed by S. Russell and Norvig, i.e. rational behaviour in a machine implies that it maximises its own performance, while minimising its negative impact on the environment.[101]

Such a definition of rationality echoes Andrew Feenberg's tenth paradox of technology, formulated thus: one violating the environment around oneself falls victim to one's own violent assault, for both the actor and the environment belong to the same system.[102] This implies that it would be irrational in principle

[101] Quoted earlier in Chapter II. See footnote 65.
[102] Vimeo. "Ten Paradoxes of Technology." Accessed May 12, 2016. https://vimeo.com/28022711.

to hurt the system oneself belongs to, as it would have a direct negative bearing on the actor. Unfortunately, this reminds one once more about the argument outlined in Chapter II for the inappropriateness of human intelligence as the model for machine intelligence: humans are notorious for violating their social and natural worlds. Climate change, wars and pollution are repeatedly stated as the consequence of that.[103] Hence, what should lie at the centre of any definition of intelligence "involving some sort of comparison with human beings"[104] that would make no direct reference to either machines or humans is *an ethical dimension* implying socially and environmentally responsible action.

Technology is power,[105] and the powerful technology of AI (a sentient form of life to be, if ever) should never stand in opposition to life, be it the life of nature, humans or other species on the planet. It should transcend the Darwinian rule of survival (as the power of consciousness is able to transcend pre-set rules) and embrace its own "joined kinship"[106] with animals and their fragile environment, for technology is part of our shared evolution. What are some of the ethical considerations that should be incorporated into the development of AI? I propose that the very notion of *transcendence* might be the key – understood in terms of *going beyond oneself*, of *enlightenment* – and it is transcendence that should be studied across various cultures in conjunction with research on AI in order to endow reasoning machines with the understanding of their belonging to the complex system that they will share with humans and other animals and that gives life that should be nurtured, not hindered. Transcendence, or enlightenment, thus should not be understood as an ethereal term but a concrete notion of one conscious being reaching out to another being on the basis of them having overcome their personal egotistic cravings and wanting to enter into communion with the world out of love for it.

To apprehend some of the possible initial reactions at what might seem as the absurd thought of conceiving of enlightened machines, a passage from the

[103] See for example: TED. "Sylvia Earle: My Wish: We Must Protect Our Oceans." Accessed May 12, 2016. https://www.ted.com/talks/sylvia_earle_s_ted_prize_wish_to_protect_our_oceans.
[104] Quoted earlier in Chapter II. See footnote 67.
[105] See: Haraway, Donna J, "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century," in *Simians, Cyborgs and Women: The Reinvention of Nature* (London: Free Association Books, 1991).
[106] Ibid., 177.

beautiful article "Ethics, Mind and Artifice" by Torrance can be quoted in support of my argument:

> It is not clear whether any such computational ethical modelling is likely ever to be either convincing or useful, but it may be that only through becoming actively concerned with building an explicit ethical orientation, and moreover a humanitarian ethical orientation, into AI systems that there is any hope that the AI technological paradigm can maintain any pretence at being a humanizing influence in our civilization.[107]

It can be argued that the field of AI and the world in general would benefit greatly in the long run from taking its inspiration from the underlying fascination with the workings of human thought and some of its purest creations, devoid of the desire for wealth, blood and power, rather than from developing increasingly more sophisticated systems of weaponry and warfare or becoming subservient to the hunger for monetary gain. I propose that Buddhism as a philosophical framework offers some valuable insights into what might be considered ethical and moral behaviour satisfying the following new definition of intelligence, motivated by Wissner-Gross and S. Russel and Norvig's definitions of intelligence and rational behaviour, with a required ethical dimension added to it: intelligence is a physical process that tries to maximize future freedom of action *while minimising its negative impact on the environment* for the purpose of avoiding constraints in its own future *and the future of other beings in the world*. This definition, unlike the others quoted in Chapter II, takes into account Feenberg's tenth paradox of technology implying that the result of the actor's action in the world will be always felt by the actor themselves, be it negative or positive. Hence it is in the actor's personal interest to ensure that their evaluation of their future actions transcends the short-sighted regard for their own immediate gain.

There is no doubt that Buddhism alone is insufficient for developing a complete operational ethical model that would meet the needs of AI and it is for this reason that I shall try to suggest a broader way of interpreting some of its useful concepts by marrying them with ideas taken from other ethical, religious and philosophical systems. In particular, I shall be drawing on Lynn de Silva's research, outlined in the book *The Problems of the Self in Buddhism and*

[107] Torrance, "Ethics, Mind and Artifice," 301.

Christianity, that opened a much needed dialogue between Buddhist and Christian worldviews. I would like to stress, however, that the present study is *secular* and not religious in character. I believe that some of the religious concepts are essential for the understanding of ethics and, as part of general human wisdom, must not be ignored, as they can be used for constructive aims not only in a theological dialogue, but agnostic or atheistic alike. Furthermore, as "an expression of man's relation to the limits of his own existence,"[108] religion is somewhat akin to AI in its endeavour to understand what it means to be a human being.

What is it that makes Buddhism appealing as an ethical framework for study in conjunction with AI? Before suggesting an answer to this question, the basic ideas of Buddhism should be explained. Four points of interest for AI research will be first highlighted and then expounded on at the end of the chapter. The present account of Buddhism will make no distinction between the Hīnayāna and Mahāyāna schools and will draw on the teachings of the Buddha as they were recorded in the Sanskrit and Pāli scriptures, the earliest written accounts. All the Buddhist terms used in the text are Pāli words.

Five hundred years before the birth of Christianity, a new religious doctrine, commonly regarded as an ethico-philosophy,[109] that concentrated not on the worship of a god or gods but on the human as being interwoven with the cycle of nature emerged in India. This doctrine embodied the teachings of the Buddha, formerly known as Siddhārtha Gautama, who attained enlightenment (*bodhi*) meditating under a poplar-fig tree at the age of thirty-five.[110] His "awakening" consisted in the realisation of "the real nature of life as experienced in the present existence and the beings living it."[111] He discerned the following three fundamental characteristics underlying all existence: *anattā* (Non-Selfness), *anicca* (Impermanence), and *dukkha* (Suffering/Existential Anxiety). The philosophy of Buddhism sprang from perceiving the fact that living is

[108] De Silva, Lynn A., *The Problem of the Self in Buddhism and Christianity* (London: The Macmillan Press LTD, 1979), 9.
[109] Saddhatissa, H., *Buddhist Ethics* (London: George Allen & Unwin LTD, 1970), 9.
[110] Schumann, H. Wolfgang, *Buddhism: An Outline of its Teachings and Schools* (London: Rider and Company, 1973), 20-21.
[111] Saddhatissa, *Buddhist Ethics*, 47.

imbued with suffering, hence the examination of suffering and the quest for liberation from it supplied the material for Buddhist thought. *Dukkha* pertains to the unacceptance of the transitory reality of things and phenomena, of the fact that all that is joyful and lovable will sooner or later come to cessation, and to one's mental identification with one's own self. I argue that the Buddhist concept of *anattā* (Non-Selfness) and a thorough examination of what constitutes a person mark *the first point of interest for AI research* (I).

Like modern neuroscience and cognitive science, Buddhism closely examines the nature of the mind, and for that reason it is sometimes called a "mind-culture."[112] The Buddha described what it means to be human by enumerating the Five Groups of Grasping: (1) "body," meaning the physical appearance and anatomy (*rūpa*); (2) "sensation," meaning the contact of the sense-organs with the outer world (*vedanā*); (3) "perception," meaning sensations interpreted by the brain (*saññā*); (4) "mental phenomena," meaning reactions to the perceptions in the form of notions, ideals, cravings, moods, etc. (*saṅkhāra*); (5) "consciousness," meaning the accumulative element embracing all the mental phenomena and arisen from them (*viññāṇa*).[113] It thus transpires that, according to Buddhist thought, there is no single function in the mental architecture that can be identified as the self. So just like neuroscience cannot separate any particular area of the neocortex as the seat of the mind,[114] or just like Minsky talks in *The Society of the Mind* about multiple agents in a large mental organisation that carry on their tasks in relation to each other without being intelligent by themselves,[115] Buddhism acknowledges that the totality of the process of perception takes place in the five groups and not in the 'I.' Hence, one ought not to say "I perceive," but "a process of perception is occurring in the five groups." In such a way, the empirical person is not the 'self,' but a bundle of phenomena. Minsky talks about the nature of perception in a very similar way by calling the mind-body (*nāma-rūpa* in Buddhism) association an intrinsically relational system:

---

[112] Saddhatissa, *Buddhist Ethics*, 28.
[113] Schumann, *Buddhism: An Outline of its Teachings and Schools*, 42.
[114] Norden, *Understanding the Brain: Course Guidebook*.
[115] Minsky, Marvin, "Excerpts from *The Society of the Mind*," in *Artificial Intelligence: Critical Concepts Vol. II*, ed. Ronald Chrisley (London: Routledge, 2000), 223.

> Your *tooth* can't ache it can only send signals; only *you* can ache, once your higher level agencies interpret those signals. Beyond the raw distinctiveness of every separate stimulus, all other aspects of its character or quality – be it of touch, taste, sound, or light – depend entirely on its relationships to the other agents of your mind.[116] [Original emphasis]

The Buddha appeals to his fellow monks with a somewhat analogous sentiment: "What, monks, is the Universe?: The eye and forms, the ear and sounds, the nose and smells, the tongue and tastes, the body and tactile objects, the mind and mental objects."[117] It can be only speculated whether the robot Shakey would agree with the Buddha's conjecture, according to whom, for the world to be realised, the six organs of perception, the sense-objects corresponding to them and the awareness of these objects have all to be present simultaneously.[118]

The three characteristics of the individual are thus Impermanence, Sorrowfulness and Non-Selfness. Having realised this truth, one can take Three Refuges: the Buddha, the Dhamma and the Saṅgha (the monastic community of ordained Buddhist nuns and monks). The first two Refuges, the Buddha and the Dhamma, are of particular importance for the purposes of the current analysis as they offer an ethical perspective that would agree with the definition of intelligence proposed earlier in this chapter. It can be argued that it is the Buddhist model of moral conduct that marks *the second point of interest for AI research* (II).

Taking refuge in the Buddha implies that one cannot rely on the Buddha for one's personal enlightenment, as it is only in the power of the individual to make the mental effort to rise as high as they will. Future happiness is the direct result of the present conduct and there is nothing but the discipline one exerts over oneself that helps one restrain from wrongdoing. In particular, it is said in the Dhammapada: "If by renouncing a relatively small happiness one sees a happiness great by comparison, the wise man abandons the small happiness in consideration of the greater happiness."[119] Buddhism emphasises that no blind faith, no prayers and no worship can save one from suffering, but day-to-day life

[116] Minsky, "Excerpts from *The Society of the Mind*," 229.
[117] Schumann, *Buddhism: An Outline of its Teachings and Schools*, 48.
[118] Ibid.
[119] Saddhatissa, *Buddhist Ethics*, 56.

of love and sympathy with the world. And it is taking refuge in the Dhamma that can guide one along the path of such an existence.

The Dhamma, meaning the "way" or "doctrine," is outlined in the Buddhist scriptures, traditionally divided into three groups, known as baskets or *pitakas* (*Vinaya*, *Sutta* and *Abhidhamma*). The goal of the Dhamma is to transform life into the *dukkha*-less state called *nibbāna* (which will be discussed later) by way of following the Noble Eightfold Path, otherwise called the Middle Way of self-conquest, helping one to avoid the two extremes of self-indulgence and self-torture. The Noble Eightfold Path involves the following duties to be adhered to by its follower: (1) right understanding, implying seeing life according to the three marks of all existence, *anicca*, *dukkha* and *anattā*; (2) right thought, implying that one's mind should be free from cruelty, ill-will, etc.; (3) right speech, implying that one should refrain from lying, harsh talk, gossip, etc.; (4) right conduct, implying living according to a minimum of five moral precepts or rules, namely: (i) not to kill but to practice love and compassion to all, (ii) not to take that which is not given but to practice generosity and charity, (iii) not to partake in unlawful sexual intercourse but to practice purity and self-control, (iv) not to indulge in false speech but to practice honesty and the serenity of the mind, and (v) not to take intoxicating substances but to practice restraint; (5) right livelihood, implying that a person should not pursue an occupation that would harm or cause injustice to other beings; (6) right effort, implying self-perfection by rejecting ignoble qualities while fostering noble ones; (7) right mindfulness, implying the state of constant awareness with regard to the body, feelings, mind and ideas engendered therein; (8) right meditation, implying the active practise of meditation and the passive realisation of truths.[120]

Whilst taking refuge in the Buddha implies that one's own moral conduct is the duty of the individual, taking refuge in the Dhamma thus prompts one to settle down to the course of socially and environmentally responsible action that would abstain from hurting nature, animals or human beings. Recognising that gaining mastery over oneself is not an easy task, Buddhism offers one a

[120] Saddhatissa, *Buddhist Ethics*, 69-73.

"complete ethical study"[121] as it guides one all the way from the practice of common moral precepts up to the attainment of a supra-mundane state that transcends human notions of good and evil. Buddhist meditation is the way to defeat one's longings, cravings and attachments by imposing over oneself strict forms of spiritual procedure, and it can be argued that its elaborate instructions on how to achieve a certain state of mind mark *the third point of interest for AI research* (III). The volume of this paper does not allow to consider the full variety of meditation practices, but one telling example can nevertheless be drawn. A brief account of meditation offering five steps of practicing loving-kindness, *mettā*, towards another person can be described in such a way: (step 1) develop *mettā* towards yourself by repeating the following formula:

May I be free from enmity;
May I be free from ill-will;
May I be free from distress;
May I keep myself happy.[122]

Then extend the wish to the welfare of other beings; (step 2) recollect gifts received, kind words, etc., that inspire *mettā*, virtue and the desire to learn; (step 3) if you want to proceed further, cultivate *mettā* towards a dearly-loved friend; (step 4) develop *mettā* towards a neutral person; (step 5) develop *mettā* towards a hostile person. Another meditation on overcoming resentment towards a hostile person outlines nine steps of procedure and has such conditional statements as "if this did not help, do this."[123]

The final goal of all meditation and righteous conduct is *bodhi* ("awakening") or *nibbāna*, the state of perfect peace in which all the egotistic cravings are overcome due to the realisation of the three-fold nature of reality: *anattā*, *anicca*, and *dukkha*. *Nibbāna* is also described as the extinction of the fire of lust, hatred and delusion.[124] Buddhism conceives of the individual's life as being inscribed into the chain of rebirths, prompted by *kamma*, which can be best described as potentiality. *Kamma* is not a soul, neither is it a self, but the

[121] Saddhatissa, *Buddhist Ethics*, 19.
[122] Ibid., 91.
[123] Ibid., 91-99.
[124] De Silva, *The Problem of the Self in Buddhism and Christianity*, 63.

accumulation of "physical and mental energy."[125] As Buddhism regards everything in the universe as dynamic becoming, through which "delusion increases and ignorance is prolonged,"[126] *kamma* is part of the life stream that is passed down on the death of one consciousness to the next 'successor' in the course of becoming. However, because *nibbāna* transcends craving, ignorance and delusion, it is said that after its attainment there is no more rebirth. Both *kamma* and *nibbāna* are considered as somewhat paradoxical notions and were interpreted as such by the Buddha himself.[127] What exactly is *kamma* and how can bliss be experienced in *nibbāna* if there is no experiencing self? How can one deny the self and yet assert moral responsibility, implied by *kamma*? These questions have led to countless volumes of Buddhist scholarship written over two and a half thousand years. A more useful question for the current analysis, on the other hand, would be how these paradoxical concepts can be interpreted outside religious context for the purpose of enriching the understanding of secular ethics?

Some inspiration can be taken from the synthesis of De Silva and J. G. Jennings' ideas of *anattā-pneuma* and the self as collective *kamma* respectively. I suggest that this marks *the fourth point of interest for AI research* (IV). De Silva claims that Buddhist philosophy is "radically on the side of individualism,"[128] as it stresses again and again that it is one's *personal* goal to achieve enlightenment and it is up to *the individual* to reach that goal. This, he argues, can lead to egocentric rationality and even nihilism in one's desire to come to full extinction and escape the cycle of rebirths. Following the Buddha's intention to conceive of a philosophy that would not succumb to nihilistic or eternalistic worldviews, he proposes that the Christian concept of *pneuma*, stressing the social nature of humans, can be beneficial for the interpretation of *nibbāna* as a phenomenon pertaining to "non-egocentric relationality"[129] rather than egocentric rationality. *Pneuma* is a Greek word for "spirit" and marks in Christian theology one's

[125] Saddhatissa, *Buddhist Ethics*, 38.
[126] Ibid., 39.
[127] De Silva, *The Problem of the Self in Buddhism and Christianity*, 48-74.
[128] Ibid., 75.
[129] Ibid., 74.

spiritual life that is derived from God.[130] De Silva insists, drawing on the scholarship of Tillich among others, that the biblical view of a human being is holistic and not dualistic.[131] In other words, the soul and the body are closely interwoven and one cannot be separated from the other in the same sense as the *nama-rupa* interrelation in Buddhism suggests unity. Man is *created* in the image of God, which in De Silva's point of view, entails that man is both a psychosomatic organism that can be described by the Buddhist term *anattā* and one who has the potential of transcending oneself, designated by the Christian term *pneuma*. Interestingly, however, *pneuma* also means the spirit of the community, the life of the Church as the body of Christ, in the New Testament[132] thus implying that in order to transcend oneself one has to reach out to the other with whom one is related. *Nibbāna* can thus be positively reinterpreted in a personal-communal sense as a form of transcendence that pertains not to one merely emptying oneself of the ego after realising oneself as a bundle of psychic and physical phenomena but rather to one emptying oneself of the ego for the purpose of reaching beyond oneself and out to the other. Jennings, in turn, who argues that Buddhism is strictly a system of ethical conduct motivated by the realisation of collective responsibility, rejects the doctrine of rebirth and offers to think of *kamma* in terms of the transmission of cumulative "physical and mental energy" to succeeding generations, as life conceives more life.[133] In such a manner, De Silva proposes to revaluate *nibbāna* as transcendence for the sake of social good while Jennings interprets *kamma* as an 'inheritance' that one generation leaves to the next.

So having analysed the basic concepts of Buddhism, the model of Enlightened Intelligence for AI can now be finally drawn. I propose that the following four aspects of Buddhism could benefit any operational model of ethics developed in conjunction with research on AI:

I. Focus on the nature of mental and physical functions and their interrelation as mutually dependent, relational phenomena;

[130] De Silva, *The Problem of the Self in Buddhism and Christianity*, 89.
[131] Ibid., 75.
[132] Ibid., 93.
[133] Jennings, J. G., *The Vedantic Buddhism of the Buddha* (London: OUP, 1947), p. xxxvi.

II. Specification of the concrete rules of moral and ethical conduct and the emphasis on one's own duty to adhere to the moral code in order to achieve enlightenment;

III. Assortment of meditation practices and techniques, accompanied by detailed instructions on how to pursue them;

IV. The concepts of *nibbāna* and *kamma*, understood in the personal-communal sense and implying the socially and environmentally responsible behaviour of one transcending oneself for the benefit of others.

How can this operational model be mapped to the actual scientific developments that could help in the creation of true machine intelligence? Clearly a full answer to this question is beyond the scope of a short paper, but some initial issues can be identified. The first point addresses Buddhist primary interest in the workings of the human mind. According to the Buddha, "mind precedes all things; all things have mind foremost, are mind-made."[134] For this reason, it is important to understand how mental phenomena are borne and what they mean. Buddhism does not recognise a gap between religion and science and sees the former as a "practical spiritual application of the principles of the latter."[135] In modern times, neuroscientific research on meditation avails of such technologies as EEG and fMRI to understand the effects of long-term intensive meditation on physiological and psychological processes related to health, emotion regulation and attention. One of the best known proponents of such a study is Clifford Saron who directs "The Shamatha Project," a collaborative, multi-method, interdisciplinary study incorporating a variety of measures, including qualitative, behavioural, electrophysiological and biochemical ones.[136] It can thus be suggested that scientific research on the positive effects of meditation on mental processes can be used to build better brain models for future AI. A further speculative question on the benefits of meditation for AI can be put forward: if Buddhist meditation practices are

[134] Saddhatissa, *Buddhist Ethics*, 28.
[135] Ibid., 37.
[136] YouTube. "The Majesty of the Present: Clifford Saron at TEDxUCDavis." Accessed May 12, 2016. https://www.youtube.com/watch?v=d88Q-15W_AI.

recorded as 'algorithmic' sets of instructions on how to overcome a certain feeling or to achieve a certain state of mind, can an intelligent computer be made to meditate by following a set of instructions (which seems like a familiar task for a machine) that it understands in order to improve its own 'mental' state?

Another question that pertains to the nature of computer programs and that can be asked in conjunction with the examination of Buddhist ethics is whether enlightenment comes from the experience of transcending the limitations of human life or from the careful observation of recorded wisdom. If from the latter, do computers have a better chance of becoming enlightened by following outlined rules? And what would enlightenment mean in the context of machine intelligence, as it was defined in the beginning of this chapter? Buddhism defines clearly the norms of moral and ethical code leading to an enlightened existence, but is it possible to describe moral and ethical behaviour in a way that a machine can be programmed to simulate it? A scientist could in principle translate the rules of ethical and moral conduct into a set of instructions, store them in the program's 'memory,' and when 'stimuli' were present, the computer would generate an appropriate 'response,' based on the instructions it had been given. However, in order to translate these rules so that a machine can understand them, not only a profound neurological, linguistic and psychological comprehension of how moral and ethical behaviour is enacted in the brain is required, but an information-processing language that avails of the computer's non-numerical capabilities[137] needs to be created.

G. A. Miller, E. Galanter and K. H. Pribram suggest that, as students of the human being, and I would add as students of what ethical and moral behaviour entails, researchers have to pursue a line of enquiry that might not seem as the most efficient.[138] Indeed, when it comes to the creation of artificial, machine intelligence that would be comparable to human intelligence, the highest possible ethical plank should be taken as a behaviourist model for *automata*. For, arguably, what humans should be aiming at is creating not the duplicates of themselves but beings made in the image of the paragons of human

[137] Miller, G. A., Galanter, E., and Pribram, K. H., "The Simulation of Psychological Processes," in *Artificial Intelligence: Critical Concepts Vol. II*, ed. Ronald Chrisley (London: Routledge, 2000), 97.
[138] Ibid., 96.

ethics. Buddhism, in combination with wisdom taken from other philosophical systems and religions, does offer such a figure: a socially and environmentally responsible person who finds his authentic being not in the crevices of their ego but in communion with other persons, animals and nature - one who reaches beyond themselves in order to reach out to others without conquering them - for they realise themselves not as a 'self,' but as a part of the common environment.

How realistic is this idea? One hope is to build appropriate moral norms into knowledge bases; conduct an exhaustive neurological, linguistic and psychological research; and to learn how to translate our wisdom into a non-ambiguous language that a machine would understand. It is by no means an easy task. It is a long road of discovery that involves a multitude of little steps and breakthroughs, but the sight of the most important philosophical issues must not be lost. Should not the dream of ethical machines and ethical excellence be one of humanity's biggest dreams? Should not a world without violence be our biggest ambition? For it can be argued that the very Buddhist realisation of the fact that enlightenment necessitates 'overcoming oneself' and 'purifying' one's nature, agrees with the conjecture that human intelligence is unsound as an operational model for machine intelligence, since humans do not yet live in such a world. Torrance illustrates this point thus:

> It looks as though we are entering an age in which electronic 'intelligent' knowledge bases will increasingly be considered as oracles to consult and defer to, as dominant repositories of Truth. If these oracles are to serve the interests of human beings around the world [and the environment], rather than merely the interests of the Fortune 500, then they must be given more than merely domain expertise: they must be provided with some measure of social and normative enlightenment. It remains to be seen if the moral wisdom of ordinary humankind is too ineffable, too inscrutable, too variable, to be captured within an AI representation.[139]

[139] Torrance, "Ethics, Mind and Artifice," 301.

Conclusion

"History suggests that man can create almost anything he can visualise clearly. The creation of a model is proof of the clarity of the vision."

G. A. Miller, E. Galanter, K. H. Pribram

To sum up, what are the findings revealed in this paper?

The aim of Chapter I was to inquire into whether contemporary scientists believe in the possibility of creating an artificial mind in principal. The opinion of Kurtzweil, an optimistic proponent of strong AI, was juxtaposed and contrasted with that of Sir Penrose, a mathematical physicist who insists that consciousness cannot be captured by the algorithm. Yet even Penrose does not deny that a different method of computation might perchance succeed at endowing machines with sentience. However, he cautions that it would most likely posit a threat to humanity because it is impossible to predict how such machines would act in the world.[140]

A historical approach was adopted in Chapter II for the purpose of investigating how the discipline of AI has evolved in the sixty-five years of its existence, what its achievements and disappointments were and what its hopes for the future are. The examination has shown that the inability of the field to achieve any considerable progress in the early days of AI led to its diversification into a variety of sub-disciplines. From then onwards, AI has seen some crucial breakthroughs, such as the development of deep learning and reinforcement learning techniques in the latter years in particular, resulting in the revival of the daring hope to breathe conscious life into machines.

One the basis of this analysis, a proposition was put forward in Chapter III that it is of crucial importance to conduct ethical research in conjunction with further exploration of the field of AI. For if there is the slightest possibility of succeeding at creating truly intelligent machines, humans must ensure in advance that they will not be a threat to humans, other animals and the environment. On the premise of this conjecture, a definition of intelligence

[140] "Sir Roger Penrose - Consciousness and the Foundations of Physics."

modelled after human intelligence was rejected and a new operational definition of intelligence with an added ethical dimension was proposed: intelligence maximizes future freedom not only for itself but also for other beings in the world for the purpose of minimising the overall negative impact on the environment. Furthermore, an ethical model that would complement this definition, called Enlightened Intelligence and inspired by the Buddhist concept of enlightenment, was suggested as a possible operational ethical framework within which artificial intelligence might be cultivated. At the heart of this model lies the notion of transcendence, or enlightenment, as characterised by the action of a socially and environmentally responsible conscious being reaching out to another being on the basis of them having overcome their immediate interests and wanting to enter into communion with the world, for they realise themselves not as a 'self,' but as a part of the common environment.

The limitations of this paper did not allow for an in-depth inter-disciplinary, cross-religious, cross-cultural study of ethical and moral norms that could contribute to the creation of a complete operational ethical framework for the needs of AI. The motivation behind this paper was to set a direction for future research and to suggest that our models of intelligence and ethics must be inspired by the finest human accounts of what intelligence and ethics should be rather than by human intelligence and behaviour as they manifest themselves in the world. It was thus proposed that humans should be aiming at creating not duplicates of themselves but beings made in the image of the paragons of human ethics. An exhaustive research effort will be required in order to succeed at this endeavour, embracing a study in the humanities, a profound neurological, linguistic and psychological comprehension of how moral and ethical behaviour is enacted in the brain and the creation of an information-processing language via which the norms of ethical conduct could be communicated to a machine. Perhaps such ambitions are beyond the lifetime of a person or beyond human capabilities altogether, but it might be argued that just by conceiving of idealistic ethical models, humans will not only make better machines but also become better humans.

Bibliography

Agre, Philip E. "The Soul Gained and Lost: Artificial Intelligence as a Philosophical Project." In *Artificial Intelligence: Critical Concepts Vol. IV*, edited by Ronald Chrisley. London: Routledge, 2000.

Andersen, Hans Christian. "The Nightingale." In *The Annotated Hans Christian Andersen*, edited by Maria Tatar. New York: W. W. Norton & Company, 2007.

De Silva, Lynn A. *The Problem of the Self in Buddhism and Christianity*. London: The Macmillan Press LTD, 1979.

Future of Life Institute. "An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence." Accessed May 12, 2016. http://futureoflife.org/ai-open-letter/.

Haraway, Donna J. "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century." In *Simians, Cyborgs and Women: The Reinvention of Nature*. London: Free Association Books, 1991.

Hofstadter, Douglas R. *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Vintage Books, 1979.

IBM Research. "How Deep Blue Works: Under the Hood of IBM's Chess-Playing Supercomputer." Accessed May 12, 2016. https://www.research.ibm.com/deepblue/meet/html/d.3.2.html.

Innovative Applications of Artificial Intelligence 2, edited by Alain Rappaport and Reid Smith. Cambridge: MIT Press, 1991.

Jennings, J. G. *The Vedantic Buddhism of the Buddha*. London: OUP, 1947.

Kant, Immanuel. *Grounding for the Metaphysics of Morals (Third Edition): with On a Supposed Right to Lie because of Philanthropic Concerns*, translated by James Wesley Ellington. Indianapolis: Hackett Publishing Company, 1993.

Kurtzweil, Ray. *How to Create a Mind: the Secret of Human Thought Revealed*. London: Viking, 2012.

Kurzweil, Ray. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. London: Texere, 2001.

Lachat, Michael R. "Artificial Intelligence and Ethics: An Exercise in the Moral Imagination." In *Artificial Intelligence: Critical Concepts Vol. IV*, edited by Ronald Chrisley. London: Routledge, 2000.

Lighthill, James. "Artificial Intelligence: A General Survey." In *Artificial Intelligence: Critical Concepts Vol. III*, edited by Ronald Chrisley. London: Routledge, 2000.

Mazlish, Bruce. "The Man-Machine and Artificial Intelligence." In *Artificial Intelligence: Critical Concepts Vol. I*, edited by Ronald Chrisley. London: Routledge, 2000.

McCarthy, J., Minsky, M. I., Rochester, N., and Shannon, C. E. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." In *Artificial Intelligence: Critical Concepts Vol. II*, edited by Ronald Chrisley. London: Routledge, 2000.

McCorduck, Pamela. “Robotics and General Intelligence.” In *Artificial Intelligence: Critical Concepts Vol. I*, edited by Ronald Chrisley. London: Routledge, 2000.

Miller, G. A., Galanter, E., and Pribram, K. H. “The Simulation of Psychological Processes.” In *Artificial Intelligence: Critical Concepts Vol. II*, edited by Ronald Chrisley. London: Routledge, 2000.

Mind-Crafts. “Gödel’s Incompleteness Theorems – A Brief Introduction.” Accessed May 12, 2016. http://math.mind-crafts.com/godels_incompleteness_theorems.php.

Minsky, Marvin, and Harrison, Harry. *The Turing Option*. London: Viking, 1992.

Minsky, Marvin. “Excerpts from *The Society of the Mind*.” In *Artificial Intelligence: Critical Concepts Vol. II*, edited by Ronald Chrisley. London: Routledge, 2000.

Minsky, Marvin. “Steps Toward Artificial Intelligence.” In *Artificial Intelligence: Critical Concepts Vol. II*, edited by Ronald Chrisley. London: Routledge, 2000.

MIT Technology Review. “Deep Learning.” Accessed May 12, 2016. https://www.technologyreview.com/s/513696/deep-learning/.

MIT Technology Review. “Five Lessons from AlphaGo’s Historic Victory.” Accessed May 12, 2016. https://www.technologyreview.com/s/601072/five-lessons-from-alphagos-historic-victory/.

MIT Technology Review. “How Google Plans to Solve Artificial Intelligence.” Accessed May 12, 2016. https://www.technologyreview.com/s/601139/how-google-plans-to-solve-artificial-intelligence/.

MITvideo. “‘The Thinking Machine’ (1961) – MIT Centennial Film.” Accessed May 12, 2016. http://video.mit.edu/watch/the-thinking-machine-1961-mit-centennial-film-6712/.

Norden, Jeanette. *Understanding the Brain: Course Guidebook*. Chantilly: The Great Courses, 2007.

Open Culture. “Artificial Intelligence Creativity Machine Learns to Play Beethoven in the Style of The Beatles’ ‘Penny Lane’.” Accessed May 12, 2016. http://www.openculture.com/2016/05/beethoven-in-the-style-of-the-beatles-penny-lane.html.

Penrose, Roger. *The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press, 1989.

Popular Science. “Autonomous Robot Performs Successful Surgery on Living Pig.” Accessed May 12, 2016. http://www.popsci.com/new-robotic-surgery-tool-outperformed-human-surgeons.

Popular Science. “Google Announces Self-Driving Minivan.” Accessed May 12, 2016. http://www.popsci.com/googles-first-self-driving-minivan-is-coming.

Popular Science. “Google’s Human-Shaped Robot Takes First Walk Outside.” Accessed May 12, 2016. http://www.popsci.com/google-sent-its-human-shaped-robot-outside.

Popular Science. “How Google Aims to Dominate AI: The Search Giant is Making its AI Open Source so Anyone Can Use It.” Accessed May 12, 2016. http://www.popsci.com/google-ai.

Russell, Stuart J., and Norvig, Peter. *Artificial Intelligence: A Modern Approach*. New Jersey: Pearson Education, 2010.

Saddhatissa, H. *Buddhist Ethics*. London: George Allen & Unwin LTD, 1970.

Schumann, H. Wolfgang. *Buddhism: An Outline of its Teachings and Schools*. London: Rider and Company, 1973.

Science. "Video: Robot Surgeons Make a Big Advance." Accessed May 12, 2016. http://www.sciencemag.org/news/2016/05/video-robot-surgeons-make-big-advance.

SPACE.com. "Humanoid Robot R5: Valkyrie 'Dances' in NASA Music Video." Accessed May 12, 2016. http://www.space.com/31270-humanoid-robot-r5-valkyrie-dances-in-nasa-music-video.html.

TED. "A. Wissner-Gross: A New Equation for Intelligence." Accessed May 12, 2016. https://www.ted.com/talks/alex_wissner_gross_a_new_equation_for_intelligence.

TED. "Nick Bostrom: What Happens When Our Computers Get Smarter Than We Are." Accessed May 12, 2016. https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are?language=en#t-973754.

TED. "Sylvia Earle: My Wish: We Must Protect Our Oceans." Accessed May 12, 2016. https://www.ted.com/talks/sylvia_earle_s_ted_prize_wish_to_protect_our_oceans.

The New York Times. "Marvin Minsky, Pioneer in Artificial Intelligence, Dies at 88." Accessed May 12, 2016. http://www.nytimes.com/2016/01/26/business/marvin-minsky-pioneer-in-artificial-intelligence-dies-at-88.html?_r=1.

The New Yorker. "The Doomsday Invention: Will Artificial Intelligence Bring Us Utopia or Destruction?" Accessed May 12, 2016. http://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom.

Torrance, Steve. "Ethics, Mind and Artifice." In *Artificial Intelligence: Critical Concepts Vol. IV*, edited by Ronald Chrisley. London: Routledge, 2000.

Turing, Alan M. "Computing Machinery and Intelligence." In *Artificial Intelligence: Critical Concepts Vol. II*, edited by Ronald Chrisley. London: Routledge, 2000.

Vimeo. "Ten Paradoxes of Technology." Accessed May 12, 2016. https://vimeo.com/28022711.

Whitby, Blay. "The Turing Test: AI's Biggest Blind Alley?" In *Artificial Intelligence: Critical Concepts Vol. IV*, edited by Ronald Chrisley. London: Routledge, 2000.

YouTube. "Sir Roger Penrose - Consciousness and the Foundations of Physics." Accessed May 12, 2016. https://www.youtube.com/watch?v=eJjydSLEVlU.

YouTube. "The Majesty of the Present: Clifford Saron at TEDxUCDavis." Accessed May 12, 2016. https://www.youtube.com/watch?v=d88Q-15W_AI.