DONG JUN (JUSTIN) KIM

ARTIFICIAL INTELLIGENCE AND CRIME: WHAT KILLER ROBOTS COULD TEACH ABOUT CRIMINAL LAW

Submitted for the LLB (Honours) Degree

LAWS523: Issues in Sentencing and Penology

Faculty of Law

Victoria University of Wellington

2017

Abstract

Criminality and punishment have always been applied to human beings. However, the technological field of artificial intelligence ('AI') is becoming impressively sophisticated. Machines that utilise AI ('AI entities') may soon be able to commit actions which, if committed by humans, would be considered criminal. This paper poses a hypothetical fact scenario to explore whether, and how, existing criminal law should respond to such AI entities. This paper concludes that existing criminal laws are ultimately a bad fit for AI. First, regulating AI entities becomes complicated by the conceptual difficulties in defining AI. Secondly, existing party liability mechanisms, such as corporate liability, are unsuitable for non-humans. Thirdly, criminal liability has always assumed that the offender is human, meaning that AI entities cannot satisfy the mens rea element of criminality. Finally, the purposes of sentencing are so deeply rooted in society that its application to non-humans would be inappropriate. AI entities ultimately show that criminal law and social expectations are inextricably linked. This paper accordingly raises two talking points: the role of criminal law going forward, and whether AI entities will ever be accepted into the wider society.

Keywords: artificial intelligence, crime, liability, sentencing, human

CONTENTS

I Introduction

"*Natura abhorret vacuum*."[1]

Artificial intelligence ('AI') is a field of science which outlines the human attempt to build intelligent entities.[2] Machines that use AI technology will be called "AI entities" in this paper. AI entities were designed to aid and depend on humans. However, from defeating champion Go players[3] to outperforming the best lipreaders,[4] technological advances have resulted in AI entities surpassing humans in many aspects of life. AI entities may soon be able to commit actions which, if committed by humans, would be considered criminal. The question is whether, and how, criminal law should respond.

Developments in the AI field thus suggest that it is worth re-examining New Zealand's existing criminal law framework. Are only humans subject to criminal law? Can AI entities commit crimes? Can they be punished? This paper explores the relationship between criminal law and AI entities through a hypothetical fact scenario. The hypothetical raises four issues: determining a legal definition of AI (in Part III), determining which legal mechanism to use to find parties liable (in Part IV), whether AI entities themselves can be found criminally liable (in Part V), and whether AI entities can be punished (in Part VI). This paper concludes that criminal law's societal nature makes liability and sentencing a bad fit for AI entities. However, this paper's discussions will help clarify the criminal law's role in society generally. Accordingly, this author calls for a conversation on how AI entities may fit into the wider society.

---

1 François Rabelais *Gargantua* (France, 1534) at 22 ("Nature abhors a vacuum").

2 Stuart Russell and Peter Norvig *Artificial Intelligence: A Modern Approach* (3rd ed, Prentice Hall, New Jersey, 2010) at 1.

3 Jon Russell "After beating the world's elite Go players, Google's AlphaGo AI is retiring" *TechCrunch* (online ed, San Francisco Bay Area, 27 May 2017).

4 Yannis Assael and others "LipNet: End-to-End Sentence-level Lipreading" (paper presented to International Conference on Learning Representations, Toulon, April 2017).

II The Situation: A New Species of Potential Criminals

A Overview

1 The role of criminal law

There is a tension in the role of criminal law. On the one hand, criminal law punishes unlawful acts as determined by the state. Criminal law somewhat responds to society's views, as criminal actions affect the wider society. Sir William Blackstone stated that criminal law protects "the person injured by every infraction of the public rights *belonging to that community*".[5] He notes that crime, being a public wrong, speaks to not only the injured individuals but also "the very being of society".[6] Richard Fuller also argues that most Western criminal codes contain a societal "moral minimum": "those values which we hold most sacred and least dispensable are elevated by public opinion to the status of protection".[7] Therefore, criminal law has inherent ties to natural law by encompassing "core existential truths" about what it means to be human and how to live "responsibly in our societies".[8]

However, criminal law does not always follow the anachronistic whims of societal beliefs. Fuller recognises that the values of the majority affects the criminal law's universal applicability, as "the problem of criminal law in action reduces to the problem of conflicting moral values held by different groups and classes in the community".[9] Therefore, it is insufficient to define criminal behaviour as an act that is *considered* criminal in the opinion of many.[10] Hart notes that those who accept that criminal law enforces a majority's view of morality must also admit that "its imposition on a minority is

---

5 William Blackstone *Commentaries on the Laws of England* (9th ed, reissue, 1978) vol 4 Of Public Wrongs at 2 (emphasis added).

6 Blackstone, above n 5, at 5.

7 Richard Fuller "Morals and the Criminal Law" (1942) 32 J Crim Law Criminol 624 at 628.

8 Sandra Jacobs "Natural Law, Poetic Justice and the Talionic Formulation" (2013) 14 Political Theology 661 at 662.

9 Fuller, above n 7, at 624.

10 Fuller, above n 7, at 639.

justified".[11] Thus, legal positivism adopts the idea that morality alone cannot determine what is criminal.

2 The thinking machine rises

Both natural law and legal positivism have deep ties to humanity: natural law speaking directly to human morality, and legal positivism assuming that laws are "commands of human beings".[12] Humans are unique in that we use tools to simplify life.[13] As our thought processes became more complex, so did our scientific knowledge and tools. Some machines became so complex that they now interact with humans and assist in decision-making.

AI entities are prevalent in 2017, including Apple's virtual assistant Siri (carrying out smartphone actions through voice commands),[14] flight control systems in planes (automating laborious tasks such as maintaining altitude),[15] and self-driving cars (driving autonomously through sensors and GPS).[16] Machine learning algorithms also lie at the heart of social media websites (pushing individualised information to users),[17] and video games (simulating strategies and opponents).[18] AI entities are machines that can think,[19] allowing them to carry out deeply complex tasks.[20]

Ethical judgments are such complex tasks. Take the trolley problem. You are a trolley driver; while operating the trolley, you see five people tied to the tracks ahead. You step

11 HLA Hart *Law, Liberty and Morality* (Oxford University Press, Oxford, 1963) at 81.

12 HLA Hart "Positivism and the Separation of Law and Morals" in *Essays in Jurisdiction and Philosophy* (Oxford University Press, Oxford, 1983) 49 at 57.

13 Gabriel Hallevy *When Robots Kill: Artificial Intelligence Under Criminal Law* (Northeastern University Press, Lebanon (NH), 2013) at 13 (*When Robots Kill*).

14 Matthew Hutson "Our Bots, Ourselves" *The Atlantic* (online ed, Washington DC, March 2017).

15 "Chapter 4: Automated Flight Control" Federal Aviation Administration <www.faa.gov> at 4-2.

16 Danielle Muoio "Tesla's new Autopilot is getting a big update this weekend – here's everything you need to know" *Business Insider* (online ed, New York City, 16 June 2017).

17 Mark Smith "So you think you chose to read this article?" *BBC News* (online ed, London, 22 July 2016).

18 Georgios Yannakakis "Game AI Revisited" (paper presented at the Proceedings of the 9th conference on Computing Frontiers, Cagliari, May 2012) at 2.

19 *When Robots Kill*, above n 13, at 13.

20 Russell and Norvig, above n 2, at 1.

on the brakes; they fail. There is a lever that switches to a different set of tracks, but there is another person tied to those tracks. If you do nothing, five will die; if you pull the lever, one will die.[21] The issue is whether (and how) AI entities would be culpable if it uses its programming to make this decision. Our hypothetical involves a similar scenario.

If the trolley driver was a human, they may be held criminally liable. The situation is complicated if the trolley was on autopilot. There is no New Zealand case which involves an AI committing a crime. There is a gap in the current criminal law here. However, a re-examination of the existing legal framework is also necessary. An exploration of the latter issue will help fill any gaps in criminal law regarding AI regulation. Therefore, while this paper explores the applicability of criminal law on AI entities, it also explores the societal and philosophical foundations of criminal law and sentencing using AI entities as a vehicle.

Criminal liability has historically required a degree of human involvement in the criminal action. However, technological capabilities are expanding exponentially;[22] the gap between humans and machines is closing. Existing criminal law frameworks that draw a clear distinction between humans and AI entities are likely to be challenged to the point where there may be calls for AI entities to be held legally accountable.[23] Criminal law's applicability is entering unchartered waters, as its applicability is being questioned in relation to AI entities independent of humans is being called into question.

The tension between legal positivism and natural law is important for AI entities. Until now, criminal law has applied to legal persons who were always linked to humans.[24] The tension between positivism and natural law is inherently linked to human morality and behaviour. However, AI entities are no longer "mere data depots", but beings that appeal

---

21 Judith Thomson "The Trolley Problem" (1985) 94 Yale LJ 1395 at 1395.

22 Ray Kurzweil *The Singularity is Near: When Humans Transcend Biology* (Viking Press, New York City, 2005) at 24.

23 Jack Beard "Autonomous Weapons and Human Responsibilities" (2014) 45 Georgetown J Int Law 647 at 662.

24 This idea will be discussed in Part IV of this paper.

to mankind's "innate receptiveness to anthropomorphi[s]ed machines".[25] Because the law is being asked to deal a novel scenario,[26] the issue is whether criminal law can be used to regulate AI entities, allowing it to deal with a fact scenario involving an AI entity.

3 Regulating criminal AI entities

AI entities committing crimes is a novel concept. However, Judge Frank Easterbrook once warned that there could be no "Law of the Horse", as any area of legal study must "illuminate the entire law".[27] There cannot be new laws for every new area of life; traditional legal principles should be applied instead, and any effort to collect various legal principles into a "Law of the Horse" is "doomed to be shallow" and "miss unifying principles".[28] Judge Easterbrook applied this reasoning to the fast-moving field of cyberspace, and argued that the Internet should be regulated by property and trade law.[29] Similarly, AI's novelty alone is an insufficient justification for creating AI-specific laws; existing legal principles of criminality and sentencing may be adequate.

Lawrence Lessig, however, claimed that cyberspace was different. He argued that the law was only one part of regulation.[30] Using cyberspace to describe the real and physical world, Lessig argued that all behaviour is regulated by four constraints:[31]

(1) *Laws*, ordering people to behave in certain ways and threatens punishment upon failure;
(2) *Social norms*, which are societal rules that communities decide when and whether to enforce with punishment upon failure;
(3) *Markets*, regulating the price of things, and thus what people are free to do; and
(4) *Architecture*, describing the physical space around us.

---

25 Ignatius Ingles "Regulating Religious Robots: Free Exercise and RFRA in the Time of Superintelligent Artificial Intelligence" (2017) 105 Geo LJ 507 at 515.

26 Ingles, above n 25, at 516.

27 Frank Easterbrook "Cyberspace and the Law of the Horse" [1996] 207 U Chi Legal F 207 at 207.

28 Easterbrook, above n 27, at 207.

29 At 215–216.

30 Lawrence Lessig "The Law of the Horse: What Cyberlaw Might Teach" (1999) 113 Harv Law Rev 501 at 506 ("The Law of the Horse").

31 "The Law of the Horse", above n 30, at 507.

Lessig argued that cyberspace only differed with the physical world in architecture.[32] Cyberspace has no physical boundaries, and is only governed by manmade code. Code is malleable and therefore capable of being embedded with societal values and beliefs.[33] The flexibility of code, however, also means that it can be regulated. How cyberspace's architecture is regulated will impact all other constraints, and therefore affect the interactions and behaviours of all individuals within that architecture. Lessig's conclusion is that "more than law alone enables legal values, and law alone cannot guarantee them";[34] everything reflects societal values, and all constraints will be affected accordingly. Given the omnipresence of the Internet,[35] and the bevy of Internet-specific laws and legal services that have followed,[36] Lessig's argument has won the day.

While his theory is economic, Lessig notes that criminality is no exception to the constraints.[37] The United States regulates illicit drugs by using *criminal law*, using those laws to seize all drugs at the border to reduce supply (increasing its *market* price), and affecting the structural *architecture* of drugs to make them more harmful and less palatable.[38] Lessig notes that the United States has failed to regulate drugs as they failed to address the *social norms* that reinforced drug abuse.[39]

This paper explores the tension between the application of existing rules (Easterbrook) and the need to take a more nuanced view depending on the circumstances (Lessig) when it comes to regulating crime-committing AI entities. This concept will be explored through the hypothetical.

---

32 "The Law of the Horse", above n 30, at 506.

33 "The Law of the Horse", above n 30, at 548.

34 "The Law of the Horse", above n 30, at 549.

35 Kurzweil, above n 22, at 28.

36 See for example the Harmful Digital Communications Act 2015; Copyright Act 1994, ss 79–81A and 128; and Litigation Support & Discovery Management – E-Discovery Consulting <www.e-discovery.co.nz>.

37 Lawrence Lessig "The New Chicago School" (1998) 27 J Legal Stud 661 at 669 ("The New Chicago School").

38 "The New Chicago School", above n 37, at 669.

39 "The New Chicago School", above n 37, at 669; citing Tracey Meares "Social Organization and Drug Law Enforcement" (1998) 35 Am Crim L Rev 191.

B Hypothetical[40]

Matthew is lives in Karori. He works in Wellington CBD, so he needs a car. Matthew is good friends with technology tycoon William Billingham, head of Intuitive Technologies Ltd ('ITL'). Billingham gifts Matthew the Marvellous Intuitive Technological Transporter ('MITT'), a self-driving car.

MITT was developed by ITL. ITL is a large software company boasting more than 3,000 software engineers worldwide. ITL outsources MITT's hardware to a manufacturer, Marvellous Transport Ltd ('MTL'). MTL is known for its affordability, but has a history of skimping on safety features.

MITT has a front-mounted sensor, allowing it to look ahead and scan for information. The vehicle is controlled by a Logic Module, which receives and interprets information. The Logic Module was programmed on Android, an open source operating system. However, the Logic Module's code also includes proprietary elements which ITL included to ensure that the driver would be protected at all costs. ITL's Head Programmer, Eva Shoelace, said that she was "very proud of the Logic Module's ability to make complex and contextual ethical judgments". After interpreting information through the Logic Module, MITT communicates to Matthew through a voice synthesiser. While MITT can drive autonomously, MITT also has a steering wheel, brakes, and an engine, also allowing Matthew to drive. MITT is a supercomputer on wheels with character; Matthew describes MITT's personality as "sarcastic, but intelligent and well-meaning".

One morning, Matthew and MITT were driving down Glenmore Street. MITT was self-driving. Wellington High School students were visiting the Botanical Gardens that morning. MITT only notices the students crossing the road at the last second. It turns into the footpath. While this averted injury from Matthew and the students, 38-year-old Amy

[40] This hypothetical was inspired by the ethical debate surrounding self-driving cars. Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan "The social dilemma of autonomous vehicles" (2016) 352 Science 1573.

Phillips was killed on impact. MITT knew that Ms Phillips was on the footpath, but knew that this was the safest course of action. Ms Phillips' partner, Andrew Radish, is devastated.

Matthew knows that New Zealand's Accident Compensation scheme bars civil proceedings against him.[41] However, the police still want to bring criminal charges. Unfortunately, they are unsure as to who should be held responsible. Who should the criminal law punish: Matthew, ITL and/or MTL as companies, the employees and/or directors of those companies, or even MITT?

III The Issue of Definitions: What is AI?

It is important to define something before attempts are made to analogise said thing with other legal mechanisms. Before we try to apply existing laws to our hypothetical, this Part looks to the challenges in defining "AI" before determining a working definition for the purposes of this paper.

A Challenges in Defining AI

1 Lack of scientific consensus

There are three major difficulties in trying to find any definition of AI. First, there is a lack of consensus in the scientific community. Matthew Scherer notes that there is not even a consensus among AI experts, let alone a working definition that can be used for regulatory purposes.[42] Computer scientists Stuart Russell and Peter Norvig have gathered how AI has been historically defined, and have found that they are machines that can:[43]

- Think humanly;
- Think rationally;
- Act humanly; and
- Act rationally.

---

41 Accident Compensation Act 2001, s 317(1).

42 Matthew Scherer "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies" (2016) 29 Harv JL & Tech 353 at 359.

43 Russell and Norvig, above n 2, at 2.

A program can "think like a human" if it can emulate cognitive science.[44] This can be done in three ways: emulating human introspection, psychological thought experiments, and brain imaging. The analysis of these three methods is called cognitive modelling, and its goal is to replicate human thinking outside a human being.[45] A machine can then "act like a human" if it can provide evidence of human cognitive thinking through communication, passing the Turing Test (where a human interrogator "cannot tell whether the written responses come from a person or a computer").[46] In contrast, a program can "think rationally" by using the "irrefutable reasoning process" of logic;[47] a program then "acts rationally" by acting out what is the most logical in the physical circumstances and environment.[48]

Although scientists have historically defined AI under all four categories, Russell and Norvig's categories are contradictory. This is something that the authors recognise, who state that it is important for the reader to note that human beings are "are not perfect", and it is therefore necessary to distinguish the human from the rational.[49] It is difficult to ascertain which category (or categories) of definition should be adopted in general terms, let alone for a regulatory purpose.

2 Technology changes with time

Secondly, the definition of technology (and therefore what amounts to "AI") is constantly changing alongside societal expectations. Few would consider the ballpoint pen a piece of technology in 2017. However, the ballpoint pen was a revelation in post-World War II United States as it did not leak nor smear at high altitudes (a weakness of fountain pens).[50] Similarly, technology will almost certainly develop to a point where modern smartphones will be viewed as primitive relics of the past.

---

44 At 3.

45 Russell and Norvig, above n 2, at 3.

46 Russell and Norvig, above n 2, at 2.

47 For example, "Socrates is a man; all men are mortal; therefore, Socrates is mortal." Russell and Norvig, above n 2, at 4.

48 Russell and Norvig, above n 2, at 5.

49 Russell and Norvig, above n 2, at 1.

50 James Ryan and Leonard Schlup *Historical Dictionary of the 1940s* (ME Sharpe, London, 2006) at 40.

AI technology is no exception. There is now a term for the phenomenon where an AI system is no longer recognised as AI because it has reached mainstream use called "the AI effect".[51] Researchers have complained about this phenomenon, but it nonetheless reflects the view that society views technology differently throughout the ages.

3 AI = artificial + intelligence?

The final issue can be found within the term itself: "artificial" and "intelligence" are both words that are difficult to pin down from a legal perspective.

(a) "Intelligence"

First, Scherer notes that "[t]he difficulty in defining AI lies not in the concept of artificiality but rather in the conceptual ambiguity of intelligence".[52] This is because humans have been the only beings that possess "intelligence".[53] Therefore, the standard for intelligence has always been human intelligence; "we cannot yet characteri[s]e in general what kinds of computational procedures we want to call intelligent … we understand some of the mechanisms of intelligence and not others".[54] However, some AI entities now possess capabilities that rival (and even surpass) human capabilities. Since humans decide what is intelligent, this also supports the idea that the definition of AI is (at least partially) a societal issue.

Due to the difficulty of defining intelligence, some have sought to define AI has a machine who thinks and acts according to "clearly specified goals".[55] Stephen Omohundro argues that AI machines thus have four elements (or "drives") that allow it to achieve its goals: self-preservation, the ability to act efficiently, the ability to acquire resources, and the ability to act creatively. Omohundro's requirements fall within the "acting rationally"

---

51 Jennifer Kahn "It's Alive!" *Wired* (online ed, San Francisco, 1 March 2002).

52 Scherer, above n 42, at 359.

53 At 359.

54 John McCarthy "What is Artificial Intelligence?" (12 November 2007) Stanford University Computer Science Department <www-formal.stanford.edu>.

55 Stephen Omohundro "The Nature of Self-Improving Artificial Intelligence" (21 January 2008) Self-Aware Systems <www.selfawaresystems.com> at 7.

category, and are somewhat helpful in orienting towards a general definition of AI. However, this goal-oriented definition remains a hindrance. "Goal" is just as difficult to define as "intelligence", as "goal" is synonymous with "intention".[56] Unfortunately, "intention" is just as difficult to define without looking at outward evidence of that intent. A person's intention to act is usually inferred by the surrounding circumstances; a person intends the probable consequences of his or her actions.[57]

Furthermore, because AI entities are machines, "[w]hether and when a machine can have intent is a … metaphysical question" which requires a philosophical exploration of what it means to be self-aware.[58] This question is beyond the scope of this paper; however, it is sufficient to say that neither "intelligence" nor "goals" are objective, fixed, or stationary.

(b) "Artificiality"

Defining artificiality is also an issue. It is true that something is artificial if it was created by humans. However, this raises the question of where the line is between artificial and non-artificial. Take the scenario where an AI entity replicates itself. Would the products of those AI entities be considered artificial?

The line between artificial and non-artificial is further complicated when that boundary is on the same being. Perhaps something that is mostly artificial should be considered artificial. However, Professor Stephen Hawking suffers from Amyotrophic Lateral Sclerosis, and is incapable of moving and communicating without using special computer software.[59] Despite effectively being kept alive and communicating through artificial tools, it would be incorrect to label Stephen Hawking as artificial.

---

56 Scherer, above n 42, at 361.

57 Mitch Eisen "Recklessness" (1989) 31 Crim LQ 347 at 369; and *Director of Public Prosecutions v Smith* [1961] AC 290, (1960) 44 Cr App R 261 at 287.

58 Scherer, above n 42, at 361.

59 João Medeiros "Giving Stephen Hawking a voice" *Wired* (online ed, San Francisco, 2 December 2014).

B Determining a Working Definition

Obstacles aside, this paper presents a working definition of AI entities based on the evidence we have before us today. A working definition is better than no definition, especially for the purposes of discussing AI's role in the criminal law framework. First, any evidence of thinking should be external. As discussed above, the concepts of "intelligence", "intent", and "goals" are abstract formulations of a person's mind. However, once there is outward evidence of those things, it becomes possible to impute an individual's intelligence.

Secondly, machines need a degree of independence from human input to be considered as AI entities. AI entities must have a degree of autonomy. This recognises that "inputted data and programming prior to a particular operation will not necessarily result in a specific outcome in response to any given set of circumstances".[60] It is unhelpful to label AI entities as "just a programmed machine", as the programming of sophisticated AI entities are too similar to "the combination of [human] biological design and social conditioning".[61] If "the hand of human involvement in machine decision-making" is so far disconnected,[62] perhaps criminal liability and sentencing rules are relevant to AI entities.

Accordingly, this paper adopts the following definition: "AI entities" refer to machines that can perform tasks:[63]

(1) with a degree of *independence* from humans; and
(2) that *humans* would consider requiring a reasonable degree of *intelligence*.

This definition of AI falls under the "acting rationally" category. It is not, nor does it intend to be, comprehensive. The difficulties in defining "AI" rises from the fact that any

---

60 Beard, above n 23, at 651.

61 Ugo Pagallo "What Robots Want: Autonomous Machines, Codes and New Frontiers of Legal Responsibility" in Mireille Hildebrandt and Jeanne Gaakeer (eds) *Human Law and Computer Law: Comparative Perspectives* (Springer, Dordrecht, 2013) 47 at 61.

62 David Vladeck "Machines Without Principals: Liability Rules and Artificial Intelligence" [2014] 89 Wash Univ Law Rev 117 at 121.

63 This definition is similar to the one that Scherer adopts. Scherer, above n 42, at 362.

definition remains firmly rooted to subjective human factors, such as intelligence. This supports the idea that AI entities are firmly tied to societal views and social norms, which will be helpful in our analysis of criminal law below. The independence element of this definition will also bring in a level of objectivity. Therefore, this paper's definition should serve as a helpful baseline to analyse the foundations of criminal law.

IV The Issue of Liability: Who Should Be Responsible?

If Judge Easterbrook's theory holds true, it may be that existing legal mechanisms could be applied to fact scenarios involving AI entities. This section seeks to explore Judge Easterbrook's thesis for AI entities and criminal liability. While New Zealand does not have AI-specific laws, it may be able to use existing legal mechanisms to find individuals criminally liable. Section A explores the use of existing principles in corporate liability, vicarious liability, parental liability, and slavery. Section B explores Gabriel Hallevy's Virtual Models of Responsibility.

A Similar Existing Mechanisms

1 Company law and corporate criminal liability

If AI entities are too primitive in their current form, thereby making it impossible to hold AI entities themselves criminally liable, perhaps it could be argued that its owners (Matthew) or creators (ITL and MTL) should be held liable instead. This view draws similarities between existing corporate criminal liability. A company is a legal entity that carries out business. In New Zealand, a company has "full capacity to … do any act".[64] Despite being closely connected to its shareholders and controlled by its directors, a company's legal personality is separate from that of any human's.[65] This means that the company itself is independently capable of suing, being sued, and signing contracts. These fundamental aspects of a company are comparable with our working definition of AI entities, as a company is separate from its constituent directors and employees (*independence*) and has the capacity to undertake contractual obligations (*intelligence*). Therefore, a company can be analogised to an AI entity, while people responsible for the

64 Companies Act 1993, s 16(1)(a).

65 *Salomon v A Salomon & Co Ltd* [1896] UKHL 1, [1897] AC 22.

company (directors and shareholders) can be analogised to an AI entity's owners or creators.

The comparison between the company structure and AI entities is important as companies themselves can also be found criminally liable. Academics have drawn comparisons between companies and AI entities; Hallevy points out that people were initially sceptical as to how criminal liability would apply to companies, but the answer ended up being "simple and legally applicable".[66] Thus, given that terminologies in criminal law have adopted into incriminate companies, Hallevy argues that the same should be done for AI entities in the 21st Century.

It is worth briefly exploring how criminal liability attaches to companies. In New Zealand, the Crimes Act 1961 states that a "company, and any other body of persons" meets the legal definition of a "person" who can be found criminally liable.[67] If a criminal act has been committed by a company, the court has two options: to "lift the corporate veil" and find the individual actor criminally culpable (despite the fact that the company is its own separate legal entity),[68] or to use the rules of attribution to determine which individual person's actions amounted to that of the company's.[69] Determining which rule to apply depends on the purpose and context of the statute in question.[70]

A brief explanation of both methodologies will be helpful. When the court decides to "lift the corporate veil" to hold a shareholder liable for a company's actions, they are recognising two things: that the company is its own legal entity, and that there are people behind that company structure who is operating the company in fact.[71] The corporate veil

66 Gabriel Hallevy "Virtual Criminal Responsibility" (2010) 6 Orig Law Rev 6 at 22–23.

67 Crimes Act 1961, s 2.

68 *Chen v Butterfield* (1996) 7 NZCLC 261,086 (HC) at 5.

69 *Meridian Global Funds Management Asia Ltd v Securities Commission* [1995] 3 NZLR 7 (PC) [*Meridian*].

70 *Meridian*, above n 69, at 12–13.

71 The process identifies "the real nature of a transaction and the reality of the relationships created." *Attorney-General v Equiticorp Industries Group Ltd (in statutory management)* [1996] 1 NZLR 528 (CA) at 541.

can only be lifted if "special circumstances exist indicating that it is a mere facade concealing the true facts".[72] Therefore, the process of lifting the veil therefore recognises that, despite having its own legal structure, a company is a facade for the individual persons running the company itself. Given that parent companies can be shareholders for the purposes of lifting the corporate veil,[73] there is an argument to "lift the AI veil" of MITT to hold either ITL or MTL liable due to their involvement in the death of Ms Phillips in our hypothetical.

The rules of attribution (set out in the *Meridian Global Funds* case) are rules that tell the court "what acts were to count as acts of the company".[74] The phrase "directing mind and will of the company" is used to determine whose actions count as a company's actions.[75] The rules of attribution recognise that, despite it having its own legal personality, a company's actions are inherently tied to its directors and employees' actions (whoever is considered as being the "directing mind and will of the company"). Similarly, there is the argument that the "directing mind and will" of MITT could be either ITL for developing the Logic Module (programmed to protect the driver "at all costs"), or Matthew for being its owner and controller.

Both the corporate veil and rules of attribution show that "corporate activity is always the product of human agency".[76] The subjective states of mind can "only sensibly be said to be found within individuals" rather than in the abstract concept of a company's mind.[77] Corporate criminal liability and AI liability are not readily comparable. The main issue can be found in our working definition of AI entities, which involves independence. While companies have a degree of *legal* independence from humans in the form of separate corporate personality, they do not have *factual* independence from humans. Companies may be independent in the sense that they often have systems where anyone can be

---

72 *Chen v Butterfield*, above n 68, at 5; citing *Woolfson v Strathclyde Regional Council* [1978] UKHL 5.

73 *Smith, Stone and Knight Ltd v Birmingham Corporation* [1939] 4 All ER 116 (KB).

74 *Meridian*, above n 69, at 11.

75 *Meridian*, above n 69, at 11; citing *Lennard's Carrying Co Ltd v Asiatic Petroleum Co Ltd* [1915] AC 705 (HL).

76 Stephanie Earl "Ascertaining the Criminal Liability of a Corporation" [2007] 13 NZBLQ 200 at 200.

77 Earl, above n 76, at 207.

replaced; as Lord Reid recognises, a company acts "through living persons, though not always one or the same person".[78] However, a crucial element of our working definition of AI is its ability to act independently from humans. Both the corporate veil and the rules of attribution show that a company cannot commit any act, let alone a crime, unless a person attributed to that company commits that act.

There are two challenges to this argument. The first is omissions; if a company can be held criminally liable for its actions, a company can also be held liable for its non-actions. Therefore, there is an argument that a company's failure to do something is not necessarily attached to any specific person. However, the above reasoning applies to corporate omissions. It is the "relevant state of mind" of the company that has committed an act (or failed to do so), and no company can establish said state of mind without a person within the ranks of that company to do it for them.[79]

The second is health and safety laws, which involves having specific mechanisms in place to prevent injuries in the workplace.[80] A faulty piece of machinery at a factory could injure someone; that faulty machine is not a human. In New Zealand, the Health and Safety at Work Act 2015 ('HSW Act') labels all businesses and non-volunteer workplaces as a person conducting a business or undertaking (PCBU).[81] PCBUs have the primary duty of care, meaning that businesses have the primary responsibility for the health and safety of its workers and volunteers.[82] However, the same reasoning applies here; health and safety laws are designed to protect workers, and hold individual persons accountable upon any failure. This is supported by the definition of PCBU which includes any type of business whether or not it is for profit,[83] meaning that the corporate personality hurdle does not always exist. Furthermore, the HSW Act recognises that PCBUs require human assistance in meeting its primary duty of care. The HSW Act states that all PCBUs have officers, who

78 *Tesco Supermarkets Ltd v Nattrass* [1972] AC 153 (HL) at 177.
79 Earl, above n 76, at 206.
80 Health and Safety at Work Act 2015 [HSW Act].
81 HSW Act, s 17.
82 HSW Act, s 36.
83 HSW Act, s 17.

are the PCBU's senior ranking members.[84] An officer has the duty of due diligence to ensure that the PCBU complies with its primary duty of care.[85] Both omissions and health and safety regulations highlight the fact that corporations and businesses are innately connected to humans.

Hallevy argues that existing mechanisms can be easily transferred due to the analogous nature between corporations and AI entities. This is untrue; as Ingles notes, "[c]lassifying [AI entities] within the current spectrum of legal personhood is like trying to cup fine sand in your hands".[86] The ability to extend human criminal liability to companies comes from the fact that humans are a common denominator of both individuals and companies. If corporate liability were extended to AI entities, criminal liability is being applied to something that is distinct and separate from any human involvement. Whether this should be the case, and whether the wider society would find this acceptable, is a larger societal question beyond the scope of this paper. In any case, corporations and AI entities are not necessarily analogous.

2 Vicarious liability

Like corporate liability, AI entity's creators or owners could be held vicariously liable. Vicarious liability is a common law doctrine that holds a superior responsible for a subordinate's actions.[87] It is related to corporate liability, as it allows actions of an individual employee to be attributed to the company as long as the *Salmond* test is satisfied: the unlawful act was done "within in the scope of employment".[88] However, in *Bazley v Curry*, McLachlin J also applied the *Salmond* test to a non-profit organisation.[89] The Judgment states that, while it could be unfair to impose strict vicarious liability to an organisation that exists to benefit the community, it is equally unfair to not hold an

84 Section 18.
85 Section 44.
86 Ingles, above n 25, at 517.
87 *Bazley v Curry* [1999] 2 SCR 534 at [1].
88 *Bazley v Curry*, above n 87, at [6]; citing RFV Heuston and RA Buckley *Salmond and Heuston on the Law of Torts* (19th ed, Sweet & Maxwell, London, 1987).
89 At [47].

organisation accountable for the actions that its members have committed;[90] in *Bazley*, the offence was child abuse. Given that vicarious liability is based on harm, it could be argued that vicarious liability should be applied in a similar way to AI entities to hold Matthew, ITL, and/or MTL criminally liable.

Unfortunately, vicarious liability has its limitations when applied to AI entities. First, vicarious liability's role in criminal law is controversial. It can be over-inclusive as it can hold a superior criminally liable for the actions of a rogue subordinate.[91] Criminal sanctions remove significant freedoms from individuals, so evidence in criminal proceedings must be admitted cautiously.[92] Therefore, the standard of proof in criminal law is "beyond reasonable doubt", meaning that there is an "honest and reasonable certainty left in [the jury's] mind about the guilt of the accused".[93] Given the rebuttable presumption that superiors tend to have in a vicarious liability situation, most Commonwealth nations have rejected vicarious liability (including New Zealand).[94] Vicarious liability has little practical value in criminal law generally.

Secondly, even if vicarious liability is adopted, there is an issue as to who should be held liable: the AI entity's owner, its user, the software programmer, or the hardware manufacturer. This is a problem unique to AI entities, as this logistical matter was a non-issue with humans. It could be argued that elements of vicarious liability could be applied on a fact-by-fact basis depending on which party is most blameworthy. However, this runs contrary to the no-fault nature of vicarious liability.[95] If vicarious liability is to be applied to AI entities, all the parties must be found liable. However, the overwhelmingly complex makeup of AI projects raises "fundamental logistical difficulties that were nor present in earlier sources of public risk"; it would be untenable to hold every party liable for having

---

90 At [50].

91 Maxwell Smith "Corporate Manslaughter in New Zealand: Waiting for a Disaster?" (2016) 27 NZULR 402 at 403.

92 Blackstone, above n 5, at 358.

93 *R v Wanhalla* [2007] 2 NZLR 573 (CA) at [49].

94 Smith, above n 91, at 404.

95 *Bazley v Curry*, above n 87, at [1].

a part in the creation of the AI entity.[96] In the hypothetical, ITL, MTL and Matthew would all be held strictly liable, as opposed to just one party being responsible for MITT. This is not only undesirable from a logistical perspective, but could also have a chilling effect on future AI development. If ITL and MTL knew that they could be held vicariously liable not only for their own actions but also for their customers', they will be reluctant to develop AI entities. Chilling effects on developments were already an issue with vicarious liability.[97] If even more parties are involved, the issue would be amplified for AI entities.

3 Parental liability

Holding the owners of AI entities liable is similar to parental liability in some jurisdictions. Parental liability involves holding parents liable for their children's actions, even if the parents did not commit a criminal act. Parental liability is another form of vicarious liability,[98] meaning that it is worth exploring this law's applicability to AI entities and their owners.

The governing legislation for youth offending in New Zealand is the Oranga Tamariki/Children's and Young People's Well-being Act 1989 ('OT Act'). One aim of the OT Act is to promote the wellbeing of children and young persons,[99] ensuring that young persons can "develop in responsible, beneficial, and socially acceptable ways".[100] Section 283 empowers the New Zealand Youth Court to impose specific responses to proven youth offending, with these orders being divided into groups by level of restrictiveness (Group 1 responses being least restrictive).

Group 2 responses empowers the Youth Court to impose financial penalties on parents or guardians. The Youth Court can order reparations by ordering any parent or guardian of a young person under the age of 16 to pay the person who "suffered the emotional harm or

96 Scherer, above n 42, at 372.

97 Ann Barry "Defamation in the Workplace: The Impact of Increasing Employer Liability" (1989) 72 Marquette Law Rev 264 at 265–266.

98 Naomi Cahn "Pragmatic Questions About Parental Liability Statutes" (1996) Wis L Rev 399 at 410.

99 Oranga Tamariki Act 1989/Children's and Young People's Well-being Act 1989, s 4 [OT Act].

100 Section 4(f).

the loss of, or any damage to, property", or restitution to the same person.[101] The leading case on parental financial penalties is *Police v Z*, which involved a young person who had a "history of offending and behavioural difficulties".[102] The decision held that, in considering whether reparation should be ordered against the parents (or guardians) of an offender, "the focus is not on the level of culpability of the offender or the punishment appropriate to the crime committed".[103] As reparations are not punitive, the focus is on the connection between the offence and the loss or harm caused to the victim.[104] Therefore, although parental fault is a relevant consideration, it was not a requirement to determine whether a reparation order should be made against the parent or guardian.[105] Reparations under the OT Act are thus similar to vicarious liability because the parents are being held strictly liable.

Parental liability is a better legal analogy than company law and vicarious liability because AI entities are comparable to young persons; AI entities are designed to have an innate capacity to adapt and improve.[106] Omohundro argues that all biological systems follow the same categorised stages, and argues that the ideal AI system is one capable of self-improvement.[107] Self-improvement fits comfortably with this paper's working definition of AI. Given that one aim of the OT Act is to assist in development of delinquent young persons,[108] the use of parental liability is arguably a good fit for finding criminal liability of AI system owners.

Ultimately, however, parental liability is incompatible with AI entities. First, reparations is the only remedy available against parents under the OT Act. However, a monetary response is not always a satisfactory response to a crime. Take the hypothetical;

---

101 Section 283(g).
102 *Police v Z* [2008] NZCA 27, [2008] 2 NZLR 437 at [5].
103 At [24].
104 At [25].
105 At [32].
106 Omohundro, above n 55, at 5.
107 Omohundro, above n 55, at 13.
108 OT Act, s 4(f).

Ms Phillips has died, so a wholly monetary response such as reparations is unlikely to be of assistance.

Secondly, and most significantly, parental liability is rooted in the idea of the family unit. The family unit is a "major influence in the presence or absence of youth offending", and the relationship between the parent and the child is paramount.[109] Poverty and weak relationships between parents and children are likely to lead to damaged families, which in turn increases the likelihood of youth delinquency.[110] The importance of the family unit is also noted in *Police v Z*, where the court noted that "[t]he imposition of a parental reparation order is not, of itself, threatening to the stability/strength of the family group",[111] suggesting that the preservation of the family unit is one reason for why these orders exist. While it is difficult to tell whether it will be possible to program human emotions in the future, it is difficult to imagine AI entities and their owners sharing a familial relationship in 2017. Furthermore, it is even more difficult to imagine that owners neglecting their AI entities is the same as parents neglecting their children, as AI entities are unlikely to be affected in the same way as human children. The foundations of youth justice are rooted in family relationships, and such a scheme cannot be readily applied to AI entities.

4 Slavery

Some authors have taken the vicarious liability doctrine to new levels by likening AI entities to slaves in the United States and Ancient Rome. Hallevy notes that vicarious liability was rooted in the ancient slavery laws, and slave masters were criminally liable for offenses committed by their subjects.[112] The reason for this was that masters should enforce the criminal law among their own subjects; a failure to do so meant that the master deserved to be held criminally liable.[113] Hallevy's slavery analogy is arguably more persuasive than direct comparisons to vicarious liability, given that the master's subjects were treated as property; computers are personal property, after all.

---

109 Raymond Arthur "Punishing Parents for the Crimes of their Children" (2005) 44 How LJ 233 at 237.

110 At 239. Arthur's view appears to be the consensus on youth delinquency; see Nessa Lynch *Youth Justice in New Zealand* (2nd ed, Thomson Reuters, Wellington, 2016) at 252.

111 *Police v Z*, above n 102, at [31](a).

112 *When Robots Kill*, above n 13, at 65.

113 Above n 112.

However, Hallevy's traditional slavery model raises major difficulties when applied directly to AI entities. First, slavery has similar logistical difficulties to vicarious liability regarding liability. When a slave committed a criminal act, the master would be held personally liable;[114] no other parties were involved. Unlike with the liability of parents of youth offenders, however, children are not an issue with the slavery model; the child of a slave was often also a slave, meaning that liability would always come back to the master.[115] Nonetheless, the slavery model only allows for Matthew to be held liable in our hypothetical as MITT's master, even though there are convincing arguments the Logic Module was designed by ITL, or that MITT made a decision independently of Matthew's input (given that MITT was on self-driving duties).

Secondly, the application of slavery laws gives this author pause. Slavery is a sensitive area of history and cannot simply be applied directly to modern day law. New Zealand is no exception, having been involved in the Pacific slave trade in the 1870s.[116] The status of slavery had a "severe stigma" in 19th Century New Zealand, and this likely remains the case today.[117] Slavery laws have not been in effect for decades, and this author is concerned about any domino effects that may arise from using such laws by analogy (a non-company employer could use similar reasoning for an employee, for example). Academics like Hallevy were "so preoccupied with whether or not they could that they didn't stop to think if they should";[118] any consideration of slavery must be tread very carefully.

Pagallo suggests that AI entities are more suited to the Ancient Roman slavery model. He cites the Ancient Roman mechanism of *peculium* which grants limited liability to

---

114 Above n 112.

115 See for example United States. Colette Guillaumin "Race and Nature: The System of Marks" in E Nathaniel Gates (ed) *Cultural and Literary Critiques of the Concepts of "Race"* (Routledge, Abington (UK), 1997) 117 at 120.

116 Interview with Scott Hamilton, Pacific researcher (Wallace Chapman, Sunday Morning, National Radio, 27 November 2016).

117 Andrew Vayda "Maori Prisoners and Slaves in the Nineteenth Century" (1961) 8 *Ethnohistory* 144 at 146.

118 Michael Crichton and David Koepp *Jurassic Park* (Universal Pictures, California, 1993).

slaves, allowing them to maintain a degree of freedom but remaining as the head of the household's property.[119] Pagallo argues for an analogous "digital peculium", whereby AI entities are given a degree of rights and responsibilities and are thus "guaranteed by their own portfolio".[120]

The Roman slavery model lessens the logistical issue suffered by the analogies found in Hallevy's slave law and vicarious liability. It does not defer strict liability to a party that contributed to its creation and development or a party that owns it. Rather, criminal liability attaches to the AI entity itself. This line of reasoning fits more comfortably with this paper's working definition of AI entities, as it somewhat recognises a degree of independence and intelligence of AI entities.

However, it is this failure of recognition of AI entities' independence and intelligence that remains as a limitation for the Roman slavery model. The stigma of slavery law being applied in modern day society remains as a limitation for using slave analogies, and potential domino effects remain.

B Virtual Models of Criminal Responsibility

Like in the analysis in Section A, Hallevy uses existing legal mechanisms of criminal liability for all parties involved to present three new "Virtual Models of Responsibility" to find the liability of the relevant parties. The models work together and apply on a case by case basis.[121]

1 Hallevy's models

The first model is the Perpetration-by-Another Virtual Responsibility Model.[122] This model assumes that AI entities do not possess human attributes.[123] AI entities are treated as an innocent agent, even if the AI entity committed the act. Hallevy identifies two

119 Pagallo, above n 61, at 59.
120 Above n 119.
121 Hallevy "Virtual Criminal Responsibility", above n 66, at 21.
122 Hallevy "Virtual Criminal Responsibility", above n 66, at 11.
123 At 11.

candidates as the perpetrator-by-the-other: the programmer and user.[124] This model is inappropriate if the AI entity committed a criminal offense based on its own accumulated experience or knowledge, or if the AI's software was not designed to commit that specific offense but did so anyway.[125] Under this model, MITT is treated as nothing more than a car; Matthew and/or the ITL programmers would stand trial. Unfortunately, this model misses out on the fact that MITT acted independently of Matthew, and overlooks MTL's hardware involvement.

The second is the Natural-Probable-Consequence Virtual Responsibility Model.[126] If the AI entity commits a criminal offense which is the natural probable consequence of its programming or use, the programmer and/or user will be found criminally liable. This model assumes deep involvement of programmers and/or users in AI entities' daily activities, but neither intended to commit a crime. Matthew and ITL's programmers could be liable: Matthew by activating MITT's self-driving, and ITL because they programmed MITT to protect its driver "at all costs".[127] While the creator of a dangerous machine should take some responsibility for their creation's actions, the complexity of creating AI means that this model is too simplistic.

The Direct Virtual Responsibility Model finds AI entities capable of satisfying both mens rea and actus reus without relying on a human person.[128] The applicability of this model is based on how AI entities can fulfil the requirements of criminal liability and whether AI entities and humans should be distinguished.[129] Given MITT's sarcastic personality and ability to make an ethical decision, it could be argued that MITT has the capacity to form mens rea.[130] This model best recognises the fact that AI entities act independently from humans, but it must be assisted by the other models of responsibility for it to be effective.

---

124 At 11.

125 At 12.

126 At 13.

127 Bonnefon, Shariff, and Rahwan, above n 40.

128 Hallevy "Virtual Criminal Responsibility", above n 66, at 16.

129 At 16.

130 AI entities' ability to form mens rea is discussed in Part V.

2 Critique

While acting as a helpful starting point, the models outline the complications in applying the requirements of criminal law to AI entities. First, Hallevy's models fail to recognise the complex processes of how AI entities are built.[131] This limitation is similar to the issues identified with vicarious liability, parental liability, and slavery in Section A. Hallevy argues that "*a* programmer" can be found liable.[132] However, technological developments are collaborative and polymorphic.[133] Take the hypothetical; while it is possible to discern Eva Shoelace as the Head Programmer, it is unclear what her role was within the Logic Module project. ITL also employs at least 3,000 programmers worldwide. Programming an AI entity requires several groups of engineers. While certain acts can be attributed to the heads of each company, attributing each line of code and task to individual programmers is a monumental task. Hallevy also fails to note that AI is not just software-based; he fails to consider the criminal liability of hardware manufacturers. In our hypothetical, MTL would not be charged under any of Hallevy's three models. We know that MTL is known for skimping on safety features, and this may have contributed to the loss of Ms Phillips' life.

Secondly, AI code can be open source. Open source software is software where the original creator "surrender[s] all … rights granted by copyright", allowing anyone to study, change, and distribute it.[134] Some argue that open-sourcing AI promotes effective peer review.[135] Others have suggested that open sourcing AI code should be a legal requirement.[136] This further complicates pinning down liability. In a complex open-source program, there are thousands of people involved in the creation of its code, some of whom will be anonymous.[137] In our hypothetical, ITL's Logic Module runs on Android, which is

---

131 Scherer, above n 42, at 371.

132 Hallevy "Virtual Criminal Responsibility", above n 66, at 11.

133 Beard, above n 23, at 651.

134 Andrew St Laurent *Understanding Open Source & Free Software Licensing* (O'Reilly Media, Sebastopol (CA), 2004) at 4.

135 See for example OpenAI, a non-profit organisation that aims to open-source all AI code. Cade Metz "Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free" *Wired* (online ed, San Francisco (CA), 27 April 2016).

136 See for example Scherer, above n 42, at 399.

137 Beard, above n 23, at 651.

an operating system based on the open source Linux kernel.[138] Thousands of people helped build the foundation of the Logic Module, again adding to the idea that finding liability for a specific task will be monumental.

Finally, Hallevy's models are not futureproof. Hallevy proceeds with the assumption that his models will apply in the courtroom *today*, even though modern AI remains primitive.[139] Hallevy then assumes that technology evolves by applying yet-to-be-developed attributes to existing items, which is not always the case. As Charney notes, the current capabilities of AI entities "do not reach the legal standard of awareness and volition that our criminal law requires".[140] Current law cannot necessarily apply to future technology that does not yet exist.

Hallevy's models, alongside existing analogous legal mechanisms, are a good starting point for the applicability of criminal law to such scenarios. Given the exponential pace of technological developments closing the gap between humans and technology, and the idea that the law ought to evolve alongside economic norms,[141] it is not out of the question to say that AI entities will be able to be subject to the criminal law one day. However, Hallevy's models cannot be applied in their current form. His models showcase the difficulty of applying the existing criminal law framework to a scenario where an AI entity has committed the criminal act. It follows that there must be a re-examination of the existing criminal law framework.

---

138 See Sam Williams "The GNU General Public License" in *Free as in Freedom: Richard Stallman's Crusade for Free Software* (O'Reilly Media, Sebastopol (CA), 2002).

139 Rachel Charney "Can Androids Plead Automatism? A Review of *When Robots Kill: Artificial Intelligence Under the Criminal Law* by Gabriel Hallevy" (2015) 73 U T Fac L Rev 69 at 70.

140 Charney, above n 139, at 71.

141 Simon Deakin "Legal Evolution: Integrating Economic and Systemic Approaches" (June 2011) Centre for Business Research, University of Cambridge Working Paper <www.cbr.cam.ac.uk>.

V The Issue of Mens Rea: Thinking Machines?

A Overview

Part IV focused on how parties connected to an AI entity could be held liable. However, the rapid development of technology means that the fault of AI entities cannot go ignored. Accordingly, this Part of the paper explores the mens rea requirement of criminal liability, and whether AI entities themselves can satisfy that element.

On the orthodox view of liability, the Crown must prove the existence of two elements for truly criminal offences: the factual element (actus reus) and fault element (mens rea).[142] Factually speaking, most things can satisfy the actus reus element; a bear can commit the factual element of manslaughter or murder by mauling a person. Similarly, most AI entities today can satisfy the external element of a crime; a worker was killed by an automatic hydraulic arm in 1981.[143]

In contrast, mens rea asks whether the individual who carried out a criminal act was at fault. Fault is what makes an action truly criminal; even if a bear can satisfy the actus reus element of murder, it is incapable of forming its mens rea requirement.[144] Similarly, MITT committed the act that killed Ms Phillips in our hypothetical; Matthew engaged the machine on self-driving mode, and both ITL and MTL were far-removed from Glenmore Street on the day of the crime. Therefore, the question must lie with an AI entity's mens rea.

This Part of the paper deals with two issues: whether humanness is a requirement for the mens rea standard, and whether existing formulations of mens rea can extend to AI entities.

---

142 Jeffrey Gurney "Crashing into the unknown: an examination of crash-optimization algorithms through the two lanes of ethics and law" (2015) 79 Alb L Rev 224 at 240.

143 "Trust me, I'm a robot" *The Economist* (online ed, London, 8 June 2006).

144 Hallevy "Virtual Criminal Responsibility", above n 66 at 17.

B Humanness as a Requirement?

In Part III, this paper sought to define AI. In contrast, this paper did not seek to define "human" or "person". Again, on the orthodox view of criminal liability, the mens rea and actus reus are the only requirements. Criminal courts do not often undertake an inquiry as to whether the accused is a human being; this is often self-evident upon arrest. However, AI entities are not human. AI entities thus raise a question as to whether "humanness" is a requirement for criminal liability.

Criminal liability did not always require the accused to be a human; animal trials were conducted in 15th Century Europe. If an animal killed a person, that animal would be granted trial.[145] The animal itself was often held criminally liable. Ecclesiasts argued that the focus was on the result, and "pigs and locusts who harmed man must alike stand trial in the interests of universal justice".[146] A similar argument could be made for AI entities; if AI entities can harm to people (like MITT did in the hypothetical), they must stand trial. In animal trials, however, it was the owner that argued the animal's innocence,[147] and which ran counter to "commonly accepted conceptions of natural justice" given that animals are beings that are unable to defend their innocence.[148] Indeed, the practice no longer exists.

The modern legal landscape is different. The phrase "mens rea" itself is a shortened Latin maxim of criminal liability: *actus non fit reus nisi mens sit rea*, which literally translates to: "no external conduct, however serious or even fatal its consequences may have been, is ever punished unless it is produced by some form of [guilty mind]".[149] One of the roles of mens rea is for the prosecutor to show the defendant's specified mental state necessary to make their action criminal.[150] The mens rea standard is contingent on the mental state of the defendant, and it is doubtful that this requirement can be applied to non-humans.

---

145 Esther Cohen "Law, Folklore and Animal Lore" (1986) 110 Past Present 6 at 10.

146 Cohen, above n 145, at 19.

147 Cohen, above n 145, at 11.

148 Cohen, above n 145, at 15.

149 Winnie Chan and AP Simester "Four Functions of Mens Rea" (2011) 70 CLJ 381 at 381.

150 Chan and Simester, above n 149, at 382.

The definition of "person" in the Crimes Act 1961 supports this view. While the Act does not expressly state that only humans are subject to the Act, its wording and scheme suggest that humanness is a prerequisite of criminal liability. The Crimes Act defines "person", "owner", and "other words and expressions of the like kind" to include corporations, public bodies, the Crown, and "any other bodies of persons".[151] Section 4 defines the scope of "person" for the purposes of the Crimes Act. One such example is corporations. However, this paper has already discussed the challenges with applying the rules of attribution and the corporate veil in relation to AI entities in Part IV. Companies themselves can also be held criminally liable, but this paper has already discussed the strong links between corporations and humans (also in Part IV). This indicates that all definitions of "person" for the purposes of criminal liability will directly involve humans, whether individually or in groups.

The scheme of the Crimes Act also supports this view. The Act specifies its jurisdiction for certain crimes over New Zealand citizens, those who are "ordinarily residents", or is a body corporate or corporation sole incorporated under New Zealand law.[152] The Act specifies the procedure for arresting individual persons.[153] Finally, many of the criminal offences in the Act are worded to indicate humanness as a requirement as indicated by (for example)[154] the allowance of reasonable parental force being permitted by the Act,[155] the fact that it is possible to conspire to commit a crime with his or her spouse or civil union partner, and the use of the words "every one" throughout the Act.[156] Such wording indicates that criminal liability is a bad fit for AI entities, who do not fall under any of these categories of "person".

---

151 Section 2.

152 Section 7A(a).

153 Sections 30–38.

154 See for example ss 57, 61, 72, and 73.

155 Sections 59 and 60.

156 Section 67.

Thus, humanness appears to be an unofficial requirement for criminal liability. The mens rea standard was designed with humans in mind, and thus humanness appears to be a prerequisite in criminal liability.

C Types of Mens Rea

Assume that the unofficial humanness requirement for liability does not apply, and AI entities can stand trial in the criminal courts. This section discusses whether AI entities can meet the different standards of mens rea found in different criminal offences: intention, recklessness, and negligence.

1 Intention

Intention displays the highest form of culpability in a criminal and finds liability based on an individual's choice.[157] It was held in *Director of Public Prosecutions* ('DPP') *v Smith* that intention is to be viewed from the subjective state of mind of the accused.[158]

If an individual's subjective state of mind is reviewed by the courts, it must also be asked whether that individual's motives are relevant in the mens rea inquiry. However, the law has long held that the mens rea inquiry was limited to the specific intention to commit a criminal action, upholding the "orthodox theory" of mens rea.[159] In *Chandler v DPP*, it was held that a group of protestors obstructing an airfield for a political purpose was nevertheless a wrongful intention, limiting the scope of "purpose" to direct intention.[160] The same conclusion was reached in *DPP v Smith*, a case involving a man who bribed a mayor to place himself in a better position to expose the mayor's corruption; the act of bribery was itself wrongful, and the defendant had unequivocally intended to bribe the

---

157 Michael Moore "Intention as a Marker of Moral Culpability and Legal Punishability" in RA Duff and Stuart Green (eds) *Philosophical Foundations of Criminal Law* (Oxford University Press, Oxford, 2011) 179 at 179.

158 *Director of Public Prosecutions v Smith*, above n 57, at 287–288.

159 Whitley Kaufman "Motive, Intention, and Morality in the Criminal Law" (2003) 28 Crim Justice Rev 317 at 317.

160 *Chandler v Director of Public Prosecutions* [1962] UKHL 2, (1962) 46 Cr App R 347 at 371.

mayor.[161] This reasoning is persuasive; those caught committing crimes could simply say that they were acting for a righteous motive to prevent liability.[162]

Given this narrower scope, AI entities can arguably form intentions. While AI entities cannot yet form complex underlying motives, they could form simpler processes akin to direct intention. As discussed in Part III, there is an entire area of computer science dedicated to simulating human thought processes in software form called cognitive modelling. Take Artificial Neural Networks, which are computer communication systems that mimic the electrical neural signals found in animal brains.[163] It could even be argued that it is easier to prove an AI entity's intention because it is possible to physically point at their mental element through a line of code. This is not possible with humans, as it is impossible to decipher human thought in the same way. Hallevy thus argues AI entities can form a limited form of mens rea.[164]

However, a significant part of criminal liability is predicated on the idea that defendants are aware of their actions. Take the action of raising a hammer: the level of awareness will be lower for the routine work of a blacksmith compared to doing so with intent to crush someone's skull.[165] The fact that criminal liability is used as a last resort indicates that the behaviour must have been as a result of "seriously anti-social attitude" of the offender.[166] Despite AI entities having sophisticated programming which draws parallels with "the combination of [human] biological design and social conditioning",[167] modern AI entities still act through logical programming. AI entities are automatons; for them, there is no difference between the blacksmith and a skull crusher. This is because AI entities do not

---

161 *R v Smith* [1960] 2 QB 423, (1960) 44 Cr App R 55 at 62.

162 *Chandler v Director of Public Prosecutions*, above n 160, at 384–385 per Lord Devlin.

163 See Warren McCulloch and Walter Pitts "A Local Calculus of the Ideas Immanent in Nervous Activity" (1990) 52 Bull Math Biol 99.

164 *When Robots Kill*, above n 13, at 64.

165 Mordechai Kremnitzer "Is the Subjective Mental Element Superfluous?" (2008) 27 Crim Justice Ethics 78 at 80.

166 Kremnitzer, above n 165, at 81.

167 Pagallo, above n 61, at 61.

recognise that latter is antisocial and the former is not; they see a hammer as a tool that can be raised.

Furthermore, the orthodox theory of mens rea is a regulator of social values. Intention is a marker of serious culpability which is thought to be at the root of human agency.[168] Intention is subjective in nature; criminal law, being grounded in moral blameworthiness, finds liable those who make a subjective and conscious choice to cause harm.[169] The orthodox theory stresses this idea by telling society that "a good end does not justify wrongful means".[170] Furthermore, the orthodox theory makes room for motive to be determined at the prosecutorial and sentencing stages, addressing society's moral concerns.[171] Therefore, mens rea is never devoid of discussions of morality and society. If the orthodox theory of mens rea is adopted, there is also a supposition that the person with the requisite intent is a member of society. However, it is difficult to picture MITT as a member of society in 2017; it is a car with a speech module.

2 Recklessness

Recklessness involves an offender being irresponsible. The test for recklessness was discussed in the English Court of Appeal decision *R v Stephenson*, which held that an offender is reckless when he or she "carries out the deliberate act appreciating that there is a risk that damage to property may result from his act".[172] England flirted with an objective reckless standard with decisions like *R v Caldwell* and *Elliott v C*.[173] However, such decisions were criticised as it often led to unjust results where (for example) the defendant is young or educationally sub-normal;[174] indeed, criminal liability was attached to a 14-year-old girl with low intelligence in *Elliott v C*. The issue was settled in 2003 through *R v G*, which held that an offender is reckless if it was unreasonable to take a risk "in the

168 Moore, above n 157, at 180.

169 Eisen, above n 57, at 347.

170 Kaufman, above n 159, at 327.

171 Kaufman, above n 159, at 330.

172 *R v Stephenson* [1979] QB 695, [1979] EWCA Crim 1.

173 *R v Caldwell* [1982] AC 341, (1981) 73 Cr App R 13; and *Elliott v C (A Minor)* [1983] 1 WLR 939, (1983) 77 Cr App R 103.

174 David Ibbetson "Recklessness Restored" (2004) 63 CLJ 13 at 13.

circumstances known to [them]".[175] The subjective approach was upheld in New Zealand.[176]

Because of its subjective standard, however, recklessness suffers the same incompatibilities as intention when applied to AI entities. First, the recklessness test requires the offender to have appreciated the risk that they could cause damage or harm. It does not make sense to then apply the subjective test of recklessness to a being who is incapable of subjectively appreciating the risks it could create in the same way that a human could.

Secondly, the recklessness standard is a regulator of social and moral values, as it amounts to a taking of risk that could cause significant harm. Again, any subjective standard of mens rea has its grounding in moral blameworthiness, as the criminal law is used to find culpable those who choose to cause harm.[177] Much like with intention, recklessness is a subjective choice of the offender as they are aware of the risk of harm but disregards it. It is doubtful that MITT, a self-driving car programmed to speak, belongs in the society that the mens rea standard seeks to rebuke its moral blameworthiness.

3 Negligence

The final standard of mens rea, negligence, is an objective one. Due to its objective nature, negligence is technically not a form of mens rea. However, negligence assumes that the offender has shown disregard for others by failing to consider risk of harm created by his or her conduct.[178] Negligence should nonetheless be distinguished from recklessness; the latter requires the offender to be conscious of the risk taken, while the offender can be held criminally liable under negligence even if he or she was not conscious of their actions causing risk.[179]

---

175 *R v G* [2003] UKHL 50, [2004] 1 AC 1034 at 41.

176 *R v Tipple* CA217/05, 22 December 2005 at [27].

177 Eisen, above n 57, at 347.

178 Andrew Ingram "The Good, the Bad, and the Klutzy: Criminal Negligence and Moral Concern" (2015) 34 Crim Justice Ethics 87 at 89.

179 Ingram, above n 178, at 106.

Negligence is used as a standard in several criminal offences in New Zealand. An offender will be criminally liable for failing to meet a legal duty by action or omission if that action or omission is a "major departure from the standard of care" expected of a reasonable person to whom that legal duty applies.[180] This standard, known as gross negligence, applies where there is a statutory duty laid out in ss 151–157 of the Crimes Act.[181] The same standard also applies to the ill treatment or neglect of children and vulnerable adults,[182] and for unlawful act culpable homicide under ss 160(2)(a) and (b) if "the unlawful act relied on requires proof of negligence or is a strict or absolute liability offence".[183] Other serious offences apply the lower standard of ordinary negligence. Sexual violation is one example.[184] The Crimes Act does not specify a mental element for sexual violation; the offender commits rape or other unlawful sexual connection if the offender commits a sexual act "without believing on reasonable grounds" that the victim consented to that connection.[185]

Whether gross or ordinary, the standard for criminal negligence in New Zealand is objective; it is irrelevant what the offender thought at the time of the offence. It could therefore be argued that the AI entities should be held accountable under criminal negligence. Even if AI entities are not self-aware, an objective standard of mens rea does not require awareness from the offender. One function of objective standards is to protect society from certain dangerous behaviours, whether or not that the offender was aware that they were carrying out such actions.[186]

However, it is this objective nature of negligence that makes it a bad fit for AI entities. Objective standards in criminal law play an essential part in articulating the limits of individual freedom; an absence of objective standards would mean it would be "impossible

---

180 Crimes Act 1961, s 150A(2).

181 Sections 150A(1)(a).

182 Crimes Act 1961, s 195; and Law Commission *Review of Part 8 of the Crimes Act 1961: Crimes Against the Person* (NZLC R111, 2009) at [28].

183 Crimes Act 1961, s 150A(1)(b); and *R v Powell* [2002] 1 NZLR 666 (CA) at [35].

184 Bruce Robertson (ed) *Adams on Criminal Law* (online looseleaf ed, Brookers) at [CA128.05].

185 Crimes Act 1961, ss 128(2)(b) and 128(3)(b).

186 Eisen, above n 57, at 370.

to distinguish wrongful acts from accidents or from cases of justification".[187] Further, the purpose of objective standards is to send a message to members of society to take extra care and attention when carrying out certain actions.[188] Thus, the objective nature of negligence (whether gross or standard) is a regulator of moral and social values, much like the orthodox theory of intention.[189] Again, it is difficult to picture AI entities like MITT to be members of society to which social expectations can be placed upon.

D Summary

The mens rea requirement is not a good fit for AI entities and criminal liability. The subjective nature of intention and recklessness is incompatible with current AI entities as they do not have the capacity for awareness. The objective nature of negligence is incompatible with current AI entities as the objective standard sets moral boundaries on human behaviour. The mens rea requirement is contingent on the human mind, and AI entities do not have that level of capacity.

VI The Issue of Sentencing: Punishing AI Entities

Part V concluded that existing mechanisms in criminal liability were incompatible with AI entities. Assume nonetheless that AI entities can be found criminally liable; the next question is whether AI entities can be punished for their criminal actions. This Part discusses why the state punishes humans, and asks whether the same justifications can apply to AI entities and the hypothetical. In New Zealand, the purposes of sentencing are found in s 7 of the Sentencing Act 2002. This Part of the paper will explore each in turn.

A Purposes of Sentencing and AI Entities

Because all other parties in our hypothetical are human, and extensive literature has been written as to why humans are punished, this Section of the paper focuses on the purposes of sentencing AI entities themselves. Section 7 of the Sentencing Act states how

187 Lord Irvine of Lairg "Intention, Recklessness and Moral Blameworthiness: Reflections on the English and Australian Law of Criminal Culpability" (2001) 23 Sydney L Rev 5 at 17.

188 Eisen, above n 57. at 370.

189 Ingram, above n 178, at 88.

the state can justify sanctioning criminally offenders. This Section of the paper deals with each purpose found in s 7 and apply them to AI entities.

1 Accountability

One purpose of sentencing in New Zealand is to "hold the offender accountable for harm done to the victim and the community".[190] Retributive justice is one of the most widely known theories of justice as it is found in ancient religious texts. The Book of Exodus states that "if any mischief follow, then thou shalt give life for life, eye for eye, tooth for tooth, hand for hand, foot for foot".[191] Similar verses are found in the Qur'an,[192] and countries that enforce Sharia law continue to apply this principle literally.[193] This approach supports the idea that criminal punishments should involve suffering.[194]

However, the retributive principle (*lex talionis*) was never meant to be taken literally, as the principle is (and historically has been) a "measured and proportionate response to punishable conduct by a member of the community".[195] This view has support from the Kantian principle of reciprocity, as the purpose of sentencing is to "restore the 'moral equilibrium' or relationships of justice which existed prior to the offence".[196] This idea is expressed in *R v Sargeant*, which states that retribution requires a degree of societal input, and that society has a role in "show[ing] its abhorrence of particular types of crime" through sentencing.[197] Although courts cannot impose the principle of retribution on public opinion alone, "courts must not disregard it",[198] thereby distinguishing the state's retributive punishments from revenge and vigilantism. Lawton LJ is convincing as he supports the idea that society has a "moral equilibrium" in which all its members are affected.

---

190 Sentencing Act 2002, s 7(1)(a).

191 The Bible, Exodus 21:23–24 (King James Version).

192 Qur'an, 2:178.

193 See for example Iran. "Court orders Iranian man blinded" *BBC News* (online ed, London, 28 November 2008).

194 Nicola Lacey *State Punishment: Political Principles and Community Values* (Routledge, London, 1988) at 17.

195 Morris Fish "An Eye for an Eye: Proportionality as a Moral Principle of Punishment" (2008) 28 Oxford J Legal Stud 57 at 61.

196 Lacey, above n 194, at 23.

197 *R v Sargeant* (1974) 60 Cr App R 74 (CA).

198 *R v Sargeant*, above n 197.

The retributive purpose of criminal punishment is not a good fit for AI entities. First, the moral equilibrium is difficult where humans are not involved. Nicola Lacey states that the moral equilibrium is best described as the offender forfeiting a set of rights equivalent to those which he or she has violated, and returning to political society on fair terms with the law-abiding once a proportionate amount of rights has been forfeited.[199] This is at the heart of retributive theory, and sits comfortably with a historical view of *lex talionis*.[200] However, it is unlikely that AI entities are members of society, and it is even less clear whether that will ever be possible. MITT is a self-driving car; despite being able to have sarcastic retorts at the ready, it is unlikely that the inhabitants of Karori or Greater Wellington would consider MITT a member of its society.

Secondly, retribution is dependent on inflicting an unpleasant punishment which is proportionate to the level of the offence. Humans are capable of suffering, as humans can feel pain. Pain is a kind of emotion, and emotion remains difficult to define outside of producing examples.[201] However, even working definitions can prove to be a hurdle when applied to non-human AI entities. Take Michel Cabanac's definition of emotion: a "mental experience with high intensity and high hedonic content".[202] A mental experience is a distinct process of "thinking humanly", and thus falls under human intelligence. As discussed above, however, it is difficult to define what intelligence is.[203] Because AI entities are built on man-made code, humans will not only need to define intelligence, but also program intelligence from scratch. AI entities are currently incapable of feeling emotions (including pain). Accordingly, retribution is difficult to apply to AI entities. As Hallevy puts it, "[p]unishing machines, including highly sophisticated AI robots, by retribution would be the same as kicking a car."[204]

---

199 Lacey, above n 194, at 22.

200 Fish, above n 195.

201 Michel Cabanac "What is emotion?" (2002) 60 Behav Process 69 at 69–70.

202 Cabanac, above n 201, at 80.

203 Part III, Section A3. See also McCarthy, above n 54.

204 *When Robots Kill*, above n 13, at 133.

Finally, the revenge-retribution distinction does not make sense for AI entities. Again, AI entities are unable to feel pain. Hallevy notes that this fact is critical, as revenge and vigilantism is "assumed to cause more suffering to the offender than would the official punishment".[205] If AI entities cannot experience emotions, the distinction between revenge and retribution is meaningless. This again supports the idea that retribution is societal and tied to humanness.

2 Interests of victims vs promoting responsibility

Sentencing should look to promote a sense of responsibility for the harm caused (s 7(1)(b)), and provide for the interests of the victim (s 7(1)(c)).[206] These two purposes should be read together. Neither purpose is not found in traditional literature on the philosophy of punishment;[207] however, as the Court of Appeal stated in *R v Tuiletufuga*, "vindication of the law is inherent" in these statutory purposes of sentencing.[208]

Given s 7(1)(c), it could be argued that one purpose of sentencing is to minimise harm rather than on the individual who committed the crime. This is supported by s 7(2), which states that no one purpose should be given more weight than any other.[209] This purpose can be traced back to JS Mill's harm principle, which states that restrictions on freedom (including criminal sanctions) can only be justified to prevent harm to others.[210] Mill argues that the only reason to interfere with a person's actions is if they commit harm on others, whether directly or indirectly.[211] Mill intended for his principle to apply to society generally;[212] therefore, the principle can apply to the coercive use of criminal law.[213] This view is also in line with s 7(1)(c), which focuses on the impact of the offence on the victims

---

205 *When Robots Kill*, above n 13, at 133.

206 Sentencing Act 2002, ss 7(1)(b) and (c).

207 Robertson, above n 184, at [SA7.02].

208 *R v Tuiletufuga* CA205/03, 23 September 2003 at [23].

209 Sentencing Act 2002, s 7(2).

210 JS Mill *On Liberty* (eBook by Batoche Books, Ontario, 2001) at 13.

211 Mill, above n 210, at 13.

212 Mill's harm principle was to have a broad application, "whether the means used be physical force in the form of legal penalties, or the moral coercion of public opinion." Mill, above n 210, at 13.

213 See John Stanton-Ife "What is the Harm Principle For?" (2016) 10 Crim Law and Philos 329 at 330.

rather than on the offender. Therefore, the impact of the crime is an important factor in weighing up a sentence.

At face value, the harm principle arguably supports the inclusion of AI entities in sentencing. This is because the principle does not focus on the offender but on the impact of the offender's actions. Mill's harm principle does not explicitly mention wrongfulness, and does not qualify that the harm done must be illegitimate or immoral.[214] This formulation of the harm principle is arguably a good fit for AI entities. AI entities are already more than capable of committing the actus reus of various crimes in 2017. In our hypothetical, MITT has satisfied the actus reus element of manslaughter when MITT drove onto the footpath and hit Ms Phillips. Whether Ms Phillips' death was caused by a self-driving car or a careless driver, the impact and result of both actions are arguably the same: Ms Phillips is dead, and Mr Radish is devastated.

However, Mill's formulation of the harm principle is not the only driving force in sentencing. While Mill does not expressly mention wrongfulness when defining the harm principle, Mill's formulation was also intended to have a broad societal effect. The harm principle was formulated to support freedom of expression because silencing opinions is a decision that amounts to "robbing the human race".[215] Mill also argues that allowing the expression of all opinions will support the search for "living truth" (as opposed to "dead dogma"), which is also something that will benefit society at large.[216] Thus, even though the curtailment of individual freedom can only be justified if it harms others, the harm principle exists for a broader societal benefit. From a sentencing perspective, the offender cannot be removed from the equation; Mill's formulation of the harm principle does not allow for it.

Viewed another way, it is worth comparing punishment and compensation. Simple compensation restores any loss that an individual has directly suffered because of the

214 Stanton-Ife, above n 213, at 334.

215 Mill, above n 210, at 19.

216 Mill, above n 210, at 34.

action.[217] Given that AI entities can cause harm to people (like in our hypothetical), compensation is an option for AI entities (for example, through their owners). In contrast, AI entities themselves likely fall beyond the scope of punishment and sentencing as Mill's harm principle was intended to apply to the wider society. This author stated above that the impact and result of Ms Phillips' death may be the same, whether she was killed by a self-driving car or through a careless driver. However, this is unlikely to be true; Mr Radish will have different interests in each scenario. In the latter scenario, Mr Radish would likely want to hold the careless driver accountable. In the former, however, he may feel less strongly about holding Matthew accountable, given that he had nothing to do with the final act that killed Ms Phillips (aside from putting the car on self-driving mode). It is likely that Mr Radish wants nothing more than financial compensation from MITT rather than to punish it; again, anything more would amount to nothing more than "kicking a car".[218]

The harm principle must therefore work in tandem with the offence principle. The offence principle was formulated in response to the harm principle, and argues that wrongfulness (or what society considers to be "wrong") should also be a factor when deciding to curtail freedoms.[219] Stanton-Ife argues that criminal law speaks with a "distinctively moral voice", and sentences should at least consider the moral offence that society has been presented with.[220] While murder and rape can be quantified (for example by rate), Stanton-Ife also states that the inherent wrongfulness of either act must also be a positive reason to punish.[221] This author agrees. Given the societal nature of criminal law, sentencing procedures should ideally consider the level of harm and wrongfulness in tandem. Indeed, the harm principle alone is not determinative in New Zealand sentencing. As stated in *Tuiletufuga*, heavy sentences cannot be imposed on criminals for the sole purpose of meeting the interests of the victim; Parliament could not have intended

---

217 Harvey McGregor "Compensation versus Punishment in Damages Awards" (1965) 28 Mod L Rev 629 at 629.

218 *When Robots Kill*, above n 13, at 133.

219 Stanton-Ife, above n 213, at 331; citing Joel Feinberg *Offense to Others* (Oxford University Press, Oxford, 1985).

220 Stanton-Ife, above n 213, at 340.

221 Above n 220.

sentences to be shaped solely based on how vindicated the victim would feel.[222] The harm principle cannot be the only reason why a sentence is given; s 7(1)(c) is a single factor, and must be seen together with the offence principle in s 7(1)(b), alongside all other purposes of sentencing. This supports the idea that the harm and offence principles should remain excised and work together to produce a sentence.

This is further supported by the fact that both principles are also found in restorative justice. John Braithwaite defines restorative justice as a procedure where all individuals affected by an injustice can discuss how they have been affected by the injustice, and look to how any harm from that injustice can be repaired.[223] The focus is not only on the impact of the offence, but also on the individual who committed the offence. As Braithwaite states, restorative justice places active responsibility on offenders by making them discuss what circumstances may have led to the offence.[224] Restorative justice's emphasis on active responsibility reflects civic participation in a democratic society, and Braithwaite argues that it is this aspect of restorative justice that makes it an effective regulator of crime.[225] Andrew Ashworth supports this approach, emphasising that the stakeholders of all crimes are "the victim, the offender and the community".[226] This view is also encapsulated in the law through the Canadian decision *R v Gladue*, which outright states that the principles of restorative justice reflect "the needs of the victims" (s 7(1)(c)) and "the community, as well as the offender" (s 7(1)(b)).[227]

Therefore, the sentencing purposes found in ss 7(1)(b) and (c) do not apply well to AI entities, especially if the two purposes are read in tandem. Looking again at our fact scenario, a self-driving car has never been involved in a crime in New Zealand. It is also difficult to say that AI entities like MITT are a part of the fabric of New Zealand society in 2017. In contrast, ss 7(1)(b) and (c) are societal in nature. Both purposes are also found in

---

222 *R v Tuiletufuga*, above n 208, at [23].

223 John Braithwaite "Restorative Justice and De-Professionalization" (2004) 13 Good Soc 28 at 28.

224 At 28.

225 At 28–29.

226 Andrew Ashworth "Responsibilities, Rights and Restorative Justice" (2002) 42 Brit J Criminol 578 at 578.

227 *R v Gladue* [1999] 1 SCR 688 at 71.

the relationship-based restorative justice procedure. AI entities must therefore be capable of forming relationships with other humans and belong in the wider community before such purposes can apply; it is doubtful that MITT is capable of either at this moment in time.

3 Denouncing the conduct

Another purpose of sentencing is to denounce the offender's conduct.[228] Denunciation is a public condemnation of an offender by having his or her wrongdoing and its repercussions on the wider society described to them.[229] Denunciation is a form of education for the offender; Lowenstein notes that, although judicial denunciation did not necessarily have a single underlying theoretical or philosophical policy that underpinned its process, a common thread in the English and Danish jurisdictions was re-educating the offender as to what they did wrong.[230]

Like retribution, judicial denunciation is rooted in "centuries of theological and philosophical concern with sin, authority and punishment".[231] As mentioned in *R v Sargeant*, public denunciation of specific criminal behaviour can be considered a part of retribution.[232] This view was echoed in New Zealand, as a sentence must (at least in part) "reflect the value which the Court, acting as the voice of the community, places on the right in question".[233] Denunciation's community-based nature shows its overlaps with promoting a sense of responsibility and catering to the victim's interests.

It could be argued that the machine learning capabilities of AI entities means that it is possible to tell them what is right and wrong by programming them after the action has taken place. This draws parallels with a judge in the court denouncing the conduct of an offender. However, like with the purpose of promoting of responsibility, it is difficult to

228 Sentencing Act 2002, s 7(1)(e).

229 Max Lowenstein "Towards an understanding of judicial denunciation: Relating theory to practice by comparing the perceptions of English and Danish lower court judges when sentencing minor theft offenders" (2012) 13 Criminol Crim Just 21 at 21.

230 At 33.

231 Michael Marcus "Sentencing in the Temple of Denunciation: Criminal Justice's Weakest Link" (2004) 1 Ohio St J Crim L 671 at 674.

232 *R v Sargeant*, above n 197.

233 *R v Albury-Thomson* (1998) 16 CRNZ 79 (CA) at 85.34.

envision AI entities to be a part of the human society – the very group that this sentencing purpose was targeted towards. It is therefore not possible to tell MITT that the decision to run over Ms Phillips was a wrong one, and have MITT reflect over what that means.

4 Deterrence

Another aim of criminal punishment is to deter offenders.[234] A prominent Court of Appeal decision held that deterrence was "one of the main purposes of punishment", which is to protect the public by making it clear to "the offender and to other persons with similar impulses that, if they yield to them, [will be met] with severe punishment".[235]

There are two types of deterrence: general deterrence[236] and individual deterrence.[237] Individual deterrence aims to deter an individual offender from carrying out that offence in the future because of the unpleasant effects of the punishment experience.[238] Lacey argues that individual deterrence is a way of meeting social goals through the means of threatening unpleasant punishments on existing offenders;[239] the difference being that individual deterrence is for people who have already offended, whereas general deterrence is for all potential offenders considering committing a similar crime.

Like with denunciation, however, individual deterrence is unlikely to be effective for AI entities given that AI entities are not capable of feeling intimidated. Intimidation is a primary tool used by individual deterrence. However, it is again difficult to see AI entities being able to experience conscious thought (and therefore emotion).[240] Feeling intimidated is an emotion also known as fear, which is one of the examples that Cabanac lists in his paper.[241] This remains as a fatal factor in applying a sentencing purpose based on fear and intimidation to a being that cannot form emotions.

---

234 Sentencing Act 2002, s 7(1)(f).
235 *R v Radich* [1954] NZLR 86 (CA) at 87.15.
236 Lacey, above n 194, at 28.
237 Lacey, above n 194, at 32.
238 Lacey, above n 194, at 32.
239 Lacey, above n 194, at 29 and 32.
240 Section A1.
241 Cabanac, above n 201, at 70 and throughout.

General deterrence imposes penalties on an offender, treating him or her as a "means to ends" to deter others from carrying out similar conduct.[242] This purpose aims to put a "great deal of downwards pressure on levels of offending".[243] In New Zealand, general deterrence is often used to punish behaviour that is seen to be societally wrong, for example drug dealing[244] and immigration fraud.[245] If there are accessories to a crime, the Court of Appeal has held that deterring principal offenders alone is insufficient, and that those who "play subsidiary roles" must also be deterred.[246] Because it sends a message to the wider society about how certain behaviour is punished, general deterrence is linked to denunciation.[247]

This author doubts whether AI entities can be subject to general deterrence. First, like retribution and individual deterrence, general deterrence depends on potential offenders feeling intimidated. Further complicating this idea is that general deterrence is not specific to the individual offender. This not only requires AI entities to feel emotions, but also empathy. The purpose of general deterrence is to make potential offenders intimidated by placing a punishment on an individual offender (i.e. something that the individual offender is intimidated by). Therefore, AI entities not only need to be capable of having an emotional experience, but they must also be able to understand that other beings can feel emotions. If it is doubtful that MITT can feel emotions as at the time of writing, it is even less likely that MITT can feel empathy.

Secondly, general deterrence is an inherently societal punishment. In *R v Vhavha*, William Young P stated that immigration fraud is something that requires general deterrence in order to uphold "a firmly maintained border, the effective investigation and

---

242 Lacey, above n 194, at 29.

243 *R v Vhavha* [2009] NZCA 588, [2010] BCL 109 at [40] per William Young P dissenting.

244 *R v Terewi* [1999] 3 NZLR 62 (CA) at [13].

245 *R v Vhavha*, above n 243, at [22]–[23] per Chisholm and Priestley JJ and [41] per William Young P dissenting.

246 *R v Terewi*, above n 244, at [26].

247 *R v Coe* (1997) 15 CRNZ 387 (CA) at 391.25.

prosecution of immigration offences and a robust criminal justice system".[248] In *R v Terewi*, Blanchard J held that those who may be considering cultivating and dealing drugs would not be deterred if they are likely to escape imprisonment.[249] In *IRD v Song*, Mallon J held that general deterrence and denunciation were important sentencing purposes for punishing the crime of bribery.[250] These cases, alongside several others, outline the idea of sending an intimidating message to the wider society: do not carry out such conduct, or you will be subject to community detention,[251] home detention,[252] or imprisonment.[253] Much like denunciation, however, AI entities are unlikely to understand such ramifications, nor is it likely that we would consider AI entities to be members of society. It is true to say that we, as a society, do not want MITT to be using its self-driving mode to deliver methamphetamine. However, it is not true that general deterrence through the courts is a suitable mechanism through which this message is delivered.

5 Incapacitation

A major purpose of sentencing is to protect the wider society by taking offenders out of the public sphere to limit or eliminate their opportunities to reoffend.[254] Incapacitation aims to reduce the total number of offences and therefore the number of risks to the public at large.[255] In *R v Leitch*, it was held that the protection of society was a "fundamental purpose of sentencing", and that harsh sentences needed to reflect this purpose.[256]

Given that the primary rationale behind incapacitation is to protect society, and is therefore not entirely dependent on the offender, it could be argued that this purpose suits AI entities. The focus is on the actions of the offender rather than on the offender as an individual, and incapacitation as a purpose looks to remove the actions by restricting the offender's movement and freedoms. This is entirely possible with AI entities; in our

248 *R v Vhavha*, above n 243, at [42] per William Young P dissenting.
249 *R v Terewi*, above n 244, at [15].
250 *Inland Revenue Department v Song* HC Wellington CRI-2008-485-158, 10 February 2009 at [28].
251 *Inland Revenue Department v Song*, above n 250, at [33].
252 *R v Vhavha*, above n 243, at [25].
253 *R v Terewi*, above n 244, at [36].
254 Sentencing Act 2002, s 7(1)(g).
255 Lacey, above n 194, at 33.
256 *R v Leitch* [1998] 1 NZLR 420 (CA) at 428.4.

hypothetical, all it involves is putting MITT into a locked garage and using a wheel clamp. By taking either or both of those actions, MITT has been incapacitated and society is protected from further instances of being run over. The purpose has been met, and the criminal sanction on AI entities is justified.

However, it could equally be argued that incapacitation brings with it a social aspect. Incapacitation amounts to removing an offender's freedoms because they have caused harm. Therefore, incapacitation is Mill's harm principle in action yet again.[257] Like with victim's interests, the harm principle is intended to better society. Hallevy argues that incapacitation is an "expression of disappointment" on an individual offender after other purposes such as deterrence have been unsuccessful, meaning that the court must resort to incapacitation.[258] This author is convinced by this argument. Take the bear who mauled someone in Part V. Much like how it is possible for that bear to commit the actus reus element of manslaughter, it is equally possible to lock up said bear in a cage so that the risk of being mauled to death is removed. However, there is a difference between locking up a dangerous bear and imprisoning a rapist to protect members of society from sexual assault; the former is the removal of a direct source of danger, while the latter is also the sternest statement on a member of society who has failed to live up to society's expectation of having a safe and welcoming civilisation. Criminal incapacitation performs both functions, and it is doubtful that AI entities like MITT have an expectation from society for the purposes of the latter function.

6 Rehabilitation and reintegration

Finally, one aim of sentencing is rehabilitating the offender and reintegrating them back into society.[259] It aims to treat offenders through various means, ranging from counselling to psychotherapy.[260] For clarity, rehabilitation refers to an offender's "attitudes, values,

[257] Mill, above n 210, at 13.
[258] *When Robots Kill*, above n 13, at 136.
[259] Sentencing Act 2002, s 7(1)(h).
[260] Lacey, above n 194, at 30.

cognitive and problem-solving skills", while reintegration is about strengthening an offender's law-abiding tendencies.[261]

Like the name suggests, reintegration into society is a societal goal. Accordingly, it is not likely to be a good fit for AI entities. AI entities are not members of human society at the time of writing. MITT has never been a part of the Karori or Wellington Central community, so it does not make sense to say that MITT will be reintegrated back into either group as a law-abiding citizen.

In contrast, rehabilitation is more a promising sentencing purpose for AI entities. Courts have used rehabilitation to reduce a sentence if the offender has made genuine efforts to recognise and address the causes of delinquency. In *R v Hill*, the Court of Appeal was wary of increasing a home detention sentence given that the offender had shown a "real commitment to change and is working toward that in specific and realistic ways".[262] The same court in *R v Rawiri* held that judges should "strive to avoid a custodial sentence where there is a genuine prospect of rehabilitation" for the purposes of the Misuse of Drugs Act 1975.[263] Such statements indicate that a major purpose of rehabilitation is to recognise and address the roots of delinquency rather than to rebuke the offender for having committed a societal wrong. Such a view is compatible with AI entities through machine learning and programming.[264] Take MITT: in theory, ITL could rehabilitate MITT (and future self-driving cars) by reprogramming them in so that they do not kill in any circumstances.[265]

B Summary

This Part of the paper discussed the rationales behind sentencing in New Zealand, and has argued that all those purposes are deeply rooted in society. Accordingly, all but one of the sentencing purposes sat uncomfortably with AI entities. On the one hand, it could be

---

261 Robertson, above n 184, at [SA7.07].

262 *R v Hill* [2008] NZCA 41, [2008] 2 NZLR 381 at [39].

263 *R v Rawiri* [2011] NZCA 244, (2011) 25 CRNZ 254 at [22].

264 *When Robots Kill*, above n 13, at 135.

265 Because MITT made an ethical decision, however, there are issues as to whether it needs rehabilitation at all. All that needs to be said is that MITT's tendencies to kill (whatever the circumstances) have been rehabilitated, and that an exploration of ethics is beyond the scope of this paper.

possible to rehabilitate AI entities through reprogramming and machine learning. However, it does not make sense to hold AI entities accountable, nor is it possible to deter AI entities from carrying out a certain act, given that they are incapable of experiencing fear or intimidation. It is not possible to promote a sense of responsibility to AI entities as responsibility is associated with forming and exploring complex relationships. It is not possible to denounce an AI entity's conduct, as a consensus on what is to be denounced comes (at least partly) from community participation. Finally, it does not make sense to incapacitate an AI entities, at least from a criminal law perspective, as the purpose of criminal incapacitation is to provide a stern statement on who has failed to meet society's expectations of living in a safe and hospitable environment.

The law of sentencing (and criminal law generally) does not exist in a vacuum. In *State Punishment*, Lacey argues that criminal punishment is an act of social practice within a community, "geared towards the pursuit of … a plurality of the community's central goals and values".[266] A societal view of criminal punishment draws parallels with Lessig's theory, as it shows an interplay between the constraints of law and social norms. Both Parliament and the Supreme Court of New Zealand have endorsed this view when enacting the principles and purposes of sentencing in the Sentencing Act 2002; not only were the provisions enacted for the benefit of judges, but they also serve to "foster greater awareness of the public concerning the complexity of what has to be considered in the sentencing task".[267] It follows that criminal law is firmly tied to the wider society; unfortunately, AI entities are not tied to either.

VII Conclusion

> "In an all-inclusive system encompassing … the extra-human world, [criminal trials] reveal man's view of his place within the universal scheme as well."[268]

---

266 Lacey, above n 194, at 200.

267 *Hessell v R* [2010] NZSC 135, [2011] 1 NZLR 607 at [42].

268 Cohen, above n 145, at 35.

While Judge Easterbrook's analogy to the Law of the Horse is helpful in keeping the law grounded, existing criminal law is ill-equipped to deal with AI entities. Lessig's four constraints hold true in relation to AI entities. In our hypothetical, MITT's action was one that could amount to manslaughter or murder. Both are regulated within New Zealand; the government wants to regulate the killing of people by banning it altogether, so it is illegal (through *criminal law*) to kill people.[269] The government also affects the *architecture* of society, as well as the *market*, by (for example) requiring a firearm licence to purchase a gun (which requires the licence holder to pass a safety test) which limits gun access.[270] Finally, the government addresses *social norms* by using advertising campaigns that discourage dangerous driving.[271] The hypothetical shows that the relationship between all four constraints are essential in regulating behaviour. The law may affect the marketplace, physical architecture, and social norms, and the latter three constraints also have an impact on how the law works. Criminal law is no exception; in fact, given the societal and moral roots of criminal law,[272] the relationship between the law and other constraints are even more significant.

Killer robots show that the criminal law is a system that is rooted in what it means to be a member of society. This paper explored this issue through existing mechanisms of criminal liability and the purposes of sentencing. While this paper discussed criminal law in relation to currently non-existent advanced machines, similar issues can be raised with people. Why does the state get involved in punishing individuals who have committed certain wrongs? It is because all individuals before the criminal courts belong in the society of New Zealand, equally and throughout. If the criminal system does not reflect this idea, punishments lose all power for some members and becomes disproportionately controlling for others. Criminal sanctions should apply equally to Māori men, Pākehā women, and everyone else, based on the offence committed.

---

269 Crimes Act 1961, ss 167–181.

270 Arms Regulations 1992.

271 See for example the "If you drink and drive, you're a bloody idiot" campaign. Reece Hooker "Greg Harper, mastermind behind 'Drink Drive, Bloody Idiot' ads passes away" *The Age* (online ed, Melbourne, 15 January 2017).

272 Blackstone, above n 5.

As it stands, MITT is unlikely to be brought before the criminal courts. The police would instead likely look to MTL as manufacturers, ITL as programmers, or Matthew as the owner. Criminal law was specifically designed with such people in mind; it was made by and for human beings within the wider society. Despite its sophisticated circuitry, MITT lacks the capacity to form mens rea. It lacks the remorse to be punished. Most importantly, it lacks the humanity to fall under the umbrella of criminal law. While AI entities may yet be subject to criminal law in the future, criminal law is, and historically has been, a bad fit. Whether the societal nature of criminal law should change is a philosophical question beyond the scope of this paper, and is one that requires rigorous debate about the role of AI systems in our society.

Considering whether AI entities are subject to criminal law is, ironically, not just a question of law. Rather, it is a complex question that requires consideration of two things: first, how AI technology develops in the future, and thereby AI being able to form conscious thought (thus being able to form mens rea beyond doubt, and being punished); and secondly, how society will view such developments in technology, and whether the wider public will accept criminal sanctions being placed on a non-human entity (if humans were to accept them). Both questions are deeply technical and social issues beyond the scope of this paper. However, it is certain that existing criminal mechanisms were not designed with such technological developments in mind. This confirms that criminal law is not mere regulation of behaviour, and any changes to criminal must be explored with conscious consideration of both questions.

With technology moving at breakneck speed, both conversations need to take place soon. However, one must note that the two conversations are intertwined. Neither criminal law nor society exist in a vacuum; in fact, they are inseparable.

VIII Bibliography

A Cases

1 New Zealand

Attorney-General v Equiticorp Industries Group Ltd (in statutory management) [1996] 1 NZLR 528 (CA).

Chen v Butterfield (1996) 7 NZCLC 261,086 (HC).

Hessell v R [2010] NZSC 135, [2011] 1 NZLR 607.

Inland Revenue Department v Song HC Wellington CRI-2008-485-158, 10 February 2009.

Meridian Global Funds Management Asia Ltd v Securities Commission [1995] 3 NZLR 7 (PC).

Police v Z [2008] NZCA 27, [2008] 2 NZLR 437.

R v Albury-Thomson (1998) 16 CRNZ 79 (CA).

R v Coe (1997) 15 CRNZ 387 (CA).

R v Hill [2008] NZCA 41, [2008] 2 NZLR 381.

R v Leitch [1998] 1 NZLR 420 (CA).

R v Powell [2002] 1 NZLR 666 (CA).

R v Radich [1954] NZLR 86 (CA).

R v Rawiri [2011] NZCA 244, (2011) 25 CRNZ 254.

R v Terewi [1999] 3 NZLR 62 (CA).

R v Tipple CA217/05, 22 December 2005.

R v Tuiletufuga CA205/03, 23 September 2003.

R v Vhavha [2009] NZCA 588, [2010] BCL 109.

R v Wanhalla [2007] 2 NZLR 573 (CA).

2 United Kingdom

Chandler v Director of Public Prosecutions [1962] UKHL 2, (1962) 46 Cr App R 347.

Director of Public Prosecutions v Smith [1961] AC 290, (1960) 44 Cr App R 261.

Elliott v C (A Minor) [1983] 1 WLR 939, (1983) 77 Cr App R 103.

Lennard's Carrying Co Ltd v Asiatic Petroleum Co Ltd [1915] AC 705 (HL).

R v Caldwell [1982] AC 341, (1981) 73 Cr App R 13.

R v G [2003] UKHL 50, [2004] 1 AC 1034.

R v Sargeant (1974) 60 Cr App R 74 (CA).

R v Smith [1960] 2 QB 423, (1960) 44 Cr App R 55.

R v Stephenson [1979] QB 695, [1979] EWCA Crim 1.

Salomon v A Salomon & Co Ltd [1896] UKHL 1, [1897] AC 22.

Smith, Stone and Knight Ltd v Birmingham Corporation [1939] 4 All ER 116 (KB).

Tesco Supermarkets Ltd v Nattrass [1972] AC 153 (HL).

Woolfson v Strathclyde Regional Council [1978] UKHL 5.

3 Canada

Bazley v Curry [1999] 2 SCR 534.

R v Gladue [1999] 1 SCR 688.

B Legislation and Regulations

Accident Compensation Act 2001.

Arms Regulations 1992.

Companies Act 1993.

Copyright Act 1994.

Crimes Act 1961.

Harmful Digital Communications Act 2015.

Health and Safety at Work Act 2015.

Oranga Tamariki Act 1989/Children's and Young People's Well-being Act 1989.

Sentencing Act 2002.

C Books and Chapters in Books

William Blackstone *Commentaries on the Laws of England* (9th ed, reissue, 1978) vol 4 Of Public Wrongs.

Joel Feinberg *Offense to Others* (Oxford University Press, Oxford, 1985).

Colette Guillaumin "Race and Nature: The System of Marks" in E Nathaniel Gates (ed) *Cultural and Literary Critiques of the Concepts of "Race"* (Routledge, Abington (UK), 1997) 117.

Gabriel Hallevy *When Robots Kill: Artificial Intelligence Under Criminal Law* (Northeastern University Press, Lebanon (NH), 2013).

HLA Hart *Law, Liberty and Morality* (Oxford University Press, Oxford, 1963).

HLA Hart "Positivism and the Separation of Law and Morals" in *Essays in Jurisdiction and Philosophy* (Oxford University Press, Oxford, 1983) 49.

RFV Heuston and RA Buckley *Salmond and Heuston on the Law of Torts* (19th ed, Sweet & Maxwell, London, 1987).

Ray Kurzweil *The Singularity is Near: When Humans Transcend Biology* (Viking Press, New York City, 2005).

Nicola Lacey *State Punishment: Political Principles and Community Values* (Routledge, London, 1988).

Nessa Lynch *Youth Justice in New Zealand* (2nd ed, Thomson Reuters, Wellington, 2016).

JS Mill *On Liberty* (eBook by Batoche Books, Ontario, 2001).

Michael Moore "Intention as a Marker of Moral Culpability and Legal Punishability" in RA Duff and Stuart Green (eds) *Philosophical Foundations of Criminal Law* (Oxford University Press, Oxford, 2011) 179.

Ugo Pagallo "What Robots Want: Autonomous Machines, Codes and New Frontiers of Legal Responsibility" in Mireille Hildebrandt and Jeanne Gaakeer (eds) *Human Law and Computer Law: Comparative Perspectives* (Springer, Dordrecht, 2013) 47.

François Rabelais *Gargantua* (France, 1534).

Stuart Russell and Peter Norvig *Artificial Intelligence: A Modern Approach* (3rd ed, Prentice Hall, New Jersey, 2010).

James Ryan and Leonard Schlup *Historical Dictionary of the 1940s* (ME Sharpe, London, 2006).

Andrew St Laurent *Understanding Open Source & Free Software Licensing* (O'Reilly Media, Sebastopol (CA), 2004).

Sam Williams "The GNU General Public License" in *Free as in Freedom: Richard Stallman's Crusade for Free Software* (O'Reilly Media, Sebastopol (CA), 2002).

D Journal Articles

Raymond Arthur "Punishing Parents for the Crimes of their Children" (2005) 44 How LJ 233.

Andrew Ashworth "Responsibilities, Rights and Restorative Justice" (2002) 42 Brit J Criminol 578.

Ann Barry "Defamation in the Workplace: The Impact of Increasing Employer Liability" (1989) 72 Marquette Law Rev 264.

Jack Beard "Autonomous Weapons and Human Responsibilities" (2014) 45 Georgetown J Int Law 647.

Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan "The social dilemma of autonomous vehicles" (2016) 352 Science 1573.

John Braithwaite "Restorative Justice and De-Professionalization" (2004) 13 Good Soc 28.

Michel Cabanac "What is emotion?" (2002) 60 Behav Process 69.

Naomi Cahn "Pragmatic Questions About Parental Liability Statutes" (1996) Wis L Rev 399.

Winnie Chan and AP Simester "Four Functions of Mens Rea" (2011) 70 CLJ 381.

Rachel Charney "Can Androids Plead Automatism? A Review of *When Robots Kill: Artificial Intelligence Under the Criminal Law* by Gabriel Hallevy" (2015) 73 U T Fac L Rev 69.

Esther Cohen "Law, Folklore and Animal Lore" (1986) 110 Past Present 6.

Stephanie Earl "Ascertaining the Criminal Liability of a Corporation" [2007] 13 NZBLQ 200.

Frank Easterbrook "Cyberspace and the Law of the Horse" [1996] 207 U Chi Legal F 207.

Mitch Eisen "Recklessness" (1989) 31 Crim LQ 347.

Morris Fish "An Eye for an Eye: Proportionality as a Moral Principle of Punishment" (2008) 28 Oxford J Legal Stud 57.

Richard Fuller "Morals and the Criminal Law" (1942) 32 J Crim Law Criminol 624.

Jeffrey Gurney "Crashing into the unknown: an examination of crash-optimization algorithms through the two lanes of ethics and law" (2015) 79 Alb L Rev 224.

Gabriel Hallevy "Virtual Criminal Responsibility" (2010) 6 Orig Law Rev 6.

David Ibbetson "Recklessness Restored" (2004) 63 CLJ 13.

Ignatius Ingles "Regulating Religious Robots: Free Exercise and RFRA in the Time of Superintelligent Artificial Intelligence" (2017) 105 Geo LJ 507.

Andrew Ingram "The Good, the Bad, and the Klutzy: Criminal Negligence and Moral Concern" (2015) 34 Crim Justice Ethics 87.

Lord Irvine of Lairg "Intention, Recklessness and Moral Blameworthiness: Reflections on the English and Australian Law of Criminal Culpability" (2001) 23 Sydney L Rev 5.

Sandra Jacobs "Natural Law, Poetic Justice and the Talionic Formulation" (2013) 14 Political Theology 661.

Whitley Kaufman "Motive, Intention, and Morality in the Criminal Law" (2003) 28 Crim Justice Rev 317.

Mordechai Kremnitzer "Is the Subjective Mental Element Superfluous?" (2008) 27 Crim Justice Ethics 78.

Lawrence Lessig "The Law of the Horse: What Cyberlaw Might Teach" (1999) 113 Harv Law Rev 501.

Lawrence Lessig "The New Chicago School" (1998) 27 J Legal Stud 661.

Max Lowenstein "Towards an understanding of judicial denunciation: Relating theory to practice by comparing the perceptions of English and Danish lower court judges when sentencing minor theft offenders" (2012) 13 Criminol Crim Just 21.

Michael Marcus "Sentencing in the Temple of Denunciation: Criminal Justice's Weakest Link" (2004) 1 Ohio St J Crim L 671.

Warren McCulloch and Walter Pitts "A Local Calculus of the Ideas Immanent in Nervous Activity" (1990) 52 Bull Math Biol 99.

Harvey McGregor "Compensation versus Punishment in Damages Awards" (1965) 28 Mod L Rev 629.

Tracey Meares "Social Organization and Drug Law Enforcement" (1998) 35 Am Crim L Rev 191.

Matthew Scherer "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies" (2016) 29 Harv JL & Tech 353.

Maxwell Smith "Corporate Manslaughter in New Zealand: Waiting for a Disaster?" (2016) 27 NZULR 402.

John Stanton-Ife "What is the Harm Principle For?" (2016) 10 Crim Law and Philos 329.

Judith Thomson "The Trolley Problem" (1985) 94 Yale LJ 1395.

Andrew Vayda "Maori Prisoners and Slaves in the Nineteenth Century" (1961) 8 *Ethnohistory* 144.

David Vladeck "Machines Without Principals: Liability Rules and Artificial Intelligence" [2014] 89 Wash Univ Law Rev 117.

E Reports

Law Commission *Review of Part 8 of the Crimes Act 1961: Crimes Against the Person* (NZLC R111, 2009).

F Internet Resources

1 News articles

"Court orders Iranian man blinded" *BBC News* (online ed, London, 28 November 2008).

"Trust me, I'm a robot" *The Economist* (online ed, London, 8 June 2006).

Reece Hooker "Greg Harper, mastermind behind 'Drink Drive, Bloody Idiot' ads passes away" *The Age* (online ed, Melbourne, 15 January 2017).

Matthew Hutson "Our Bots, Ourselves" *The Atlantic* (online ed, Washington DC, March 2017).

Jennifer Kahn "It's Alive!" *Wired* (online ed, San Francisco, 1 March 2002).

João Medeiros "Giving Stephen Hawking a voice" *Wired* (online ed, San Francisco, 2 December 2014).

Cade Metz "Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free" *Wired* (online ed, San Francisco (CA), 27 April 2016).

Danielle Muoio "Tesla's new Autopilot is getting a big update this weekend – here's everything you need to know" *Business Insider* (online ed, New York City, 16 June 2017).

Jon Russell "After beating the world's elite Go players, Google's AlphaGo AI is retiring" *TechCrunch* (online ed, San Francisco Bay Area, 27 May 2017).

Mark Smith "So you think you chose to read this article?" *BBC News* (online ed, London, 22 July 2016).

2 Other

"Chapter 4: Automated Flight Control" Federal Aviation Administration <www.faa.gov>.

Litigation Support & Discovery Management – E-Discovery Consulting <www.e-discovery.co.nz>.

Simon Deakin "Legal Evolution: Integrating Economic and Systemic Approaches" (June 2011) Centre for Business Research, University of Cambridge Working Paper <www.cbr.cam.ac.uk>.

Interview with Scott Hamilton, Pacific researcher (Wallace Chapman, Sunday Morning, National Radio, 27 November 2016).

John McCarthy "What is Artificial Intelligence?" (12 November 2007) Stanford University Computer Science Department <www-formal.stanford.edu>.

Stephen Omohundro "The Nature of Self-Improving Artificial Intelligence" (21 January 2008) Self-Aware Systems <www.selfawaresystems.com>.

G Other Resources

The Bible (King James Version).

Qur'an.

Yannis Assael and others "LipNet: End-to-End Sentence-level Lipreading" (paper presented to International Conference on Learning Representations, Toulon, April 2017).

Michael Crichton and David Koepp *Jurassic Park* (Universal Pictures, California, 1993).

Bruce Robertson (ed) *Adams on Criminal Law* (online looseleaf ed, Brookers).

Georgios Yannakakis "Game AI Revisited" (paper presented at the Proceedings of the 9th conference on Computing Frontiers, Cagliari, May 2012).

Word count: the text of this paper (excluding abstract, table of contents, non-substantive footnotes, and bibliography) comprises exactly 14,776 words.