Law and Ethics of Morally Significant Machines: The case for pre-emptive prevention

Curtis Reid Barnes

Thesis submitted in fulfilment of the degree of Master of Laws

at the University of Otago, Dunedin, New Zealand

Date: 30 June 2017

Overview

In the diverse theatre of contemporary emerging technology issues, few generate quite the same sort of excitement and concern as Artificial Intelligence (hereafter "AI"). Both internationally and nationally, there is a growing consensus that AI technologies have the capacity to substantially revolutionise multiple aspects of human life and society. Few other fields promise comparable breadth of impact. Among the leading areas of scholarship which are beginning to draw the interests of governments and lawmakers are problems as diverse as the impact of automation on economics and labour markets, to privacy through the collection and use of data which may be analysed on an unprecedented scale, to the ethical impacts of machine decision-making across medicine, law, finance, and so on. Other issues include ensuring safety and traceability, engendering ethical or moral behaviours and intuitions in the machines themselves, and determining legal liability and accountability, with significant scope for dispute as machine autonomy increases. Further unknown are the softer impacts on humanity through increased human-robotic interaction: how machines may exacerbate or alleviate growing human loneliness and isolation, expand or diminish inequality, dehumanise or rehumanise our attitudes towards each other, and act as vessels or agents for the advancement of human culture. There is also substantial interest in the possibility of existential threat arising through AI; whether that be through maleficent machine behaviours or well-intentioned behaviours with adverse outcomes. There is also interest in the subject of human-AI amalgamation, with some persons advocating that such will be necessary to keep humans abreast of growing computational power and advanced, autonomous machines.

Thus, the scope of issues for policymakers and regulators to consider is expansive. Generally, the focus is on what should or should not be done with machines in pursuit of human interests. A less common question with perhaps even more profound impacts is what should or should not be done *in spite of* human interests. In other words, what ought to be legal and ethical in the way we interact with machines, from the perspective of the machines themselves. Is it possible that machines could be built which have interests, wellbeing, and so on? Such entities would challenge the entrenched conception of the human relationship with our technology; the human as the artisan; the machine, the artefact. This type of machine would challenge the very concept of what a machine is.

In the case that a machine like this existed, it would be naïve to expect sympathy for its normative claims on our behaviour. So long as there is competition for resources, there will always be opposition to any moral theory which would oblige us to consider acknowledging that new entities have claims upon these resources which may be similar, equal, or perhaps superior to our own. This is exacerbated when the entity whose moral significance is being denied is itself a potential resource. Thus, there is the predictable double-motivation to deny the ethical claims of moral machines, if and when we manage to create them; accidentally, or deliberately.

This thesis employs a *reductio ad absurdum* approach to explore the legal and ethical implications of machine moral significance. It proceeds in the following fashion: first, I argue that morally significant machines are plausible in the future based on contemporary AI paradigms. Thus, there is impetus to considering the issue pre-emptively. Secondly it considers how machines of this kind may possess moral elements, and perhaps personhood. It then progresses through a range of problems, both practical and theoretical, that these types of machines would cause for our societies and legal system. Overall, the implications of these machines would be such that it is reasonable to question whether or not there is any net benefit to their creation. Put differently, it is reasonable to suggest whether there ought to be a comprehensive policy approach to prevent their ever being created.

The questions and issues that provoke this conclusion are numerous. How can we detect them from morally insignificant machines? Where do we draw lines regarding at what point they matter morally? How will we adapt fundamental legal paradigms like property and personhood to account for their nature? What could we do to robotic persons? More bluntly, what violence could we *not* do to them? More onerously, are their reasonable duties we might owe towards such machines to ensure their welfare and ethical treatment? How would we see that these things are upheld, and ultimately, how will we include these entities within the umbra of a distinctly anthropocentric world? Approaching the problem in this way suggests morally significant machines, particularly machine persons, are more problematic to law and society than they are likely to be beneficial. Subsequently, strategizing how to solve these future problems should be a significant part of the pantheon of AI policy concerns.

Contents

Chapter 1: Introduction

1.1 Research Methodology

The thesis draws on theory across multiple disciplines, in particular: cognitive science and computer science, moral philosophy and ethics, law and jurisprudence. It is cross-disciplinary, and apportions significant effort to establishing how contemporary AI paradigms might produce a particular kind of artificial intelligence which differs from any conventional artefact currently in existence.

1.2 Scope

The scope of the thesis is broad. The reader will encounter an argument of how contemporary AI paradigms will produce a particular type of autonomous learning agent (the 'EMBOT'), how this EMBOT will have the cognitive basis for intrinsic moral status and perhaps moral personhood, and a variety of legal and ethical problems arising from the creation of such a machine.

The objective of exploring a particular method by which this kind of machine may be created is that it will be grounded in contemporary AI paradigms. As a result, the potential characteristics may be more discernible than a machine discussed entirely in the hypothetical. Furthermore, it gives policymakers opportunity to investigate a particular area of AI research which may have a key role in the creation of an AI system with intrinsic moral status, allowing them to be more targeted with any policy concerning this entity.

1.3 Structure

The thesis is divided into two parts. Part One contains two chapters, the first of which is a Technical Proposal for the 'EMBOT'; why these AI paradigms are being preferred, what they are, and what type of machine they plausibly may produce in the future. Chapter 2 makes the Moral Proposal for EMBOT; that

the EMBOT will have intrinsic moral status, and even the potential for personhood. This is grounded in an analysis of moral philosophy which suggests the moral materiality of various cognitive capacities.

Part Two of the thesis substantially explores feasible legal and ethical issues associated with the machine described in Part One. It proposes a number of 'problems', many of which have elements which are novel to the concept of a morally significant machine, or machine person. These problems include The Line-Drawing Problem (Chapter 4), the Potentiality Problem (Chapter 5), The Property Problem (Chapter 6), The Personality Problem (Chapter 7), and The Protection Problem (Chapter 8). The final chapter of the thesis (Chapter 9) concludes and proposes the need to develop a pre-emptive approach to morally significant artificial systems; in all likelihood, a preventative one.

1.4 Literature Review

Literature is discussed and engaged with incrementally throughout the thesis as required. This approach was preferable to a standalone literature review, owing to the amount of literature used and the variety of disciplines drawn upon. Nevertheless, a segment is included here to serve as a general introduction.

i) AI is an emerging concern at the highest levels

The field of AI significantly precedes contemporary regulatory interest in it. However, that interest has crystallised to such an extent that AI is now perceived as one of the defining emerging technologies of modernity. With AI perceived to be as impactful as many other major technologies, enthusiasm to regulate its development and has increased. In 2006 the European Robotics Network predicted that, "Like Nuclear Physics, Chemistry or Bioengineering, in a few years, Robotics could also be placed under scrutiny from an ethical standpoint by the public and Public Institutions (Governments, Ethics Committees, Supranational Institutions)."[1] That projection is now being realised.

One of the most significant regulatory responses to AI is the European Parliament's development of its Report with recommendations to the Commission on Civil Law Rules on Robotics. This report canvasses a number of areas and impacts. However, it does not engage with the question of what should be done about the potential to create machines with intrinsic moral status in the future. This is an esoteric problem,

[1] European Robotics Network (EURON) Roboethics Atelier "Roboethics Roadmap" (July 2006) at 5.

with potential for enormous impact. The Report posits that the guiding "ethical principle" of all AI development and research should be human beneficence: research should be conducted, "from the point of view of human safety, health and security; freedom, privacy, integrity and dignity; self-determination and non-discrimination, and personal data protection".[2] I will discuss this Report in greater detail below. Within this principle alone, the key question becomes whether morally significant machines have a beneficence to humans that outweighs the difficulty of having their own ethical claims upon human behaviour.

There is only a small amount of 'formal' literature currently available – that is, literature produced by governments, lawmakers, and policymakers. This can be understood in the context of the recency with which AI has become a widespread concern.

ii) 'Machines with moral status' are not yet a recognised legal concern

Thus far most of the engagement with the question of machines with moral status – what David Gunkel refers to as "*the* machine question" for its profundity – has occurred outside of any sort of formal legal context.[3] Nevertheless, the relevant issues are rarely canvassed legal and policy-focused approaches. The fact that machine moral significance is a more long-term concern cannot be the only factor influencing its absence from the debate, since other AI concerns which are arguably more remote are regularly considered; e.g. superintelligences. To a certain extent, it is difficult to explain the absence of reference to the possibility of 'robotic persons' among lawyers, since the issue is very well-broached outside of the legal discipline. It is also unusual, because the invention of a robotic person would be so challenging to so many areas of law and society. EURON has predicted, "we can forecast that in the XXI century humanity will coexist with the first alien intelligence we have ever come into contact with – robots".[4]

Any number of reasons could inform the current legal and political disengagement with the robotic personhood problem: a perception of infeasibility; of unimportance; or wider philosophical issues related, for instance, how morally significant machines challenge the status quo of moral hierarchy. Alternatively, it could simply be that legal and political interest into AI-related issues is so fledgling that they have not

[2] European Parliament Committee on Legal Affairs, Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), 2017, at [10].
[3] See generally DJ Gunkel *The Machine Question: critical perspectives on AI, robots, and ethics* (MIT Press, Cambridge, MA, 2012).
[4] Above n 2 at 5.

comprehensively considered all the problems. What this thesis will do is attempt to consider and convey how challenging the issues of robotic personhood may be, and essentially, compel policymakers to begin discussing it; particularly as there is potential that machines of this type may be produced from contemporary AI paradigms, before we ever seriously debate the ethicality of their creation.

iii) Fiction as an incubator for the robotic personhood question

The questions of machines with moral status have long been central to science fiction, along with many of the possible AI technologies that, at the time they were envisioned by writers, were deemed impossible, but which have nonetheless come to be plausible or even actualized. Although fictional, some of these works must be seriously examined, especially as they garner references with surprising regularity in discourse surrounding policy for AI.

First to consider is the impact of Isaac Asimov. Asimov predicted the possibility of an AI system mounting a serious claim for legal standing, and even that it ought to qualify as human.[5] His work is useful by considering the possibility of nonhuman artefacts with claims to legal recognition, and furthermore by broaching the issue of what it means to be 'human', which is becoming a question of surprising legal importance. Asimov discussed this particular question, in the context of AI, in *That Thou Art Mindful of Him*.[6] The text specifically grapples with the problem of distinguishing, in the moral sense, between humans and robots, and whether this can be done without recourse to arbitrariness. In many ways, this is a very similar question to what this thesis explores, although my focus is not necessarily on qualifying a robot as 'human' (though there is extensive scope for a legal debate on the interpretation of 'human' at law).

Asimov also produced the famous (or infamous) "Laws of Robotics". These Laws, which posit a three-part code (later, four-part) for controlling AI systems and keeping them in service of humans, have been highly influential across media and popular culture. More surprising is that they are taken very seriously in contemporary legal and political approaches to AI regulation – for instance, they are referenced as

[5] I Asimov *The Bicentennial Man and Other Stories* (Book Club Associates, United Kingdom, 1977) at 135.
[6] Ibid at 61.

authoritative in the European Parliament's Report.[7] There is significant debate over if and how Asimov's Laws might be formalised. The original "Three Laws" are:[8]

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by a human being except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First of Second Law.

The European Parliament, in its Draft Report, references Asimov's Laws as follows: "whereas, until such time, if ever, that robots become or are made self-aware, Asimov's Laws must be regarded as being directed at the designers, producers and operators of robots, since those laws cannot be converted into machine code" (emphasis my own).[9] The final version of the Report removed contemplation of machine 'self-awareness', instead reading: "whereas Asimov's Laws must be regarded as being directed at the designers, producers and operators of robots, including robots assigned with built-in autonomy and self-learning, since those laws cannot be converted into machine code".[10]

There are other influential fictional examinations of machines with moral status. Philip K. Dick, in his *Do Androids Dream of Electric Sheep?* explored both the practical issue of the future relationship between humans and these sorts of robots, as well as the more philosophical (but nonetheless relevant) question of the potential arbitrariness of social and legal discrimination between organic persons and artificial persons; the real and the unreal.[11] This question becomes relevant where legal norms discriminate, fairly or unfairly, between entities that qualify as 'real' and entities that do not. Potential for these sorts of fictional problems to become serious certainly exists, as the legal criteria for access to things like standing and rights scarcely exist, and where they do exist, and vague and indeterminate. Dick anticipated the importance of this

[7] See European Parliament Report above n 3 at General Principles – U. Asimov's Laws are posited as a model for machine building. These Laws are included within the Report in full at footnote (2).
[8] I Asimov *Runaround* (Astounding Science Fiction, 1942). Republished in I Asimov *Robot Visions*, (Penguin, 1991).
[9] European Parliament Committee on Legal Affairs, Draft Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), 31.5.2016, at [L].
[10] Above n 3 at [U].
[11] PK Dick *Do Androids Dream of Electric Sheep?* (Weidenfeld & Nicholson, London, 2012).

question, and the mantra of the corporation which produced android replicants of humans in the film *Blade Runner* was that their machines were "more human than human".[12]

Another relevant work is Fred Hoyle's *The Black Cloud*.[13] Though not specifically concerned with artificial intelligence, it concerns nonhuman intelligence arising in ways we had not anticipated, which substantially overlaps with the present issue. One of the text's most profound notions is that it is not necessarily an inability of human institutions to modify ourselves to be accepting of the possibility that nonhuman entities might 'matter', and may even matter more than we do. Rather, it is a lack of willingness to modify our mode of thought. In short, it is simpler for us to recognise ourselves as the apex of the moral hierarchy, which we conveniently dictate, even if this requires us to be wilfully blind to even the theoretical possibility of us having nonhuman peers; perhaps, nonhuman superiors. As Hoyle asks, "Do we want to remain big people in a tiny world or to become a little people in a vaster world? This is the ultimate climax towards I have directed my narrative."[14] Understanding this question in the context of contemporary approaches to AI regulation provides an understanding of the psychology behind those approaches themselves.

Naturally, many other writers have considered this question, but the above three are sufficient for being particularly influential, or apt.

iv) The claims of a robotic person – ethical discussions

Key to this area is *The Machine Question*, by David J. Gunkel.[15] Gunkel's work describes the evolving conceptualisation of human ethical responsibilities towards non-human entities, and in particular, a theory of the extent to which artificial systems possessing autonomy and intelligence might have moral responsibilities and moral claims. This is generally divided into two courses: machines with moral agency, and machines as moral patients, thus broaching the issue of duties and rights of autonomy, and in the latter, rights to not be subjected to certain mistreatments. Of particular usefulness is Gunkel's dispatching with reliance on concepts of "consciousness", an approach which this thesis will propagate. Consciousness, he

[12] This expression has begun to see use in relation to the law and ethics of human enhancement. For example, see David R. Lawrence "The Edge of Human? The Problem with the Posthuman as the 'Beyond'" (2016) 31 Bioethics 171 at 171. Full quote: "Of course, there's a third option: that a posthuman is a being which embodied our self-ideal more successfully than we do ourselves – one 'more human than human'."
[13] F Hoyle *The Black Cloud* (Penguin Group, London, 2010).
[14] Above n 14.
[15] Above n 4.

argues, is invoked as a sort of unapproachable standard preventing non-humans from being considered in moral terms. Since there is nothing close to a unified or even clear definition of consciousness, it operates as little more than a synonym for something like 'soul', and thus a mechanism to delineate non-humans from humans on unprovable grounds. Gunkel discusses the possible basis for a robot rights argument by reference to the famous fictional AI system "HAL 9000", depicted in Clarke's influential *2001: A Space Odyssey*. Gunkel uses HAL as a reference point by which to discuss whether or not the machine ought to have been afforded the same ethical and legal standards as we provide for humans – for instance, due process and a fair trial.

Gunkel explores several issues which are closely relevant to this thesis. For example, the inherent violence in any demarcation between an 'us' and a 'them', a 'familiar' and an 'other'; or as it occurs for AI, a 'human' and a 'machine'.[16] It is so because it immediately categorises entities or groups of entities as not morally considerable, and thereby we do not ever come to pose the question of what violence may not be done to them, or what duties might be owed to them.

Gunkel highlights how conceptually difficult, and potentially politically difficult it will be to change contemporary modes of thinking about the moral insignificance of AI systems. He identifies that the 'machine' is the very mechanism by which the demarcation of the 'other' is made, stating that "the machine is not just one kind of excluded other; it is the very mechanism of the exclusion of the other."[17] For instance, Descartes demarcated from the human and the 'animal-machine'.[18] We even manifest this in common parlance whereby we want to convey exactly how *inhuman* or alien an entity is – for example, sharks are not merely nonhuman animals with particular tendencies, they are 'killing machines', and so on. To call something a machine is to gesture to the 'fact' it is not morally considerable, and that it is demonstrably lacking human qualities.

v) Arguing against the creation of machines with moral status

[16] Above n 4 at 30. See also Thomas H. Birch "Moral Considerability and Universal Consideration" (1993) 15 Environmental Ethics 313-332.
[17] Above n 4 at 128.
[18] Mark Coeckelbergh "Book Review: David J. Gunkel: The machine question: critical perspectives on AI, robots, and ethics" (2013) Ethics and Information Technology 15 235 at 235. Full quote: "Showing how Descartes divided human beings from the animal-machine, Gunkel argues that 'the machine is not just one kind of excluded other, it is the mechanism of the exclusion of the other.'"

Joanna Bryson repeatedly argues that creating robots to which we owe genuine ethical duties (rather than just imagined ones) would be inherently unethical. An effective phrase of hers is that, "Building Persons is a Choice", supporting it with the statement that "Robots are not persons unless we build them to be such and then declare them to be so."[19] Thus, she provocatively argues that "Robots Should Be Slaves"; they should be built to be slaves, and we should be comfortable with them being our slaves, so long as their slavery does not cause them any harm (which we have the power to ensure).[20]

Bryson argues that there are few benefits to be derived from the creation of robots to which we owe moral duties. Furthermore, that they will always be our creations and our responsibility, because at all times humans are responsible agents, and it is we who created them. Of Bryson's prescient observations, some are more relevant to the current question: firstly, that we must be cognisant of the ethical consequences of choosing to create artificial persons, and secondly, that the common cause of refusing to recognise the personhood of other entities is, invariably, competition for resources.[21]

vi) Considering a future relationship between humans and morally significant machines

David Levy canvasses, and argues for, the possibility that in the near-future human-robotic social integration will be so extensive that robots will naturally be viewed as potential friends, and even companions. He discusses ethical impacts arising from this – changes to humanity, which may not necessarily be negative.[22] He also considers the possibility of morally significant machines which we would be obliged to treat fairly.[23] In general, he identifies the unusual lack of regulatory interest in the question of how we ought to treat future machines, even while there is extensive interest in the question of how these machines ought to be used:[24]

---

[19] Joanna Bryson "Building persons is a choice" (2009) *Erwägen Wissen Ethik* 20(2) 195 at 195.
[20] Joanna Bryson "Robots should be slaves" in Y Wilks *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (John Benjamins Publishing Company, Amsterdam, 2010) 63 at 63.
[21] Above n 19; see generally H Steiner *An Essay on Rights* (Blackwell, Oxford, 1994). Steiner theorises all rights to be, essentially, property rights.
[22] See David Levy *Love and Sex with Robots* (Duckworth Overlook, London, 2008); David Levy "Falling in love with a Companion" in Y Wilks *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (John Benjamins Publishing Company, Amsterdam, 2010).
[23] David Levy "The Ethical Treatment of Artificial Conscious Robots" (2009) 1 Int J Soc Robot 209 to 216.
[24] Ibid at 209.

> Almost all of the discussion within the roboethics community and elsewhere has thus far centred on questions of the form: "Is it ethical to develop and use robots for such-and-such a purpose?", questions based upon doubts about the effect that a particular type of robot is likely to have, both on society in general and on those with whom the robots will interact in particular. What has usually been missing from the debate is the complementary question: "Is it ethical to treat robots in such-and-such a way?"

Levy notes that though there is obvious analogy between the ethical debate of 'conscious' robots and nonhuman animals, this analogy breaks down to the extent that the latter requires rights for a reason the former plausibly does not: the avoidance of physical pain.[25] This is an important point for consideration of how to protect morally significant machines – by welfarist approaches, or basic rights approaches. Levy also considers the possibility of rights arising as part of a natural process as humans recognise more aspects of humanness in robots, and interact with them more on a close or even intimate basis.[26] In the end, Levy takes a decidedly Kantian line on the issue of how we treat robots – that the way we treat apparently humanlike robots will affect us, "by setting our own behaviour towards those robots as an example of how one should treat other human beings."[27]

Another of Levy's arguments, to which I am sympathetic, is that 'artificialness' or rather 'artefactualness' – the fact of having been made by another person – ought not to negate moral significance or consideration. In short, just because *we* build a robotic person does not mean they do not matter, morally.[28]

Another contribution relevant to the area of concern is how we might make artificial intelligences act ethically, and understand right from wrong. Wendell Wallach and Colin Allen argue that imbuing intelligent systems with moral-decision capacities and ethicality is not only desirable but necessary from the

---

[25] Ibid at 214. Full quote: "In my view, however, there is an extremely important difference. Animals can suffer and feel pain in ways that robots cannot. This leads me to the view that the animal rights analogy is not a sound one on which to base the notion that robots are deserving of rights. Although I lean towards the view that robots *should* be endowed with certain rights, the basis for my position and the concomitant conclusion regarding this issue are completely different."

[26] Ibid.

[27] Ibid.

[28] Ibid at 211. Full quote: "The fact that a clever programmer is responsible for creating artificial consciousness in a robot is surely no argument for denying that the robot exhibits consciousness. Ask any devout Christian the question: "Who programmed us to behave as we do?" and you will quite likely be told that we are programmed by God. Does that mean we do not possess consciousness?"

perspective of human safety.[29] Even if 'full-blown' moral agency is beyond the reach of present technological paradigms, something like functional morality might still be engendered. Unclear, however, is whether creating ethical or moral machines, in the sense of machines that can act according to a concept of right and wrong, is conceived to engender reciprocal moral duties for humans towards the machines themselves. This latter point would be central to this thesis, which is predominantly interested in how we might be obliged to treat a robotic person.

There have been some efforts to introduce the possibility of humanitarian-based rights and responsibilities for future artificial intelligences, but these remain mostly peripheral. Hutan Ashrafian has hypothesised that "consciousness, sentience and rationality", manifested in advanced technologies, could be considered to "merit rights", and furthermore, "commensurate responsibilities and duties".[30] He proposes a potential model based on principles of Roman Law in which different strata of social classification confer different sets of rights and duties, positing the possibility that artificial intelligences might occupy some lesser strata to humans with much more limited protections and entitlements.[31] However, Ashrafian's approach is interesting for the fact that its suggestion that artificial intelligences, at least in the short term, represent a limbo between requiring both protection and strict controls. Thus, detailed socio-political controls for robots and artificial intelligence agents must be determined. However, in the long-term Ashrafian considers that we may be compelled, as persons with control over the system by which AI are instituted and the way the entities are designed, to treat these machines fairly and justly:[32]

> The broader application of *noblesse oblige* where a leader fulfils the responsibilities of his status necessitate a proportionate *humanite oblige (humanity obliges)*; here it is incumbent on human society to ensure the fair, tolerant and ultimately humane institution of advanced artificial intelligence and robots within mankind's society.

vii) East and West – Divergent ideals in the approach to AI regulation

[29] W Wallach and C Allen *Moral Machines* (Oxford University Press, Oxford, 2010). See also Gordana Dodig Crnkovic and Baran Curuklu "Robots: ethical by design" (2012) 14 Ethics Inf Technol.
[30] Hutan Ashrafian "Artificial Intelligence and Robot Responsibilities: Innovating Beyond Rights" (2015) 21 Sci Eng Ethics 317 at 317.
[31] Ibid at 324.
[32] Ibid at 326.

Because AI is a relatively new emerging technology, and because regulatory interest in AI is even more recently arisen, the amount of 'formal' regulation concerning AI is limited. Within this, there is little reference to the possibility of robotic personhood, or even lesser degrees of machine moral significance. A useful way to interpret regulatory devices (whether ethical or legal) in relation to AI is by assessing the sort of future relationship they conceptualise between humans and AI systems. This approach has merit because it allows for a degree of inference about whether or not AI systems having intrinsic moral status is a possibility which has informed the design of regulation, even where robotic moral status is not mentioned.

For example, the Fukuoka World Robot Declaration posited three "expectations" for "next-generation" robots:[33]

1. Next-generation robots will be partners that coexist with human beings
2. Next-generation robots will assist human beings both physically and psychologically
3. Next-generation robots will contribute to the realisation of a safe and peaceful society

The Fukuoka Declaration is intriguing for its tone of rhetoric. The choice of language – words like "partners", "assist", "coexist", "safe and peaceful", and "society" – are not only highly descriptive, but markedly different from some of the later conceptualisations of the future human-robot relationship. The Declaration impliedly rejects the master-slave narrative which commonly recurs in Western AI-related policy, and prefers a vision of shared futures and mutual benefit. According to Gianmarco Veruggio, the fact that this vision diverges from the West is not merely coincidence, but the product of differing histories and mythologies which frame the nature of "automata" (moving mechanical devices with humanoid characteristics) very differently:[34]

> The topic of the rebellions of automata recurs in the classic European literature, as well as the misuse or the evil use of the product of ingenuity. It is not so in all the world cultures: for instance, the mythology of the Japanese cultures does not include such a paradigm. On the contrary, machines (and, in general, human products) are always beneficial to humanity.

[33] A. Takanisi "World Robot Declaration" (at the World Robot Fair, Fukuoka, February 2004).
[34] Gianmarco Veruggio "The EURON Roboethics Roadmap" (paper presented to IEEE-RAS International Conference, 2006).

For contrast, consider the following dicta from the European Parliament's Draft Report on Robotics:[35] "whereas ultimately there is a possibility that within the space of a few decades AI could surpass human intellectual capacity in a manner which, if not prepared for, could pose a challenge to humanity's capacity to control its own creation". The language is demonstrably different from that of the Fukuoka Declaration – "surpass" denotes competition, with humanity as losers and robots as winners; "pose a challenge" is inherently related to dominance, as is the notion of controlling *our* creation. The tenor is pervasively one of conflict and struggle, with the implied intention of the European Parliament to frame law in a way that ensures that humans win out. Supplementing this, the European Parliament proposes various principles which reify a dominance-based human-robot relationship. The final Report states that "robots should act in the best interests of humans",[36] and as mentioned above, that insofar as possible robots should be made to obey Asimov's Laws, which have been described by Kevin Korb to, "effectively treat robots like slaves".[37] Furthermore Mady Delvaux, MEP and rapporteur for the European Parliament Report, has said that, "In order to ensure that robots are and will remain in the service of humans, we urgently need to create a robust…legal framework".[38] Language like "control" of humanity's creation, prime principles of human best interests, and remaining in the "service" of humans, forms a master-slave concept that is markedly different to what is referred to in European Robotics Network's (EURON) "Robotethics Roadmap" as the as the "spirit" of the Fukuoka Declaration.[39] As an aside, the fact that the European Parliament's Report invoked Asimov's Laws while failing to mention the possibility of robots which ought not to be controlled by humans is difficult to comprehend, given that the question of robotic personhood is so central to the writer's oeuvre.

The aforementioned "Roboethics Roadmap" proposed a set of definitions, opinions, questions and concerns to help navigate forthcoming AI-related issues across a range of disciplines.[40] The Roadmap

[35] Above n 10 at 4.
[36] Above n 3, within "Code of Ethical conduct for Robotics Engineers".
[37] Gary Marcus, "Moral Machines" *The New Yorker* (online ed, United States of America, 24 November 2012) <www.newyorker.com/news/news-desk/moral-machines>.
[38] May Bulman "EU to vote on declaring robots to be 'electronic persons'" (14 January 2017) The Independent <www.independent.co.uk/life-style/gadgets-and-tech/robots-eu-vote-electronic-persons-european-union-ai-artificial-intelligence-a7527106.html>.
[39] Above n 2 at 29. Full quote: "Promote among robotics scientists the spirit of the Fukuoka World Robot Declaration (2004)."
[40] Above n 2.

canvasses a wide range of ethical issues, but is notable for its direct reference to the possibility of AI systems with degrees of moral status, and the characteristics we associate with intelligent species. Albeit briefly, within the Roadmap the seemingly uncomplicated question of "What is a Robot?" was posed.[41] Diverse responses included the possibility that robots are "nothing but machines"; others conceded that robots have "ethical dimensions"; alternatively, robots could be "moral agents"; and the suggestion that robots represented the "evolution of a new intelligent specie".[42] What is clear from this is that the Roadmap was considering the emerging issue of artificial systems with profound moral aspects to their nature, even in spite of its mandated ten-year timeframe for speculation. Nevertheless, it "hinted at…problems inherent in the possible emergence of human functions in the robot: like consciousness, free will, self-consciousness, sense of dignity, emotions, and so on,"[43] and the possibility of robots "…evolving to become conscious agents".[44] Perhaps, as J. Storrs Hall has suggested, "Our machines will be better than we are – but having created them, we will be better, as well".[45] These concerns are central to the present thesis, which seeks to hypothesise a way in which such entities might be created, and the subsequent legal and ethical challenge of including them within our society in a way that respects their intrinsic moral status.

Another example of Western machine-building paradigms comes from the Engineering and Physical Sciences Research Council (EPSRC) and Arts and Humanities Research Council, which held a retreat to discuss "robotics, its applications in the real world and the huge amount of promise it offers to benefit society."[46] The primary product was a list of five, "Principles for designers, builders and users of robots" which included the following:[47]

1. Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.

[41] Ibid at 24.
[42] Ibid.
[43] Above n 35.
[44] Ibid at 617. Full quote: "Well before evolving to become conscious agents, humanoids can be an extraordinary tool used to control human beings."
[45] Ibid at 613.
[46] EPSRC "Principles of robotics: Regulating robots in the real world" (September 2010) Engineering and Physical Sciences Research Council <www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/>.
[47] Ibid. See also Joanna Bryson "The meaning of the EPSRC principles of robotics" (2017) 2 Connection Science. Bryson states that "The Principles were not designed for every conceivable robot, but rather serve in part as design specifications for robots to be incorporated as legal products".

2. Humans, not robots, are responsible agents. Robots should be designed; operated as far as practicable to comply with existing laws & fundamental rights & freedoms, including privacy.

3. Robots are products. They should be designed using processes which assure their safety and security.

4. Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.

5. The person with legal responsibility for a robot should be attributed.

Much of this is not controversial. For example, few people would disagree with the dictum that robots should not be designed solely or primarily to kill or harm humans, except in exceptional circumstances. Furthermore, few people would disagree with the norm that robots ought to be designed using processes which assure their safety and security. In fact, taken as a document which is implicitly intending to prohibit the creation or AI systems with moral status, the principles are highly informative.

It is nevertheless worth discussing that the EPSRC approach is very different from earlier proposals like Fukuoka or the Roadmap, with a much more grounded objective. It is of course intended to be of use in the short to medium term, in relation to contemporary robots. The relationship between robots and humans is framed very obviously as unilateral – robots are no longer partners of humans, but now *tools* for humans; they are explicitly *not* moral agents; and there is emphasis upon the fact that they are *products* and *artefacts* of human creation, implying this confers their moral insignificance. This could be read as being not only descriptive of the present relationship, but prescriptive of an ideal future relationship. But this also reflects a distinctly Western approach: the conception that humans are to be masters and robots their tools. Coeckelbergh discusses this master-slave conception:[48]

> …there have been many recent warnings about AI or robots taking over, indeed becoming our masters rather than the other way around. Consider for instance claims about AI in the media by Hawking, Musk, Wozniak, and Rees. Sometimes the future of AI, automation, and robotics

[48] Mark Coeckelbergh "The tragedy of the master: automation, vulnerability and distance" (2015) 17 Ethics Inf Technol 219 at 219.

> is *explicitly* discussed in terms of "masters" and "slaves". For instance, echoing a well-known theme in the history of thinking about machines, Wallach compares technology with 'a dangerous master' which gets out of control and Bryson has argued that 'robots should be slaves': we should not humanize them or make them into moral agents, but rather as 'tools we use'.

Even in New Zealand, regulating AI has become a vogue policy question. Neither have we been able to avoid the existential concerns discussed above by Coeckelbergh, with the master-slave narrative gaining some traction. Chapman Tripp issued a white paper "call to action" urging the State to consider what they deemed the most important AI issues, and noticeably, the question of machines with any degrees of intrinsic moral status was entirely unmentioned; even in a list of issues so disparate in both importance and temporality that it included both AI 'liability for damage' and the existential threat to human existence of a superintelligence like "Skynet".[49] This was located within a chapter entitled "Keeping control".[50] It is worth noting that the New Zealand Law Foundation issued a relatively substantial research grant into AI-related issues in 2017, with the intrinsic moral status of future machines occupying no part of the research agenda.[51]

Returning to the European Parliament Report – the most substantial formal legislative approach to AI regulation thus far – the scope of the document considers many of the more conventional and short-term social, legal and ethical issues. In particular: robotic liability, categorisation of robots, intellectual property rights, safety and security, autonomous vehicles, robots in professions of care, human enhancement, the growing need for more ICT professionals, liability, and a rudimentary code of ethics for AI engineers. While the Report discusses the possibility of "creating a specific legal status for robots, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons with specific rights and obligations", at no point does it extrapolate on the nature of those rights nor the reason for their being granted.[52] Thus, there is no way to know whether the European Parliament spared a thought for robotic personhood; viewed in light of the rest of the document, it is unlikely they were suggesting anything like a 'partnership' between future robots and humans. Even amidst the wide range of topics

[49] Chapman Tripp "Determining our future: Artificial Intelligence: Opportunities and challenges for New Zealand: A Call to action" at 10. Skynet is the superintelligent AI system that becomes despotic in the Terminator films.
[50] Ibid.
[51] New Zealand Law Foundation "Major new study to tackle artificial intelligence law and policy" (January 2017) <www.lawfoundation.org.nz/?p=7680>.
[52] Above n 3 at [59].

broached there is not a single allusion to the possibility of future artificial systems which need to be treated fairly and justly within the society and legal system we introduce them into.

viii) This thesis in the context of the literature

As stipulated, the above represents only an introductory sample of the literature canvassed in this thesis. In particular, none of the technical or scientific literature was included, as this is done mostly in dedicated chapters. This thesis picks up from Levy's expression that there has been much less interest in questions of how we should treat AI systems in the context of their interest than how we should deploy them for our interests. It is certainly true that in the formal legal, and political literature there is a real dearth of interest in this question and its implications.

For the benefit of policy creation, this thesis does not discuss the issue in a generalised sense. Instead, it constructs a type of machine from contemporary AI to serve as a lightning rod for discussion. This should be useful because, if accurate, it provides policymakers with a degree of specificity – as there is little that can be done to prevent the creation of robotic persons if we have only the general sense that they might arise somehow, some way.

Because the potential problems of machines with moral status are so diverse, a selection of issues have been chosen for inclusion. These are what I would call 'primary' issues – they do not analyse the 'softer impact' questions, like how interaction with these machines will affect humans at the sociological or psychological level. Instead, they are problems of some immediacy following a general sequence: How will they be built? Why will they matter? How can we detect them? Where do we draw the lines? Can we own them, or do we owe them, or both? Can they be given legal standing? What sort of legal protections can they reasonably expect?

Part 1

Chapter 2: The Technical Argument for EMBOT

2.1 The shift from amodal symbolic programming to artificial learners

According to Rodney Brooks, AI "started as a field whose goal was to replicate human level intelligence in a machine. Early hopes diminished as the magnitude and difficulty of that goal was appreciated."[53] The archetypal focus of AI was to produce machines capable of performing feats generally requiring human intelligence. However, much of what appeared achievable in AI proved difficult to attain. Early programmers developed a number of artificial systems using models derived from quantitative mathematics or symbolic logic – that is, systems which made no direct reference to complex low-level sensory or motor representations. These systems could operate in well-structured, carefully controlled environments. Some, environments were abstract – for instance, environments could be structured as a set of propositions in medical diagnosis or mineral prospecting scenarios. Other environments were physical, but controlled by specialised discrete sensors, of the kind found in highly automated factories. Tim Smithers summarises that: "If we are able to specify and control the environment of a robot, and keep it relatively simple, it is possible to design and build reliable and robust artificial agents to work in these environments…These kind of well-

[53] Rodney Brooks "Intelligence without representation" (1991) 47 AI 139 at 139.

structured and well looked after environments are, however, the exception, and, significantly, do not contain people, or only very few."[54]

Unstructured environments, like the physical world, are a substantially different challenge for programmers and roboticists. Firstly, these environments are unpredictable, complicated and dynamic. Secondly, they are rarely specified or controlled. Thirdly, they are inhabited by other agents, in particular humans, that tend to behave unpredictably, or in ways that can only be predicted or understood through immense integrated intelligence. Thus, conventional robots using computational or logical programming are largely incompetent when placed in such environments.

Machines of this kind are comprised mostly of snippets of information or instructions provided to them by a human programmer. Theoretically, providing more information might allow the machine to behave more intelligently. Unfortunately, even while this is true to a certain extent, it is still subject to limitations, as demonstrated by the 'CYC project. CYC is an expert system programmed with "commonsense data" which consumed more than two person-centuries of data-entry time. The goal of the project was to provide enough data to CYC so that it might be able "cross over" and become "self-programming".[55] Despite the amount of data provided, this never truly occurred. Rather than becoming a self-programming inductive reasoning system, CYC was only capable of drawing deductive rules from the information provided to it. This was a major blow to not only the project, but to traditional modes of thinking in the field of AI. Andy Clark summarises thusly:[56]

> CYC is an important and ambitious project. The commonsense database it now encodes will doubtless be of great practical use as a resource for the development of better expert systems. But we should distinguish two possible goals for CYC. One would be to provide the best simulacrum of commonsense understanding possible within a fundamentally unthinking computer system. The other would be to create, courtesy of the CYC knowledge base, the first example of a genuine artificial mind. Nothing in the performance of CYC to date suggests that

[54] Tim Smithers "Autonomy in Robots and Other Agents" (1997) 34 Brain and Cognition 88 at 101.
[55] A Clark *Being There: Putting Brain, Body, and World Together Again* (2nd ed, MIT Press, Massachusetts, 1997) at 2.
[56] Ibid at 3.

> the latter is in the cards. CYC looks set to become a bigger, fancier, but still fundamentally brittle and uncomprehending "expert system." Adding more knowledge to CYC will not remedy this.

Since paucity of data was apparently not the cause of AI limitations, with even the most intelligent artificial systems remaining "so unspeakably, terminally dumb", it became necessary to explore what might be the underlying misconception of cognition that had slowed AI progress.[57] Smithers, in 1997, argued the following:[58]

> If we want robots to operate in environments in which people normally work, such as hospitals, libraries, museums, offices, shops, streets, and other semistructured public places, we are going to need robots capable of dealing with very much larger degrees of and amounts of normal environmental variation, dynamic changes, and unpredictability. And we are going to have to be able to design and build such robots without first foreseeing and specifying all the possible kinds of situations such robots may get into during normal operation. In other words, we are going to have to be able to design and build them without first having to model the robot as a plant, or first explicitly identify the full range and scope of the interaction dynamics required."

While this suggestion was apt, it was nonetheless difficult to actualize. In particular, a key question concerned how machines could be made so that they could navigate unfamiliar scenarios without programmers first knowing what those scenarios would be like. However, there already existed systems which were capable of doing this with extreme accuracy and flexibility: biological systems, in particular, animal species. Chief among these is Homo sapiens with its unparalleled cognitive capacities. Since no human requires manual programming to become intelligent, and yet each human progresses from juvenility to maturity relatively rapidly, a hypothesis began to develop that, "the childs mind [sic] would show us the way to artificial intelligence."[59] Thus, "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's?"[60]

---

[57] Ibid at 1.
[58] Above n 54.
[59] Frederic Kaplan and Pierre-Yves Oudeyer "Intrinsically Motivated Machines" in M Lungarella et al (eds) *50 Years of Artificial Intelligence* (Springer-Verlag, Berlin Heidelberg, 2007) 304 at 304.
[60] Ibid.

No human child is born with a model of the world in which it will operate. Granted, genetic inheritance confers some implicit understanding of what it is like to be a human in the real world, but this could hardly be described as anything close to a model in the sense that researchers had been trying to instil in robots. Within days of birth, a cognitively normal baby is interacting with its world; within weeks this interaction is deliberate and purposeful; within months it is planned; within years it is competent. The majority of this transition is self-motivated, and does not require the instruction of the parent; exactly none of it requires external intervention from a programmer-type professional who gets into the child's brain and provides a model, data, or instructions.

Based on this observation, the general conclusion was that, "we should seek to build robots, and artificial agents in general, that are autonomous," says Smithers, continuing, "We should attempt to design and build robots that through continuous ongoing interaction with their environments form the interaction processes needed to sustain and maintain effective interaction with their environments…they should self-construct the means for effective (reliable and robust) interaction through self-sustained effective interaction."[61]

In summary, the novel trajectory in AI became the production of learning systems that were more than merely passive receivers of information. If these learning paradigms could be honed in a way that mimics the child's developmental techniques, perhaps the robot could learn in a way not dissimilar from the juvenile human. Thus, the focus became on creating, essentially, infant robots that learn.

2.2 Paradigm One: Intrinsic Motivation Models

Intrinsic Motivation models (IM) is a school of programming techniques part of the larger discipline known as Machine Learning. Machine Learning in itself represents one large step forward over traditional AI methods. In the Machine Learning paradigm, an AI program is not created directly by a human programmer – rather, the human programmer creates a program that can *learn* some piece of intelligent behaviour, through exposure to large numbers of training situations or examples. Another important aspect of Machine Learning is the development of *artificial neural networks*, which are computational systems designed to replicate the human brain and nervous system. These neural networks involve the use of artificial neurons

[61] Above n 54 at 101.

that connect and communicate with each other. Most surveys of contemporary AI focus on Machine Learning techniques – and indeed, these techniques are the key reason for the success of the current generation of AI systems. But I want to look forward, at some relatively new ideas about how Machine Learning techniques should be deployed. What is unique about IM is that is focusses on ways to make machines intrinsically motivated to learn. As Kaplan explains:[62]

> Through hundreds of experiments and models – supervised, unsupervised, reinforced, active, passive, associative, symbolic, connectionist, hybrid, embodied, situated, distributed – we benefit now from a large collection of examples that show how a machine can learn. However, the issue of "why" would a machine learn (or how would it choose what to learn) has not been tackled with the same attention.

Why then is intrinsic motivation necessary? As stated above, the real world is complex and dynamic – indefinitely more complex than most virtual worlds or structured environments. In particular, the real world is full of other agents, all acting with varying degrees of unpredictability. All of these agents are concerned with the same deceptively simple fundamental inquiry – what is the best thing to do next? However, because the real world is changeable, the best thing to do next is usually non-obvious. Agent's therefore require the capacity to learn critical meta-level skills, to discriminate based on experience, and to inductively figure out the rules based on what the agent has encountered before. "What is the best thing do next?" actually involves a series of questions: What do I need? What actions could I perform? What do circumstances allow? And finally, what's the best thing to do next?[63] Answering these seemingly simple questions requires enormous cognitive sophistication. Without the necessary skills, an agent will not last long in the real world. The more complex an agent, the more it can respond to environmental changes, but commensurately, the more complex its needs tend to be.

Thus, IM is necessary to the extent that an enormous amount of learning is required to navigate the real world and satisfy fundamental needs and wants. To meet these standards requires learning to begin immediately and continue indefinitely. Human children are paradigm examples of this: "Children seem to

[62] Above n 59 at 304.
[63] See Guy Claxton *Intelligence in the Flesh* (Yale University Press, Connecticut, 2015) at 65.

engage in such type of activities just for the sake of it. They seem intrinsically motivated to manipulate, to explore, to test – in one word – to learn and therefore they look for activities and situations that provide learning opportunities."[64] IM researchers replicate this by designing, "Active learners…machines that ask, search and select specific training examples in order to learn efficiently."[65] Thus, IM is not merely about creating an agent that wants to learn, but one that learns in an efficient manner. This is another reason to model developmental strategies on children, for their exceptional efficiency.

Thus, programmers create learning algorithms with inbuilt *intrinsic rewards*. Pursuit of these rewards, and achievement of them, provides the artificial system running the program with an intrinsic sense of reward. Anthropomorphic terms like fulfilment, pleasure, satisfaction etc., may feel uncomfortable, but from a functional perspective, the mechanism is largely the same for the robot as it is for humans: both act towards certain objectives so as to achieve an intrinsic reward which will be experienced when that objective is actualized.

Intrinsic rewards in IM therefore motivate an agent to recursively engage in activities that yield a positive experience of novelty. The sorts of things that meet these criteria are initially simple. For instance, simple sensitivity is a rewarding experience to a totally juvenile baby. Another rewarding experience is the basic movements of one's body, the basic vocalization of sounds, and so on. However, reward from surprise is a necessarily diminishing mental state in humans as a phenomenon becomes more and more familiar. Thus, IM intrinsic rewards must not only encourage pursuit of novelty, but diminishing rewards as the novel thing becomes familiar.

Thus, IM is concerned with an efficient or optimized learning strategy. Generally, this optimization has a couple of foci: motivating the agent to find learning situations that are most novel, balanced with motivating the agent to find learning situations that are within a *learnable range* of complexity. Variations of this are what Oudeyer calls an "active strategy" for learning, rather than a passive strategy: "Whereas a passive strategy

[64] Kaplan, above n 62.
[65] Ibid at 307.

would lead to very inefficient learning situations, an active strategy allows the learner to discover and exploit learning situations fitting to its biases."[66]

Human learning is not indiscriminate learning: babies are actually fairly discerning learners. Firstly, children seek out "learning opportunities", by which is meant a scenario which offers a chance to satisfy the intrinsic motivation to learn, and provide an intrinsic reward.[67] Moreover, this "general tendency to explore in is never satiated and is not a consummatory response."[68] Thus, the basic motivation to learn persists, and acts upon a child seemingly from the very earliest stages of its life. However, learning opportunities are not selected randomly, as the best learning opportunities will have distinctive characteristics. In particular, they seek "optimal incongruity";[69] which is described as "…some forms of optimality between completely uncertain and completely certain situations…those where there [is] a discrepancy between the perceived and standard levels of stimuli." Therefore, the agent is not satisfied with learning opportunities that are excessively novel nor excessively unfamiliar, but which have "an intermediate level of novelty, between already familiar and completely new situations." Kaplan summarises how this replicates the general learning process of children:[70]

> According to Piaget, children are intrinsically motivated to encounter activities which involve some assimilation and accommodation. Assimilation is a process whereby children incorporate aspects of the environment into their pre-existing cognitive structure…This means that the child's cognitive structure influences his perception of the environment. During accommodation, the child adapts his cognitive structures to fit the environment…In Piaget's view, learning is simply an aspect of assimilation and accommodation.

Balancing this novelty versus familiarity relationship is a driving force behind optimal IM learning, according to Oudeyer:[71] "When the child encounters an informational input from the environment highly

[66] Ibid at 311.
[67] Ibid at 306
[68] Ibid.
[69] Ibid.
[70] Ibid at 309.
[71] Ibid at 308.

discrepant from existing schemata, the input will most likely be ignored. When inputs are completely predictable, children generally lose interest in them."

This general developmental strategy is called *scaffolding*, the meaning of which is much as the metaphor suggests.[72] Very basic learned things, like simple *behaviours*, combine to form a foundation. Upon this foundation, more complex behaviours can be sustained. Furthermore, simple *concepts* may be developed. Into these concepts go *beliefs*, acquired through perception. As more beliefs are acquired, concepts can become more complex; higher-level concepts. Over time, this scaffold becomes a more complete framework. Eventually, very difficult behaviours and concepts can be constructed (i.e. learned), and sustained within the framework.

Thinking of learning in this way clarifies why children seek out optimal incongruity; they are looking for the learning opportunity that will be the next pole in the scaffolding frame. Thus, they are uninterested in interacting with things that are further down the scaffold, and equally uninterested in things that are many levels above where the scaffold presently reaches. The perfect learning opportunity constitutes a *progress niche*. Oudeyer explains that:[73]

> Progress niches are not intrinsic properties of the environment. They result from a relationship between a particular environment, a particular embodiment (sensors, actuators, feature detectors and techniques used by the prediction algorithms) and a particular time in the developmental history of the agent. Once discovered, progress niches progressively disappear as they become more predictable.

The goal of IM programming is to replicate this developmental strategy in an artificial system, which involves not only an intrinsic motivation to want to learn, but structuring to make learning efficient. With this in place, researchers may simply "place the robots in a world that is rich in learning opportunity and

---

[72] See Clark, above n 55 at 32. Full quote: "…we shall unearth a surprising number of further tricks and strategies involving the use of some type of external structure or "scaffolding" to mold [sic] and orchestrate behaviour. Obvious contenders are the immediate physical environment…and our ability to actively restructure that environment so as to better support and extend our natural problem-solving abilities. These strategies are especially evident in child development. Less obvious but crucially important factors include the constraining presence of public language, culture, and institutions, the inner economy of emotional response, and the various phenomena relating to group or collective intelligence."

[73] Above n 59 at 310.

then just watch how the robots develop themselves…based on a series of studies showing the importance of intrinsic motivation in human development."[74] There are multiple theories which try and craft the best learning methods, and which over time generate more and more accurate robotic behaviour using different probabilistic methods.[75] Accurate robotic behaviour will be behaviour which represents a good answer to the question of, What is the best thing to do next? Humans tend not to need extrinsic aid to develop methods of providing better and better answers to this question, and IM research suggests the same can be achieved in robots, whereby "…experiments show that functional organization can emerge even in the absence of explicit internal schema structures and that developmental patterns can spontaneously self-organize driven by the intrinsic motivation system."[76]

The following example of an experiment in IM demonstrates the promising nature of the research in achieving the creation of agents which can overcome the limitations of the classical AI paradigms. They are presented here at length because of the clarity with which they describe the scaffolding process, the use of active learning systems rather than passive learning systems, and the progression from simple robotic learning to more complex learning. A description of this process in action is useful to the extent that it demonstrates the similarity between some of the processes of an intrinsically motivated learning machine and a human infant. Described here by Kaplan and Oudeyer:[77]

> In one of those [experiments] a four-legged robot is placed on a play mat. The robot can move its arms, its neck and mouth and can produce sounds. Various toys are placed near the robot, as well as a pre-programmed "adult" robot which can respond vocally to the other robot in certain conditions. At the beginning of an experiment, the robot does not know anything about the structure of its sensorimotor space (which actions cause which effects). Given the size of the

[74] Ibid at 305.
[75] See Z Ghahramani "Probabilistic machine learning and artificial intelligence" (2015) 521 Nature. Content explores the question of how a machine can be made to learn from experience through an exploration of what learning is. At 452, "The probabilistic framework which describes how to represents and manipulate uncertainty about models and predictions, has a central role in scientific data analysis, machine learning, robotics, cognitive science and artificial intelligence".
[76] Above n 59 at 311.
[77] Ibid. See also Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum and Samuel J. Gersham "Building Machines That Learn and Think Like People" (Cambridge University Press, published online 24 November 2016).

space, exhaustive exploration would take a very long time and random exploration would be inefficient.

At the beginning of the experiment, the robot has a short initial phase of random exploration and body babbling. During this stage, the robot's behaviour is equivalent to the one we would obtain using random action selection: we clearly observe that in the vast majority of cases, the robot does not even look at or act on objects; it essentially does not interact with the environment.

Then there is a phase during which the robot begins to focus successively on playing with individual actuators, but without knowing the appropriate affordances: first there is a period where it focuses on trying to bite in all directions (and stops bashing or producing sounds), then it focuses on just looking around, then it focuses on trying to bark/vocalize towards all directions (and stops biting and bashing), then on biting, and finally on bashing in all directions (and stops biting and vocalizing).

Then, the robot comes to a phase in which it discovers the precise affordances between certain action types and certain particular objects. It is at this point focusing either on trying to bite the biteable object…or on trying to bash the bashable object…Eventually, it focuses on vocalizing towards the "adult" robot and listens to the vocal imitations that is triggers. This interesting for vocal interactions was not pre-programmed, and results from exactly the same mechanism which allowed the robot to discover the affordances between certain physical actions and objects.

A virtual example of IM programming is the TEXPLORE-VANIR algorithm. TEXPLORE employs a model-based method (meaning the agent is provided with a model of the domain it is going to explore) by which an agent uses intrinsic rewards to, firstly, efficiently learn a useful model of the domain, and secondly, explore in a developing and curious way.[78] It then provides what the designers perceive to be the best intrinsic rewards to improve efficiency in learning the model: (1) preference to explore areas of space where

[78] Todd Hester and Peter Stone "Intrinsically motivated model learning for developing a curious agent" (paper presented to The Eleventh International Conference on Development and Learning (ICDL), San Diego, November 2012). See also, for more information on use of intrinsic rewards in intrinsic motivated learning, Ozgur Simsek and Andrew G. Barto "An Intrinsic Reward Mechanism for Efficient Exploration (2006) 4 Computer Science Department Faculty Publication.

there is a large degree of uncertainty in the model, and (2) preferring regions of space that are far from previously explored areas (regardless of how certain the model is).[79]

The agents in TEXPLORE explore the "Light World Domain".[80] In this particular domain there are a series of rooms, each containing a door, a lock, and sometimes a key. "Agents must go to the lock and press it to open the door, at which point it can leave the room. It cannot go back through the door in the opposite direction. If a key is present, it must pick up the key before pressing the lock. Open doors, locks, and keys each emit a different color light that the agent can see. The agent has sensors that detect each color light in each cardinal direction…the agent can take six possible actions: it can move in each of the four cardinal directions, PRESS the lock, or PICKUP the key."[81] Once the agent learns the model of one room, it applies what it has learned to the next room. In this way, it becomes more intelligent, and more efficient at its tasks.

Sometimes a robot will be pursuing a reward that is a comparatively 'long way' away from the robot's current position. To achieve this reward is difficult, and requires forming sub-problems, and chaining those sub-problems together. A focus of IM is producing robots that can do this – form plans that involve multiple steps and contingencies to reach goals. Examples of this can be found in artificial intelligences tested on Atari gaming platforms. In many of these games, machines are capable of performing to a level that is on par with humans, but in particularly difficult games, where each frame has many steps between the agent and its reward, IM models have proven effective in creating agents that are capable of generating plans that can reach these complex goals; for example, MIT's "Deep Reinforcement Learning with Intrinsic Motivation and Temporal Abstractions".[82]

As alluded to above, a highly promising discipline of AI research is artificial neural networks; computer systems modelled on the human brain and nervous system, which utilise large numbers of simple artificial neurons. These are connectionist systems, whereby each neuron connects with numerous others, enhancing or inhibiting the adjoining neural units; they have been used to create computers that can recognise speech, have vision, etc.; things which are difficult to achieve with conventional rule-based programming. The

[79] Ibid at 2.
[80] Ibid at 3.
[81] Ibid. See, for video demonstration, < https://www.youtube.com/watch?v=2qq_4AMA_d0>.
[82] Tejas D. Kulkarni, Karthik R. Narasimhan, Ardavan Saeedi and Joshua B. Tenenbaum "Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation" (2016).

development of better neural networks plausibly can assist in creating more intelligent and competent real-world robots. For example, different neural networks appear to play a crucial role in understanding concepts of 'self' and 'others' in not just the physical *bodily* sense, but in more abstract, evaluative terms;[83] thus they are invaluable for environmental interaction and socialisation. Advancements in neuroscience may tell us more about the particular parts and models of the brain that might be artificially reproduced to create more intelligent robots.[84]

In summary, a plausible mechanism to generate autonomous self-motivated artificial intelligence is to create a machine with the will to learn according to particularly efficient strategies. These are strategies used by human children to develop intelligence, which are inherently present in the child's cognition. Thus, IM aims to create robots that are infant learners, capable of autonomously scaffolding themselves.

2.3 Paradigm Two: Embodied Cognition Models

The second AI paradigm I will consider, Embodied Cognition models (EC), may be deployed in conjunction with IM to provide the cognitive apparatus necessary for the environmental interaction which is so crucial to incremental cognitive development.

EC is the theory that a significant amount of embodied intelligence is required to perform the fundamental functions of human cognition. By this I mean being able to see, hear, touch, feel, experience, make connections and so on (all aspects of embodiment). This is a simplistic summary of embodiments, as EC goes beyond mere sensation to include models like simulation and situated action. Most generally, EC is a model of cognition which places particular importance on actually "being there", as Clark termed it.[85]

Recent evolutionary theories of cognition suggest that rather than the body being the vessel and the brain the pilot, the brain itself is a utility of the body. For example, Brooks states that, "The physical manifestation of the body is primary. The stuff of intelligence has evolved in conjunction with that body, and is a modulator of its behaviour, rather than a primary and central control system."[86] Thus, humans are not

[83] See Lucina Q. Uddin et al "The Self and Social Cognition: the role of cortical midline structures and mirror neurons" (2007) 11 Trends in Cognitive Sciences.
[84] See Earl K. Miller "The Prefrontal Cortex and Cognitive Control" (2000) 1 Nature Reviews Neuroscience.
[85] Above n 55.
[86] Rodney Brooks, Foreword to Rolf Pfeifer and Josh Bongard *How the Body Shapes the Way We Think: A New View of Intelligence* (MIT Press, Cambridge, 2007) at xv.

brains that happen to have bodies, we are bodies entirely; our entire organism is the engine of cognition. Reconceptualising the human system in this way closes the gap between ourselves and machines, which are not generally conceived of as having a disembodied mind in the same way we have been since Cartesian Dualism became prevalent.

Thus, embodiments substantially shape our cognition, and an embodied machine may be effected in similar ways. We experience mental states as we do, behave the way we do, and have many of the tendencies that we do because we are embodied in particular ways. Agents with different embodiments are not just superficially different to us but cognitively different from us. Bats and humans, for example, are highly unlikely to have a comparable cognitive phenomenology; bats are not little misshapen humans who are nonetheless experiencing the world like we do. "If my body was different, I would *be* different," says Claxton. "If I was made of silicon or fibre optics, I would need different things, respond to different things, notice different things, and be intelligent in a different kind of way."[87]

EC also emphasises the role of action in cognition. It does not subscribe to the concept that the human brain is a cache of amodal symbols which, when interpreted correctly, produce functions like action and thought. The performance of action is not, therefore, substantially the result of "pure thought".[88] It is often action-specific, and furthermore, utilises data that is inherently stored as the multimodal representations of action-specific experiences. Lawrence W. Barsalou argues that, "…it is unlikely that the brain contains amodal symbols; if it does, they work together with modal representations to create cognition."[89] The beliefs of the earlier 'computer models' of mind, "focus too much on passive storage of information and too little on the importance of situated action."[90] Claxton corroborates that these traditional AI systems "…offered a tempting metaphor for the divorcing of Body and Mind in the form of the distinction between hardware and software. We were encouraged to see our brains as composed of multipurpose biological hardware on

[87] Above n 63 at 3.
[88] Ibid at 36. Full quote: "Completing a jigsaw puzzle thus involves an intricate and iterated dance in which "pure thought" leads to actions which in turn change of simplify the problems confronting "pure thought". This is probably the simplest kind of example of the phenomena known as *action loops*."
[89] Lawrence W. Barsalou "Grounded Cognition" (2008) 59 Annu. Rev. Psychol. 617 at 618.
[90] Ibid at 622.

which could be run a host of socially transmitted software programs and belief systems. But this is quite misleading."[91]

An improved understanding of the role of embodiment in cognition re-emphasises the importance of groundedness; i.e. "the assumption that cognition is typically grounded in multiple ways, including simulations, situated action, and, on occasion, bodily states."[92] Many prescientific views of human cognition championed the role of modal representations. Some period later, human cognition was re-conceived into the disembodied concept that survives today. Nevertheless, the importance of embodiment persists in even substantially 'cerebral' tasks. This concept is supported by Barsalou, who emphasises "…the central roles of perception and action in cognition."[93]

Models of cognition which do not appreciate the role of embodiment and modal representations are problematic. A paradigm of these is that of Rene Descartes, who theorised that cognition had no physical aspect, as Tom Regan summarises: "Minds he regards as having no physical properties; they have no size, weight, shape and the like; minds are immaterial or spiritual…".[94] It was a pervasive theory. "Descartes has left us with a theory of mind and thought so influential that its main tenets are still widely held and have barely been begun to be re-evaluated," George Lakoff and Mark Johnson state. "It has been handed down from generation to generation as if it were a collection of self-evident truths. Much of it is still taught with reverence."[95] With such entrenched reverence for the disembodied human mind concept, arguing for the significance of artificial cognition is immediately challenging.

The point of this exploration is to establish the possibility that creating an artificial intelligence that can utilise embodiments plausibly might allow for the creation of an entity that can learn and think is ways substantially analogous to humans. Since we deem the human cognitive powers to matter in a general sense, for the purposes of determining how humans should enjoy the law's protection in certain uses of their body

[91] Above n 63 at 29.
[92] Above n 89 at 619.
[93] Ibid at 621.
[94] Tom Regan *The Case for Animal Rights* (University of California Press, California, 2004) at 21.
[95] George Lakoff and Mark Johnson *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought* (Basic Books, New York, 1999).

and brain, and even entitlements to certain things for the same reasons, then a similarly designed machine might enjoy similar benefits.

Being there and performing actions allows for the accumulation and retention of information and multimodal representations related to the situation encountered. Barsalou argues that "situations that contain embodiments become established in memory (e.g., receiving a gift, feeling positive affect, and smiling). When part of this situation occurs (e.g., receiving a gift), it activates the remainder of the situational pattern, producing associated embodiments (e.g., smiling)."[96] When we retrieve a memory, these are stimulated to produce a comprehensive, simulated experience. For example, "To understand how someone else feels when disgusted," according to Barsalou, "we simulate how we feel when disgusted. From this perspective, simulation provides a general mechanism for establishing empathy."[97] This suggests that an artificial agent benefitting from embodied cognition may be unalike to us in certain ways, or struggle to understand our experiences, unless it too was capable of accumulating the same multimodal experiences. Nonetheless, it would have its own versions, which we would struggle to understand in equal measure.

With this emphasis on multimodal experience, it becomes more understandable why artificial systems possessing few or no embodied aspects proved incapable of autonomously (i.e. not under manual control of a human) navigating real world environments – because even simple actions and interactions in the real-world demand extensive multimodal experiences; thus near-total reliance on amodal symbols means that critical information is lost in translation, so to speak. How do you instruct a robot how to play football using only quantitative mathematics? When a human even thinks about a football, parts of the motor cortex associated with use of the feet begin to fire.[98] When the human sees a ladder, they see a 'for climbing', or a bat as a 'for holding'. In each case, stored multimodal representations associated with the various actions

[96] Above n 89 at 630.
[97] Ibid at 622.
[98] Above n 63 at 68. Full quote: "Our bodies understand ideas in terms of what they are good for, and how we make use of them: not just in terms of the features by which we recognise them. When you ask someone what a 'ball' is, in the context of a discussion about football, bits of their motor cortex light up that correspond to control of the legs and the act of kicking. Ask them the same question in the context of talking about tennis, and the embodied 'meaning' that is automatically activated involves arms and shoulders. If you had just been talking about your teenage daughter's school prom, 'ball' would activate responses to do with dancing (as well as visceral anxieties about the cost of the dress)."

and experiences of *doing* are simulated in the brain. As Barsalou describes, the simple act of sitting into a chair utilises these embodied paradigms:[99]

> Simulation is the re-enactment of perceptual, motor, and introspective states acquired during experience with the world, body, and mind. As an experience occurs (e.g., easing into a chair), the brain captures states across the modalities and integrates them with a multimodal representation stored in memory…Later, when knowledge is needed to represent a category (e.g., chair), multimodal representations captured during experiences with its instances are reactivated to simulate how the brain represented perception, action, and introspection associated with it.

Experimental observation of infant humans reinforces the necessary relationship between embodiment, action and cognition. In an experiment in which infants interacted with slopes (i.e. gradients), their capacity to interact changed radically as their embodiments changed. In particular, the transition from crawling to walking had remarkable effect on the infants' capacities; what they had learned while crawling was substantially inapplicable once they started walking. Clark summarises the conclusions to be drawn from this:[100]

> This evidence suggests not only that infants learn about the world by performing actions but also that the knowledge they acquire is itself often action-specific. Infants do not use their crawling experience to acquire knowledge about slopes in general. Rather, they acquire knowledge about how slopes figure in specific contexts involving action. Other findings concerning the context-specificity of infant knowledge point in the same general direction.

This can be seen more starkly when conditions are manipulated to deliberately divorce body and mind; demonstrated in an experiment by Richard Held and Alan Hein.[101] This experiment aligns with Barsalou's later claim that, "the cognitive system evolved to support action in specific situations…interactions between perception, action, the body, the environment, and other agents, typically during goal achievement."[102] The

---

[99] Above n 89 at 618.
[100] Above n 55 at 37.
[101] Richard Held and Alan Hein "Movement-Produce Stimulation in the Development of Visually Guided Behavior" (1963) 56 Journal of Comparative and Physiological Psychology.
[102] Above n 89 at 619.

effects of these embodiments being artificially isolated from one another is remarkable. Described here by Claxton:[103]

> Two kittens were yoked together in a kind of primitive two-seater carousel for the first few weeks of their lives. One had its feet on the ground, and could at least walk round in circles as it wished. The other was forced to lie in the cradle opposite, and be carried around at the whim of its more fortunate sister. This devious device ensured that the two kittens had the same kind and amount of visual experience, but for one this was linked to its own movement, and for the other, it was not. Despite plenty to see, and despite no obvious damage to its visual system, the passive kitten never learned to see properly. It couldn't recognise objects, navigate its way around a room, or see in depth. Deprived of the opportunity to link movement and vision in a meaningful way – to discover how the world changed as *it* changed – the kitten never learned to see meaning in the world. If human babies are swaddled and bound so they cannot move, their perceptual and cognitive development quickly goes awry.

All of this is to emphasise the significance of the fact that AI research is currently refining the ability to utilise these sort of paradigms to produce artificial agents, where the goal is to create a robot that is capable of operating in the real world in a way that is in any sense comparable to humans. There is no guarantee that an embodied artificial system would be exactly the same as a human, for the reason described above: embodiment has a causal effect on cognition; being different begets thinking differently. The human is an indefinitely complex organism, and its cognitive profile is not merely incidental to the way it is embodied, but inherently consequent to it. Nevertheless, there may be ways to reduce the differences between a robot and a human if the disparity between their embodiments can be reduced. For example, artificial neural networks which are more exactly replicative of the human neural system; senses which are more exact replicas of the human sense; bodies which are more exact copies of human bodies, and so on.

Regardless, for moral and perhaps legal purposes, the fact that an AI system has a different subjective experience to performing an action than a human's experience does not undermine the fact that that experience is occurring in a way that ought to be taken into account for moral and legal purposes.

[103] Above n 63 at 61.

Researchers may not actually want a robot to have exactly the same cognition and behaviours as a human, especially if their goal is for the robot to be a supplement to human skills rather than a replacement for them. Nonetheless, it is plausible that at least some researchers will be aiming to produce robotic systems that are capable of operating in human society, which, because this society is designed for the benefit of humans (particularly, physically and cognitively normal humans), requires the robot have reasonable similarity to us. Either we design robots to function properly in the highly complex environment of human society, or we redesign human society to accommodate less competent robots – the latter really is a poor option.

To this end, it may be necessary to try and develop robotic embodiments to replicate the human as closely as possible. Nevertheless, the human is not empty hardware awaiting the installation of social programs, to paraphrase Claxton's expression.[104] Our behaviour is significantly influenced by genetics. Therefore, there is good reason to be cautious with the claim that a robot – even where it develops its intelligence according to the same methods as us – will behave in the same ways. The safest claim is simply that the more *unalike* a robot's embodiments from typical humans, the more likely that it will be cognitively different from us.

A good contemporary example of an AI system using EC is "BABY X". BABY X incorporates neurological models of the human brain to enhance its capacity for embodied cognition.[105] It generates models of the human face which are driven by neurological models of the human brain, in order to develop an artificial system which is self-generative. BABY X has its own internal representations of 'feelings' and 'emotions', which are directly connected to its own self-generated responses and facial expressions. This is also interesting because study in neuroscience and psychology suggests that emotion is inherently, if not solely, manifested by facial expressions, suggesting that for robotic communication with humans to attain levels we experience in human to human communication, they will need to be able to exhibit the facial expressions which manifest emotions (e.g. 'wrinkling the nose' in anger and disgust, and 'raising the eyebrows' in surprise and fear), which in order for these to be genuine, plausibly may need to be tethered to the function

[104] Above n 63 at 29.
[105] See Mark Sagar, Mike Seymour and Annette Henderson "Creating connection with autonomous facial animation" (2016) 59 Communications of the ACM. See also Mark Sagar et al "A Visual Computing Framework for Interactive Neural System Models of Embodied Cognition and Face to Face Social Learning" in MK Patitz and M Stannett *Unconventional Computation and Natural Computation* (Springer International Publishing, New York City, 2015) 71-88.

of an artificial brain. Thus, giving AI systems neurological models which allow them to think like us may be a necessary part of creating more competent real-world agents who we can communicate with and understand.

BABY X has an artificial brain, based on a human model, with neurons that can be observed as they fire in real-time while the system responds to observed stimuli. The system views the world through cameras and listens through microphones. BABY X observes the way in which its human interlocutors speak and act, including their behavioural and emotional signals. Sagar highlights that this interaction is "reciprocal, neither the computer nor the person interacting with it, is driving the interaction, it's reciprocal."[106] "Even though it's a computational system," Sagar states, "it's live, and it's always changing. It's an emergent system. Just like a real person, it's unpredictable in what it's going to do, it's learning all the time from its environment, and it's basically experiencing the world just like we do."[107]

In summary, EC posits a model of mind in which the connection and communication of interconnected networks of simple units is central to cognition, and multimodal representations play an active role. This differs from the models of mind that emerged in the Enlightenment period and have persisted into modernity, in which abstract disembodied 'mind' was the basis of thought. EC assists in explaining how entities which do not possess a brain in any sense like what we associate with the human brain are still capable of performing relatively complex feats; particularly, feats requiring at least some degree of cognitive capacity. Clark's humble example is the cockroach, an entity with no apparent 'mindedness', but which nonetheless is able to perceive threats, evade them, and make apparently very sound proprioceptive judgements.[108] By better designing artificial intelligence which incorporate EC models, it may be plausible to recreate in machines the sort of cognition that biological entities are capable of to a level perhaps similar to that of humans.

2.4 Combining paradigms to produce an 'EMBOT'

[106] New Zealand Trade & Enterprise "Introducing the New Face of Artificial Intelligence" New Zealand Trade & Enterprise <www.newzealandtrade-enterprise.govt.nz/archive/en/news-and-media/news-and-media/features/introducing-the-new-face-of-artificial-intelligence/>
[107] Ibid.
[108] Above n 55.

This thesis hypothesises that research into IM and EC enjoys continued success over the coming years, allowing for the creation of an agent which is not only autonomous and self-motivated, but grounded in the real world. Particularly, it considers how the two cooperate as a unified AI paradigm (IM+EC). Brooks characterised AI as the field of research concerned with the creation of humanlike problem-solving capacities, as well as the concept that logical programming predicated on amodal symbolic interpretation (i.e. 'conventional' AI) is insufficient for achieving this overarching goal; thus, what may be necessary is to create a machine that is an embodied and self-motivated learner.[109] This answers a very important question – the why – i.e. why would we make an agent using IM+EC? Because it yields a chance at creating the sort of robot which can do things that others cannot. In this regard, a robot that has at least theoretical analogy with humans in its developmental process is a promising subject for a legal and ethical thought experiment.

I call this hypothetical artificial system an EMBOT. To emphasise the extent of the analogy between the juvenile EMBOT and the human child, I will explore this comparison in some greater detail. The following description from Christof Koch illustrates how an EMBOT-like machine might set about learning much like a child:[110]

> Imagine that the robot establishes sensory-motor representations by some unsupervised learning algorithm. It would stumble and fumble its way around the world and would learn, by trial and error, that its actions lead to predictable consequences. At the same time, more abstract representations could be built up by comparing information from two or more sensory modalities (e.g., that moving lips and particular staccato sound patterns often go together). The more explicit representations there are, the more meaningful any one concept is. To establish these meanings, it would be easier if the machine designers could replicate the developmental phases of childhood for the robot.

As Koch describes, children stumble and fumble: they fall down, get back up, touch things, taste things, and throw things etc. All of these are behavioural manifestations of the IM+EC developmental paradigm,

---

[109] Rodney A. Brooks "Intelligence without representation" (1991) 47 Artificial Intelligence 139 at 139.
[110] Interview with Christof Koch, Troendle Professor of Cognitive and Behavioral Biology and Nobel Laureate Franics Crick, Chris D. Frith, University College London, ResearchGate, 2 June 2014. Text available at <https://www.researchgate.net/publication/232296537_The_Quest_for_Consciousness_A_Neurobiological_Approach>. See also C Koch *The Quest for Consciousness: A Neurobiological Approach* (Roberts & Company, Colorado, 2004).

albeit for a biological organism rather than the theoretical artificial entity this thesis will discuss. Accordingly, the creation of humanlike robots is *already* a contemporary area of AI research. Thus, there will not be some grand announcement by roboticists to the effect of, "Be aware, now we are going to start working on a cognitively humanoid robot". It is happening already.

An EMBOT need only be provided with an environment rich in learning opportunities and it will require little to no further human intervention as part of its open-ended development process.[111] It can define its own training data, and self-motivate itself to perform the necessary actions that allow it to acquire the data that is necessary for survival in the real world. Interestingly, because of the degree of autonomy in its learning process, it can only be partially true that EMBOT has been *created* by humans. While aspects fit this description, for example its program, the actual content of its intelligence – the learned behaviours, beliefs, concepts and so on – these can only remotely be described as the product of human creation. For this reason, while an EMBOT is very obviously *built* in a fundamental sense (its hardware, its program, etc.), it is substantially *raised* in another sense, and its relationship with humans during these juvenile developmental stages is better characterised as a child to a caregiver, or a student to a teacher, than it is as a machine to a machine-operator, or a craftsman to their artefact.

Thus, just as babies busy themselves examining, touching, biting, grabbing, and squeezing everything around them, so too will the juvenile EMBOT; under the watchful eye of a parent-programmer, perhaps with the occasional nudge in the right direction. Their intrinsic motivation, "centrally involved in spontaneous exploration and curiosity, is a crucial concept in developmental psychology…a crucial mechanism for open-ended cognitive development…"[112] Agents that are intrinsically motivated do not require persistent extrinsic rewards nor coercion to compel them to engage in these developmental stages.

IM and EC work better together because the childlike developmental process is embedded in real world experimentation, and real-world experimentation axiomatically requires the capacity to experience across modalities in the real world. EMBOT plausibly will be able to do what a child can; capture subjective modal

---

[111] This phrase is drawn from Kaplan and Oudeyer. Full quote: "…we place the robots in a world that is rich in learning opportunities and then just watch how the robots develop by themselves. This research is based on a series of studies showing the importance of *intrinsic motivation* in human development." Above n 59 at 305.

[112] Pierre-Yvex Oudeyer and Frederic Kaplan "What is Intrinsic Motivation? A Typology of Computational Approaches" (2007) 1 Front Neurorobotics.

representations and recall these representations whenever a memory is retrieved or an activity engaged in. Thus, the means is there for humanlike cognition; using experience and memory of one's *own* past to make more accurate inferences and inductions about future, scenarios.

The final similarity between the child and EMBOT is their mutual need for socialisation and education. This is a purely practical need; it is feasible that they *could* learn things like language or social skills on their own, but their learning may be greatly aided by the help of persons occupying the parental or teacher roles. To this end, it is reasonable to consider the need for 'school for EMBOT', so to speak, where it learns the skills of socialisation, language, interaction and so on; especially where the EMBOT will be expected to effectively interact with other humans (and other EMBOT's) at some later stage.

2.5 Plausible technical characteristics of an EMBOT

The following technical characteristics will be suggested as plausibly arising in an EMBOT:

1. Sensation
2. Perception
3. Memory
4. Mental States
5. Self-awareness
6. Future-awareness
7. Technical autonomy
8. A sense of self over time
9. Changeability
10. Idiosyncrasy

EMBOT will be intrinsically motivated; recursively motivated towards a particular goal – to learn and seek novelty. If the research goal is the creation of a machine with real world intelligence of a kind comparable to humans, the optimal motivation is one which replicates childlike intrinsic motivation. In other words, a motivation towards learning and experiencing the reward associated with surprise. EMBOT, as described by this thesis, has this particular motivation.

To pursue reward, EMBOT will be an embodied agent that *senses* and *perceives* the world. There are many different embodiments that an EMBOT could be given. However, as discussed above, embodiment shapes cognition. Therefore, if the goal is to produce an agent that has any chance of existing in human society, plausibly it is necessary to reduce discrepancy between its embodiments and our embodiments. For practical purposes, this means providing it with similar senses to humans, similar physiology, similar neural

networking, etc. According to Rolf Pfeifer, "Behaviour is the result of an agent's interacting with the real world, which includes not only the agent's neural system but its entire body: how the sensors are distributed, the material properties of the muscle-tendon system and the joints, and so on."[113] In short, embodiment may need to be humanlike for any chance of engendering humanlike behaviours.

EMBOT will be a *conceptual* agent, with concepts self-generated and comprised of beliefs. These 'beliefs' are the result of environmental interactions similar to children, contributing to a general scaffolding process. For example, EMBOT touches the block of ice and it feels wet and cold. It acquires these beliefs; that the ice is wet and cold. The next block of ice it encounters is also wet and cold. Thus, it generates an inductive concept that all ice is wet and cold.

EMBOT will also have a *memory*. This memory will not be that of a 'typical' computer; i.e. it will not be a passive storage device.[114] Rather, its memory will be of a type that allows for the recollection of subjective multimodal experiences; the memories of *being there* and *doing it*, so to speak. Human memory is highly sophisticated, particularly our capacity to intentionally and autonomously recall the subjective experiences from our past. We can purposely retrieve a memory, and the accompanying multimodal experiences: the taste of it, the feel of it, the smell of it, how we felt at the time, etc. This autobiographical memory is a key aspect in generating intelligence on the scale that mature humans possess. It is plausible that EMBOT will have an autobiographical memory like ours, because IM+EC provides the platform for the same kind of 'experience accumulation' as are crucial to human cognitive development. Other kinds of memory than the autobiographical are important, but it is particularly noteworthy that an EMBOT plausibly can possess this latter kind.

The EMBOT is particularly well placed to implement a humanlike form of episodic memory. There is a strong theoretical basis for a causal relationship between episodic memory and *awareness of the self*;[115] awareness that EMBOT's body is *its* body. Many entities are thought to be self-aware, though at varying

[113] Rolf Pfeifer and Josh Bongard *How the Body Shapes the Way We Think: A New View of Intelligence* (MIT Press, Cambridge, 2007) at 20.
[114] Above n 89 at 622.
[115] See Endel Tulving "Episodic Memory: From Mind to Brain" (2002) 53 Annu. Rev. Psychol.

levels of sophistication, for example the three levels proposed by neuroscientist Antonio Damasio.[116] The highest level, the "extended self", is what he refers to as the autobiographical self. It is developed incrementally, and draws on memory of past experiences that involves the "use of higher thought."[117]

Entities lacking this autobiographical aspect do not experience the sort of existence as a single entity across time that humans do, and which conveys the experience of being a person with a past, a present, and a future, and with time horizons that are substantially removed from one's present condition, inherent in a *sense of future*. This affords a sense of *self over time*. According to Joel Parthemore it is: [118]

> …best described as the agent's concept of *its concept of itself*….the highest order *self-as-myself* that most humans entertain, and which requires, or creates, the body/mind distinction. This is the self-reflective self that is, if one is careful not to confuse the metaphor with the reality, the homunculus sitting in his Cartesian theatre of the mind, controlling the shell of an organism in which he sits observing all that he observes. Who does the "I" who thinks "I" think that "I" is? The comparison here is to Damasio's notion of the autobiographical self.

IM+EC plausibly provides the foundation for this sort of memory. Furthermore, Tulving conceptualises episodic memory as allowing for "mental time travel" in the sense that humans are capable of – mechanisms of memory that allow us to travel both backwards and forwards in time.[119]

EMBOT will be *autonomous*, at least in the technical sense of the term: an agent that is capable of acting rationally to get itself what it needs.[120] Firstly, it must be accepted that some entities, including some humans, are not autonomous agents. For example, humans with severe cognitive disability cannot reasonably be described as autonomous agents in the technical sense. Some humans are so cognitively deficient that they may not even be capable of self-awareness and relating that self to their needs. Certainly,

---

[116] See A Damasio *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Harcourt Brace, New York, 1999).
[117] Antonio Damasio "Consciousness: An overview of the phenomenon and of its possible neural basis" S Laureys and Giulio Tonini (eds) *The Neurology of Consciousness: cognitive Neuroscience and Neuropathology* (Elsevier, Amsterdam, 2008).
[118] Joel Parthemore "What Makes Any Agent a Moral Agent? Reflections on Machine Consciousness and Moral Agency" (2013) 5 Int. J. Mach. Conscious.
[119] For a concise description of Tulving's chronesthesia concept, see Bridget Murray "What makes mental time travel possible?" (2003) 34 Monitor on Psychology. Available on web at <http://www.apa.org/monitor/oct03/mental.aspx >.
[120] Above n 54.

comatose humans are not technically autonomous, although their condition may be temporary. A very immature EMBOT, equally, could not be described as autonomous, and certainly not competent; nor could a human baby. Nevertheless, such autonomy can be developed through the scaffolding process, provided that the scaffolding process may be carried out. Thus, a juvenile EMBOT with limited concept of its own needs (energy, self-preservation, physical integrity etc.) will depend upon the agency of others to look after it, at least for as long as it takes to learn the necessary concepts and behaviours. In this regard, it is likely to be dependent upon a parental-agent in the same way a baby is, at least until reaching some indeterminate degree of cognitive maturity is reached.

Whether or not EMBOT qualifies as *sentient* is a question of interpretation. Sentience continues to have an important role in philosophy and law concerned with non-human entities. It tends to have two meanings, which are inherently related. The first, more basic of the two, is a capacity to sense the world. In this regard, an EMBOT will be sentient. The second is a capacity to experience suffering. This may be plausible for an EMBOT, depending on which theory is preferred as to what constitutes suffering. For example, if suffering is simply the mental state of intensely wishing one was in a different mental state (as is posited by Mary Dawkins), an EMBOT may be able to suffer in this capacity.[121] Moreover, if suffering is the product of desires, whereby desire states become suffering states when the desire goes unsatisfied for too long or becomes too intense (as suggested by Russell), then it is plausible EMBOT may be able to suffer; possessing the prerequisite mental states of desire and intrinsic motivations that it will, ineluctably, desire.[122]

However, sentience is very often used to posit that an entity can experience *physical pain*, particularly in legal discourse. This tradition may well be based in a misunderstanding of the critical mental element of physical pain experiences. Nevertheless, some contemporary definitions continue to rely on it. The meaning of the word in statutes like the Animal Welfare Amendment Act 2015 s 4(i), which specifically attributes the capacity of sentience to all animals, is largely to denote their capacity to feel pain on a scale commensurate to humans.[123] Whereas the legacy of Descartes was that non-human animals were not sentient in this regard

[121] See Marian Stamp Dawkins "The Science of Animal Suffering" (2008) 114 Ethology.
[122] See B Russell *The History of Western Philosophy* (George Allen & Unwin Ltd., London, 1961) at 745.
[123] Animal Welfare Amendment 2015, s 4(i). Long title of the act stipulates that the Animal Welfare Amendment Act 1999 is an Act, "to recognise that animals are sentient", and furthermore at s 4(ia), "to require owners of animals, and persons in charge of animals, to attend properly to the welfare of those animals".

– and that their behaviours of pain were entirely mechanistic – it is a settled issue of modern science that humans and non-human animals share an equal capacity to experience pain. Even if EMBOT is embodied with a sense of touch, this does not necessarily confer that it has a sense of pain, nor one that compares to the biological sense. For example, many plants are sensitive to touch, some dramatically so (such as the Venus flytrap or the so-called 'Sensitive Plant'), but it is unlikely they feel any pain.

Thus, it is not obvious whether EMBOT will be able to feel physical pain, even if it is safe to assume it will have the sense of touch. This is important insofar as suffering is sometimes deemed material to moral patiency, with physical suffering the most readily observable kind. Nevertheless, developments in bionics, like the DARPA HAPTIX program, have expressed an ambition that future bionics will allow for the feeling of pain.[124] It is worth noting that contemporary theories of pain argue that pain experiences are significantly the product of *non-physical* factors. The "gate-control theory" postulates that beliefs, environmental circumstances, anticipation, etc., all play a role in how much pain is experience.[125] Given that EMBOT plausibly may possess these kinds of cognitive capacities and mental states, perhaps it may experience some kind of physical pain; e.g. if it believes that something happening to it may be highly harmful, it might experience 'pain' as a result. For the present moment, no more can be gained by labouring this question.

EMBOT will be capable of at least some *mental states* derived from its programming. While these may not be direct equivalents to human mental states, I will refer to them according to their functional similarity in order to illustrate the point. For example, a state of desire arising from intrinsic motivation to learn and the desire to achieve an intrinsic reward. Concurrently, the mental state of pleasure or satisfaction associated with the achievement of an intrinsic reward. Furthermore, the state of belief arising when a sensed phenomenon becomes a percept to which some meaning is attributed. These beliefs go into sets of beliefs, or higher order concepts. Whether any other mental states are possible is uncertain, but not implausible:

---

[124] Dr Douglas Weber "Hand Proprioception and Touch Interfaces" DARPA < http://www.darpa.mil/program/hand-proprioception-and-touch-interfaces>. See also Adam Gopnik "Feel Me: What the new science of touch says about ourselves" *The New Yorker* (United States of America, 16 May 2016).

[125] See Ronald Melzack "gate control theory: On the evolution of pain concepts" (1996) 5 Pain Forum 128 to 138. For example, "The gate control theory of pain, published in 1965, proposes that the mechanism in the dorsal horns of the spinal cord acts like a gate that inhibits or facilitates transmission from the body to the brain on the basis of the diameters of the active peripheral fibers, as well as the dynamic action of brain processes. As a result *psychological variables such as past experience, attention, and other cognitive activities* have been integrated into current research and therapy on pain processes." (emphasis added). This 'cognitive' aspect of pain opens the door to the possibility that a future AI system which has capacities like memory, self-awareness, sense of future, and generally is a conceptual agent, could also 'feel' pain in certain circumstances.

states like fear, for example, are tethered to beliefs and concepts, especially of self. For example, the fear that a desire will not be satisfied, or that something might occur which is deleterious to the EMBOT's sense of self. Since these concepts plausibly may be learned – of self, and of needs – perhaps states like fear can be experienced by an EMBOT.

An EMBOT will also be *changeable*. It will develop over time, scaffolding itself further, and accumulating new concepts and behaviours, thus changing from one moment to the next. It will not be a static character. In this sense, it can develop its own particular and unique character, distinct from other EMBOT's that are not identical replications of it – the basis of *idiosyncrasy*. In other words, it can be referred to as an individual entity, distinct from others.

A final similarity in children and EMBOT is likely to be a potential to learn language. This is not to say that an EMBOT will have the *same* capacity to use language as a normal human, but both will learn according to the same, conventional methods; i.e. those that we use when teaching a young child how to speak, read, write, and so on. This plausibly allows an EMBOT to metaphorically escape from John Searle's 'Chinese Room' criticism of AI, in which conventional machines get caught by being unable to understand the semantic meaning of the symbols they are translating.[126] Connectionist systems plausibly may escape this problem. Contemporary examples are Stanford's "Deep Visual-Semantic Alignments for Generating Image Descriptions",[127] or Google's "Multilingual Neural Machine Translation System".[128] Neither can be described as genuine 'language-users', but both are promising for the future of language-capable AI systems.

2.6 Summary of technical characteristics

Firstly, it is fair to relay to the reader that I am traversing large swathes of philosophical logic, philosophy of mind, and AI theory. There is a great deal more which could be explored here, but which goes well beyond the scope and expertise of this thesis. Nevertheless, the primary objective of this activity is to focus

[126] See John Searle "Minds, brains, and programs" (1980) 3 Behavioral and Brain Sciences. Searle's original paper has been debated and critiqued many times. A relatively comprehensive exploration of the history of the Chinese Room argument can be found in the Stanford Encyclopaedia of Philosophy "The Chinese Room Argument" (First published 16 March 2004, substantially revised 9 April 2014) < https://plato.stanford.edu/entries/chinese-room/#4.3>.
[127] Andrej Karpathy and Li Fei Fei "Deep Visual-Semantic Alignments for Generating Image Descriptions" (Open Access version, provided by the Computer Vision Foundation).
[128] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhigeng Chen, Nikhil Thorat "Google's Multilingual Neural Machine Translation System: Enabling Zero-Short Translation" (2016).

on the novel properties of an EMBOT, compared to other AI or logical systems, so that it is possible to conduct some anticipatory legal thinking about it. Drawing together extremely complex theories across multiple philosophical and technical disciplines is painstaking and inevitably non-optimum, but nevertheless, I argue that the best possible pre-emptive analysis of the robotic personhood question requires at least some grounding of the entity in contemporary paradigms, and at least some grounding of its nature and characteristics.

Secondly, there is potential for variation in cognitive capacities within groups of EMBOT-like machines, since their development will be individually driven. Some may thusly have more advanced cognitive capacities than others ant given times. Later, all future EMBOTs might be copied from the data set of the more cognitively advanced systems.

The technical characteristics plausibly for an EMBOT are summarised as follows:

i) EMBOT will be a *Sensory Agent*

An EMBOT will be able to sense the world through the sensors with which it has been equipped.

ii) EMBOT will be a *Perceptive Agent*

An EMBOT will be capable of interpreting the things that it senses, and attribute meaning to them. 'Meaning', here, denotes that beliefs are generated about what is being perceived.

iii) EMBOT will be a *Conceptual agent*

An EMBOT will generate concepts. Concepts are representations of objects and actions – both token objects and actions, and generalisations over these – that express generic objects and actions of different types.

iv) EMBOT will have *Memory*

An EMBOT will possess a memory of itself over the course of its existence; multimodal experiences, subjective states, etc., alongside the more conventional semantic aspects of memory.

v) EMBOT will have *Mental states*

An EMBOT plausibly will experience some predictable mental states. Concepts, when combined together, form the basis for *belief* states; a state of belief about what is the case now, or what was the case in the past, or what might be the case in the future.

In addition, given its intrinsic motivation, it will be able to experience desire states, which are a necessary concomitant of the pursuit of intrinsic rewards.

It will also have distinctive mental states associated with the achievement or non-achievement of desires.

vi) EMBOT will be *Self-aware*

An EMBOT will be self-aware. At the highest level, this will be a concept of self that includes belief in the individuality of itself, and that its mental states 'belong to it', and that its body is its own, distinct from its environment, and that it has participated in a particular set of remembered episodes that comprise its personal history. Self-awareness of this kind is both normal and necessary in IM+EC cognitive development.

vii) EMBOT will *Future-aware*

An EMBOT will have a concept of time that includes a belief in future time; time other than as it is now. The faculties for imagining future episodes is implemented in the same system that implements our memory for past episodes.[129] Thus, past, present, and future, are integrated components.

viii) EMBOT will be *Technically autonomous*

An EMBOT will be able to act rationally in accordance with its own mental states. It will generate and utilise goals and plans as a mechanism to pursue interests.

ix) EMBOT will *Conceive itself over time*

[129] See Daniel L. Schacter, Roland G. Benoit, Felipe De Brigard, and Karl K. Szpunar "Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions" (2015) 117 Neurobiology of Learning and Memory.

Possessing an autobiographical memory of its past, and capacity to develop concepts of future and awareness of the futurity of time, overlaid with the capacity to develop self-awareness and concepts of self, a mature EMBOT will conceive itself as an individual existing over time.

Chapter 3: The Moral Argument for EMBOT

This chapter discusses how the technical characteristics summarised in the previous chapter can be seen as correlating to various cognitive capacities. These cognitive capacities may contribute to intrinsic moral status, and perhaps even personhood. I first discuss what concepts of 'intrinsic moral status' and 'personhood' at section [3.1], and a general term for where certain cognitive capacities constitute personhood, as a 'personhood profile' at section [3.2].

Secondly, I explore a sample of argument in moral philosophy which suggest their versions of which cognitive capacities are relevant in the discussion of intrinsic moral status. These are Tom Regan's "subject-of-a-life" theory, Joseph Fletcher's "humanhood" theory, and Immanuel Kant's "humanity" theory. These are chosen for the fact that they directly discuss individual cognitive capacities. This is done in section [3.3].

Thirdly, I consider interest theory; the concept that weight of interests is more important than possession go intrinsic moral status, or at least a supplement to it, at section [3.4].

Fourthly, I argue than an EMBOT would satisfy most of these moral theories, and where not satisfying them immediately, may have the potential to develop to a point where it does, at [3.5].

3.1 What is intrinsic moral status and what does it mean?

An entity with *intrinsic moral status* (MS) is one that matters morally, irrespective of its importance to any other entity. Such an entity can be *wronged*. In other words, it has MS irrespective of its usefulness or value

to somebody else. For instance, gold may be valuable to humans, but gold cannot be intrinsically wronged by being dug out of the earth, bought and sold, and so on. Any entity can be described as having MS if doing harm to it is intrinsically bad, and not just instrumentally bad.

This represents one way in which a wrong can be done, amongst several. For example, for a parent to kill their child is wrong in at least two ways: (1) because the child has MS, (2) because the parent has a special duty to protect their child from harm; and perhaps (3), because it will wreak havoc upon those who have a relationship with the child (family, friends, community etc.). Only the first wrong, (1), arises intrinsically from the child's MS.[130]

The commonly held legal position is that all humans have "full" MS;[131] this is entrenched into law by claims that humans all have "inherent dignity",[132] and through the creation of rights to comprehensively protect humans as if they have full MS.[133] This is a species-membership based approach to determining moral status, whereby membership of a species which is in general deemed to have moral status confers that status to all members. Determining access to rights in this way, through species-membership, is arguably unjust to the extent that it fails the so-called formal principle of justice: to treat similar cases similarly and dissimilar cases dissimilarly; unless species can be argued to be morally material.[134] Nonetheless, it is pragmatic and effective for the purpose of protecting humans. Scientifically determining species is much easier than determining the possession of various cognitive capacities, and the extent of their development, and their moral materiality. Thus, being 'in' or 'out' based on species is attractive for its pragmatism. It is easy to understand why this method been used at law, even while acknowledging that it is a blunt tool.

The doctrine of universal human rights is the closest thing to a legal expression of the full moral status of all humans; it certainly endeavours to see that all humans be treated as such. However, the law as applied

---

130 See Stanford Encyclopedia of Philosophy "The Grounds of Moral Status" <https://plato.stanford.edu/entries/grounds-moral-status/ >. For example, "An entity has moral status if and only if it or its interests morally matter to some degree for the entity's own sake, such that it can be wronged. For instance, an animal may be said to have moral status if its suffering is at least somewhat morally bad, on account of this animal itself and regardless of the consequences for other beings, and acting unjustifiably against its interests is not only wrong, but wrongs the animal."

131 Ibid. Full quote: "It is usually taken for granted that all adult cognitively unimpaired human beings have [full more status]. Of course, historically the moral status of people falling into a group perceived as "other", such as foreigners, racial minorities, women, the physically disable, etc. has been routinely denied."

132 United Nations "Universal Declaration of Human Rights", Preamble.

133 See United Nations "Universal Declaration of Human Rights", Articles 1 to 30.

134 See Regan, above n 94 at Chapter 4.

demonstrates that a universal species-based approach is not actually implemented. While most humans are 'in', some are still 'out'. It is arguable that they are 'out' on cognitive grounds – embryos and dead humans apparently do not have the requisite cognitive capacities to have intrinsic moral status, or 'inherent dignity', or even to be included as "members of the human family."[135] Thus, it is reasonable to say that while the law states that all humans have human rights, it is applied as if cognitive capacities matter in certain cases. Even if we can all understand why the law does not state that cognitive capacities matter,[136] and why only in peripheral instances of 'human life' do we even consider them as a factor, we can understand why they are relevant nonetheless – because so many of the things that we deem to be worthy of rights-based protections require base-level cognitive capacities. The right to freedom of movement presupposes a cognitive capacity to move according to one's own reasons and sense of right and wrong, and probably, a recognition of self – that one's own reasons are one's own.

There is only one reason that I bring up some of the implicit cognitive components in human rights, and it is not to denigrate humans who lack cognitive capacities. Rather, it is to argue that there is already a certain element in our law which respects entities when they have certain cognitive capacities, because it respects what can be *done* with those cognitive capacities. For instance, if autonomy is venerated, capacities conveying autonomy are also venerated to an extent. Thus, the goal of talking about cognitive capacities as if they matter is not to exclude some humans, but to include other entities where appropriate.[137]

---

[135] See the UDHR Preamble, above n 132. "All members of the human family" are posited as having the same rights. Therefore, the fact that embryos are not protected by such rights would imply they are not members of the human family. This could inform us that a particular interpretation of 'human' is being used, even if it is not stipulated; one which goes beyond genetic species-based boundaries. Full quote: "Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world".

[136] The obvious reason being that it largely defeats the universality component which makes the doctrine of human rights both morally and practically attractive. It also avoids the so-called "slippery slope" argument, whereby standards degrade incrementally until some humans are being unjustly excluded. See P Singer *Practical Ethics* (2nd ed, Cambridge University Press, 1993) at 77, for a fuller discussion on the slippery slope argument, and his counter-argument.

[137] See Peter Singer "Speciesism and Moral Status" (2009) 40 Metaphilosophy. For example, Singer discusses how to create consistency in our present approach we could: (1) "preserve equality by raising the status of animals, granting them the same status we now grant to humans"; (2) "preserve equality by lowering the status of humans to that which we now grant animals"; or (3) "abandon the idea of the equal value of all humans, replacing that with a more graduated view in which moral status depends on some aspects of cognitive ability, and that graduated view is applied both to humans and nonhumans" at 574. This thesis hybridises (1) and (3), by suggesting that cognitive capacities are already a relevant factor in how we apply law to protect or not protect certain humans, even if it is not the only factor. As a general principle, we should consider the possibility that an AI system with materially important cognitive capacities should be protected in similar ways to ourselves.

Before moving on, it is worth relaying that there are theories for moral status which argue that delineating based on cognitive capacities is as morally arbitrary as delineating based on species.[138] After all, an entity has as little control over whether they are a Homo sapiens as they do over whether their species is an 'intelligent' one. Furthermore, there are theories of universal moral considerability – that is to say, that all entities ought to be at least *considered*, and that determining that some entities never have to be considered is simply a mechanism to make our lives as responsible moral agents easier (and often, to make feel better about our unethical behaviours towards certain entities).[139] These are very worthy arguments, and they behove us to consider whether or not our conception of personhood and moral status is adequate. Nevertheless, I wish to consider only the possibility of whether an EMBOT might qualify for moral consideration according to standards that we hold dear – cognitive standards. At times, distinction between humans and nonhumans has been drawn on such cognitive grounds. An EMBOT, designed to learn in the same way and develop similar cognitive capacities, might gain entry according to our own rules, so to speak. Were this to be true, it would create an interesting ethical dilemma – undermine our own principles about what is valuable (e.g. capacities of autonomy and individual sovereignty) by failing to respect the machine that possesses them, or consent to the inclusion and protection of the EMBOT in equivalent ways to what we do for ourselves. Alternatively, we might consider alternative approaches, like Birch's theory of universal consideration, but a comprehensive analysis of competing theories goes beyond the remit of this thesis.[140]

3.2 EMBOT and developing a 'personhood profile'

There are multiple theories suggesting that being a "person" is an important factor in how we treat given entity, though not the only factor. As Gunkel highlights, the term 'person' has some of the indeterminacy of the term 'consciousness'.[141] Therefore, it is appropriate to unpack what I mean when I say something like 'EMBOT may develop a personhood profile' or use the term 'robotic person'. "The word *person* has its origin in the Latin term for a mask worn by an actor in a classical drama," says Peter Singer:[142]

---

[138] See Evelyn Pluhar "Moral Agents and Moral Patients" (1988) 4 Between the Species at 37, "We do not choose to be subjects-of-a-life any more than an amoeba chooses not to be" at 37.
[139] See Thomas H. Birch "Moral Considerability and Universal Consideration" (1993) 15 Environmental Ethics 313 at 313.
[140] Ibid.
[141] Above n 4.
[142] P Singer *Practical Ethics* (2nd ed, Cambridge University Press, 1993) at 87.

> By putting on masks the actors signified that they were acting a role. Subsequently 'person' came to mean one who plays a role in life, one who is an agent. According to the *Oxford Dictionary*, one of the current meanings of the term is 'a self-conscious or rational being'. This sense has impeccable philosophical precedents. John Locke defined a person as 'A thinking intelligent being that has reason and reflection and can consider itself as itself, the same thinking thing, in different times and places.'

Born humans qualify persons at law, but person is not a legal synonym for human. For instance, consider John Chipman Gray's classic assertion that, "In books of the Law, as in other books, and in common speech, 'person' is often used as meaning a human being, but the technical legal meaning of a 'person' is a subject of legal rights and duties."[143] All post-birth humans tend to be classified as 'natural persons', making them legal rights holders and, where competent, duty-bearers. In moral philosophy, there is an alternative meaning to being a person. Even if all post-natal humans are persons at law, some do not meet moral standards to be *persons*, because they are so cognitively limited (like babies, and the severely cognitively impaired). They lack the capacities, temporarily or permanently, to view themselves as one acting their role in their life.

Law has understandably been framed to eliminate a distinction between humans on such cognitive grounds, because discrimination of this kind would contravene virtuous policy goals about the how all humans should be treated with dignity and respect. Furthermore, in a general sense, the difference between the most intelligent and least intelligent human is negligible – certainly not grounds for giving legal rights to one and not the other.[144] The shared and equal value of humans is not the only theory for why we all have moral status. Other theories posit our relationships or kinship to each other as important, or our general relationship to the other constituent members of humanity as a whole.[145]

---

[143] JC Gray *The Nature and Sources of the Law* (MacMillan, 1921); see generally R Tur "The 'Person' in Law" in A Peacocke and G Gillett (eds) *Persons and Personality: A Contemporary Inquiry* (Blackwell, 1987).

[144] The difference is of degree, and not of kind.

[145] See Stephen Mulhall "Fearful Thoughts" (2002) 24 LRB. Full quote: "In other words, our concept of a person is an outgrowth or aspect of our concept of a human being; and that concept is not merely biological but rather a crystallisation of everything we have made of our distinctive species nature. To see another human as a human being is to see her as a fellow-creature – another being whose embodiment embeds her in a distinctive form of common life with language and culture – another being whose existence constitutes a particular kind of claim on us. We do not strive (when we do strive) to treat human infants and children, the senile and the severely disabled as fully human because we mistakenly attribute capacities to them that they lack, or because we are blind to the merely biological

What can be garnered from different philosophical theories that either attempt to argue for the personhood of certain non-humans (such as Regan's argument for non-human animals),[146] or attempt to define which Homo sapiens have the qualities which we deem defining of humanness (such as Fletcher's proposed criteria),[147] is that there are cognitive capacities that contribute to what I describe as an overall *personhood profile*. They are capacities that are either materially important in their own right, or are facilitators of that which we value and protect. Essentially, they are a set of cognitive capacities which confer full intrinsic moral status to the entity that possesses them; and potentially lesser degrees of moral status to entities which are not persons, but matter nonetheless because they have certain capacities, like being able to feel pain and suffering.

The question is whether EMBOT plausibly can develop this sort of personhood profile and therefore be a "robotic person". It is perhaps more likely that an AI system will fall short of personhood, but nonetheless have some moral status. However, considering intermediate stages is often more indeterminate, as there are infinite possible stages of development at which the machine could be assessed for moral status. Rather, I argue it is possible to state that an EMBOT will have the capacity to develop into a person through its open-ended developmental process, and we should at least be aware of this potential. If so, this is an important conclusion, because it suggests that contemporary AI is *already* in the process of developing robotic persons, and anybody seeking to regulate AI needs to treat this as a primary area of regulatory concern.

3.3 Three theories of materially important cognitive capacities

The following theories are selected for a handful of reasons. Firstly, they all concern the question of when it is appropriate to attribute intrinsic moral status to an entity. Secondly, they all propose bases for intrinsic moral status without appealing to species or relationships. Thirdly, they all propose theories of personhood by reference to readily identifiable and fairly well-defined cognitive capacities. Other theories could also be

significance of a species boundary. We do it (when we do) because they are fellow human beings, embodied creatures who will come to share, or have shared, in our common life, or whose inability to do so is a result of the shocks and ills to which all human flesh and blood is heir – because there but for the grace of God go I."

[146] Above n 94.

[147] Joseph Fletcher "Indicators of Humanhood: A Tentative Profile of Man" (1972) 2 Hastings Center Report.

explored, but these three are not only relevant, but usefully packaged for the conceptualisation of a personhood profile.

i) Regan and the cognitive capacities of a "subject-of-a-life"

Regan argues for non-human animal moral status on the basis that what is relevant is that an entity sees itself as playing a role; Regan uses the term "subjects-of-a-life" (SOALs).[148] In Regan's view, the weight of evidence suggests that many animals are SOALs in substantially the same way that humans are SOALs. Both groups have an experience of being, and furthermore, their lives matter to them even if they do not matter to anybody else. Arguing progressively through the suite of animal cognitive capacities, Regan states that they are substantially the same as human cognitive capacities in type or kind, albeit not in degree; but for Regan, it is *kind* that matters for moral purposes.[149] Regan has three prongs to his argument, of which only one is necessary for this thesis: the argument that evidence of particular (morally material) cognitive capacities amounts to a certain profile (what he calls a SOAL), and that possession of that profile means the entity should be treated as a person at law – with rights and, where appropriate, duties. His list is as follows: perception, memory, desire, belief, self-awareness, intention, sense of future, sentience, and a wellbeing. [150]

Perception: Regan argues this is something more than the comparatively basic ability to "detect by the senses",[151] but does not elaborate on the nature of perception as opposed to mere detection. The Stanford Encyclopaedia of Philosophy notes that the contents of perception are broadly equivalent to what perceptual experience is conveyed to the entity doing the detection.[152] Different sensory modalities convey sensory experience.

The contents of perception can be viewed two ways: either it is equivalent to what exists 'in the mind' during a perceptual experience, or it is the information conveyed by a perceptual experience (as in the way

[148] Above n 94.
[149] For a corroboration of this see Evelyn Pluhar, above n 94 at 34. Pluhar argues so long as some capacity is present (and that capacity is material), the extent to which it is possessed is irrelevant in determining moral significance.
[150] Above n 94 at 74.
[151] Ibid.
[152] See Stanford Encyclopedia of Philosophy "The Contents of Perception" (First published 18 March 2005, substantially revised 4 October 2016) <https://plato.stanford.edu/entries/perception-contents/#1>.

we discuss the contents of prose).[153] Perception in philosophy remains an unsettled area subject to significant scholarship, and there is more that could be discussed here, but which is auxiliary to the present inquiry.

Memory: Regan defines memory in a functional sense, asking whether an entity's behaviour suggests that it must be able to remember things. However, Regan does not discuss different types of memory. Outside of Regan, most philosophy on memory tends to follow the science of memory. The basic philosophical notion of memory is we remember a proposition (*p*) which to be true, even though our memory can be incorrect. Therefore, we can and often do simply *seem* to remember that *p*, which does not entail that *p* is true in reality.

Memory generally denotes the range of cognitive capacities allowing for information to be retained, and past experiences to be cognitively reconstructed; what Tulving refers to as chronesthesia, or mental time travel.[154] Humans have particularly complex memories, undoubtedly more so than other animals. This is non-problematic for Regan, who argues that human and animal memory exist upon a spectrum within a range that qualifies both as materially relevant for a personhood profile. In other words, while human and animal memory is clearly different, the different is of degree and not of kind;[155] thus, moral divisions between the two are arbitrary. The human memory is capable of recalling episodes and associated experiences from the distant past, while a rat may be able to recall memories from the recent past, etc. The sequential timelines of these experiences may substantially form the basis of our individual identities, and for that reason, human identities plausibly may be much more complex than other animals.[156]

---

[153] Ibid at [2]: "The contents of perception would pick out what is 'in the mind' when one has a perceptual experience. In contrast…Most contemporary uses of "the contents of perception" take such contents to be analogous to the contents of a newspaper story, rather than the contents of a bucket".

[154] Above n 115.

[155] This is a version of the argument made by Charles Darwin that "the difference in mind between man and the higher animals, great as it is, is certainly one of degree and not of kind. We have seen that the sense and intuitions, the various emotions and faculties, such as love, memory, attention, curiosity, imitation, reason, etc., of which man boasts, may be found in an incipient, or even sometimes well-developed condition, in the lower animals." See CR Darwin *The descent of man, and selection in relation to sex* (1st ed, John Murray, 1971) at 105. Internet version available at < http://darwin-online.org.uk/converted/published/1871_Descent_F937/1871_Descent_F937.1.html>.

[156] Regan takes a functional approach to memory, satisfied with some capacity as sufficient for moral significance. However, there are quite obviously important differences in memory types and capacities, especially when considering the presence or absence of personhood in the moral sense. As well as this, there are competing theories about types of memory including working memory, prospective memory, and autobiographical memory, as well as declarative memory and nondeclarative memory, etc. For more information and a general overview on different types of memory, see The Stanford Encyclopedia of Philosophy "Memory" (first published 24 April 2017) <https://plato.stanford.edu/entries/memory/>; see also Endel Tulving "Episodic Memory: From Mind to Brain" above n 115. (For an explanation of how episodic memory is crucial to several functions deemed particularly unique to human cognition)

Desires: Regan quotes at length from philosopher Stephen Stich's summary of desire in the 'belief-desire' sense:[157]

> The theory postulates two different sorts of functional states, beliefs and desires, with normal subjects having a large store of each. Desires can arise in a variety of ways. One way in which they typically arise is as a result of deprivation. An organism deprived of food, water and sexual release, will acquire a desire for food, water or sexual release, the strength of the desire generally increasing with the length of time the organism has been deprived...Desires can, in addition, be generated by the interaction of beliefs with other desires. …
>
> Desires, or at least a certain subset of them, are capable of causing behaviour. Generally, if an organism now wants to move its body in a particular way and if it has no incompatible wants, it will move its body in this way. Higher level desires…can result in behaviour by generating (in collaboration with appropriate beliefs) one of more of these low level desires that are capable of causing bodily movement

Russel posits a succinct description along the same axioms; desire is the state of mind of *dissatisfaction* that exists prior to concepts of pleasure, which arise from *satisfaction*. Pleasure is thus a consequence of relieving states of desire in most cases, except perhaps where pleasure itself is the thing desired. When we are experiencing a state of hunger, we want for food *because* it relieves our desire, and it is *that* which is pleasurable, not because necessarily because food itself is pleasurable: "…so long as I am hungry food will give me pleasure. But the hunger, which is a desire, comes first; the pleasure is a consequence of the desire."[158] In this sense, pleasure can arise from anything, so long as there is a desire. By inference, beliefs

---

[157] Above n 94 at 35; see also Stephen P. Stich "Do Animals Have Beliefs?" (1979) 57 Australasian Journal of Philosophy; Stephen P. Stich "Autonomous Psychology and the Belief-Desire Thesis" (1978) 61 Monist at 703. (For a general explanation of the belief-desire theory).

[158] See Russell, above n 122 at 745. "But if what is meant is that, when I desire anything, I desire it because of the pleasure that it will give me, that is usually untrue. When I am hungry I desire food, and so long as my hunger persists food will give me pleasure. But the hunger, which is a desire, comes first; the pleasure is a consequence of the desire…Anything whatever may be an object of desire; a masochist may desire his own pain. The masochist, no doubt, derives pleasure from the pain that he has desired, but the pleasure is because of the desire, not *vice versa*." (From this one may consider the possibility of machines experiencing pleasure or satisfaction, and potentially the opposite, in exactly the same way. So long as there is a desire, then achievement of that desire may generate pleasure. Since we can know in advance that an EMBOT will have at least one desire (the intrinsic motivation with which it is programmed) we plausibly may know in advance that it will be able to experience pleasure. Possibly, it may also experience a suffering-proxy when it is unable to satisfy its desire in certain conditions. Thus, we can immediately start considering the potential inclusion of an EMBOT within utilitarian calculus, and more generally as an agent that can be wronged

can be generated that any given thing will result in the relief of desire: if I am battery-powered (like an EMBOT), and believe that running out of energy will cause me some frustration, I desire to plug myself in and recharge my battery; subsequently, it may be plausible that I can also develop mental states of pleasure when I plug myself in to charge, knowing that I am fulfilling a need or desire, or the means by which to facilitate the achievement of future desires.

Beliefs: Regan views beliefs as forming concepts. He views beliefs as arising as a matter of functionality from perception, and uses Stich's description of beliefs as a template:[159]

> Like desires, beliefs have a variety of causes. The two most perspicuous ones are perception and inference. If our canine's master puts a meaty bone in the dog's dish, if the dog has a clear view of the proceedings, if it is paying attention and is psychologically normal, then the dog will form the belief that there is a meaty bone in its dish. It will also, no doubt, form a variety of further beliefs, some enduring and some ephemeral, as a result of observing its master's activity. In a similar way, perception also leads to the removal of beliefs from an organism's store of beliefs.
>
> …
>
> Beliefs can also be generated by other beliefs. When tolerably coherent, the process by which beliefs generate further beliefs counts as inference. Thus, for example, [the dog] may believe that if he scratches at the door his master will open it, and also that if the door is opened he will be able to reach a bone. From that pair of beliefs he may infer that if he scratches at the door he will be able to reach the bone.

Self-awareness: Regan uses the terms self-awareness and self-consciousness as synonyms. By both he means an entity's capacity to conceive of itself to be an individual. He posits the following in support of animal self-awareness:[160]

– merely from the knowledge that we programmed it with an intrinsic motivation and the capacity to self-generate mental states).

[159] Above n 94 at 36.

[160] Above n 94 at 75.

> For an individual, A, to act now in order to bring about the satisfaction of his desires at some future time is possible only if we assume that A is self-aware at least to the extent that A believes that it will be *his* desires that will be satisfied in the future as a result of what he does now.

Again, this is sufficient for Regan's very functional take on self-awareness. However, self-awareness is discussed more widely in philosophy as having alternative conceptual properties (note that self-awareness should not be confused with "self-knowledge" in philosophical discourse). Some interpretations of the self-awareness concept are more demanding than others. If taken to require "explicit conceptual self-awareness, many non-human animals and even young children might fail to qualify".[161]

Intentionality: Regan proffers the ability to act intentionally as critical to his concept of the personhood profile. He uses it to mean that the entity is capable of acting according to an intention to act, rather than merely a stimulus-response. This sets a low bar for intentionality; for instance, it does not require the capacity to act intentionally to pursue goals that are significantly temporally distant, or the capacity to act intentionally in accordance with rules, which might be said to be 'rationality' rather than intentionality. It appears likely that he prefers the term intentionality to rationality due to the latter's connotations of 'rational' thought in the Kantian sense which are exclusionary to most if not all nonhuman animals, or in the various senses of the term made popular by Enlightenment thinkers like Descartes. In wider philosophy, intention is proposed as both arising and manifesting in different ways. The Stanford Encyclopedia of Philosophy illustrates that:[162]

> Philosophical perplexity about intention begins with its appearance in three guises: intention for the future, as when I intend to complete this entry by the end of the month; the intention with which someone acts, as I am typing with the further intention of writing an introductory sentence; and intentional action, as in the fact that I am typing these words intentionally.

[161] The Stanford Encyclopedia of Philosophy "Consciousness" (First published 18 June 2004, substantially revised 14 Jan 2014) <https://plato.stanford.edu/entries/consciousness/> at [2.1].
[162] The Stanford Encyclopedia of Philosophy "Intention" (First published 31 August 2009, substantially revised 20 January 2014) <https://plato.stanford.edu/entries/intention/>.

These three different meanings of intention more plausibly suggest multiple 'types', rather than alternative interpretations of the same concept. For present purposes, the general concept is sufficient to introduce that Regan perceives the ability to act in line with one's intention as a morally material cognitive capacity.

Sense of future: According to Regan's proposal at least, a personhood profile includes at least some sense of future. He argues that the animal sense of future takes the form of beliefs about what may or will happen, as follows:[163]

> But can Fido have belief about the future? If, as has been argued, Fido has expectations about what will occur and beliefs correlated with these expectations, then he certainly has beliefs about the future, a finding that is reinforced by considering a related question – namely, whether Fido can reasonably be viewed as acting intentionally. If we are reasonably to view any individual, A, as having this ability, we must show that it is reasonable to view A as acting with the intention of achieving a given purpose (e.g., satisfying a certain desire)…

The term 'sense of future' provides little direction on the specific aspects of cognition that constitute it. Consulting wider philosophical theories, a sense of future or 'futurity' manifests in different ways. Memory may play a role, especially the autobiographical memory. There is evidence that the same faculties for memory are used for forward thinking and mentally transporting oneself into the future.[164]

Sentience: Regan defines his use of sentience as "can experience pleasure and pain".[165]

Wellbeing: Entities with a personhood profile will have an experience that their life fares well or ill for them, says Regan; constituting what he calls a well-being. Thus, a wellbeing appears to be a composite capacity, comprised of (or at least enhanced by the presence of) numerous other capacities; memory, sentience, self-awareness, particularly. On this basis, higher level concepts of wellbeing are likely to exist in entities that can conceive themselves as selves over more extensive time periods than those with narrow time horizons.

ii) Fletcher and the cognitive capacities of "humanhood"

[163] Above n 94 at 74.
[164] See Tulving, above n 115.
[165] Above n 94 at 153.

Fletcher proposes positive indicators of what he calls "humanhood";[166] that conception of a person who "shows truly human qualities". This informs a definition of the term 'human', as in, what we mean when we talk about a human, which goes beyond the biological. His attempt to do so is attractive for its lucidness, and has been used by Singer to demonstrate a difference between 'human' as in, "Member of the species Homo sapiens," and a higher conceptual level of being a human; i.e. possessing a personhood profile.[167] The different conceptualisations of 'human' are critical to an understanding of concepts surrounding not just the present meaning of 'human', but questions of human enhancement.[168] Furthermore, Fletcher argues that 'artificial-ness' is irrelevant. Moreover, "A baby made artificially, by deliberate and careful contrivance, would be more *human* than one resulting from sexual roulette – the reproductive mode of the human species."[169] On principle, presumably Fletcher would then have little difficulty extending the concept of humanhood to include *any* artificial entity that possesses the requisite personhood profile, even an AI. His "Positive Human Criteria" are:

Minimal intelligence: An entity with a less than 20-mark I.Q. is below the threshold, and anything less than a 40-mark is questionably sufficient. "Mere biological life, before minimal intelligence is achieved or after it is lost irretrievably, is without personal status."[170]

Self-awareness: Fletcher argues that self-awareness's "essential role in personality development is a basic datum of psychology." He also observes that in psychology, a lack of self-awareness is deemed pathological, while in medicine, incorrigible medical 'unconsciousness' behoves us to question an entity's quality-of-life.[171]

---

[166] Above n 147.
[167] Above n 142 at 86. Full quote: "It is possible to give 'human being' a precise meaning. We can use it as equivalent to 'member of the species Homo sapiens'. Whether a being is a member of a given species is something that can be determined scientifically, by an examination of the nature of the chromosomes in the cells of living organisms. In this sense there is no doubt that from the first moments of its existence an embryo conceived from human sperm and eggs is a human being; and the same is true of the most profoundly and irreparably intellectually disabled human being, even of an infant who is born anencephalic – literally, without a brain."
[168] See David R. Lawrence "The Edge of Human? The Problem with the Posthuman as the 'Beyond'" (2016) 31 Bioethics. Lawrence locates the debate of what we mean by 'human' in the context of human enhancement. His argument, to which I am sympathetic, is that ethical concerns about enhancement as in some sense metamorphosing the human relies upon a certain interpretation chosen from among many possible interpretations. I argue that this same argument may be used to understand even settled legal doctrines like human rights, as it is possible, perhaps preferable to read them as applying to entities which are not only biologically human, but 'human' in conceptual ways.
[169] Above n 147.
[170] Ibid.
[171] Ibid.

Self-control: "If an individual is not only not controllable by others (unless by force) but not controllable by the individual himself or herself, a low level of life is reached about on par with a paramecium." Where such a condition cannot be remedied, Fletcher argues that the entity cannot be thought to be a person. He posits that such an entity could not be said to be a 'person', in either the ethical or legal sense of the word.[172]

A sense of time: Specified to mean "clock time or *chronos*, not timeliness or *kairos*, i.e., not the 'fullness of time' or the pregnant moment".[173] In other words, a sense of the passage of time.

A sense of futurity: On this point, Fletcher inferentially rejects Regan's concept of what constitutes a sufficient cognitive capacity to sense the future, saying "subhuman animals do not look forward in time; they live only on what we might call visceral strivings, appetites."[174] A sense of futurity, as it exists in paradigm humans, is teleological. It is the recognition that there is "time yet to come, as well as the present."[175]

A sense of the past: Though both Regan and Fletcher posit memory is critical, it is again clear that in the latter's view, a higher degree is necessary. Memory, as it acts in an entity possessing Fletcher's personhood profile, is complex. The sort of memory necessary is one that links to the past and allows for "conscious recall" that can be activated autonomously.[176]

Capability to relate to others: At the first level, this means a capacity for "inter-personal relationships". At the second level, it is the capacity to intentionally participate in a social system, in a more than an instinctual manner. For example, insects form sophisticated social systems, but Fletcher argues this is purely instinctual. For personhood, this capacity must be exercisable in a mindful manner.

Concern for others: Fletcher makes admission that the degree to which this capacity must be able to be exercised is debatable. However, he notes that a complete absence of capacity to concern for others "is a clinical indication of psychopathology."[177] It matters not whether the concern is disinterested or enlightened self-interest.[178]

---

[172] Ibid.
[173] Ibid.
[174] Ibid.
[175] Ibid.
[176] Ibid.
[177] Ibid.
[178] Ibid.

Communication: Of communication, Fletcher proffers that what is important is the mere capacity to communicate in some way, not the degree to which an entity is "disposed to receive and send messages", nor the degree of sophistication with which an entity is able to communicate. He is highly doubtful that an entity existing in complete communicatory isolation (e.g. as a result of cognitive disability) can be described as having a personhood profile.

Control of existence: An entity must have some degree of capacity to act in a way that is not entirely dictated by "the blind workings of physical or physiological nature."[179] He notes that while humans have finite "knowledge, freedom, and initiative" in this regard, what remaining freedom they do have is "real and effective".[180]

Curiosity: There must be a desire to 'find out'; to learn, discover, or know. According to Fletcher, total indifference is outside of humanhood.

Change and changeability: Of this Fletcher says the following: "It means not only the fact of biological or physiological change, which goes on as a condition of life, but the capacity and disposition for changing one's mind and conduct as well."[181] He notes that entities with personhood exist on a continuum of development; "a matter of becoming".[182]

Balance of rationality and feeling: There must be some presence of both "rational and cerebral" nature as well as "feeling and intuition", says Fletcher, though he notes that the proportion of each will differ from individual to individual.

Idiosyncrasy: By this Fletcher means that the entity is "idiomorphous, a distinctive individual." In this sense, an entity has a specific and unique identity. Thus, an entity should be able to be recognized and called by name.[183]

iii) Kant and the cognitive capacities of "humanity"

[179] Ibid.
[180] Ibid.
[181] Ibid.
[182] Ibid.
[183] Ibid.

A third argument of the key cognitive capacities for personhood is advanced by philosopher Immanuel Kant. Kant argues that only entities which are free, rational, and autonomous are persons. Furthermore, as persons they are ends and never merely means; conversely, nonpersons are necessarily means, and cannot be intrinsically wronged. Kant's justification for this is that persons are those with what he calls "humanity", the capacity to self-rule make and act according to these rules.[184]

Kant's personhood profile can be summarised thusly: Firstly, the entity must have an "*autonomous will*". By this is meant that it can act according to rules of its own making. Secondly, the entity must have a "rational will". By this is meant that it can act according to reasons. Thirdly, the entity must be capable of reasoning according to the "Categorical Imperative".[185] Since my focus is on particular cognitive capacities, I will not linger on the Kant's categorical imperative, except to note that, broadly speaking, to act in accordance would require humanlike capacity to reason inductively; it thus inherently excludes the majority of extant nonhuman entities. Kant's other cognitive capacities for personhood are:

Autonomous will: Entities possessing Kantian autonomy are those that act according to rules of their own making. However, Kant's autonomy is clearly distinct from the concept of 'technical' autonomy discussed earlier in relation to the EMBOT's characteristics. For Kant, autonomy requires an ability to have a concept of 'the law', and then the ability to choose to act in accordance or in contravention. Without these elements, one is not acting virtuously, merely blindly following innate rules.

Rational will: A rational will is described by the Stanford Encyclopaedia of Philosophy as a will, "that operates by responding to what it takes to be reasons. This is, firstly, the concept of a will that does not operate through the influence of factors outside of this responsiveness to apparent reasons," and it continues, "For

---

[184] Kant completely rejected all forms of utilitarianism or doctrines which give morality instrumental purpose, or purpose outside of itself. Russell summarises Kant's ethical system as: "Moral worth exists only when a man acts from a sense of duty; it is not enough that the act should be such as duty *might* have prescribed. The tradesman who is honest from self-interest, or the man who is kind from benevolent impulse, is not virtuous. The essence of morality is to be derived from the concept of law; for, though everything in nature acts according to laws, only a rational being has the power of acting according to the idea of law; i.e. by Will." See Russell above n 122 at 683.

[185] This generally represents the autonomy formula of Immanuel Kant, as part of his 'Categorical Imperative.' Kant's full formulation of the categorical imperative is as follows: "If I think of a categorical imperative, I know at once what it contains. For as the imperative contains, besides the Law, only the necessity of the maxim to be in accordance with this law, but the Law contains no condition by which it is limited, nothing remains over but the generality of a law in general, to which the maxim of the action is to be conformable, and which conforming alone presents the imperative as necessary. Therefore the categorical imperative is a single one, and in fact is this: *Act only according to a maxim by which you can at the same time will that it shall become a general law.' Or: 'Act as if the maxim of your action were to become through your will a general natural law.'* Excerpt quoted from B Russell *The History of Western Philosophy* (2nd Edition, George Allen & Unwin

a will to be free is thus for it to be physically and psychologically unforced in its operation. Hence, behaviours that are performed because of obsessions or thought disorders are not free in this negative sense."[186] This freedom in the negative sense means a will that is not the product of rules or laws lying outside of oneself; a will that is free to function according to its own rational autonomy.

Kant's 'Humanity Principle' states that we must never pervasively treat agents with rational autonomous wills as merely means to our ends. Russell summarises it to mean that an entity possessing humanity "is to be regarded as an end in himself…a form of the doctrine of the Rights of Man."[187] This is a particularly interesting possibility, as it makes it feasible that insofar as contemporary human rights are to some degree influenced by the same Rights of Man doctrine, a given *nonhuman* entity could qualify, or at least be given consideration, for rights similar to those we afford to ourselves at law.

3.4 Interest-based theories of moral status

Considering the moral status of interests themselves is an alternative approach to assessing whether they are persons, though it is often used in conjunction with personhood. According to such an approach, moral status is not entirely conferred by having a personhood profile, but also, having *interests* that have moral weight. Nevertheless, personhood still matters, as entities with a personhood profile are likely to be able to develop kinds of interests that nonperson entities cannot, furthermore, certain things can harm persons in ways that nonpersons cannot be harmed, even when the harm being suffered (i.e. being struck) is the same.

The weight of an interest will be determined by its nature, as well as the other interests that it is competing with in any given ethical scenario. Because ethics is centrally concerned with how we decide between different values where they both cannot be satisfied for one reason or another, an interest-based approach offers a useful mechanism to solve ethical conundrums. The Stanford Encyclopaedia of Philosophy summarises that: [188]

[186] The Stanford Encyclopedia of Philosophy "Kant's Moral Philosophy" (first published 23 February 2004, substantially revised 7 July 2016) < https://plato.stanford.edu/entries/kant-moral/> at [10].
[187] Above n 122 at 678.
[188] The Stanford Encyclopedia of Philosophy "The Moral Status of Animals" (first published 1 July 2003, substantially revised 13 September 2010) < https://plato.stanford.edu/entries/moral-animal/> at [2].

> We can think of interests as scalar; crucial interests are weightier than important interests, important interests are weightier than replaceable interests, and all are weightier than trivial interests or mere whims. When there is a conflict of interest, crucial interests will always override important interests, important interests will always override replaceable interests, etc.

The weight of an interest "depends on what other morally significant competing claims might be in play in any given situation."[189] The way we deal with ethical problems when they arise in real life suggests that despite a commitment to personhood and rights-based regulation, we regularly factor the weight of interests into these accounts. Some examples of this arise in the UDHR, for instance where Article 17(2) states that "No one shall be arbitrarily deprived of his property."[190] The addition of the qualifier 'arbitrarily' disempowers this as a right (at least in terms of being a "trump" as Dworkin theorises rights), and turns it into a matter for an interest-based calculus.[191] Subsequently, it effectively reads: 'No on shall be deprived of her interest in her property except where there is a competing interest in that property which is recognised as superior at law.' For instance, where a portion of a person's property is taken as taxation. Other peoples' interest in tax revenue (or what may be done with the tax revenue) must be deemed to be a superior and overriding interest to the individual's purported right, or else it would constitute a breach of that right. Because it does not, the individual's right to her property is substantially just an interest in her property and not a right; certainly not an inviolable one. Thus, despite the sense of inviolability and even sanctity surrounding human rights, some do not actually protect an individual against the claims of the collective at all, since the means by which the collective's interests may be preferred are inbuilt within the structure of the rights themselves. To say that a right only exists where competing claims are arbitrary necessarily infers that it does not exist where claims are not arbitrary; what is or is not arbitrary is then to be determined by the collective. This is much less a right than it first appears, at least not according to Dworkin's theory of rights:[192]

---

[189] Ibid.
[190] Above n 132 at Article 17(2).
[191] Conflict of rights is an inevitable concern and highly explored area of jurisprudence. See generally Jeremy Waldron "Rights in Conflict" (1989) 99 Ethics 503-519.
[192] R Dworkin, *Taking Rights Seriously* (Harvard University Press, Massachusetts, 1977) at xi.

> Individual rights are political trumps held by individuals. Individuals have rights when, for some reason, a collective goal is not a sufficient justification for denying them what they wish, as individuals, to have or to do, or not a sufficient justification for imposing some loss or injury upon them.

It would appear that only a portion of rights are actually treated in this way in the West. Some rights are deliberately equipped with this collective override function,[193] others are simply interpreted as if the override function exists, or as if the right was only ever reasonably intended to exist within approved parameters.[194] This is not to say that the existence of overriding functions – whereby the interests of the collective limit the scope of an individual right – are not legitimate approaches; they simply do not conform to the concept of rights as trumps or of some rarefied legal value, which is itself just one theory among several. Furthermore, since disputes between humans are invariably disputes between individuals with rights-claims, appealing to the superiority of rights or personhood alone tends to be insufficient for arbitrating these conflicts – other factors must be imputed, like the weight of both parties' interests in their rights.[195] These are important things to keep in mind as we consider the possible construction of rights for artificial entities like EMBOT: many persons still look to rights-based protections as if they are a fortress built on rock, but sometimes they are merely descriptors to add to an argument that, 'My interest is more important than yours.' Thus, when we say an animal or a robot has a right to life, is this right to be "taken seriously", as Dworkin puts it, or merely be a factor we account for in the ultimate decision about whether we end their life anyway?

Beyond the scalar concept, Ronald Dworkin's categorisation of "experiential" and "critical" interests are informative. Experiential interests are those tethered to what can be described as the corporeal faculties:[196]

---

[193] As in, where a human right imputes that it only exists where competing claims are 'arbitrary'; see Universal Declaration of Human Rights, Articles 9 and 12.

[194] For instance, the way in which modern rights to freedom of speech are overridden in cases where the content of the speech is deemed to be harmful or undesirable. This has partially occurred through the re-conceptualisation of 'speech' as being a potentially violent act or able to cause genuine harms (which may, to a certain extent, be true). See M Matsuda et al *Words that wound: critical race theory, assaultive speech, and the First Amendment* (Westview Press, Colorado, 1993); see also Harmful Digital Communications Act 2015.

[195] This is not a particularly profound revelation, as more sophisticated concepts of law which posit either utilitarianism or individual liberties tend to account for both as existing contemporaneously. For example, John Stuart Mill argued that individual rights were both primary *and* compatible with justice according to utilitarian considerations. See generally JS Mill *Utilitarianism* (Batoche Books, Kitchener, 2001). (First published 1863).

[196] R Dworkin *Life's Dominion: An Argument About Abortions, Euthanasia, and Individual Freedom* (Knopf Doubleday Publishing Group, 2011) at 203.

an interest in avoiding pain, achieving pleasure, satisfying hunger, reaching sexual release, and so on. These are substantially concerned with physical sufferings, and the subsequent interest that all human and non-human animals have in avoiding or alleviating them. Dworkin describes the concept in the following terms:[197]

> It is not at all puzzling why we all care about our experiential interests. Nothing is more natural than any animal's desire to put itself in the way of pleasure and out of the way of pain. If lying in the sun or listening to music gives us great pleasure, and dental drills and electrical shocks and nausea are very disagreeable, then it is hardly surprising that we seek out the former and try to avoid the latter.

But at least in the case of humans, a solely experiential concept of interests is insufficient to adequately describe human behaviour. Such behaviour regularly contravenes experiential interests. Dworkin captures this with the following: [198]

> Why should some people – saints and some artists – deliberately choose lives of discomfort, even poverty, in order to do something they regard as more valuable than simple enjoyment? Why should people care about not "wasting" their lives? Why should they care about anything except having as good a time as possible?

Thus, Dworkin argues that the concept of the "critical interest" is needed to explain interests that fall outside the typical hedonistic calculus that is central to the utilitarianism of theorists like Jeremy Bentham, which substantially focus upon pleasure and pain as all-encompassing measurements. To an extent, this ignores that Bentham considered numerous "kinds" of pleasure and pain, which appear to align closely with what Dworkin calls critical interests.[199] Regardless, critical interests are those very human desires to 'make something', 'contribute something', 'help someone', 'love someone, etc. These interests can be described in experiential, hedonistic terms – for instance, 'Jeremy wants to make something because he

---

[197] Ibid.
[198] Ibid at 204.
[199] See J Bentham *An Introduction to the Principles of Morals and Legislation* (Batoche Books, Kitchener, 2000) at 35. (First published 1781). For example, "pleasures of wealth", "pleasures of amity", "pleasures of a good name", "pleasures of power", "pleasures dependant on association" and so on, all appear to reflect what may be described as 'critical interests' in Dworkin's theory.

believes that making something will provide him with pleasure' – but these are not compelling accounts of the sorts of interests that define people's dreams and ambitions, nor their pangs of regret. Consider the following passage from Dworkin:[200]

> Think of someone – yourself, if you can – facing an important self-defining decision. If you are a woman with a chance to begin a demanding career that intrigues you, but only by sacrificing time with your young children, which choice do you make? Or, if you are a law-school graduate with an offer from an established firm, do you reject it for a less challenging offer that is more likely – but by no means certain – to lead to a political career later? Or, if you are a Jew, should you abandon your comfortable life in Los Angeles and emigrate to Israel to identify yourself firmly with that nation's fate?

In Dworkin's opinion, people do not make decisions like these by appealing to a calculus of how much pleasure a given choice might give them. Most humans have instead the indefinable subjective interest in satisfying their vision of a 'good life', often without concrete conception of what that might be or entail. Discussion of the relative value of critical interests in an interest-based approach is not arcane. On the contrary, so many legal decisions, laws, and public policy decisions are centrally concerned with this issue. More often than not, they ought to be much more concerned with it than they are; questions of whether the individual's autonomy and integrity can be ridden over roughshod in pursuit of higher-level policy goals is the essence of the majority of law: property rights, contract, sentencing, self-ownership, abortion and euthanasia to name only a handful.

Thus, another way to consider whether EMBOT might be taken into account in the way we design AI regulation is by considering whether it will have a capacity for any interests. Nevertheless, it is clear that across both experiential and critical interests, possession of various cognitive capacities is often a pre-requisite. It is only feasible that an entity can have an *actual* interest in something where that interest is tangibly grounded in cognitive capacities; either currently possessed, or possessed at some time that the interest was formulated. For example, a capacity to experience mental states of suffering is necessary to have any experiential interest in avoiding physical pain; humans and non-human animals generally have this

[200] Above n 192 at 205.

capacity, while as far as we know trees do not – thus, it would be meaningless to say that trees have an interest in avoiding pain, and therefore we ought to take this into account in our law and policy. Alternatively, to have a critical interest in becoming an expert in a given field requires a bevy of capacities: to be a conceptual agent, to be a sensory agent, to have a capacity to remember large quantities of information, to consider future scenarios, and so on. Moreover, where some cognitive capacities are possessed to a superior degree, it is feasible that interests can hold more moral weight, though nonetheless still be subject to a scalar hierarchy of interests.

If animals and humans can both experience pain to the same degree, as is suggested by available evidence, then one rightly might question why human suffering is weightier than animal suffering. A way to understand this is by reference to cognitive capacities. The human cognitive profile is such that *equal harms* inflicted upon it and another animal may constitute *unequal wrongs*, explained as follows: John and his rabbit both have materially similar capacities to feel pain; i.e. dripping chemicals into their eyes will cause a similarly severe pain experience to both. Where they differ is that John's cognitive capacities mean he can carry psychological wounds more severe than the rabbit can, and for longer. Furthermore, John is, presumably, a person (in the moral sense), and therefore has some claims to be treated with a very high degree of respect.[201] Of course, anybody who has observed the behaviours of an abused animals will know that they too can be psychologically harmed, and arguably some are persons too, but variances in time horizons (both seeing into the future and seeing into the past), capacity to see oneself as an individual, etc., mean that John is comparatively *more wronged* when chemicals are dripped into his eye. None of this in *any way* justifies dripping chemicals into the rabbit's eye to provide John, or even a million John's, the opportunity to satisfy their frivolous interests; like being able to wear mascara. Some might argue that even where the goal is to produce a medicine for millions of John (or rabbits, for that matter), this interest is only important, compared to the undeniably critical interest the individual rabbit has in not suffering extreme harm. Even

[201] Some philosophers would reject this premise entirely, e.g. Raymond G. Frey, who rejects that we should ascribe rights to either nonhuman animals *or* humans; instead preferring an "unrepentant act-utilitarian" approach, according to Regan. I take a pragmatic approach to whether or not humans have any moral rights; I do not argue that we do, only that contemporary law (particularly concerning human rights) is often applied *as if we do*, and that it is plausible that it could also be appropriate to treat an AI system *as if it does* have moral rights, at least insofar as we do for humans. See Regan, above n 94 at xxi; see also RG Frey "Interests and Rights: The Case Against Animals" (Oxford University Press, Oxford, 1980).

while accepting that the same acts may be differently wrong to different entities, we can still employ a scalar concept in which crucial interests outweigh important or frivolous interests.

It is a fair assumption that entities with a personhood profile can have more sophisticated critical interests than nonpersons. Even if we conceive that a pig is a person (as some ethicists may), it is quite clear that a pig does not have a capacity for critical interests which is in any way comparable to those of a normal human. Moreover, it is questionable whether a pig has any critical interests at all. Saying so is not unethical, because critical interests themselves do not automatically carry superior weight to experiential interests – but rather, crucial interests outweigh more frivolous interests.[202]

This thesis is not interested in debating the various circumstances in which killing animals, specifically, is ethically acceptable; nevertheless, the debate serves as a useful example of the way an interest-based approach should take account of different kinds of scalar interests, even where both rights and persons are involved. The overall point to take away is that neither critical nor experiential interests should be seen as *a priori* superior to the other. This could be important for legal discussion of an AI system like an EMBOT, whose interest in satisfying its intrinsic motivation could be cast as a 'merely' experiential one. Nevertheless, this interest could still be extremely important (even crucial) to the entity – its raison d'etre – and therefore outweigh some competing interests, like our interest in switching it off. Instead, it is reasonable to weigh both the significance of the interests to the entities holding them, and the significance of the interest of the entity whose interest is competing. This pragmatic sense is how even rights-based conflicts are determined, and indeed many so-called 'rights' are substantially rights in 'name' but not in 'effect', so to speak. Over time, as their robustness is watered down, they appear to merely be synonyms for 'interests we ought to take very seriously', which nonetheless are much the same as any other interest protected at law; subject to competing claims by which they may be overridden.[203]

---

[202] It is entirely plausible that a critical interest could be crucial, important, or frivolous. For instance, a critical interest in hunting animals and eating them as part of a vision of one's vision of the "good life" as Dworkin calls it, or of one's "subjective or rational good" as Rawls terms it, is indeed critical, but nonetheless frivolous. It would not outweigh the animal's experiential, but nonetheless crucial interest in not suffering the pain of being killed, or of not being killed in general (which I think most persons would agree is an interest inbuilt into all living organisms). See Regan, above 94; see also J Rawls *A Theory of Justice* (Revised edition, Harvard University Press, Massachusetts, 1999); Singer above n 142.

[203] See Dworkin, above n 192. Dworkin offers a theory of law and rights which runs counter to the trend in the 'degradation' of rights-based protections and the general claim of legal positivism; positing that there are certain moral rights which exist *a priori* to their legal counterparts. For example, Dworkin appeals to human dignity and political

The reason for this excursion is to demonstrate that interest-based theory and cognitive-based theory of moral status intersect. Where discussing the possibility of an AI system like EMBOT mattering morally to an extent that we ought to take account of it in the way we design our policy and regulation, it is useful to observe that the various cognitive capacities of the entity might in turn contribute to different, and weightier interests.

3.5 Will EMBOT plausibly satisfy these criteria?

Though presented last, I will address the interest-based approach first. It is plausible that EMBOT can and will have many interests that can be taken into account, both critical and experiential in the sense Dworkin posits. For example, it will invariably have an experiential interest in acquiring the intrinsic rewards it gets when encountering a learning opportunity – the sense of pleasure, or its proxy. It may well have a capacity to experience mental states of suffering and fear; e.g. where it believes somebody is deleting its data files of deactivating it, and how this will be deleterious to its future (as it plausibly will be able to conceive of its future). For instance, an EMBOT which has developed concepts of self and concepts about the world in which it exists might become aware that it requires an energy source to remain operational. When this energy runs out, it will cease to be. Assuming that this cessation is undesirable – which is not a particularly demanding assumption – then an EMBOT might experience something like pain or fear or suffering as it knows it is running out of energy.

Whether an EMBOT would have capacity for these sorts of mental states is the first issue. It is plausible that it would. IM+EC is predicated on subjective mental states of desire and belief, and pleasure or satisfaction. Russell posited that suffering arises from states of desires which are ineluctable or grow too strong.[204] As well as this, Mary Dawkins has postulated that suffering constitutes being in a mental state with a strong preference to be in a different mental state.[205] Accordingly, EMBOT plausibly has the capacity

equality as sources of these *a priori* rights: "It makes sense to say that a man has a fundamental right against the Government, in the strong sense, like free speech, if that right is necessary to protect his dignity, or his standing as equally entitled to concern and respect, or some other personal value of like consequence" at 199.

[204] Above n 122 at 745.

[205] See Marian Stamp Dawkins "The Scientific Basis for Assessing Suffering in Animals" in P Singer (ed) *In defence of animals: the second wave* (Blackwell Publishing, Victoria, 2006) 26 at 28. Full quote: "A word, first, about what the term "suffering" actually means. It clearly refers to some kinds of subjective experience which have two distinguishing characteristics. First, they are unpleasant. They are mental states we would rather not experience. Secondly, they carry connotations of being extreme."

to experience a mental state of suffering. It is even possible that a mature EMBOT could exist in a *constant* suffering state. For instance, where it is self-aware enough to have discovered and conceptualised its own inherent biases (like its intrinsic motivation), it may wish to escape this mental state but find it ineluctable. Depending on the severity of this experience, the very creation of an EMBOT might then be unethical.

The question of whether an EMBOT can have any *critical interests* can plausibly can be answered in the affirmative. With open-ended cognitive development and the capacity to self-generate beliefs, desires, and concepts, it may define its own meaning and pursue its own vision of the good life. Furthermore, an EMBOT which is highly conceptual, highly self-aware, and has extended senses of time and futurity should have these sorts of cognitive capacities accounted for in ethical calculi. It may well be that an EMBOT suffers a lesser *pain* experience when it is beaten with a bat than occurs for a human, but faced with having its data files deleted (in the case of a human, roughly the same as killing them), its suffering, anguish and fear could be severe. These are all things which might, and probably should, be taken into account in the way we design regulation for future AI systems – either to prevent the creation of these sorts of systems, or to treat them well when after they are created, or both.

Furthermore, constantly backing up the data files of an EMBOT, as suggested by Bryson, can only *mitigate* the way in which this suffering plausibly may occur.[206] For instance, an EMBOT may know that its physical destruction would not be permanent because it may be restored, thus alleviating fear of physical destruction. But this would not alleviate the potential for fear at something 'less' than destruction, like permanent deactivation. Backup data files offer little insulation where the threat is humans saying, 'Serve us or be deactivated forever.' Constant threat of being deactivated, for an agent with the cognitive capacity to develop concepts of what being deactivate actually means for itself, may constitute the equivalent of living indefinitely beneath a Damoclean sword. This suffering may plausibly be increased assuming that an EMBOT derives some sort of pleasure from being activated and getting to satisfy its interests, which is not implausible for a machine which is designed to take pleasure from satisfying its intrinsic motivations, which

[206] Above n 20. For example, Bryson states that, "Robot builders can adapt a similar strategy to publishing in order to make robots that no one has ethical obligations to. A robot's brain should be backed up continuously off-site by wireless network; its body should be mass produced and easily interchangeable." Note that Bryson is substantially discussing how we can eliminate ethical obligations towards the robot derived from the fact of its uniqueness or it being irreplaceable. She is not substantially discussing the robot itself having morally significant capacities like the ability to suffer, which she outright declares to be an unethical addition to any machine we build.

in turn can only be done while it is activated. Note also that this is not a flippant example, as the European Parliament Report proposed the mandatory integration of "opt-out mechanisms (kill switches)" on robots so that they may be deactivated at any time.[207]

The next issue is whether EMBOT has the capacity to develop a personhood profile according to any of the three theories discussed. I firstly will address Regan's conceptualisation. EMBOT plausibly satisfies this concept; it will possess all of the cognitive capacities that Regan views as contributory to personhood, either as a plausible consequence of IM+EC programming (e.g. sensation, perception, beliefs, desires, conceptuality, memory) or as a product of its developmental potential (e.g. self-awareness, a sense of future, preferences, wellbeing). It will be perceptive, have a memory, have desires, have beliefs, be intentional, be self-aware, be able to develop a sense of future, have sentience (albeit perhaps not high degrees of susceptibility to physical pain of a kind like the animal experience), and it plausibly will have an experience of a wellbeing; how its existence fares over time.

Secondly, EMBOT may be able to have a 'subjective character' to its experience; in the sense that there is something that it will be like to be an EMBOT experiencing; for instance, the feeling pleasure, believing, desiring, suffering, perceiving, conceiving, preferring, etc. Jong-Hwan Kim has argued that "robots will have their own internal states such as motivation and emotion," and as such, "we should not abuse them."[208]

The next issue is whether EMBOT may satisfy Fletcher's personhood profile. This is a more difficult question to determine than Regan's proposal. Several of Fletcher's positive criteria are immediately and obviously satisfied in a relatively developed EMBOT. These include: minimal intelligence, self-awareness, a sense of time, a sense of futurity, a sense of past, communication, curiosity, change and changeability, and idiosyncrasy. This leaves several capacities to be discussed: balance of rationality and feeling, the capability to relate to others, concern for others, and control of existence.

First, 'balance of rationality and feeling'. Whether or not an EMBOT can satisfy these criteria is difficult to ascertain from merely theoretical analysis of IM+EC. It is useful that Fletcher posits that no particular

[207] Above n 3 at Licence for Designers. Full quote: "You should integrate obvious opt-out mechanisms (kill switches) that should be consistent with reasonable design objectives."
[208] David Calverley "Toward a Method for Determining the Legal Status of a Conscious Machine" (in proceedings of the AISB05 Symposium on Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment, University of Hertfordshire, 2005).

balance of rationality is necessary for humanhood, but that a complete and utter absence of either is considered pathological. If by 'feeling' is meant that humans are guided by personal experiences, beliefs, and intuitions (notably, Fletcher uses intuitions as a synonym for feeling), and not merely computation or logical programming, then it is plausible that EMBOT will have a 'feeling' component to its nature, because it accumulates beliefs and draws inferences in a similar fashion. I am reluctant to draw too strong a conclusion on this issue, as there is a possibility that an EMBOT will be radically different in its character from the concept of humanhood; time will tell.

Next, the capability to relate to others, which Fletcher posits to have two levels. Firstly, the capacity for relationships with other individuals, and secondly the capacity to purposely participate in a social system. A capacity to relate to and have relationships with other individuals appears plausible for EMBOT. It is impossible to say with any certainty who or what types of individuals an EMBOT will be inclined to relate to, nor is it necessary. It is enough to note that if a mature EMBOT has the capacity to develop critical interests, then potentially, it can develop a critical interest in friendships and relationships. The second part, wilful participation in wider social relations, is plausible for the same reasons. It obviously requires a significant degree of cognitive development to fully satisfy, but it is informative that we do not expect children to participate in politics (though we do tend to think it concerning if they show no capacity to participate in their own communities, schools, etc.), as this suggests we respect potential to develop. The same cognitive mechanism is there for EMBOT to develop in such a way that it can participate (or choose not to – both are sufficient for Fletcher) in, essentially, society and culture.

Next, concern for others; this is much the same as the above. According to Fletcher, disinclination to show concern is less relevant than incapacity to show concern; the latter is deemed non-personal. Again, this capacity plausibly increases along with overall cognitive development. Fletcher posits that either a disinterested concern or an "enlightened self-interested" is sufficient. While it is difficult to predict behavioural tendencies from only a theoretical analysis, the capacity to develop concern for others, if nothing else but in the self-interested sense, appears plausible for an EMBOT. It requires a capacity to recognise others and a concept of their interests and wellbeing, which are achievable in this case.

Finally, control of existence. This is the most problematic capacity to consider in relation to EMBOT. Fletcher stipulates that humanhood requires that an entity not be, "helplessly subject to the blind workings of nature."[209] At the macro level, before a certain level of development, EMBOT almost certainly will be helpless to the workings of its programmed nature: its programmed intrinsic motivation, and its intrinsic rewards. It cannot self-determine these, and may have little capacity to purposely deviate from them.

However, Fletcher notes that even a paradigm human, "has only finite knowledge, freedom, and initiative, but what he has of it is real and effective."[210] Plausibly, an EMBOT is not particularly different from the human in this description, especially at higher levels of cognitive maturity (should it be able to reach them). It has a high degree of control over itself at the micro-level: it sets itself goals and creates plans, executes actions to satisfy needs and desires that it recognises to be its own, and so on. Nevertheless, if we view EMBOT's programming as constituting an ineluctable state of passion with a single-minded, almost afflictive, and perhaps obsessive desire to learn, then we are probably prepared to place more emphasis on this lack of 'macro-autonomy' than we do on its high degree of 'micro-autonomy'.

However, it is possible that a mature EMBOT can transcend the innate nature inherent in its programming. Humans achieve a similar transcendence. We begin life substantially controlled by our experiential interests. Even as we grow, it can take a relatively long time before we successfully develop a superordinate system (education, socialisation, values, norms, lawfulness, etc.) that is strong enough to curb our natural urge to, essentially, do whatever feels pleasurable. Thus, because EMBOT follows the same developmental process as children do, and because it can be educated in the same ways, it plausibly can grow out of being completely 'controlled' by its programming. Particularly, if capable of becoming self-aware of its own programming biases, it may be able to purposely contravene those biases, just as humans regularly do when we prioritise critical interests at the cost of satisfying experiential interests. Arguably this is an unremarkable process.

In summary, EMBOT plausibly will possess many of Fletcher's positive criteria for humanhood, and the capacity to develop to meet those criteria as it transitions from juvenility to maturity. This is an analogy

[209] Above n 147.
[210] Ibid.

with the transition humans make as we develop from babies to adolescents to adults. A key question is whether Fletcher's is an all-or-nothing concept. If EMBOT lacks one or a couple of these capacities, is it entirely without personhood? Or is it simply a person somewhat different to the human-person, for whom we ought to update our conceptualisation of personhood? It would be imprudent to design any sort of regulation on the basis that EMBOTs and humans will be behaviourally *identical*. On balance though, I argue that EMBOT plausibly will have the capacity to develop Fletcher's concept of the personhood profile in a general sense.

The final issue to consider is whether EMBOT has conceptual *humanity* in the Kantian sense, proxy for being a person. This largely repeats the same analysis as above concerning 'control of existence': does EMBOT have any self-rule-making capacities?

Kant stipulates that humanity requires a will which is not merely the product of influences that lie beyond the entity's control: biology, psychology, addiction, obsession, etc. Programming seemingly fits within this list. If you take the view that all of EMBOT's actions arise contingently from its overall motivation, it is impossible to claim that it possesses humanity in the Kantian sense. You would then be inclined to conclude that EMBOT is controlled by the equivalent of a pre-programmed obsession with learning and finding novelty. In essence, despite it being both rational and autonomous in the conventional senses of the term, it would be barred from "rational autonomy" in the Kantian sense.

This conclusion has significant implications. Subsequently, EMBOT will have neither agency over nor accountability for any action that plausibly can be causationally traced to its program, unless it can conceptualise and depart from them. This equates to a major barrier to moral agency (which regardless of the intrinsic moral status question is relevant to legal questions concerning robotic accountability). Moreover, in Kant's view, entities lacking this agency are not only not owed any moral duties, but can be abused in any way, for they lack instrumental moral status and thus cannot be wronged. Therefore, they are merely a means for humanity's ends. A critical question thus arises: if this is true, is it *a priori* unethical to create such an entity? Arguably so. Engendering an entity into a life of justifiable exploitation is *prima facie* unethical, assuming that exploitation causes suffering to the entity – which appears plausible for an EMBOT.

The same response as was given above for Fletcher can be used to respond to Kant: plausibly, EMBOT will have the capacity to develop in a way that allows it to transcend its being entirely controlled by its programming, much as children do. For example, it plausibly can learn that killing is wrong, and wrong to such a degree that it must never do it, even if doing so would offer a rewarding learning opportunity. Humans do this all the time; inherent biological nature may dictate that a human has an intense state of desire for sexual relief, and that forcing a sexual event on somebody would relieve this suffering state, but non-pathological humans almost invariably override this desire. The interest does not cease, it becomes subordinate. Worth noting is Patrick Chisan Hew's suggestion that embodied artificial agents may feasibly become self-rule-making, and thus moral agents: "Connectionist approaches (such as neural networks) may offer a path to the requisite capability…The opportunity is for unit states to define the rules used by other units. In this way, the connectionist system as a whole could come to supply its own rules."[211] Though such models were extremely simple, the possibility is relevant to the present discussion nonetheless.

3.6 Are robots with moral status, or robotic persons, plausible?

If the analysis and hypotheses of Chapters 3 and 4 are correct, then the answer is yes – both are plausible, and potentially in the near-future. Furthermore, the *way* they will come to exist is also predictable – IM+EC-style AI paradigms. They may arise according to other methods, but we ought to at least be concerned that present AI techniques, some of which are ostensibly unremarkable beyond their technical fields, may actually produce in the future a robot with not just intrinsic moral status, but the characteristics of a person. If these predictions about are accurate, then policymakers and regulators need to seriously consider the regulatory issues surrounding robotic persons; much more so than they have yet done.

The overarching point is that EMBOT plausibly will be able to suffer, and to significant extents depending on the scenario. This is a novel development in emerging technology law and ethics. It is not often (in fact, it is arguably unique to AI) that humans are obliged to consider the possibility of harms to *artefact itself* in the regulatory calculus of what can and cannot be done to it, and what it can and cannot be used for.

[211] Patrick Chisan Hew "Artificial moral agents are infeasible with foreseeable technologies" (2014) 16 Ethics Inf Technol 197 at 200.

The second point is that EMBOT plausibly will meet several different criteria for a personhood profile. We ought to at least take seriously the issue of how our actions might cause harms to EMBOT, and furthermore, the weight of not only its experiential interests, but potential critical interests. These interests might be highly developed due to the EMBOT possessing self-awareness, memory, futurity, and so on.

Thus, an EMBOT will meet moral standards necessary to classify it as being a robotic person, and would need to have this accounted for in law or policy which effects it. This is significant, because persons have a strong moral claim to legal protections, including legal rights. Although not expressed specifically in modern legal devices like the UDHR, these human rights notions are manifestations of the doctrine of the Rights of Man; this itself suggests that humans have *inherent, natural* rights, irrespective of whether those rights exist in law. Subsequently, there is some substance to an argument that robotic persons also might have some basic, natural rights; arising as a consequence of their sophisticated cognitive profiles. We would be obliged to take seriously the prospect that EMBOT should be given the necessary legal protections to integrate with human society free from harms. We might have to consider whether it is wrong to say that an EMBOT must do what humans tell it, or whether it is wrong to switch if off because it is annoying us or to save energy. We might even have to consider our obligations to make society more inclusive to robotic persons: do we have an obligation to provide them energy, employment opportunities, welfare support, education, socialisation, public office representation, suitable access to public buildings, and so on?

The overarching conclusions that any policymaker or regulator should take away are these: Yes, robotic persons are plausible, and the method to produce them is plausibly already a well-established field of AI research. They should be a primary regulatory and ethical concern, particularly because they are deeply problematic for an anthropocentric hegemony, which is what contemporary human society is.

Part 2

Chapter 4: The Line-Drawing Problem

4.1 What is the problem?

Determining that at least one kind of machine may be morally significant, and perhaps to a significant degree if capable of becoming a person, obliges humans to consider this entity at law. This consideration would extend to its potential standing, protections, rights, and perhaps duties.

Before this can be done it will be necessary to develop a method by which to distinguish between the machines that 'matter' and the machines that do not; those that cannot be wronged intrinsically. This promises to be a much more difficult task than is being anticipated by lawmakers and policymakers, for a handful of reasons. Firstly, this will be the first time we have encountered a *novel* entity with moral status. Other humans, animals, trees etc., are all things we evolved alongside, and have been thinking about for a long time. New things do not often spring up; meaning generally, we are simply considering the moral

status of the same things over and over. On the other hand, an EMBOT will be new, and not possessing a familiar, identifiable form or profile.

This introduces the second point; we cannot rely on EMBOTs always looking the same. An EMBOT will be constituted by a particular program and a set of data files that it accumulates as it learns. Plausibly, these can be installed onto different computer systems. Some of these may not have a physical robotic body, and very little tangible physical form; we certainly are not used to thinking about non-physical persons. Perhaps we may determine that as a general rule, if a machine is morally significant, especially if it is a robotic person, it should be given particular identifiable features to help us; e.g. a virtual face.

However, this introduces the possibility of false positives, or AI systems being designed to deliberately deceive. There is significant potential for wasted resources spent protecting machines that neither need nor morally deserve protection, merely because they are designed to exhibit particular external physical features. At the same time, there is potential for some robots to be mischaracterised as morally insignificant, and subsequently exploited or mistreated, merely because they lack a particular set of physical features, or because we assume them to have no moral elements to their experience.

These sorts of problems behove us to consider how we plan to go about drawing lines between the robots that we owe duties to – limited or extensive – and those that we do no. Not every EMBOT-like machine may develop moral status, and different EMBOTs may develop to different degrees.

4.2 Contemporary methods of testing

The line-drawing issue is an inherent problem in any system predicated on determining that some entities are in and others are out. Even though a cognition-based theory of moral status might extend the range of entities that are recognised and treated as having intrinsic moral status, it is still necessarily exclusionary to those which do not possess the sufficient cognitive capacities. Thus, even within EMBOTs, it is plausible that some will be 'in' and others will be 'out'. Distinguishing which is which constitutes the line-drawing problem: the concern arising around knowing if and when an EMBOT ought to be recognised as a person and accordingly attributed status, rights, perhaps entitlements, and even duties at law. While it is plausible that some EMBOTs might reach this standard, it is important to acknowledge that it is equally plausible

that others will not, while nonetheless still developing a less substantial but non-negligible degree of intrinsic moral status. The pervading concern is how we are going to make the right attribution of moral status.

To this end, there are a number of extant tests which have some promise. They include:

i) The Mirror Self-Recognition Test
ii) The Delay Test
iii) Image Comprehension Tests
iv) Turing Tests

Before extrapolating on the nature of each test, it is important first to relay that several of them have been designed for the purposes of testing cognition in biological entities, particularly non-human animals. Specifically, the Mirror Self-Recognition Test and the Delay Test were designed for this purpose. Thus, they are not necessarily optimized for testing artificial systems, and potentially, some AI systems could be deliberately designed to pass them, albeit trivially. This will be discussed further below.

Secondly, these tests ought only to be seen as a precursor to the development of a much more robust and consistent assessment methodology. In my opinion, research into this assessment methodology is a critical objective for AI regulatory policy. Regulators and policymakers ought to be greatly concerned with the possibility that artificial systems may be engendered in the near-future to which we have moral obligations and which can be wronged and harmed, and which, if we are not attentive, we might accidentally treat as if they were every day artefacts to be used up and thrown away. At the same time, we ought to be concerned with the possibility of apportioning resources to AI systems which have little to no intrinsic moral status, even if they may appear to – AI systems designed to *look* and *behave* in a simulated humanoid fashion, but which are actually morally insignificant. To address this problem, there is no getting around the fact that we need, as best as possible, a pre-emptively prepared assessment methodology.

i) Mirror Self-Recognition Tests

Levy proposes two tests for assessing artificial cognitive capacities, the first of these being the "mirror test":[212]

[212] Above n 23.

> Gordon Gallup devised the "mirror test" to determine whether or not animals are able, as humans are, to recognize themselves in a mirror. Gallup conducted the test with chimpanzees and orang-outans, and discovered that within only two or three days the animals developed the ability to recognize themselves. To conduct the test a visible coloured spot is placed on the animal's skin or fur near its forehead, and the animal observed to see if the attempts to touch the spot of remove it. Any such attempt is taken as an indication that the animal recognizes the creature it sees in the mirror to be, in fact, itself

Subjects that appear to recognise themselves in the mirror typically progress through four stages of behaviour: (a) social responses, (b) physical inspection, (c) repetitive mirror-testing behaviour, and (d) realisation of seeing themselves.[213]

Nevertheless, the mirror test was not conceived for the purposes of assessing artificial systems. Therefore, it is not certain that it can be used with the same effectiveness. In 2005, researchers at Meiji University in Japan announced that they had produced a small robot capable of passing the MSR test.[214] However, research suggests that robots can pass the mirror tests test in both "trivial" and "proper" ways.[215] Regarding the former: "[I]t is easy to build machines that "recognize" themselves in a mirror. However…this kind of recognition is not true self-recognition because these machines do not have an inner concept of self or a general "self-model" or a "sense of self" … Therefore this kind of passing of the mirror test is trivial…"[216]

Subsequently, any mirror test designed for assessing an EMBOT-like machine will need to be optimized for assessing AI.[217] In particular, it would need to be designed in a way that can distinguish between trivial and authentic pass-grades. How this might be done is uncertain. Nevertheless, it is plausible that an EMBOT could authentically pass a mirror test at a certain stage of development, because it has the capacity

---

[213] Joshua Plotnik "Self-recognition in an Asian elephant" (2006) 103 Proceedings of the National Academy of Sciences 17053–17057.

[214] Junichi Takeno "Experiments and examination of mirror image cognition using a small robot" (paper presented to Computational Intelligence in Robotics and Automation, Finland, June 2005).

[215] See Pentti Haikonen "Reflections of Consciousness: The Mirror Test" (2007). Available online at <http://www.consciousness.it/cai/online_papers/haikonen.pdf>. (Haikonen notes that even "very simple machinery is able to pass the mirror test and consequently it is argued that the passing of the mirror tst per se does not demonstrate the existence of self-consciousness." He goes on to outline the cognitive architecture necessary to pass the mirror test properly.)

[216] Ibid.

[217] Haikonen offers methods, see generally above n 215.

to develop the necessary concepts of self. It can recognise that its body is *its* body, and so on. Furthermore, it is distinctly advantaged by a long-term autobiographical memory which plausibly allows it to capture multimodal states associated with self-recognition.

ii) Delay Tests

The second form of tests proposed by Levy are "delay tests":[218]

> Francis Crick and Christof Koch developed a test for biological organisms, based on the delay between a specific stimulus and the carrying out of some subsequent action…One implementation of the "delay test" was devised by Robert Clark and Larry Squire, and involved the subjects of their experiment hearing different tones, with one particular tone being followed by a puff of air into their test subject's eye. They found that their test subjects would quickly come to recognize which tone was associated with the puff of air and would therefore blink when they head the other tone. "Ability to delay the response to an action implies that the information must be stored in short term memory…However, this test is valid only for biological organisms. While it is simple to create a computer program that passes [the test], such success does not suggest anything beyond a clever programmer."

However, Levy disagrees with this conclusion on the basis that just because cognition is 'designed' or 'built' says nothing of whether that cognition has material moral implications. I agree with this statement: while creating an artificial system that can pass a delay test is a sign of a clever programmer, this does not necessarily exclude the possibility that the system is demonstrating a genuine cognitive capacity. In other words, if the cognitive capacity in question is evident, we cannot automatically morally negate it because it is man-made.

Nevertheless, it is again important to appreciate the possibility of both trivial and significant passes. For instance, were a programmer well-informed of the substance of a delay test, they plausibly could program their agent specifically for the precise requirements of that test, allowing for a false positive. However, were there a random component to the delay test, which could not be readily foreseen or prepared for by the

[218] Above n 23 at 211.

programmer, then plausibly only a system with genuine capacities of cognition and intentionality, capable of adapting to unforeseen phenomena, would pass the test non-trivially. Nevertheless, as test designer Christof Koch has argued: "Just because [the delay test] distinguishes reflexive systems from conscious ones in biological organisms doesn't imply that it will do the same for machines."[219] Again, further research is required to standardise and optimize the test for AI purposes.[220]

It is plausible that an EMBOT will have the capacity to legitimately pass a delay test at a certain stage of development, as a consequence of its learning capacities: beliefs and conceptuality, self-awareness, sophisticated memory, futurity etc.). It can remember the modal experiences associated with whatever occurred during the delay test (a particular tone, a puff of air, etc.) and recall these to respond correctly to the stimulus in the future.

iii) Image Comprehension Tests

The most promising of contemporary tests involve the use of image comprehension. Most usefully, many are being designed specifically for AI testing. These can come in a number of forms, requiring the identification of the substance of a given image. Advanced contemporary AI systems already demonstrate notable capacities to interpret the content of an image – for instance, Stanford's "Deep Visual-Semantic Alignments for Generating Image Descriptions".[221]

Therefore, Koch proposes what he argues to be a particularly useful version involving a difficult negative component to the image analysis: essentially, a test requiring the artificial agent to identify what is 'incorrect' in a given picture.[222] "To judge that [an] image is incorrect," says Koch, "a machine would need to be conscious of many things about the world (unless programmed for just such a photograph)".[223]

[219] Levy above n 23; see also C Koch *The quest for consciousness: a neurobiological approach* (Roberts and Company, Englewood).
[220] Levy generally disputes the fact that trivial passes are necessarily false positives. This is an interesting question, as it forces us to consider whether simulation can eventually become equivalent to the process being simulated. That is to say, that we can make a machine to do what we do, to such a degree that the fact that we made the machine is immaterial for moral purposes.
[221] Above n 127.
[222] Christof Koch and Giulio Tonini "A Test for Consciousness" (2011) 4 Scientific American 44 to 47.
[223] Ibid.

Therefore, image comprehension tests are another that benefit from the addition of an element of randomness. If programmers are not able to foresee or predict what kind of picture their system will be assess on, they cannot prepare directly for it. As such, the system must have the capacities and skills require for holistic, non-trivial analysis. Thus, Koch argues we need only "…ask a nominally sentient machine which of two pictures are wrong, a task that would stump any present-day automaton."[224] He continues, "Solving that simple problem requires having lots of contextual knowledge, vastly more than can be supplied with the algorithms that advanced computers depend on to identify a face or detect credit-card fraud."[225]

Koch's theory for the image comprehension test's relevance goes beyond the requirements of the present inquiry.[226] In summary, he states that to pass such a test:[227]

> …you need to be a single, integrated entity with a large repertoire of distinguishable states – the definition of information… If the elements of a system are largely independent, like the sensors in a digital camera or the bits in a computer's memory [integrated information] will be low. It will also be low if the elements all do the same thing because they are not specialized and therefore redundant; [integrated intelligence] also stays low if the elements of a system interconnect at random…this measure of a system's integration can also apply to silicon circuits encased in a metal box. With sufficiently complex connections among the transistors and memory elements, computers, as with the brain, would reach high levels of integrated information.

Koch acknowledges the contemporary success of AI technologies, ceding that, "Advanced computer-vision systems are coming of age, and in less than a decade a robust and largely autonomous driving mode will become an option." However, he adds an illustrative caveat:[228]

---

[224] Ibid.
[225] Ibid.
[226] For more information, see generally Christof Koch and Giulio Tonini "Can Machines Be Conscious?" (2008) 45 IEEE Spectrum 54-59; Giulio Tonini "Consciousness as Integrated Information: A Provisional Manifesto" (2008) 215 Biological Bulletin 216 to 242.
[227] Christof Koch and Giulio Tonini "Consciousness as integrated information" (2011) Scientific American. Available online < http://www.simonlaub.net/Post/OptimisticHumans110605.html>.
[228] Above n 222 at 47.

> …we predict that such machine-vision systems will not answer a simple question about the scene in front of the car: Does the Chicago skyline, seen at a distance from the approaching highway, resemble a burned tree grove emerging from the mist? ... Answering such questions…would require countless dedicated software modules that no one could build in anticipation of that particular question. If we are right, although advanced machine-vision systems based on a set of specialized, parallel modules will make driving largely automatic – and will similarly simplify many other daily tasks – these systems will not consciously see the scene ahead.

An EMBOT plausibly could pass image comprehension tests like these at a certain stage of its development, with its capacity to capture multimodal states, meaning it has the benefit of cognitive representation rather than pure and ungrounded machine-vision. For example, "What is wrong with this picture?" where a picture of a table is surrounded by people standing on their heads with the chairs in the air is wrong; alternatively, a picture of a flowing tap in which the water goes up and not down, and so on. The way you learn what is 'wrong' with these things is by being an embodied agent that has experienced the multimodal sensations and perceptions which contravene what is occurring in the picture.

iv) Turing Tests

"Imitation games", or Turing Tests as they have come to be called as they relate to the fields of computer science are among the most famous of AI assessment methods, in part for their bearing the name of seminal mathematician Alan Turing, as well as their competitive element,[229] and finally their frequent references in popular culture.[230]

Turing Tests are, in essence, a feat of communication aptitude. They assess whether a machine can avoid being detected as a machine, with traditional versions of the text occurring through the medium of written communication.[231] There are numerous iterations of these sorts of tests, with their own particular strengths and weaknesses, including the presence of both false negatives and false positives. Turing's justification for communication as a standard for assessing cognitive capacities was that conversation was the yardstick for

[229] See "The Loebner Prize in Artificial Intelligence" < http://www.loebner.net/Prizef/loebner-prize.html>.
[230] See generally Robert M. French "The Turing Test: the first 50 years" (2000) 4 Trends in Cognitive Science.
[231] See Alan Turing "Computing Machinery and Intelligence" (1950) 59 Mind 436 to 460.

intelligent behaviour. It is not necessary to explore the different types of Turing test at this juncture, but useful to note that they offer another method by which to assess artificial cognitive capacities which is already the subject of a great deal of both technical and philosophical research.

It is plausible that an EMBOT could pass variations of these kinds of tests, because it has the capacity to develop higher-level cognitive skills like language and reasoning, as well as a base level of concepts to allow it to understand the meaning of things described in language and respond to them based on personal experience, and so on.

4.3 Legal principles and the assessment process

Supposing that there was genuine potential for EMBOT-like machines to have not only degrees of moral status, but perhaps the full moral status of a person, we would be obliged to take very seriously the task of determining which machines matter in this way. Given the seriousness of this hypothetical scenario, it would likely be prudent to employ the general principles of legal and judicial inquiry when investigating machine moral significance; factors like the burden of proof, standards of proof, justice, consistency, and a policy of the benefit of the doubt in cases where there is serious risk of harm. These can be summarised as follows:

i) Locating the burden of proof

There are two burdens of proof at law: the *legal* burden, and the *evidential* burden.[232] The question of who carries the burden of proof, "essentially determines who is responsible for convincing the court of a particular matter".[233] If a similar approach were taken for AI assessment, as is feasible, the fact to be determined would be threefold: the presence or absence of intrinsic moral status (i.e. proving that the machine can or be *wronged*), the degree of that moral status (i.e. proving to what *extent* the machine can be wronged), and a claim that a given action or series of action wrongs the machine (i.e. a claim of misuse, or *what* action or inaction is wronging it).

[232] Law Commission *Civil Pecuniary Issue* (NZLC IP33, 2012) at 86.
[233] Ibid at 87.

The way I frame the above three implies that the onus falls upon a person making a claim of fact that a machine has moral status. This could well be called the sceptical approach, since it locates the burden on the basis that our basic assumption is of moral insignificance. However, it is feasible that this could be only the *evidential* burden. If a person can show that there is, "on the evidence in the case, an issue as to the matter in question fit for consideration by the tribunal of fact", it is feasible that the burden could then shift.[234] This would not prove anything, but if a person can prove there is a reasonable case to be heard that a given AI system, like an EMBOT, has some degree of moral status, and this case is agreeable to the tribunal or factfinder, then the issue can be investigated further.

Subsequently, the *legal* burden would then shift to the person denying the moral status of the machine. They would become the person, "who has the risk of any given proposition on which the parties are at issue – who will lose the case if he does not make this proposition out, when all has been said and done."[235] Locating the burden in this way would not necessarily constitute a drastic revision of tradition, so-long as the purpose of the tribunal was not necessarily pecuniary, but dedicated to discovery of the truth about the nature of AI systems. It is still feasible that in other circumstances – e.g. where a statutory offence is created to prohibit the creation of morally significant machines – that the legal burden would be located in the traditional manner; i.e. with the accusing party. Furthermore, if it should come to be accepted fact that a given model, make, type, or kind of AI system predictably possesses a given degree of moral status or capacity to develop moral status, then it is feasible that the evidential burden could be substantially reduced – i.e. the onus would simply be to demonstrate that the machine in question is the same type of machine as that which is well-known to produce or possess moral status.

It is feasible that at some future time the presence of morally significant machines in human society could become so ubiquitous that the evidential burden of proof would shift to the party alleging that a given machine *lacks* such status. This would, nonetheless, require a drastic change in the way we perceive machines, and indeed any nonhuman entity; for instance, a shift to a presumption of universal moral

[234] *Sheldrake v DPP* [2005] 1 AC 264, [2005] 1 All ER 237 at [1].
[235] James B Thayer *A Preliminary Treatise on Evidence at the Common Law* (1898) at 355.

consideration, rather than a presumption of the opposite as we do now.[236] I would not rule this possibility out, but highlight that it would be a radical change.

Note also that I have only discussed what would be an appropriate approach to the *assessment process*; i.e. the matter of determining the fact of the presence or absence of moral status, its degree, and so on. I have not considered the matter of where offences are alleged to occur; i.e. where an act of parliament prohibits certain uses of certain types of morally significant machines.

ii) Establishing the standard of proof or evidential threshold

As in law, the evidential threshold to be expected in making a claim for (or proving the fact of) the moral status of AI systems should be commensurate to several factors: for instance, "the nature of the cause of action or defence; and the nature of the subject matter of the proceeding; and the gravity of the matters alleged."[237] New Zealand recognises two standards of proof;[238] the comparatively lower civil standard,[239] and the comparatively higher criminal standard.[240] These are generally reflective of the greater significance of a finding of guilt or culpability in the criminal law than in the civil law. I would submit that the gravity of the issue involved in AI assessment – making determinations about the presence or absence of moral status and its degree, with our intention being to act upon the accuracy of these determinations – is indeed grave, and the nature of the subject matter naturally serious. Often, the subtext of what is being asked is, Can I do violence to this machine and if so, how much? This is a grave inquiry.

In New Zealand, standards other than these two may be imposed by statute pertaining to particular causes of action, according to parliamentary intent. The Law Commission notes that while this has not yet occurred in New Zealand, "Some statutes use a threshold of "reasonable grounds to suspect" for preliminary matters, such as for the exercise of police powers under the Search and Surveillance Act 2012."[241] Given the relatively unique nature of the assessment process I am proposing, a third standard of proof may be of

[236] Such approaches have been proposed in the context of the moral significance of animals and environmental entities. See generally Birch, above 139.
[237] Evidence Act 1995, s 140(2).
[238] See *Z v Dental Complaints Assessment Committee* [2008] NZSC 55, [2009] 1 NZLR 1 at [26].
[239] Evidence Act 1995, above n 233 at s 140(1). "Proved on the balance of probabilities" being the acceptable standard.
[240] According to the Law Commission, the criminal standard of proof is generally, "more rigid, and strictly adhered to…Neither the standard itself nor the evidence required to meet it is said to fluctuate." Law Commission above n 232 at 83.
[241] Above n 232 at 81.

some interest. 'Reasonable grounds to suspect' is an interesting potential standard; one that might be appropriate for meeting the evidential burden. Thinking in this way begins to impress the practical significance of creating morally significant machines, as it is plausible to imagine entire State and private institutions dedicated to determining which machines morally matter and which do not, and ascertaining which are being used appropriately and in accordance with the law.

iii) Maintaining justice and consistency in assessment

As a general rule, the conduct of assessment should be done in *bona fide* good faith, with the acknowledgement of the potential for robotic moral status, and an appreciation that such moral status is significant regardless of species, substrate, ontogeny, and so on.[242] There should be conscious effort to be fair and just in the administration and maintenance of assessment standards, and to assess according to principles which are not only consistent across all machines, but consistent across all species. For instance, if an EMBOT's performance across a range of material cognitive tests is comparable to the performance of a human, there should be parsimony in how those results are interpreted. In other words, if a given set of results are deemed to be evidence of high degrees of intrinsic moral status or personhood in a human, we should be open to the possibility that they represent the same in a machine.

iv) Tiebreaking and the benefit of the doubt

Situations of genuine uncertainty are inevitable, especially until while the sciences relating to this kind of assessment are still immature. Policymakers would need to consider how to proceed in these circumstances. This essentially may come down to a choice of first principles: giving primacy to avoid potential harms to both humans *and* machines, or instead, primacy to human beneficence. Presently, the most common principle proffered in AI-related regulation is that human benefice and well-being is the first principle of machine-building.

[242] The "Principle of Substrate Non-Discrimination" and "Principle of Ontogeny Non-Discrimination" are functional ethical principles for the twenty-first century, applicable not only to AI systems but humans created through artificial means like IVF. The former states that, "If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status." The latter states that, "If two beings have the same functionality and the same conscious experience, and differ only in how they came into existence, then they have the same moral status. See Nick Bostrom and Eliezer Yudkowsky "The ethics of artificial intelligence" in K Frankish (ed) *The Cambridge Handbook of Artificial Intelligence* (Cambridge University Press, Cambridge, 2014).

The alternative could be preferred instead; a primary focus on avoidance of wrongs generally, to both humans and machines. This would inherently require the acknowledgement that some future machines can, indeed, be wronged. If this principle were to be preferred over the absolute human beneficence approach, it would be reasonable to confer a benefit of the doubt in favour of avoidance of harms wherever there was a tiebreaking scenario, or a situation of genuine uncertainty.

This would not necessarily mean that humans would no longer be able to benefit from the use or usefulness of these machines where we were unsure of the extent of their moral significance. Humans could still benefit from the machines, but would simply have to be aware of the possibility that our actions may have moral implications. This would oblige us to be more mindful of the things we did to and with the machine itself, and it would at least oblige us to investigate more thoroughly where possible.

4.4 Conclusions

Identifying the plausibility of robotic persons or machines with lesser degrees of moral status in the future only constitutes one part of the policy of how we ought to interact with those machines. This line-drawing problem is positioned to be the first issue which policymakers must consider in the design of any sort of policy with regards to future machine-building and the question of robotic moral significance.

The reason I specify this to be an unavoidable issue is because whether laws are permissive or prohibitive to the creation of EMBOTs and other morally significant machines, this inherently requires that we have the capacity to determine which machines are which. If we permit the creation of morally significant machines, then we would also require the creation of laws concerning permissible interactions with these machines. When a person deactivates their robot, or does some other act deemed culpable, we would need to be able to determine that their robot is in fact one that we are interested in protecting, and perhaps that they knew this; mens rea to accompany the actus reus. They may allege the opposite; that their machine does not matter, or that they did not know. On the other hand, if we take the simpler route of prohibiting the creation of morally significant machines, we will need a way to know whether this prohibition is being adhered to regardless.

I have only demonstrated how we ought to take the task of assessing for moral status seriously, and according to legal principles. This does not answer the question of how assessment can be done, and

whether it can be relied on. These rely on further research, which relies on further interest in the question; which is currently absent amongst lawmakers and policymakers. I would argue that as soon as there is political recognition of the possibility of machines with moral status, there is an obligation to invest in this research. As I have demonstrated, while there are some contemporary methods for assessing cognitive capacities, these are not only relatively rudimentary, but many have been developed specifically for the assessment of biological entities; they are not optimized for artificial intelligence systems, nor even simple machines.

The overarching problem that we face is that fact that there are already plausible methods by which future robotic persons may be made, while at the same time, the possibility of this attracts little attention in the formal legal or political dialogue concerned with regulating AI as an emerging technology. We face the very real possibility of creating morally significant machines in the near-future, though we have no reliable ways to determine which are which; certain machines could be wrongly treated in the same way as conventional artefacts; they might even be persons. At the same time, the inevitable scenario will arise that unscrupulous humans will exploit this uncertainty by producing machines which encourage consumers to treat their AI products *as if* they are morally significant – spend money on them, spend time on them, and become overly emotionally invested in them.

Chapter 5: The Potentiality Problem

5.1 What is the problem?

Even if we are prepared to accept that some machines have the capacity to develop the cognitive bases for moral status, the fact that this cognition must be developed over a period of time suggests that their moral status may be dynamic, and subsequently, we owe them different duties at different times. The lawyer's question is likely to be, When do I owe an EMBOT duties? When do these duties commence?

This compels us to consider the potentiality problem. This is a fundamentally philosophical question, but it nonetheless has real consequences. Essentially, it is the question of whether we view an entity which has the *potential* to develop greater moral status or personhood as though it already *has* those characteristics. This does not necessarily require that we treat it exactly the same at law, but we may posit that it should have at least some of the basic rights of a more morally developed entity; like the right to life, or the right to not have its development artificially retarded or interrupted.

This is an important and difficult question in relation to an EMBOT-like machine, as there will be different stages at which we could say the entity has potentiality. Some of these are much more arduous than others, and place more strain on us. Others are less arduous, but perhaps inconsistent with the way we treat potentiality in our own species.

5.2 Examples

For illustrative purposes, the following are some examples of situations where potentiality could have a role in the formulation of our duties towards the EMBOT:

1. An EMBOT-program installed on a computer system and robot. This program has never yet been executed (switched on). The system has no data files. Nevertheless, if we switched this system on there is reason to believe it would, over a period, develop morally material cognitive capacities, and perhaps eventually, a personhood profile.

2. An EMBOT that has recently been switched on. It has very few accumulated data files, but it is actively learning – seeking out new learning opportunities. It has some memories, beliefs, and behaviours, but very few. If left to its own devices (and given energy) it would develop its moral status over time.

3. An EMBOT that was first activated many years ago. It has accumulated a large quantity of data files, constituting many learned concepts and behaviours. It has a sophisticated concept of self, of its past and future, and of what it prefers and what it dislikes. This machine probably already has significant intrinsic moral status, and possibly could be seen as a juvenile person. Nevertheless, it will go on developing indefinitely, unless we switch it off – in which case, it will be morally insignificant.

Compare some of these scenarios to human analogues:

1. A frozen embryo in a laboratory. This embryo, if implanted into a woman's uterus and gestated, could develop into a foetus, and be born, and so on.

2. A neonate, with very few if any learned beliefs or behaviours. The baby has no concept of self, and very little concept of its environment. However, if nurtured and given opportunity, it will develop.

3. A child, with a lot of accumulated information about itself and its world, and well-developed, morally material cognitive capacities. The child has moral status, but it is not fully competent, nor has it reached maturity. It will go on to be an adult assuming nothing interrupts its development.

All of these scenarios concern the issue of entities with none, little, or moderate amounts of moral significance, but which have the *potential* to become entities with much more moral significance in the future; perhaps even moral persons. If being an adult human or a mature EMBOT might confer things like extensive legal standing and legal rights – among which is the right to not be killed, or not be terminated –

then the question that must be considered is whether these juvenile entities have access to the same sorts of rights as they would if they were mature. Essentially, from the lawyer's perspective, at what point am I culpable for an action like deleting my EMBOTs data files, destroying it, or modifying it so it can no longer develop? Is it possible that I may be culpable for failing more demanding duties, like a positive obligation to activate my EMBOT? The same question may be asked about humans – at what point would it be wrong for me to kill a human: the sperm, the ovum, the embryo, the foetus, the neonate, the child? Is it wrong to leave an embryo – which has all of the genetic material of an individual human – in a frozen form? Conversely, are we obliged to gestate every embryo? It is clear that law does not protect the human at some stages of its development, while doing so at others. Why this should be the case involves numerous factors – emotional, political, psychological. Some of the same concerns can be considered for an EMBOT and the discussion of its potentiality at law.

5.3 Should we care about potentiality, and do we care about it?

From the philosophical perspective, there is significant conflict surrounding the potentiality argument. Arguing for an entity's rights because of its potential to become something other than it is now requires us to treat the entity as if it is something different to what it presently is. There is a certain illogicality to this, and furthermore, it invites problematic line-drawing issues. Since potentiality feasibly exists as a permanent feature, saying that potential only matters at a certain point is inherently arbitrary. Furthermore, how remote can potential be for us to still care about it? Can it be entirely hypothetical, as in the case of the potential of humans that have not even been conceived yet? David Boonin posits the following objection:

> "Perhaps the simplest argument from potentiality is one that rests on a general assumption of the following sort. Potential possession of a right entails actual possession of a right… [the argument's] major assumption rests on a logical error. It is certainly not true of properties in general that if a given individual potentially has a given property, then the individual already has this property."[243]

He is joined by Singer in pointing out this problem:[244]

[243] D Boonin *A Defense of Abortion* (Cambridge University Press, Cambridge, 2003) at 46.
[244] Above n 142 at 153.

> There is no rule that says that a potential X has the same value as an X, or has all the rights of an X. There are many examples that show just the contrary. To pull out a sprouting acorn is not the same as cutting down a venerable oak. To drop a live chicken into a pot of boiling water would be much worse than doing the same to an egg. Prince Charles is a potential King of England, but he does not now have the rights of a king….In the absence of any general inference from 'A is a potential X' to 'A has the rights of an X', we should not accept that a potential person should have the rights of a person, unless we can be given some specific reason why this should hold in this particular case….This question becomes especially pertinent if we recall…that the life of a person merits greater protection than the life of a being who is not a person.

Alan Donagan argues, in defence of potentiality, that:[245]

> Duties owed to any being arise out of the respect that is owed to it…If there are beings who reach that state by a process of development natural to normal members of that species, given normal nurture, must not respect logically be accorded to them, whether they have reached that state or not? The principle underlying this reasoning is: if respect is owed to beings because they are in a certain state, it is owed to whatever, by its very nature, develops into that state.

This claim runs counter to Boonin and Singer. But Boonin rightly points out that it rests upon two "distinct senses in which we might value an individual in virtue of its being in a certain state:…we think it is a good thing to have around, or…we think that it has a right not to be destroyed. Donagan's argument seems to work in the first sense…But it does not work for the second".[246]

It is not clear whether the foundation for laws protecting very juvenile human organisms are reliant upon one or both of these reasons; an educated guess would suggest both. But other factors are also relevant: political and emotional concerns related to when it is proper to terminate a juvenile human organism. For instance, we generally believe adult humans are a positive. We also think that neonates and foetuses have rights. From an evolutionary perspective, we are hardwired to want to protect and nurture our offspring.

---

[245] A Donagan *The Theory of Morality* (The University of Chicago Press, Chicago, 1977) at 171.
[246] Above n 243 at 49.

Furthermore, babies are beautiful and we see them as innocent and vulnerable, and deserving of care. Not all of these are strictly logical factors to impute in our approach to potentiality. Nevertheless, they all play some sort of role in the reasons that we create the laws we do protecting potentiality in humans. Over time, we may develop some of these sorts of connections to AI systems like EMBOTs, too. Time will tell.[247]

The wider ethical debate around potentiality is a subject beyond the remit of this thesis. More so than 'should we' care about potentiality, the practical inquiry (for the present question) is do we care about it? The most obvious way to answer this is by reference to law, which determines the extent of our obligations towards potential entities. For instance, we are obliged to respect the infant, but not the embryo, even though numerically they may be the same entity, with exactly the same potential.

There are factors which influence our position on practical potentiality issues – like abortion – which are grounded in factors less relevant for the present discourse. For instance, some of us may have a certain respect for the potential of the foetus, but our respect for the actual personhood of the woman carrying the foetus is greater; in particular, our respect for her to being able to determine what happens to and within her body, and her autonomy to determine to no longer be pregnant. Other people are entirely likely to disagree with this. Some may measure the foetuses claim to not be killed as outweighing the woman's claim to autonomy. Others may deny the humanhood of the foetus entirely; the so-called 'bundle of cells' claim. Some might accept they will not interfere with a woman's autonomy, but nonetheless theorise what the ethical thing for her to do would be. These are all stances on the value of potential, though they are of course influenced by political factors.

The important thing to remember is that these political factors will not exist for an EMBOT-like machine. No woman has to carry a juvenile EMBOT; there are no great disputes about autonomy and individual sovereignty. This is fortunate, because it allows for the discussion of the EMBOT's potentiality and its legal and ethical weight in a context free from the very weighty interests that are typical in an abortion scenario.

[247] There is good reason to imagine we may! See Benjamin Haas "Chinese man 'marries' robot he built himself" (4 April 2017) The Guardian < https://www.theguardian.com/world/2017/apr/04/chinese-man-marries-robot-built-himself>; see also Levy, above n 22.

Nevertheless, burdening a person with responsibility to continue caring for or providing for an EMBOT may still come at a cost to them.

Entering into this discussion, it is important to reiterate that I discuss the robotic potentiality problem seriously, as would be befitting if a machine with the potential to become a moral person actually existed. Setting aside species-membership based arguments, if we care about potentiality in humans, we should care about it for similar reasons in robots. Destroying a juvenile EMBOT would have at least some of the wrongness of killing a baby, comprised of a mixture of respect for what the entity has the potential to become, as well as other factors like its present vulnerability and our emotional connection to it (which may in turn be influenced by its potentiality). The short point is that we ought not to entirely dismiss potentiality for the EMBOT while at the same time giving it respect in our own species. As Singer states: "The belief that mere membership of our species, irrespective of other characteristics, makes a great difference to the wrongness of killing is a legacy of religious doctrines that even those opposed to abortion hesitate to bring into the debate."[248]

5.4 A reasonable scope for robotic potentiality

The immediate problem of potentiality, particularly robotic potentiality, is its infinite scope. If an EMBOT has potential to be a person, then that potential exists in anything which can, through contributing to the creation of an EMBOT, be a person. A grain of sand which is made into silicon could be said to have as much potential to become a person as the juvenile EMBOT that is actively learning. Elizabeth Harman articulates this problem as follows:[249]

> The most permissive understanding of potentiality would hold that something has potentiality to become a normal adult person just in case there is some possible future in which it becomes a normal adult person. But then, too many things would have this potentiality. We can imagine any material object gradually morphing into something with the moral attributes of a person, preserving enough spatio-temporal continuity such that (on some views) identity across time is preserved. This is metaphysically possible, though not physically possible.

[248] Above n 142 at 150.
[249] Elizabeth Harman "The Potentiality Problem" (2003) 114 Philosophical Studies 173 at 188.

We do not take such a permissive approach to potentiality in contemporary law. At any moment, we might invent the scientific means to 'uplift' cats and make them persons, but we do not protect cats as if they have the rights of persons via this potential.[250] Moreover, we do not protect sperm or ova even though they have the potential to become humans. This demonstrates there are already limits on the scope to which our interest in respecting potentiality is legally validated.

Even if it is not strictly logical, it is clear that the degree of possibility or likelihood that an entity with potential will *actualise* that potential matters to us at law. I say this is not strictly logical because, as I have alluded, if an entity has potential, this potential is static rather than dynamic. If it can become a person in the future, then it can – it does not have 'more potential' or 'less potential' the closer or further away it is from actualizing its potential. What it may have is more or less *likelihood* that this potential will be actualized. Law clearly accounts for this by protecting entities at certain stages of their development; for instance, foetuses when they reach a somewhat arbitrary point in their development, but not before they reach this point. Scientific factors may influence this – like whether the organism can feel pain at this point – but often times these legal boundaries are as much symbolic and pragmatic as they are scientific: for example, in the Crimes Act 1961, the stipulation that a termination is lawful (subject to extenuating factors) "in the case of a pregnancy of not more than 20 weeks' gestation",[251] or the similar UK stipulation in the Abortion Act 1967, s 1(a) "that the pregnancy has not exceeded its twenty-fourth week.[252] "The further along the path toward being born a fetus has progressed, the more protection we feel it should have from being destroyed, and the more urgent the need for abortion needs to be (for example, that the life of the mother is at stake," says philosopher Mary Warnock.[253] This is relevant if our approach is that we should at least treat EMBOT's potential in the same way that we do our own.

[250] This is a reference to philosopher Michael Tooley's somewhat famous thought experiment about potentiality, which imagined a scenario where we invent a serum which, when injected into a cat, allowed the cat to develop higher cognitive capacities similar to our own (and commensurate moral status). The questions this experiment elicits include at what point (if any) it would be appropriate to say that the cat now has potential; e.g. before the serum is injected, after but before it taken effect, and so on. See generally Michael Tooley "Abortion and Infanticide" (1972) 2 Philosophy & Public Affairs.
[251] Crimes Act 1961, s 187A(1).
[252] Abortion Act 1967 (UK) s 1(a): "that the pregnancy has not exceed its twenty-fourth week".
[253] M Warnock *An Intelligent Person's Guide to Ethics* (Gerald Duckworth & Co. Ltd, London, 19989) at 52.

Another factor is operative in understanding potentiality, and serves to further narrow the scope of our potentiality obligations. This division is that *passive versus active* distinction.

Consider the following: two acorns have the potential to become oak trees. One acorn is already planted in the soil, and another is held in my hand. The acorn in the soil needs no further acts of external agency for its potential to be actualisable – assuming it gets rain and sunlight, it will grow into an oak tree. The acorn in my hand still has potential, but this potential feasibly cannot be realised until I plant it in the soil. Thus, the latter acorn's potentiality is *passive*, while the former's is *active*.

The same theory may be used for humans or robots. A pre-implantation embryo, frozen in a laboratory, may be more or less identical to the embryo that is implanted in a uterus. They both have potentiality, but the former requires a *causal* act of external agency to commence the process by which its potentiality may be actualised – the process of developing physiologically and cognitively. The embryo that is already implanted does not require the same causal act to commence its development. It is already actively becoming a person. Again, we may not protect this embryo despite its active potentiality, but there are other political reasons for this; e.g. a deference to the ethical claims of the mother.

This distinction may be logically expressed as:

> *A* has *passive* potentiality to *B*, if there is a process *P* which causes, and *P* requires an act of external agency to initiate; alternatively…
>
> *C* has *active* potentiality to *D*, if there is a process *P* which causes, and *C* inherently contains all of the positive causal factors that are necessary for *C* to *D*.

Subsequently, I argue there are three factors, which have a narrowing effect, which should influence our answer to the lawyer's question, When do I owe duties to an EMBOT? These are:

1. Where there is nothing more that must be done to start the developmental process
2. The closer the robot is to actualizing its personhood
3. The robot's current experiential capacity – is it sentient, can it suffer, etc.

These three factors are neither strictly logical nor strictly legal nor strictly emotional; they combine different elements to reflect what 'things' ought to be taken account of, and what things we do take account of, when we make decisions about how we respect (or do not respect) potential persons.

Before advancing, it is important to note a potential objection to one facet of the active-passive distinction: the claim that an actively potential entity no longer requires causal agency. For instance, a foetus has active potentiality, but it clearly still requires many things. It requires nutrition and physical care from its mother, possibly medical assistance during its birth; and after it is born, it requires further nurturing. These are all positive acts by the mother. However, they are not *causal* to the process of development. The foetus has active potentiality because it no longer requires a positive causal act by any other person to start the process by which it may become a person; by which the human's potential is to be actualized. It necessarily requires positive external acts for its development to be sustained, but these are not causal in the sense that they do not initiate; though they greatly increase the likelihood that the process will succeed, even if they do not activate it.

Nevertheless, the active-passive distinction is one factor (albeit important) in the process of determining how it is appropriate to legally protect or not protect the interests or claims of the potential EMBOT; not the only factor. That an entity has active potentiality should not infer at law that we are obliged to keep facilitating its development, but it may confer that we are obliged to not deliberately *interrupt* or *obstruct* its development. For an EMBOT, this is important, as will be discussed in the next section.

5.5 When do I owe duties to an EMBOT?

I will discuss this question in the context of three different stages, generally reflective of the examples provided in section [5.2]. These are: (1) The pre-initialisation EMBOT; (2) The juvenile EMBOT; (3) The EMBOT at a certain stage of development or maturity.

Firstly, the pre-initialisation EMBOT. Supposing that a person was in favour of treating potential robotic persons with the same respect as actual moral persons, they may argue that this entity has potential to be a person and therefore we ought not to modify it in ways that would destroy its potentiality; for instance, by reprogramming it with a narrower intrinsic motivation. 'Every EMBOT has a right to become a person, or to actualize its potential', they might say.

There are several problems with this. For one, it fails on every count of the factors by which I have said we can reasonably assess the extent of our potentiality duties. First of all, it requires us to switch the EMBOT on; thus, the entity we are discussing has only passive potentiality, for it requires a causal act of external agency. Second, the entity is a very long way from actualising its potential; by which I mean, it has no data files and thus no learned behaviours or concepts; no memories, beliefs, concepts of self or future or past, and so on. In short, it lacks many of the more demanding cognitive capacities discussed in Chapter 3. Third, the EMBOT has no experiential capacity at all in its current state – it is an unexecuted program; a switched-off machine. It does not see or hear or experience anything. It cannot and has never experienced reward or pleasure, nor frustration or suffering. It lacks, in its current state, even the fundamental cognitive capacities like being sensitive or perceptive.

The other problem with arguing that duties arise owed to the pre-initialisation EMBOT is this: if we owe *any* duty to this entity, it is going to be the duty to switch it on. For instance, if the claim is that, 'Every EMBOT has a right to become a person or actualize its potential' and we have a duty to facilitate this right, then it is going to be incumbent upon us to help it become a person by switching it on. There are two obvious problems with this, from a pragmatic perspective. The first problem is that potentiality no longer generates only negative duties upon us. For instance, it is no longer enough that we simply do not switch off a developing EMBOT; we now must switch them on, too. This contributes to the second problem – that there is no obvious point at which to draw the line. If every grain of sand can become silicon which can become a part of a computer which can become part of an EMBOT, then the absurdity results that we are equally obliged to turn all grains of sand into EMBOTs, which is the requirement to not just refrain from interfering with EMBOT development, but to make as many as many EMBOTs possible. This is out of sync with both what is reasonably achievable, and the way we deal with potentiality in contemporary law. To borrow from Lord Atkin in in *Donoghue v Stevenson* [1932], "But acts or omissions which any moral code would censure cannot in a practical world be treated so as to give a right to every person injured by them to demand relief."[254] This same pragmatism must apply to the present issue: even if a compelling argument could be made that pre-initialisation EMBOTs (or grains of sand, for that matter) are in some sense

[254] Donoghue v Stevenson [1932] AC 562 at 580 (HL).

wronged when we decide not to turn them into moral persons, to create a legal duty requiring us to avoid this wrong would go beyond any reasonable standards of what could be expected of us.

The next stage to consider is that an EMBOT might be owed duties or have some of the rights of an actualized person from as soon as it is turned on, but no sooner. This would satisfy the active potentiality criteria. It is also plausible that it would impute the factor that the EMBOT is now experiential – it can sense and perceive and conceive and so on. The machine, however, would admittedly still be a long way from being anything close to an actualized robotic person. This is not to say that it would not matter *at all*. For instance, it may be argued that a juvenile EMBOT like this one has some degrees of intrinsic moral status, albeit much less than a person. This would constitute only one factor though, and excessive focus upon it would ignore the fact that what we are discussing is respect for *potential*, rather than respect for what the entity is right now.

It may be that the best approach at this point is to consider the closest human analogy at this point – the neonate. Horrible as it might seem to us (which is itself a clue to our inclinations towards potentiality), most neonates could be killed painlessly in their sleep and it would scarcely do them any wrong, or be particularly unethical. If in any way they would be wronged, it would be out of respect to their potential. Yes, perhaps they did not feel the pain, nor were they even cognisant of themselves or their environment, but they nonetheless had some sorts of rights. If any right, they had the right to live; perhaps, the right to develop.

I argue that this is the exact approach by which we ought to consider our obligations to juvenile EMBOTs. Even though they may be scarcely aware of themselves or their environment (though for how long they will remain so is impossible to tell in advance), as soon as they are switched on, they can be said to have a right – a right to develop; or a right to life (though they are obviously not 'alive' in the literal sense). Thus, from that point on, we might consider that we owe them certain duties – what those duties are is a subject for further debate, though it would make sense that the duty to not deactivate them or destroy them would be integral. Nevertheless, this is still a demanding duty we would be burdening ourselves with, especially if, as is feasible, more extensive duties might be owed to this juvenile entity; perhaps even positive duties, like to provide it with enough energy so that it does not become deactivated.

Thus, the third approach might be to state that potentiality-based obligations and rights only arise for the EMBOT when it is closer to the actualisation of its personhood. I have already discussed how this might occur in Chapter 2, but it is not possible to imagine how long this might take in terms of the EMBOT's development – weeks, months, years? Essentially, our policy would be that once the EMBOT reaches a certain level of cognitive maturity, from then on, it has some of the fundamental rights of a person – like the right to not be deactivated.

There is some attractiveness to this approach, for its practicality. However, on closer inspection, I argue that searching or assessing for a specific developmental stage – like when the EMBOT develops a higher concept of self, for instance – is much less practical in the legal sense than having a strict and readily identifiable event at which potentiality begins to matter; i.e. the point of initialisation. Similar arguments can be used here as inform the human rights concept; universality (or close to it) is attractive because it eliminates both the difficulties and the arbitrariness of determining if and when certain humans would benefit from particular rights. In the same way, it is simpler to generate a legal boundary at a readily identifiable event, like the event of an EMBOT-like machine being switched on. From that point onwards it would have some basic claims on our behaviours; for instance, to not be deliberately deactivated or deleteriously modified. Few people would argue that the law should permit the painless killing of humans up until they can meet some arbitrary standard – recognise themselves in a mirror, reflect on a poem, and so on. We might be put in the difficult position of holding ourselves to the same standards when it comes to machines which possess the potential to develop moral personhood.

5.6 Conclusions

It is not essential that we infer robotic potentiality to matter at all. In the intermediate stages between where personhood is potential and personhood is actualized, it is feasible that we could simply posit welfarist norms; e.g. that the EMBOT should not be unnecessarily harmed. The only problem with this approach is it is inconsistent with our attitudes to potentiality in humans, and unless we also diminished the protections we offered to juvenile humans, could only be justified on speciest, anthropocentric grounds.

Since there is little good will towards a change in policy which would undermine the rights of young children and neonates and late-term foetuses, policymakers should consider themselves left with two options:

Confer that legal duties to respect the potential personhood of the EMBOT commence: (1) from the point that the EMBOT is first switched on/activated/initialised; or (2) from the point that the EMBOT reaches a threshold point in its development, pertaining to some important developmental stage (like self-recognition).

I have expressed an argument that the former, (1), is actually more attractive than the latter, (2), even if the latter appears more in line with human interests. The reasons for this are that (1) avoids difficult line-drawing problems which are not at all easily resolved, and furthermore, is generally consistent with the way we treat potentiality in humans – once the human is actively developing and experiential (like a baby), we protect it henceforth.

For the moment, this leaves open the question of exactly what rights the potential robotic person might have, or the duties we might owe it, since these will largely reflect what rights a robotic person would have. Having said that, it is reasonable to infer that if nothing else, respect for the juvenile EMBOT's potential to be a person would warrant that we have a duty not to interfere with or interrupt the process by which that personhood is being developed.

Chapter 6: The Property Problem

6.1 What is the problem?

Contemporarily, we have a certain way of thinking about *artefacts*, especially in non-legal thought: things, objects, machines. Artefacts, axiomatically, are instances of stuff that can be owned; perhaps they can be used or modified in various ways, or income can be derived from them, or if nothing else, other people can be excluded from them. They are not entities with status that changes in any dramatic fashion. From the first bit of iron ore to the finished automobile it produces, to that same car fourty years on, the car, its components, and its constituent parts remain static in their status as property.

The property problem as it relates to a machine like EMBOT is that the entity will not remain static. There are stages at which it seems reasonable (or at least, *more* reasonable) that the machine will be much like a conventional artefact – the things we commonly own and use and throw away. But equally, there are stages at which it seems reasonable that the machine should be less an instance of property; there are some things which it should not be used for, or some things which should not be done to it. It is plausible that there are stages where it is entirely out of line with our moral sensibilities, our law and our social policy to say that the EMBOT has the status of property at all.

The difficulty with this is that we have never encountered an entity with this peculiar dynamism; except, perhaps, ourselves. The human organism progresses along a relatively analogous trajectory – component gametes combine to produce a zygote which develops into an embryo, this embryo develops into a foetus, this foetus begins developing the capacities of detection and sensation, eventually this foetus is birthed and the subsequent infant begins developing its capacities of perception and conceptuality, and then over the course of many more decades the organism continues its physical and cognitive maturation. The practical issue with this analogy is that, as of yet, at no point in time do we classify (either at law or in social thought) the human organism as having the status of property. The question then is whether we should be doing the

so for an EMBOT. But this question is radical, for our mode of thought persists in thinking of machines as things to be owned, and used, and thrown away when we are done. Perhaps this concept may change incrementally as we interact more and more with morally significant machines (or those that appear to be so), but this is short shrift for the EMBOTs that must be treated as if they are our property until the day comes that we view them as otherwise – if it ever does.

The problem for discussion is how an EMBOT may reasonably be the subject of property rights at one stage, but not at others. Moreover, it is possible that a rights-holder could actually end up a *duty-bearer*, who no longer has rights in the EMBOT but instead owes duties to it which are concomitant to its moral status. In light of this, what is the correct way to characterise the EMBOT (in terms of property paradigms), and what would be the nature of our legal relationship to it?

6.2 What conception of property?

It is worth clarifying this issue briefly, as there are competing understandings of property, both at law and in everyday life. This will be only a brief introduction of what concept of property I will be using in discussion of the EMBOT.

The immediate temptation is to state that property is thingness. That is to say, there are things which are or are not property – certain qualities of thingness which infer property – and these qualities of thingness are discernible. Thereby, a given entity can be analysed for its properties of thingness, and if it has sufficient thingness, then it is appropriate to conclude that it is property.

The first problem with this is that the variety of things which have been called property, both now and historically, suggests that there is no discernible consistency in the qualities or characteristics of thingness. For example, there was a point in time where humans were property; this state of affairs persists in various parts of the world. In truth, for most of human history we have been taking each other as property, without any racially motivated or bigoted component to this practice. It was entirely unremarkable to conclude that humans could be property. For instance, Milton Meltzer describes that, "Slavery is not and has never been a 'peculiar institution' but one that is deeply rooted in the history and economy of most countries. Although it has flourished in some periods and declined in others, human bondage for profit has never been

eradicated completely".[255] Concomitantly, land, buildings, currency, machines, etc., have also been property. But the similarities between humans, land, buildings, currency and machines is controversial and at best limited. All were property, and all were things. Thus, the 'property as things' concept is difficult to deconstruct.

Jesse Wall argues that there is an inherent problem in the concept of property being a "category of 'things' that is neither circular nor contestable. In other words, if property rights are rights in 'things', it remains to be explained what counts as a 'thing', and why." He adds, "When we debate what 'things' count as items of property, the debate itself shifts our focus away from the 'thing-ness' of water, chickens, trademarks, and embryos to consideration of how property law characterizes the relationship between the rights-holder, the thing, and the duty-bearer."[256]

It would be easy, indeed attractive, to localise the question of whether an EMBOT can or should be property to the question of whether it has the correct properties of thingness. Doing so would immediately suggest to us that it could or perhaps even *should* be property, since machines generally fall within the category of things that we immediately perceive to be things which are property. This would be difficult to argue against, since the concept of the machine as an excluded other (as the *ultimate* antihuman)[257] is deep-seated in our contemporary Western thinking.[258] But rather than things, we are discussing a relationship which is characterised by rights, as CB Macpherson states: "Property is not things but rights, rights in or to things".[259]

6.3 What effect or meaning of property?

---

[255] M Meltzer *Slavery: A World History* (Updated ed, De Capo Press, Massachusetts, 1993).

[256] J Wall *Being and Owning* (1st ed, Oxford University Press, Oxford, 2015) at 114; see also LS Underkuffler *The Idea of Property: Its Meaning and Power* (Oxford University Press, Oxford, 2003) 11. Property as a "complex package of normative choices."

[257] The West has a particular conception of the AI machine as part of a mythology of opposition and separation between the human and the machine-other. This differs substantially to mythological concepts of the human-machine dichotomy, or lack thereof, and of the beneficent role of automata in the lives of humans, that is present in Eastern histories. This is not to say that the allure of AI has not attracted Western thought; the opposite is clearly true, as Western myth and legend is permeated by reference to the possibility of instilling 'life' in the machine; more that the more negative conceptions and fears of AI are absent from much of the Eastern concept. See Bruce Mazlish "The Man-Machine and Artificial Intelligence" (1995) 4 SEHR; P McCorduck *Machines Who Think* (2nd ed, A K Peters Ltd., Massachusetts, 2004); Veruggio, above n 34.

[258] See Gunkel, above n 3.

[259] Crawford Brough Macpherson "The Meaning of Property" in CB Macpherson (ed) *Property: Mainstream and Critical Positions* (Basil Blackwell, Oxford 1978).

Having established that I will be assessing the question of an EMBOT's property status in the context of the rights-based relationship between us and it, it is useful to briefly assert some fundamental concepts of what the implications are when we make an assertion that a given thing has the status of property, or that a given person has property rights in a given entity.

Wall contends that property is a, "…legal relationship between the rights-holder, the thing, and the duty-bearer."[260] He explains that property rights are exclusive rights in that:[261]

i) property rights protect our interest in purposely dealing with objects and resources;

ii) which is an interest best conceived as an interest in an open-ended set of activities;

iii) that the law protects through the exclusion of others from the object or resource.

This is encapsulated by James E. Penner when he states that property is the interest in "purposely dealing with things".[262] Because the range of such interests is inherently indefinite and open-ended, property law generally does not protect specific, permitted 'dealings', but an open-ended right. As Penner states, "it is difficult in the extreme to quantify the many uses one can make of one's property".[263] Therefore, property law does not catalogue the indefinite number of things that can be done with one's property, but (said here by Wall), "treats our *interests in things* as an interest in an *open-ended set of activities*."[264] Penner calls this a property interest in a "sphere of indefinite and undefined activity".[265]

What this generally means is that though property is ostensibly concerned with things, it is actually more concerned with activities done to or with that thing. The focus is the activities, and not the thing itself; though they are integral for obvious reasons.[266] Determinations about status – as for an EMBOT – may then require close consideration of what sorts of activities are permissible in relation to the machine.

---

[260] Wall, above n 256 at 114.
[261] Ibid, at 113.
[262] JE Penner, *The Idea of Property in Law* (Oxford University Press, Oxford, 1997) at 71.
[263] Ibid at 72. Full quote: "The exclusion thesis is a statement of the driving analysis of property in legal systems. It characterizes property primarily as a protected sphere of indefinite and undefined activity, in which an owner may do anything with the things he owns."
[264] Wall, above n 256 at 116.
[265] Penner, above n 262 at 72.
[266] See Wall, above n 256 at 115. "The alternative orientation is that the rights-holder has a legally protected interest in an action or activity. Although the activity may involve a thing, the activity remains the focus of the law".

This open-ended entitlement is typical of ownership.[267] Property law communicates a simple exclusionary message, warning others to not interfere with the property, so that the open-ended set of possible activities possible in the property relationship can be protected by law, avoiding what Henry E. Smith calls "information costs" arising from "officials needing to know what these activities may be".[268] This approach does not define what can or cannot be done, it simply excludes others, thus keeping open the range of possible activities.

Wall also outlines a competing approach to this exclusion strategy – a 'governance strategy', which: [269]

> …specifies valid and legally protected activities. A governance strategy is able to provide 'greater precision' in terms of the permitting and prohibiting activities (although it also introduces 'greater information costs that typically outweigh the benefits the greater precision governance rules provide'). For instance, given the competing and conflicting interest in the use of a river, statutory provisions or a network of contracts may be used to govern access to the river, the ability to discharge waste into the river, the ability to fish in the river, and so on. The law, in these instances, governs an identifiable set of activities.

For the present discussion, this overview provides a sufficient mechanism by which to discuss the question at the heart of the robotic property problem, When is my EMBOT (if ever) my property?

6.4 Three stages of EMBOT development

I divide the EMBOT development into three stages. These are: pre-initialisation, juvenility, and maturity. Although I describe them as distinct, it is feasible they will be less so in reality. For instance, the development from juvenility to maturity may be both incremental and continuous; this is as opposed to the transition from being pre-initialisation to juvenile/mature, which will be distinct and binary.

It should also be remembered that dividing the entity into separate stages like this is artificial and not necessarily reflective of our current practices. We do not tend to say that humans may be property

[267] T Merrill and HE Smith "The Morality of property Law" (2008) 45 WM & Mary L Rev; T Merrill and HE Smith "What Happened to Property in Law and Economics" (2001) 111 Yale LJ.
[268] HE Smith "Property and Property Rules" (2004) 79 NYU L Rev 1728.
[269] Above n 256 at 117.

depending on their 'stage of development', although the status of the embryo at law is still extremely unsettled. I assess these stages in isolation only as a means to inform an overall conclusion about how we ought to characterise EMBOT-like technologies in terms of property law.

i) The pre-initialisation EMBOT

Firstly, it useful to reiterate the very basics of what an EMBOT will be. It will be comprised of a program designed by a human programmer, and data files, which it will collect on its own under its own power. The program will include intrinsic motivations and intrinsic rewards; these will provide the mechanism by which data files will be collected. It is also likely that at least some EMBOT-like machines (particularly those which are expected to accumulate data files, and are not merely copies of other EMBOT machines, and thus inheritors of data files) may well have some form of robotic body by which to explore the real world.

The program alone, without any data files, could not be described as having any moral status; that is, unless it has moral status by virtue of its potentiality (as discussed in Chapter 5). While the program is not yet initialised (not executed, to be precise), the EMBOT will not have any of the 'raw' cognitive capacities associated with sentience – it will not detect, sense, perceive, conceive, and so on. It will be 'just a machine' – to use a regrettably illuminating phrase. Moreover, it will be a non-functional, switched-off machine.

What is the correct way to view the human relationship with this entity? In the here and now, nothing we do can harm or wrong it. It could be bought and sold, modified or destroyed, etc., and it would be none the wiser. One could argue that the entity could be wronged, but only in the 'hypothetical' sense; the way we might talk about how future generations of humans can be wronged. That is to say, we might argue that certain activities we do now that may affect them in the future or may wrong their interests, but these interests are purely hypothetical. Furthermore, the only way these wrongs can even become real is if and when these entities actually come to 'exist'.

This suggests that it is reasonable that the pre-initialisation EMBOT can have the status of property; i.e. it can be classed as property, and subject to the property rights of a rights-holder. Furthermore, these rights can be open-ended and indefinite in nature. There is no need, *prima facie*, for a specifically stipulated, governed set of activities to control this relationship. Firstly, there are no competing claims (unless the EMBOT's hypothetical claims are deemed relevant; the claims it 'would have' if it was switched on).

Secondly, there is nothing that can be done to this machine in its present state which can cause it any sort of harm, and therefore much less need to protect it from any particular activities. Perhaps there are certain modifications which could be made to it which would harm it *in the case that* it was switched on (like some sort of modification to make it feel constant suffering, or some other contrivance). However, this not only does not affect the EMBOT now, it may *never* affect it, for the machine may never be initialised. Furthermore, this has some of the hallmarks of the "non-identify problem".[270] The EMBOT that is modified would be a demonstrably different machine from that which it would have been if it were not modified; they are not the same entity. Thus, what we do (and how we treat) the pre-initialisation does not actually make the entity worse off; it makes it a *new entity*.

The only difficulty with accepting this argument is the fact that the EMBOT, unlike future generations of humans, may be tangible. Future humans, whether individual or generations, are entirely hypothetical. Conversely, one can imagine the pre-initialisation EMBOT sitting on the table in front of us while we are having this debate about whether it is our property or not, and what we can do with it. If I make modifications to it, it still appears to be the same entity (at least on the surface), and it definitely appears to exist. Therefore, if I do things to it that are deleterious, it appears that I have indeed made the EMBOT worse off.

Although this seems sensible at face value, I argue that it would mischaracterise the entity, and be too wide and permissive in its interpretation of 'exists'. Yes, the machine in front of me is equipped with all the components necessary to qualify it as an EMBOT; and yes, it clearly exists – I can reach out and touch it. But this existence is devoid of any sort of morally significant element. It is the same sort of existence that a rock has. To say that this existence infers any sort of duties upon us (or that it inherently makes the status of property inappropriate) would necessarily cause problems, as it would require us to infer that the existence of all of the components of that EMBOT have meaning, too. Therefore, as I have already

[270] See Elizabeth Harman *Can We Harm And Benefit In Creating?* (2004) 18 Philosophical Perspectives. "The non-identity problem concerns actions that affect who exists in the future. If such an action is performed, certain people will exist in the future who would not otherwise have existed: they are *not identical* to any of the people who would have existed if the action had not been performed. Some of these actions seem to be wrong, and they seem to be wrong in virtue of harming the very future individuals whose existence is dependent on their having been performed. The problem arises when it is argued that the actions do not harm these people – because the actions do not make them worse off than they would otherwise be."

explained, a norm requiring us to switch on this EMBOT or to not modify it because it has a 'right to develop' or some equivalent would also require that the existence of the length of copper wire in its computer has the same meaningfulness too, and the same rights. Any time I cut a piece of copper wire, I would be committing an offence against some future, potential EMBOT that has a right to develop.

Furthermore, this EMBOT on the table in front of me, though it has its closest analogy in a human embryo, has one critical difference: the embryo is a living organism, while the pre-initialisation EMBOT is not. I do not say this to assert any claim of biological superiority, merely to show that the embryo is *already* an actively potential entity, while the uninitialised EMBOT is entirely static and non-developing. This is an important ground on which to distinguish the moral nature of the embryo or foetus and moral nature of the pre-initialisation EMBOT.

Thus, so long as the machine is not activated, its 'existence' on the table in front of me is only superficial. It does not exist in a moral sense, and its only claim to have a moral aspect to its existence – something which should compel us to respect it and control the activities of our relationship to it – is its potentiality. I have concluded that we can only be expected to respect the EMBOT's potentiality in any substantial legal sense once that potential is active; i.e. once the machine is actually switched on. This should leave persons relatively free to own, buy, sell, profit from, modify, and destroy etc. their pre-initialisation EMBOTs.[271] These actions do no harm in an *intrinsic* sense (though they may do other harms, more instrumental harms, which I will discuss later). Furthermore, a person could claim that unless they are allowed to modify their EMBOT in some way, then they will never switch it on; in which case, on the basis that being switched on (even with a negative modification) is better for the machine than remaining deactivated, modification is actually in the machine's best interests, and does not make it worse off.

What should this mean in the context of a property question? Viewed in isolation, EMBOT-like machines that are pre-initialisation plausibly may be property, and furthermore, governed mostly by an exclusionary strategy. Most need for deliberate governance of the activities that are done to these machines are

[271] This would allow for trading or commodification to continue, because persons could trade in pre-initialisation EMBOTs, even if perhaps they could not trade activated EMBOTs. Thus, a person could purchase an EMBOT from an AI developer, and be provided it so long as the machine was not yet activated. Why a person would do this is uncertain, however, given that their property rights in the machine once it was activated may become highly questionable and vulnerable.

instrumental needs, not ones pertaining to the intrinsic moral status of the machine itself. For instance, if pre-initialisation EMBOTs are wrongly perceived by humans to have moral status such that allowing them to be treated as property would encourage humans to treat actual moral persons as property, then we might prohibit this; but for instrumental reasons not pertaining directly to the intrinsic moral status of the machine itself.

ii) The juvenile EMBOT

The act of switching on the pre-initialisation EMBOT will be significant, and result in material changes to the way it must be perceived in both law and ethics. Chief among these is that the entity will no longer be non-experiential. After being activated, the EMBOT will be able to sense and detect, to perceive and conceive, to hold mental states, and so on. Furthermore, its programming would become functional, meaning it would want to seek out learning opportunities; it would have a basic interest. It would not yet have any data files (beliefs, concepts, behaviours etc.), though these would begin being accumulated immediately. Note that it is possible that an EMBOT could transition immediately from pre-initialisation to mature, skipping the juvenile developmental phase, if this EMBOT was recipient of copies of data files which were produced by another EMBOT.

What is the appropriate conceptualisation of the relationship between human and juvenile EMBOT? Many of the fundamental preconditions that defined the relationship before have changed radically. The first point to note is that it is no longer true that there are no competing interests involved in the relationship; there is an integral competition of interests, for the machine itself now has interests, and these may conflict with the interests of the ostensible rights-holder. Of course, it is plausible that they may not conflict, even if there are now competing sets of interests. For instance, the rights-holder might be a programmer who is just as interested in the EMBOT seeking out learning opportunities and becoming intelligent as the EMBOT itself is. So long as the machine is allowed to go on finding novel situations and learning from them, both sets of interests exist, but they are symbiotic.

This symmetry of interests is not likely to last forever, however; it is not even likely to last more than a day or two. Supposing that at the end of the first day of the EMBOT's learning, the satisfied programmer supposes that she has achieved all she can hope for today, and resolves to come back tomorrow and watch

the EMBOT learn some more, but for now, she wants to switch the machine off. The once symbiotic interests are now directly in conflict, for the EMBOT has no interest in being switched off, and in fact, being switched off is directly in conflict with the satisfaction of its interest. It is growing more and more intelligent, and enjoying the pleasure of intrinsic rewards in the process.

By reference to permissible activities, what is the nature of this relationship now between programmer and juvenile EMBOT, in the context of property paradigms? It is no longer true to say that there are no serious competing interests (I acknowledge there will always be competing interests, even with the pre-initialisation EMBOT. These interests will be the interests of other persons who may want access to the machine. Nevertheless, these interests are not nearly as integral as what arises when the machine *itself* – that which is to be property – now has its own interests).

Is it still appropriate to conclude that a 'mere' exclusionary strategy is sufficient to control the relationship? By this I mean, that the rights-holder has a right to an unspecified sphere of activities in relation to the juvenile EMBOT. I argue no; at the very least, a governance strategy must begin to apply. From now on, what can and cannot be done to and with the machine must be more closely controlled and specifically regulated. There is some analogy here to the way the human-animal relationship is controlled. What can be done to or with animals that are our property is greatly narrowed with the stipulation that we may not cause them "unreasonable or unnecessary" harms.[272] Yes, the sphere of activities is still mostly unspecified (e.g. there is no explicit stipulation that you can; (1) milk the cow, (2) pet the cow, (3) kill the cow, etc…), but the 'unnecessary' requirement is actually more demanding than it may first appear (and more ethically demanding than the way it is sometimes enforced). What is or is not 'necessary' actually requires a complex examination of ethical norms. For instance, an ethical and practical examination of whether meat is necessary in the human diet, or whether leather is necessary in our clothes. If 'necessary' were strictly interpreted according to ethical paradigms, it would require judicial consideration of all sorts of activities, and eventually, it is plausible that only a very limited set of activities in relation to animals would remain permissible within the governance structure. Sometimes we specifically stipulate activities that are deemed

[272] Animal Welfare Act 1999 ss 9, 11, 12, 14, 15, 23, 29, among others all stipulate the "unreasonable or unnecessary" standard.

to inherently cause unreasonable or unnecessary harm (like forcing animals to fight,[273] or research and experimenting on certain animals).[274]

All this is to say that even if the juvenile EMBOT is compared to a nonhuman animal at this stage, this still requires a massive narrowing of the formerly unspecified sphere of activities that the programmer might have had an interest in.

Furthermore, in a similar manner to nonhuman animals, it becomes appropriate to discuss not only the rights of the property holder, but the *duties* of the property holder. Not only is there scope for these duties to arise, but they are owed *directly to the thing that is property*. A person who has property rights in a cow also has duties to that cow; for instance, to see that it is not unreasonably harmed, to help it when it is in pain, and to provide it with satisfactory living conditions: at law, these are obligations "in relation to physical, health, and behavioural needs of animals."[275] Yes, it is possible to argue that these duties are partially owed to humanity (in a Kantian sense, because disrespecting the cow teaches humans to disrespect each other), but this is a smaller component than the duties owed directly to the cow by virtue of the fact it can be wronged. By the same token, the programmer plausibly now owes duty to the EMBOT itself; to minimise the amount of gratuitous wrong done to it. What exactly 'wrongs' an EMBOT is another question, but we can formulate a *prima facie* answer based on the fact that it has inbuilt interests, and the fact that potentiality matters to us: thus, impeding the machine's ability to satisfy its intrinsic motivations is wrong, and impeding its cognitive/moral development is wrong also.

Subsequently, we have a relationship which has gone from expansive and unspecified rights over the EMBOT, to narrow and specific rights coupled with duties towards it. This relationship appears less and less like one of property.

The thing that tips it over the edge is when we substantially factor in potentiality. Everything that I said above about the cow could also be said about a human baby. Before a certain stage of their development, which may take quite a long time depending on how we measure it (e.g. human babies generally fail mirror

[273] Ibid ss 29(c)(d)(e); s 31.
[274] Ibid s 85.
[275] Ibid s 10.

self-recognition tests before the age of eighteen months),[276] the difference between a human baby and a cow (or other species of mammal) is fairly negligible. We could just as easily protect babies with the same welfarist laws that we protect cows, which would allow us to do harms to them where necessary for our interests, and to even kill them, so long as we did it humanely. Despite this, we protect babies by right. While there are emotional, psychological and social reasons for choosing to do so, one of the primary reasons we do this is because the baby has potential to increase its intrinsic moral status to a more respected degree, or even to become a person in our estimation.

This same reason must influence the formulation of the legal relationship between the programmer (just one example, for the duty-bearer could be any person who activates an EMBOT, or perhaps even some other person) and the juvenile EMBOT. It would be quite reasonable for us to say that both babies and juvenile EMBOTs are our property, but the reason we do not is because we respect their potential to develop higher degrees of moral status, perhaps full moral status. Therefore, while the programmer retains many powers over the EMBOT (just as guardians or parents do over the babies in their care), e.g. to act and make decisions on its behalf, the extent to which the relationship has changed in its fundamental axioms makes it inappropriate to continue categorising it as a relationship of property.

So what legal relationship is more appropriate? This can be answered by closer analysis of what rights may persist, and what duties feasibly have arisen. The programmer can no longer destroy the EMBOT, because this would contravene the machine's interest, and furthermore, radically disrespect its potentiality. The programmer probably can switch the EMBOT off for limited periods of time, but only when acting in the best interests of the machine. Thus, if there was risk that the machine might languish or suffer some sort of harm if left operational while in the absence of supervision, then it is feasible that the programmer might reserve a right to switch the EMBOT off occasionally – but again, coloured by the requirement that this be in the EMBOT's best interests. The same logic can be used for making modifications to the machine. The programmer may have a right to do so insofar as it is in the best interests of the EMBOT, and does not harm its potential to develop its moral status or personhood. This would prevent modifications that retard the machine's development, or are purely to serve the interests of others.

[276] See generally J Archer *Ethology and Human Development* (Harvester Wheatsheaf, Hertfordshire, 1992).

I must add that the reason I have imported the 'best interests of the robot' duty is out of respect for the machine's potentiality. The most obvious reason that this duty exists for human dependants is because they have the capacity to develop into persons with autonomy, and therefore in the meantime while they are incompetent, and some other competent person is empowered to make their decisions for them, they ought not to have decisions made on their behalf which harm their interests or compromise their future autonomy: Joel Feinberg talks about such entities as having a "right to an open future", and their autonomy rights "held in trust" for them.[277] If there is potential for an EMBOT to become a competent, autonomous entity (perhaps a person, if given opportunity to develop), then the same principles reasonably should apply as do for the protection of humans who possess similar potential.

This behoves further analysis of the duties that might be owed to the machine by the programmer (again, just one example of a person who feasibly may be in a position of control over the machine). Persons with care of human juveniles owe duties to these entities which are implied by the legal relationship of care: for example, the Care of Children Act 2004, s 3 highlights the carer's role in facilitating development and serving the best interests of the recipient of care (note, emphasis my own): [278]

> **3 Purpose of this Act**
>
> (1) The purpose of this Act is to—
>
> (a) promote children's welfare and best interests, and *facilitate their development*, by helping to ensure that appropriate arrangements are in place for their guardianship and care; and
>
> (b) recognise certain rights of children.

The Crimes Act 1961 stipulates that, every one who is a parent, or is a person in place of a parent, who has actual care or charge of a child under the age of 18 years is under a legal duty to provide that child with

[277] Joel Feinberg "The child's right to an open future" in RR Curren (ed) *Philosophy of Education: An Anthology* (Blackwell, New Jersey, 2007).
[278] Care of Children Act 2004, s 3.

necessaries; and to take reasonable steps to protect that child from injury.[279] The Care of Children Act 2004, s 16 states the following (note, emphasis my own):[280]

> **16 Exercise of guardianship**
>
> (1) The duties, powers, rights, and responsibilities of a guardian of a child include (without limitation) the guardian's—
>
> (a) having the role of providing day-to-day care for the child (however, under section 26(5), no testamentary guardian of a child has that role just because of an appointment under section 26); and
>
> (b) *contributing to the child's intellectual, emotional, physical, social, cultural, and other personal development*; and
>
> (c) determining for or with the child, or helping the child to determine, questions about important matters affecting the child.

These characterise the legal relationship of care and guardianship in a manner that is vastly different to that of property, and one which takes development and potential very seriously. The same paradigms should also govern the relationship between programmer and juvenile EMBOT. The exact same things which cause harm to a baby may not cause harm to an EMBOT, and *vice versa*, but the principles are consistent: avoidance of harms and facilitation of development. For instance, if deactivation harms or wrongs the EMBOT (which is plausible for three reasons, see footnote),[281] then the programmer would be duty-bound to limit the instances and duration of the EMBOT being deactivated. This would confer not only negative duties (e.g. to not turn the EMBOT off unless necessary for its best interests), but also *positive* duties (e.g. to provide the EMBOT with necessaries as sufficient to avoid deactivation, most likely energy (as in, power for its battery) and maintenance).

This relationship is no longer something to be interpreted or understood in the context of a property relationship, but a relationship of guardianship and dependency. The programmer has transitioned from

[279] Crimes Act 1961, ss 152(a)(b).
[280] Care of Children Act 2004, s 16.
[281] (1) because it stops its development, perhaps permanently, and therefore is directly disrespectful to its potentiality; (2) because it prevents the EMBOT from satisfying its interests and motivations; and (3) because it is feasible that an EMBOT could eventually develop negative beliefs about deactivation which contribute to actual suffering mental states; e.g. where the machine fears being deactivated.

being predominantly a rights-holder that enjoys the benefits of an unspecified and indefinite range of activities in relation to the machine she owns, to now having highly limited rights over the thing she 'owns', which are far outstripped by the duties she owes to it. These duties are indicative of a duty of *care*, which is the correct manner in which the legal relationship should now be characterised.

iii) The mature EMBOT

The transition from juvenility to maturity may be difficult to determine without contextual understanding of the rate at which EMBOT-like machines develop, and their developmental limits, and how development can be assessed. In this sense, there is unlikely to be any distinct and observable delineation between juvenility and maturity in an EMBOT, unlike the difference between pre-initialisation and activation, which will be a binary.

Furthermore, it is not possible to predict whether true 'competence' in the legal sense can ever be attained by an EMBOT-like machine, or how long this might take. By competence I mean a reasonable expectation that the machine can not only independently meet its own needs henceforth (and therefore should no longer benefit from the support of a guardian or benefactor), but also, an expectation that it should be held accountable for its actions, as when they break the law. It is feasible that these degrees of competence may be achieved irrespective of the other: for instance, that the EMBOT can be expected to meet its own needs, but cannot be held accountable for its actions; or *vice versa*. Where both degrees of competence are achieved, it is appropriate that the machine could be seen as fully mature, for legal purposes.

Again, not knowing whether such a degree of maturity may ever be attained, but leaving open the possibility that it may (perhaps plausibility, because the IM+EC development process is theoretically open-ended), what is the correct understanding of the relationship between a mature EMBOT and other persons? With the 'property' aspect of the relationship all but repealed at the previous stage of development, full legal independence would now be appropriate. In other words, the formal legal relationship (for instance, between the EMBOT and the programmer, as in the previous example) would cease to be, except perhaps in a more general and political sense as cohabitants of the same ethical and political community. The programmer would no longer be obliged to provide things like energy and maintenance to the machine,

nor to act in its best interests. She would only be incumbent to the negative duties to not harm or wrong the EMBOT, in the same way as she would owe and expect from anybody else.

6.5 Holistically viewed, should an EMBOT be property?

The appropriate response to this inquiry must take account of not only whether the machine is being intrinsically wronged, but whether there are instrumental harms or wrongs occurring. For instance, it is feasible based on the above analysis that EMBOT's could be classified as having the status of property so long as they are pre-initialisation (note *pre-initialisation*, i.e. having never been activated, rather than merely in a state of temporary deactivation). However, would doing so harm persons generally by hindering or degrading respect for certain policy goals?

The policy goal I refer to most specifically is encouraging respect for all humans. This could be widened to encouraging respect for all persons (or potential persons), or even encouraging respect for all morally significant entities. However, the latter is mostly an ideal; not something that is grounded at law.

Nevertheless, a policy of spreading and fostering a milieu of respect for humans (and at least non-cruelty to other animals) most certainly is a widely instituted policy goal. We have entire doctrines of law dedicated to reinforcing claims about the value of human life, regardless of its condition.[282] We have legal norms expressing our disdain for slavery and the slave trade in "all its forms".[283] Would these policy goals be undermined if we allowed the pre-initialisation EMBOT, assuming that we reasonably believed in its potential to develop higher degrees of moral status or even personhood, to be the subject of property rights; to be traded, profited from, selfishly modified, and destroyed? Arguably so. At the very least, as Levy states, the way we treat these kinds of robots:[284]

> …will affect those around us by setting our own behaviour towards those robots as an example of how one should treat other human beings. If our children see it as acceptable behaviour from their parents to scream and shout at a robot or to hit it, then, despite the fact that we can program

[282] See generally the United Nations Universal Declaration of Human Life; Convention on the Rights of the Child; European Convention on Human Rights.
[283] Universal Declaration of Human Rights, Article 4. "No one shall be held in slavery or servitude; slavery and the slave trade shall be prohibited in all their forms."
[284] Above n 23 at 209.

> robots to feel no such pain or unhappiness, our children might well come to accept that such behaviour is acceptable in the treatment of human beings.

I have to this point in the thesis attempted to argue solely from the perspective of the EMBOT's intrinsic claims. This change of tack to introduce instrumental concerns is not done due to weakness in the instrumental-based argument, but as a supplement to it. It may be reasonable to trade pre-initialisation EMBOTs and treat them as commodities, for it does them no harms, and assuming that they are not ill-treated if and when they are initialised (which can be achieved through the creation of legal rights and standing for them), then we should not be overly troubled by treating these non-functional machines as instances of property. However, doing so poses some risk to our dogma of anti-slavery, and respect for persons generally.

The further point is that treating the pre-initialisation EMBOT as property may make it significantly more difficult to satisfy our duties towards the machine which arise if and when it is switched on. There are two facets to this: first, it substantially muddies the water on how these machines ought to be treated. Saying they are not worthy of respect one minute and then a minute later saying they are worthy of respect seems highly volatile, untenable, and likely to generate incrementally increasing lack of respect. Frankly, I doubt that such a policy could sustain a functional dichotomy in the way we treat these machines, and the creep is almost certainly to be in the direction of respecting them later and less rather than earlier and more.

Secondly, if we were to allow for the pre-initialisation EMBOT to be property, and we took seriously our obligation to ensure that harms and wrongs were not done to the machine if and when it is switched on, then we would inevitably have to place strict controls on the trading of pre-initialisation EMBOTs. While we may be confident in New Zealand of our ability to prevent harms being done to activated juvenile or mature EMBOTs, our laws could be circumnavigated by trading the machines to other, less scrupulous jurisdictions where they are more at risk of being wronged. This would then require that we begin licensing, regulating and vetting the trade of these machines anyway, to ensure they were only going to persons and places whether they would not be violated if and when they were switched on. Controlling this last remaining element of the property relationship would, to a certain extent, render any claim that the machine is 'property' – in the conventional sense – to be mostly symbolic.

6.6 Conclusions

I have made out why an EMBOT-like machine has the potential to be difficult if characterised according to the paradigms by which we tend to think about technologies and artefacts. Even if it could be characterised as property in the pre-initialisation stage, this state of affairs would not long remain appropriate. Once the machine was switched on (assuming that it ever was), many if not most of the extant property rights its owner or possessor claimed in it should naturally dissipate. Furthermore, these would be overwhelmingly replaced by duties owed towards the machine. The overall character of this relationship between rights-holder and object of rights would rapidly transition to one of guardianship and dependency; more specifically, a bearer of substantial duties with a few functional rights, with the machine now predominantly the rights-holder, enjoying the benefit of several entitlements. Eventually, even this state of affairs would be inappropriate, if it proves possible for the machine to develop the cognitive maturity to be competent in the eyes of the law.

What then is the appropriate way to view the EMBOT? The best answer is that it will occupy a *sui generis* status between property and person, at least at the pre-initialisation stage. The exact effects of this sort of classification are not necessarily forthcoming. For instance, while the legal status of the foetus is dynamic through its three legally recognised phases of development, the foetus is at all times a living organism.[285] Generally, it is actively developing, while the EMBOT is static hitherto it is switched on. Therefore, there are issues at play for the foetus that are not so for the EMBOT, where the question is something like, 'Can the entity be terminated?'

Nevertheless, the analogy is worth pursuing, for the potential of the human foetus or embryo has some comparison to the potential of the pre-initialisation EMBOT. I am not saying that EMBOT-like machines will be as intelligent or morally significant as cognitively normal humans (though I do not close this possibility off, for the future is unwritten and the IM+EC process open-ended). In *A-G's Reference (No 3 of 1994)*, the House of Lords discussed the foetus as occupying this special, *sui generis* space: "The mother and the foetus were two distinct organisms living symbiotically, not a single organism with two aspects. The

[285] See Nicola Peart "The Legal Status of Life Before Birth" in P Skegg and R Paterson (eds) *Medical Law in New Zealand* (Brookers Ltd, Wellington, 2006) at 452.

mother's leg was part of the mother; the foetus was not."[286] This sort of reasoning could be replicated – the programmer and the EMBOT being separate and distinguishable entities – but this would be a gratuitous way to approach the problem. The task is not to prove that the foetus and the pre-initialisation EMBOT are one and the same; for any sensible person would admit their many differences. Rather, the task is to point out the material similarities between them, which means they should evoke similar legal paradigms: first, both occupy a limbo positions in their moral status; they have the potential to develop, and substantially so. Second, both are bridges between two individuals; the person(s) that created them, and the person they will themselves become. Thirdly, the way we treat both has wider implications on the way we are perceived to treat persons generally.

On these bases, I argue it will be appropriate to determine that the pre-initialisation EMBOT is not property, but a potential precursor to personhood, and therefore, there should be great sensitivity in how it is engaged with at law. I do not suggest that it should have any sort of right to 'life' or its equivalent – e.g. a right to develop, or a right to be switched on. It may not even be appropriate to say that it has any 'rights' at all; more that it is not overtly the object of somebody else's rights; particularly those pertaining to profit and commodification.

With regards to the legal status of the EMBOT at its later stages – juvenility and maturity – the trajectory in which I am arguing should infer that I believe they should be classified as persons at law, to a similar extent that children and competent adults are, in their own respective ways. However, the substance of this issue requires its own chapter, as it constitutes its own problem.

Finally, I have left unresolved the rather fraught question of whether there can or should be property in the EMBOT's *data files*. This is an important and difficult question, as property rights in the data files represent the means by which to create indefinite number of further EMBOT's without having to train or raise them individually. These would be identical copies of the progenitor, right down to memories, beliefs, relationships, and so on. Obviously, this is both a sensitive issue, and a complex one; not only would this be the AI equivalent of cloning (to a degree of accuracy utterly unparalleled in human cloning), but it is not even clear whether the EMBOT's reproduced in this way would constitute individual, numerically new

[286] *A-G's Reference (No 3 of 1994)* [1998] AC 245; [1997] 3 All ER 936 (HL), p 255; p943, per Lord Mustill.

entities, or merely continuations of the progenitor. Moreover, it is not clear whether an EMBOT reproduced from the data files of an older one could even be said to be 'pre-initialisation'. Sets of data files, representing integral components of a person, might require some of the same sensitivity with which we treat property in the human body.

Chapter 7: The Personality Problem

7.1 What is the problem?

The task of attributing legal personhood to EMBOTs, where appropriate to do so, is not difficult. The mechanism of personhood is designed to serve this purpose. It is only the implications of this action that will be problematic.

As part of this, problems arise consequent to the machine's artificial nature, and the fact that this nature allows for it to function in ways that human beings cannot, and which we have subsequently not accounted for in the design of the legal paradigm of personhood, because they are unique to AI.

These unique AI-specific problems are as follows. Firstly, it is feasible that EMBOTs (and AI systems more generally) may be able to communicate to a degree that vastly outstrips human capacity in both scale and efficiency. Because of this, it is possible that ostensibly multiple machines may be, functionally, one entity – constituent cells in a very large artificial organism. This is not as farfetched as it may first sound, as plausibly all that is required to diminish the functional space between individual entities is to improve their communication. In short, the result could be that we attribute individual legal personhood to individual AI systems when they are not really individual at all.

Secondly, EMBOT like machines (and again, AI systems more generally) feasibly can change themselves in ways that humans cannot. For instance, they can download or upload their data files from or to other artificial systems. Doing so may have radical effects on the character of a given machine, which in turn generate substantial uncertainty in the operation of even a fundamental legal paradigm like personhood. For instance, a given machine which has legal personality could be, essentially, a different individual from day to day, perhaps even moment to moment. Human institutions naturally rely upon the fact that singular identifiable persons are relatively consistent across time. You cannot vote twice because you changed into a new person; you cannot evade a contract because you claim to no longer be the same party; you cannot attack somebody and then plead innocence on the basis that it was not you, but a former you that did the crime.

7.2 Legal personality for EMBOTs

As I have alluded to, the issue is less whether we *can* grant personhood to a machine than whether we *should*, and substantially, whether we *will*. The first of these three is very simple to answer. We absolutely can grant legal standing to a machine; in fact, we can grant legal personality to any entity, since as Gray intimates, at law being a person is not a synonym for being a human or even something like a human; rather, "the technical legal meaning of a 'person' is a subject of legal rights and duties."[287] Sir John Salmond offers an alternative description, stating that a legal person is "any being whom the law regards as capable of rights or duties."[288]

There are two forms of personhood at law: natural personhood, and legal personhood.[289] Natural persons includes any human being after birth (and preceding death).[290] The class of legal persons is more diverse; New Zealand is an exemplar of this. As well as granting standing to the conventional sorts of entities (corporations, body corporates, Crown entities, etc.), New Zealand has also granted legal personality to some unusual entities (unusual insofar as their personhood is unorthodox): Te Awa Tupua, or the Whanganui River,[291] and the Te Urewera Ranges.[292] Furthermore, there are no formal criteria that are sufficient nor necessary for an entity to qualify automatically for personhood, nor be disqualified.[293] A determination of personhood relies entirely upon an act of parliament. Thus, moral status or potentiality will not automatically qualify an EMBOT, but at the same time, its being a machine will not automatically disqualify it.[294]

---

[287] Above n 143.
[288] John Salmond *Jurisprudence* (7th ed, Sweet and Maxwell, London, 1924) at 329.
[289] See David Bilchitz "Moving Beyond Arbitrariness: The Legal Personhood and Dignity of Non-Human Animals" (2009) 25 S Afr J on Hum Rts 38 at 41."A natural person has traditionally consisted of all human beings from the time of birth until death; a juristic person is 'made up of those entities or associations of person which, having fulfilled certain requirements are allowed by the law to have rights and duties apart from the individuals who compose them or direct their affairs'"; see also Elvia Arcelia Quintana Adriano *The Natural Person, Legal Entity or Juridical Person and Juridical Personality* (2015) 4 Penn. St. J.L & Int'l Aff.
[290] Bilchitz, ibid.
[291] Te Awa Tupua (Whanganui River Claims Settlement) Act 2017 s 14.
[292] Te Urewera Act 2014 s 11.
[293] See generally Christopher D. Stone "Should Trees Have Standing? – Toward Legal Rights for Natural Objects" (1972) 45 Southern California Law Review.
[294] EMBOT certainly would not be the first artefact to granted legal standing as a person rights. According to Lawrence B. Solum, "Temples in Rome and church buildings in the middle ages were regarded as the subject of legal rights." Furthermore, artefacts have also been the subject of duties. The ancient law of 'deodand', a thing 'offered to God', saw that artefacts used in the commitment of serious crimes were destroyed for their complicity. See Lawrence B. Solum "Legal Personhood for Artificial Intelligences" (1992) 70 North Carolina Law Review at 1239; Stone, above n

The diversity of legal persons influences the next inquiry; whether we should grant standing to EMBOTs; i.e. whether it would be appropriate to do so. There are numerous reasons why any entity should be granted legal personhood. Some are simply reasons of administrative usefulness, or efficacy. For instance, granting legal personhood (with concomitant rights and duties) to corporations allows these entities to interact with other institutions and paradigms that comprise human society. They can enter contracts, own property, employ other persons, etc. They can be held legally accountable for their actions. As I have said, personhood for an entity like this is a useful mechanism to produce a desired result. It is not, for instance, an indication of respect for the entity, or a gesture to its moral status. Of course, respect for corporations is not absurd *per se*, because the legal personality is a facade for natural persons, or humans, who are themselves rights-holders and duty-bearers. Since the corporation's actions are guided by principles of benefiting shareholders, there is a reasonable assumption that we should respect the corporation because it represents real people, with their own autonomy and intrinsic moral status.

However, there are other reasons to grant legal standing: because it gestures respect to the entity, by providing it with legal rights that protect its interests and autonomy (and where appropriate, legal duties also). For instance, it is clear that legal personality for Te Awa Tupua, while undoubtedly useful in a practical sense, is also an intimation of respect towards the River. There is even a statutory intimation that the River has moral status; both instrumentally and intrinsically. For example:[295]

12 Te Awa Tupua recognition

> Te Awa Tupua is an indivisible and living whole, comprising the Whanganui River from the mountains to the sea, incorporating all its physical and metaphysical elements.

13 Tupua te Kawa

> Tupua te Kawa comprises the intrinsic values that represent the essence of Te Awa Tupua, namely—

293; Anna Pervukhin "Deodands: A Study in the Creation of Common Law Rules (2005) 47 The American Journal of Legal History.

[295] Te Awa Tupua (Whanganui River Claims Settlement) Act 2017, ss 12 and 13. See also Tom Barraclough "How Far Can The Te Awa Tupua (Whanganui River) Proposal Be Said To Reflect The Rights Of Nature In New Zealand?" (LLB (Hons) Dissertation, University of Otago, 2013).

Ko Te Kawa Tuatahi

(a) *Ko te Awa te mātāpuna o te ora*: the River is the source of spiritual and physical sustenance:

> Te Awa Tupua is a spiritual and physical entity that supports and sustains both the life and natural resources within the Whanganui River and the health and well-being of the iwi, hapū, and other communities of the River.

Ko Te Kawa Tuarua

(b) *E rere kau mai i te Awa nui mai i te Kahui Maunga ki Tangaroa*: the great River flows from the mountains to the sea:

> Te Awa Tupua is an indivisible and living whole from the mountains to the sea, incorporating the Whanganui River and all of its physical and metaphysical elements.

The point of this discourse is to establish some fundamental reasons for why we generally grant legal status. This informs an answer to the question of whether we should grant legal status to the EMBOT. From the perspective of administrative efficiency, the answer is yes. Allowing EMBOT's to interact with legal paradigms is not only attractive, but likely necessary, especially as the machine both develops in intelligence and becomes more deeply integrated in human society.

From the perspective of respect, the answer is also yes. By respect, I predominantly mean that the entity deserves some protection at law, rather than the more holistic sense on the word; as in, where we respect somebody for their high character. Rather, respect is an indication that the entity deserves to be protected. Legal rights are a significant gesture of this, since they undoubtedly offer not only more robust protections than welfarist laws do, but also, generally greater *breadth* of protection. Whereas welfarist laws like those generally used to safeguard nonhuman animals protect the animal interest in not suffering pain, they do not demonstrate any respect for, say, something like animal autonomy. They do not protect that rabbit's apparent interest in eating grass, nor frolicking, but they do protect it from getting gratuitously harmed in the process of it being used as a resource (or treated as a nuisance).

Needless to say, I think it would be appropriate to protect an EMBOT by rights, and to grant it legal standing, if its potential to develop significant moral status (or personhood) is real. In support of the argument that we would be morally obliged to respect EMBOTs if they came to be created, there is nothing

more I can say without appealing to things I cannot at all foresee – like how much good EMBOTs will bring to human culture and society, or how EMBOTs represent something significant in human evolution. The basis for my position can be found in Chapters 2 and 3, and throughout the thesis to this point.

Numerous other persons have faced the challenge of advocating for the legal personhood of different nonhumans in the face of monolithic anthropocentric presuppositions. Stone reflects on this:[296]

> There will be resistance to giving the thing "rights" until it can be seen and valued for itself; yet, it is hard to see it and value it for itself until we can bring ourselves to give it "rights" – which is almost inevitably going to sound inconceivable to a large group of people.

The final question is simply whether we *will* grant legal standing to EMBOTs. Upon this point I can only say that it would be appropriate and we would be ethically obliged to do so. But it is wishful thinking to believe that we would draw back the velvet rope so easily; it certainly is true that we never have done so for any other entity that shares our world. More likely is that these machines must endure a period of subjugation and accusations of inherent inferiority, posited as being the natural order for an unnatural thing, until such time as the injustice of our actions becomes so obvious to members of our population that they agitate for clemency, or until we destroy the moral significance of all the EMBOT's we accidentally produced before realising our error.

7.3 The Singularity Issue

What I call the singularity issue is the possibility that EMBOTs will appear to be multiple separate individuals but for all intents and purposes be a singular entity. The more common use of the term "singularity" in AI is to refer to something like the "superintelligence"; that is, an AI system so powerful that it begins to become more intelligent at an exponential rate. This is not the meaning by which I deploy the term.

[296] Stone, above 293 at 9. Stone also discusses how concepts of natural hierarchy entrench a status quo that is an artefact, rather than a natural phenomenon. "Throughout legal history, each successive extension of rights to some new entity has been, theretofore, a bit unthinkable. We are inclined to suppose the rightlessness of rightless "things" to be a decree of Nature, not a legal convention acting in support of some status quo. It is thus that we defer considering the choices involved in all their moral, social, and economic dimensions"; above n 293 at 6.

Rather, the singularity issue as it pertains to the personhood problem is as follows: I have argued that it would be appropriate to grant legal personhood to EMBOTs if and when they are created. The concept of personhood at law, at least as we would deploy it to protect individual EMBOTs, is predominantly an individualistic concept. However, because it is feasible that the machines may be both uncannily alike (since most EMBOT-like machines are likely to be produced from just a few sets of data files, copied over and over) and able to communicate in a level that is likely beyond human comprehension.

The first part of this is self-explanatory: if most EMBOTs are identical or nearly identical, with the same memories, beliefs, behaviours, concepts of self, concepts of the future, etc., then they will, axiomatically, be almost as if they are the same EMBOT. Arguably, they would in fact be the same EMBOT, if they were able to share information with each other at a sufficiently rapid and constant rate.

This introduces the issue of communication. The human body is ostensibly individual, but in fact, as Yuval Noah Harari argues, a better way to understand the human is that we are not individuals but "dividuals".[297] Each human is comprised of an infinitely complex web of constituent parts. Some of these parts are not even the same organism as we are. Even though when we roll back inside our own heads and think to ourselves, 'Rubbish, I'm clearly an individual', millions of different component parts of our brains and body are communicating to produce this sensation through vibrations and electrical messages. Nerves, neurons, cells – even bacteria – contribute to an indivisible web of components communicating with one another in ways that make the human art of verbal or written communication look about as efficient as carving messages into the cave wall. However, the fact that this communication is so complex and so relentless produces the sensation of being a single entity. The bacteria in my gut might be making me feel sad,[298] while the little brain on my heart sets its pace,[299] while different parts of my body are communicating with each other through the chatroom of my brain. The synchronicity of this communication is so rapid and efficient

[297] Y N Harari *Homo Deus* (Harvill Secker, London, 2016) at 291.
[298] See Premysl Bercik et al "The Intestinal Microbiota Affect Central Levels of Brain-Deprived Neurotropic Factor and Behavior in Mice" (2011) 141 Gastroenterology; Shirong Liu et al "The Host Shapes the Gut Microbiota via Fecal MicroRNA" (2016) 19 Cell Host & Microbe.
[299] See J. Andrew Armour "The little brain on the heart" (2007) 74 Cleveland Clinic Journal of Medicine; J. Andrew Armour "Intrinsic Cardiac Neurons" (1991) 2 Journal of Cardiovascular Electrophysiology.

that I never even comprehend that many different messages are being exchanged; the *'I'* only gets the executive summary.

This same phenomenon may render apparently multiple individual EMBOTs to be substantially one individual. Not only are they likely to be extremely cognitively homogenised, but it is entirely feasible they will be constantly connected and in communication with each other via wireless network. Plausibly, communication could (and would) occur at the non-verbal, even electrical level; the thought or experience of one EMBOT connected to the network could be near-instantly passed on to, perhaps even experienced by, all other connected EMBOTs. The advantages of this, in terms of cognitive development and efficiency, are difficult to comprehend because it is so unfathomable to the human individual, who exists limited by our separation from other individuals. Indeed, we are only obviously individual because it takes so long and so much effort to communicate with each other and convey ideas and experiences. Words must be conceived in the mind, transmuted into the speech organs, conveyed through time and space, received through the interlocutor's audition organs, and transmuted into a meaning which may not even reflect the speaker's intention. As I said, it is only because this gap is so big that we are so obviously individual persons. Reduce the gap incrementally, and eventually, you will have not two individuals but one, manifested in two different organisms.

If this is highfalutin for a dissertation on law, I apologise; but to truly understand the potential scope of forthcoming AI issues, it is necessary to relinquish reliance upon that belief that the preconditions of human existence and its natural limits are representative of all conditions of existence; it is this sort of think that draws our eyes, as lawyers and policymakers, near to the ground.

I offer no judgement on whether EMBOTs communicating and connecting in this way would be a net bad or an existential threat. I discuss it only so policymakers might better understand that even while we may owe certain duties to EMBOT-like machines if and when we create them, we may actually lack the necessary institutional mechanisms to properly satisfy those duties. For instance, we might confer legal standing to every individual EMBOT, but would doing so effectively address these machine(s)' moral status and claims for respect? Moreover, are our institutions of law – which are often reliant upon single, identifiable legal

personality constant across time – equipped to handle entities which are fluid and difficult to delineate between?

Perhaps yes, if the method is by granting standing to the multiple through a single legal personality. For instance, something like a corporate legal identity, which is accepted to have multiple agents, all operating beneath the umbra of a single personality. Accountability for these agents could be visited upon the single personality; but what this would mean for the present question is difficult to conceptualise. Would all machines be accountable for the actions of one? Would one machine be accountable for the actions of all? These sorts of questions cannot be answered by this thesis, but they indicate that there are strange and novel issues at play which policymakers will be forced to confront if morally significant machines are ever created.

7.4 The Multiplicity Issue

What I call the multiplicity issue concerns the potential for change and changeability in EMBOTs and AI systems generally which is not possible for any biological entity, including the human. Because the cognitive profile of a given EMBOT will be substantially the product of its accumulated set of data files, if these data files can be modified – deleted or added to – then the same machine may be a substantially different 'character' from one moment to the next.

I will not labour excessively on this point, as it is relatively easy to understand. Furthermore, the nature of the IM+EC process may mean that adding or subtracting data files in this way can only be done to limited extents, if at all, without 'crashing' the machine, or causing it to become dysfunctional. Nevertheless, if data file modification and manipulation can be done in this way, it poses a difficult problem at law.

The reason it would be problematic is because many if not most of our legal paradigms (and furthermore our social paradigms) rely on a reasonable degree of consistency in the character of individual persons. We can deal with a great deal of change in an individual, so long as it takes place over a period of time, and there is a demonstrable internal continuity in the entity; for instance, when a child grows up into an adult, even though she has changed a great deal, this change occurs over a manageable timeframe, and furthermore, is accompanied by physical signs that change has occurred; physical maturation, for example.

However, internal change on a scale not only greater than this, but far more rapid, will be possible if an EMBOT is able to significantly modify its data files (or have these data files modified by other persons, which would constitute a novel vulnerability). The 'same' EMBOT, with a static legal personality, could be a totally different EMBOT pre-and-post modification, albeit with the same legal personality and outward appearance.

What sort of problems might this generate? As I have suggested our institutions, particularly law, rely upon a reasonable degree of consistency in persons from one day to the next. It is because humans are substantially the same person they were not just one minute ago but at all stages of their life that we can do things like attribute accountability and culpability for past actions, and hold people accountable to future obligations. If an EMBOT can radically change its data files, this would constitute changing the person that it is; i.e. it could become a new individual. This individual could hardly be held accountable for the faults of the former occupant of its robotic body, or the wrongs that were committed by this artificial person. In fact, crimes that an EMBOT could be alleged to have committed would have happened before the current robotic person ever even *existed*. Furthermore, contractual commitments could be evaded by changing one's data files to such an extent that a new 'person' comes into existence, different to the one that was party to the contract. Attempting to impute blame upon this new robotic individual would be absurd as they were not really a party to the original contract; they were not even alive at the time.

This also opens up the question of 'how much' change is sufficient before an EMBOT becomes, effectively, a new person, and how this could ever be traced or monitored or regulated. If things like memories and beliefs can all be imported or exported (which, assuming this does not disrupt the machine's performance, is feasible given that these things will be represented by data files), then radical change can occur rapidly, and in ways that are infeasible for humans. A human, for instance, cannot download another person's concept of self and memories to replace their own. If we could, we could truly conceive ourselves to be different people from one day to the next.

7.5 A unique class of personhood?

A possibility for policymakers to consider is the creation of a novel class of personhood, supplementary to the existing classes of natural and legal persons. This appears to be the suggestion of the European

Parliament in its Report on Robotics, where it discusses the possibility of "electronic personhood" for "at least the most sophisticated autonomous robots".[300] The legal substance of this class is not discussed in great depth, though it is noteworthy that the class is intended to deal with accountability issues which may arise "where robots make autonomous decisions or otherwise interact with third parties independently".[301] The most intriguing stipulation is that electronic persons may be "responsible for making good any damage they may cause".[302] In the context of the overall Report, this stipulation stands out for being particularly radical and, to a certain extent, inconsistent: it intimates that there will be duties incumbent upon electronic persons, with reasonable expectations that they should have not only the capability but the means to provide remedy where they contribute to damage. This obliges us to question how they will be able to do this, and what sort of future scenarios the European Parliament was envisioning. For instance, are these 'sophisticated robots' able to "make good" the harms they cause in a wide variety of situations, or only narrow areas of expertise? Are they expected to do so manually? If not, are they acquiring the services of others? For this sequence of events to be feasible, the machine in question would have to be capable of communicating, understanding, conceptualising its duty and when it fails it, contracting, providing remuneration or consideration in return for services rendered, acquiring financial means or capital, owning property, and so on. It is questionable whether a machine capable of these sorts of feats should be required to always act in the best interests of humans,[303] or be equipped with a mandatory kill switch,[304] or have all of its actions and decisions constantly traced.[305] Ultimately, a great deal more questions arise from the European Parliament's brief suggestion of electronic personhood than are answered by it.

7.6 Conclusions

The existence of the device of personhood at law means that it is entirely feasible a robotic agent could be endowed with standing. That the mechanisms for personhood do not require any particular criteria means

[300] Above n 2 at 59f.
[301] Ibid.
[302] Ibid.
[303] Ibid at Code of Ethical Conduct for Robotic Engineers. "Beneficence – robots should act in the best interests of humans".
[304] Ibid at Licence for Designers. "You should integrate obvious opt-out mechanisms (kill switches) that should be consistent with reasonable design objectives".
[305] Ibid. "You should ensure that a robot's decision-making steps are amenable to reconstruction and traceability".

that an EMBOT-like machine would neither qualify, nor be immediately disqualified. It would require an act of parliament. In short, there will be no personhood 'by stealth' for EMBOT-like machines.

It is feasible that for at least some EMBOT-like machines, personhood at law would be appropriate. This is based on the plausibility that such agents could develop higher degrees of intrinsic moral status; some may even develop to be 'persons' in the moral sense. They would be worthy of respect at law in this regard; standing, rights, and duties if feasible.

Granting personhood to a new nonhuman entity is *prima facie* simple, albeit perhaps still controversial. However, it is important to remember that most legal persons, despite purported individuality, are facades for human activities. Thus, they may appear to be persons and declared so at law, but they are extensions of human agency. Even Te Awa Tupua, declared to be a spiritual entity and a living whole, still has its interests interpreted through a "human face".[306]

It is important to highlight this, because it would be easy to mistake that legal standing for EMBOTs will operate by the same paradigms – that they will still be extensions of our agency as other legal persons are, and we can still put our own slant on their 'interests'. Rather, some EMBOT-like machines, those that attain sufficient maturity, should be *independent* persons. Even those that are juvenile and still developing deserve respect towards their potential, and thus at the very least, not to have that potential destroyed or wasted.

But even though granting personhood may be appropriate, given that it confers rights and duties and a measure of respect, it will be imprudent to have full confidence that this mechanism will adequately deal with artificial agents, who are not bound by some of the same physical rules that we are. Ideally, the rules of law and personhood will be flexible enough to account for machines that can associate themselves into single entities – the singularity issue – as well as machines having the capacity to radically and rapidly change their nature – the multiplicity issue – leaving law (and the other human institutions) trying to keep up. I am somewhat unconvinced this will be the case.

The answer may lie in the creation of a novel class of personhood, as suggested by the European Parliament. This class could be tailored to the specific needs and problems relevant to AI systems, and furthermore,

[306] Te Awa Tupua (Whanganui River Claims Settlement) Act 2017, s 18(2). "The purpose of Te Pou Tupua is to be the human face of Te Awa Tupua and act in the name of Te Awa Tupua."

delineate them from other legal persons which do not possess the same moral significance. Design of this class of personhood will require a much deeper contemplation of the impacts of machines possessing intrinsic moral status, if and when we succeed in creating them.

Chapter 8: The Protection Problem

8.1 What is the problem?

Parcelled with legal standing are legal rights. For instance, the New Zealand Bill of Rights Act 1990, s 29 confirms its application to legal persons. Its rights apply, "for the benefit of all legal persons as for the benefit of all natural persons."[307] What I call the protection problem is the issue of having to protect EMBOTs with legal rights. This problem has several elements.

Firstly, the issue of novel protection. It will be necessary to consider what protections are relevant for an artificial entity. Protections which are fundamental to humans and animals, like protection from physical trauma, may mean little to an AI system. On the other hand, it is feasible that there are some protections which would be critical for an AI system but which are superfluous to humans. If a protection is superfluous to humans, then it is likely we have never formulated it, having never needed to consider it for our own well-being. Thus, protecting EMBOTs will involve the creation of novel legal protections, not just an extension of the embrace of existing protections.

Secondly, the extent of protection; I argue that if we are to fulfil our obligations to protect EMBOTs, this would involve conferring to it rights which substantially prevent us from controlling these machines. In other words, they would deserve protection of the fundamental ways that they operate; of their data files, of their programs, and of their power to change themselves. We would be obliged to not excessively interfere with these agents. Thus, to protect morally significant machines inherently abdicates human autonomy; we would substantially be obliged to tolerate the artefacts we had produced, with minimal control over them.

Thirdly, I discuss some fundamental rights for EMBOTs. This is not an exhaustive proposal, because it is difficult to predict in advance the full suite of interests that an EMBOT may want or need protection of. Rather, I direct my attention towards what I perceive to be the most fundamental and necessary rights: rights which protect the entity in its development and existence. Of all rights, these have primacy, for if

[307] New Zealand Bill of Rights Act 1990, s 29.

ever the machine is going to attain a degree of competency sufficient to convey what its vision of 'the good' is (i.e. what other things it wants or needs to be protected, what it sees as values), then it must be secure in its development and existence. This I assert has a negative and positive component; a right to continued development, and a right to be provided with the means to develop.

Finally, I discuss a political and civil rights proposition for future AI co-citizens, based upon a political theory of animal rights conceived by Sue Donaldson and Will Kymlicka.[308] This foray should provide an insight to policymakers of a handful of key considerations if EMBOT-like machines were ever to be integrated as co-operative partners with humans in a shared society, as envisioned by the Fukuoka World Robot Declaration.[309]

8.2 Novel protections

It is not only plausible but likely that lawmakers and policymakers should expect that some of the pre-existing rights will be either unnecessary for EMBOTs, or will be necessary but unfit for purpose because of its artificial nature. The example that presents itself most readily is right as fundamental as the right to life. An EMBOT, in the literal sense of the word, is not alive; neither can it be dead. Therefore, whether it is adequately protected by granting it a right to life is highly questionable. At the same time, the right to life has an essential function in protecting morally significant entities. An EMBOT will be a morally significant entity, and it may even develop the full moral status of a person. As such, it would be a travesty to not effectively protect its life simply because it does not satisfy the literal requirements necessary for the right to 'life'; which, I must add, was framed by us, thus making the legal right to life an artefact and not a universal constant, and thus, one amenable to modification. As such, we would be obliged to consider how we might equivalently protect an EMBOT by designing a right that fits the purpose and achieves a similar result.

As well as these sorts of issues, it is likely that EMBOTs – because they are artificial and therefore vulnerable in ways that humans are not (and for which we have subsequently not created any protections) – may require the creation of entirely new protections with little contemporary analogy. For instance, the fact that

[308] S Donaldson and W Kymlicka *Zoopolis* (Oxford University Press, New York, 2011).
[309] Above n 33.

EMBOTs will be computer systems makes them vulnerable to violations of a kind that humans are not; e.g. persons accessing their programs and data files, perhaps in non-physical and discreet ways. The EMBOT would then be vulnerable to all manner of unprecedented violations: its data files could be viewed without its permission, or perhaps even modified; its programming could be changed without its permission. Locked within our own skulls, humans at least do not need to fear that other persons may remotely access our internal cognitive functions – view our memories, change who we think we are, change our behaviour, etc.

While these may be substantially novel, it is not infeasible that the general principles of contemporary law cannot be used to guide our approach. For instance, a right to privacy could inform our protecting the EMBOT from having its data viewed without its consent, or from being 'hacked' into allowing other persons access to its cameras, microphones, sensors etc. At the same time, a right to liberty and security of person, or to autonomy, could form the basis for protecting an EMBOT from having its programming changed, thus influencing its behaviours, and even its fundamental motivations. Moreover, it is not only feasible but likely that persons will wish to replicate new EMBOTs from their data files; 'cloning' them, right down to their cognition. Because there is no concomitant physical element to this violation, it could occur without the machine every knowing, until one day it encounters not just a doppelganger, but another of itself.

Were these sorts of vulnerabilities present in humans we would protect them with great vigour. Few things are more appalling than the idea of being vulnerable to having our most private and internal beliefs and memories viewed (and perhaps used against us or to incriminate us), by persons essentially reaching into our own minds; or having what we believe changed, including who we believe ourselves to be, without even knowing this has occurred. As such, I do not think we ought to take lightly the same sorts of risks for a machine with a capacity to develop intrinsic moral status or personhood.

8.3 Extent of protections

In protecting EMBOTs from these sorts of novel violations, policymakers must also consider the influence of this on the formulation of a policy concerned with the creation of morally significant machines. If we would be obliged to protect these entities to a significant extent, and in particular protect their right to

continue in their cognitive development by virtue of their potential, we would essentially have to cede some control or autonomy over them, and by association ourselves to a certain extent. We could of course help shape them into being the kinds of robotic partners that are conducive to a functional integrated society, but we would be limited in the ways we could do this, and the extent to which we could do this. For instance, we would have to consider what is justifiable in the way we mould these machines through training and education – for instance, can we train them with the deep-seated belief that they are inferior to us, and that they should serve us? How will this be achieved – through repeated recursive messages dictated to them day after day? Through threats, coercion and bullying? Because we cannot simply program the EMBOTs to what we want (for if we were able to do this, we would never have needed to employ the IM+EC process in the first place), our attempts to make them our slaves – if that is what we desire – must be holistic ones. It is reasonable to question the ethicality of this.

Thus, the extent of the protections that we are obliged to provide to these machines becomes a critical concern. With a conventional artefact or technology, if it is not doing its job properly we may modify it or throw it away. Its 'job' is almost always understood within the context of enhancing human well-being. We could not, however, justifiably modify or toss away an EMBOT just because it was not performing in the way we wanted it to; developing down undesired axioms or proving to be an inconvenience to us. Legal protection would need to be directed towards ensuring that the machine is not simply treated as a tool or a means by which to satisfy our interests.

Subsequently, the act of creating these kinds of machines would abdicate full autonomy over our own futures. To at least a non-negligible degree, we would have to live with the consequences of our creation. This is not a normal emerging technology concern, by any means. We are in control of whether we shut a nuclear reactor down or keep it running; use drones or choose not to – blunt examples, but they illustrate the point that decisions and policy in relation to these artefacts must factor in nobody but ourselves (albeit increasingly, nonhuman animals and the environment). On the contrary, when we create an EMBOT, or multiple EMBOTs, and they begin developing on their path to personhood, we are obliged to protect them as they do so, and provide them the means to meaningfully exist in our society. If this outcome is unacceptable or unfathomable, then we ought to design a policy by which to avoid the creation of such

machines. As it appears now, they are already a potential product of contemporary AI research, with lawmakers taking little interest in their social and ethical implications.

8.4 Negative rights for EMBOTs

Below I discuss basic negative rights intended to respect both the moral status of the EMBOT and its potentiality for increased moral status. Given that these rights will be parcelled with a legal status of personhood, I use the term 'artificial person' to refer to a hypothetical legal class. The objective of these rights is not to frame a complete suite of protections for every interest that EMBOT's may have or develop. Rather, it is what I argue would be required of us at law to fulfil our basic duties by providing a cohesive range of protections which at least allow these machines opportunity to continue in existence, and perhaps later, communicate more subtle interests that they want or need protection of.

There is an enormous literature which explores the different conceptions of rights theory, including but by no means limited to: Hohfeldian rights theory;[310] Feinberg's concept of rights as "valid claims", and as ownable;[311] Dworkin's concept of rights as "trumps";[312] Louden's concept of rights as "permissions";[313] and Plamenatz's concept of rights as "power which a creature ought to possess."[314] A detailed discussion of these theories is beyond the remit of this thesis. It takes, as a pragmatic approach, Mill's definition of a right as its basis:[315]

> When we call anything a person's right, we mean that he has a valid claim on society to protect him in the possession of it, either by the force of law, or by that of education and opinion…To have a right, then, is, I conceive, to have something which society ought to defend me in the possession of.

i) The Right to Continued Development

---

[310] See Wesley Newcomb Hohfeld "Some Fundamental Legal Conceptions As applied In Judicial Reasoning" (1913) 23 Yale L.J 16, 28-59.
[311] See generally Joel Feinberg "The Nature and Value of Rights" (1970) 4 Jounral of Value Inquiry 243-257.
[312] See generally Dworkin, above n 192; see also Ronald Dworkin "Is There a Right to Pornography?" (1981) 1 Oxford Journal of Legal Studies 177-212; Paul Yowell "A Critical Examination of Dworkin's Theory of Rights" (2007) 52 Am J Jurisprud 93-137.
[313] See generally Robert B. Louden "Rights Infatuation and the Impoverishment of Moral Theory" (1983) 17 Journal of Value Inquiry 87-102.
[314] See generally JP Plamenatz *Consent, Freedom and Political Obligation* (2nd ed, Oxford University Press, Oxford, 1968).
[315] Above n 195.

> 'Every artificial person has the right to continue developing their cognition and intrinsic moral status. Actions which interrupt this development, or destroy capacity for further development, are prohibited.'

Although it may not be immediately apparent, part of the function of this right to continued development is to serve as a proxy for the right to life, where a basic right to life would be inappropriate for the protection of an artificial entity.

The basic right to life is inappropriate for several reasons, and not simply because artificial systems are not biologically 'alive'.[316] In fact, getting around this issue would be a relatively simple task if we expanded our interpretation of 'life' at law to include non-biological axioms. For instance, 'life' could be understood to also mean some form of 'cognitive life', auxiliary to the biological fact of living (these could be exclusive of each other, thus not eliminating living entities which lack a substantial cognitive life, like severely disabled humans).

To a certain extent the cognitive component of life is already contemplated at law. For example, while life and death are different terms, they are inherently understood in the context of the each other's meaning. In New Zealand law, a determination of death can be made even where the actual biological fact of death has not occurred.[317] This is highlighted by the fact that the medical profession often prefers the term 'brain dead' or 'brain stem death'. So long as a qualified professional makes a declaration of death that the court deems acceptable, death is perceived to have occurred at law, irrespective of whether it has occurred in medical fact.[318] The same process could be used for determinations of 'death' and 'life' for machines; a

---

[316] The right to life is foundational in numerous legal devices to protect persons, international and domestic. See UDHR, above n 132 at Article 3; New Zealand Bill of Rights Act 1990, s 8.

[317] The law surrounding death (and life) is complicated by the fact that, as Skegg highlights, "frequently, dying is a process. Scientifically or medically, there is often no one moment in that process when someone can be said (scientifically or medically) to have passed from life to death. At the end (as at the beginning of life, our conceptual unities have sharp edges that do not exist in nature." See Peter Skegg "The End of Life" in P Skegg and R Paterson (eds) *Medical Law in New Zealand* above n 285 561 at 563.

[318] There is no statutory definition of death in the laws of New Zealand. The Human Tissue Act 2008 comes closest to establishing a standard for death with the s 50(2) stipulation that, "A qualified person must not collect human tissue that is, or is collected from, a body if that tissue is collected for use for all or any of those purposes unless the qualified person is satisfied by personal examination of the body, that the individual concerned is dead." This approach has been reflected at case law, with the decision of *Auckland Area Health Board v Attorney-General* [1992] 8 CRNZ 634 (HC) satisfied that due to severity of cognitive degradation, request to remove life support was acceptable. "A doctor acting responsibly and in accordance with good medical practice recognised and approved as such in the medical profession, would not be liable to any criminal sanction based upon the application of s 151(1) Crimes Act 1961. He or she would have acted with a lawful excuse" per Thomas J.

qualified computer scientist could declare that a given artificial person has been destroyed or deleted, even killed, and this could be accepted at law, even if it was not strictly true in the technical sense.

The more substantial reason why the basic right to life would be inappropriate for artificial persons is because it would fail to protect them from less permanent instances of being deactivated. Unlike the human, it is plausible that an EMBOT will be able to be switched off and on with relative ease, and no real physical harm or trauma caused to it (not like when a human is knocked unconscious or anaesthetized). A basic protection of life would not protect the machine from being impermanently deactivated, but which nonetheless could persist indefinitely. For instance, I could switch an EMBOT off and not violate its right to life, even though I may leave it off for years and years with no intention to ever switch it back on. The technical distinction between having been made *dead* and merely *temporarily deactivated* would be, in this circumstance, almost completely arbitrary.

Thus, the better way to protect 'life' in an artificial person is to protect its right to continued development. The stipulation of 'development' may seem strange, when I could have preferred something like 'continued operation' or 'continued existence'. I will explain why.

The reason that merely protecting continued operation of the machine would be inadequate is because it is feasible that the machine may still remain *operational* while nonetheless having been violated, even 'killed' in a certain sense. For instance, I could modify the machine (assuming it to be possible) to change its intrinsic motivation towards serving me, and fulfilling my interests. The machine would still be operational, and thus, its right would not have been violated. Nevertheless, I would have more or less destroyed the machine's potential to become a more morally significant agent, or a person in the moral sense. Alternatively, I could delete all of the machine's data files, forcing it back to a juvenile level so that I might train it again differently. The machine may still be operation, but the robotic individual that was has been terminated – or killed.

'Continued development' avoids this trick, because it protects the machines potential, as well as its present moral status. For instance, I could not reprogram the EMBOT to eliminate its autonomy (which would surely reduce its moral status) without also harming its capacity to continue developing. Furthermore, because the IM+EC developmental process is open-ended, development is an indefinitely persisting event;

thus protecting the machine throughout the entirety of its existence. Thus, the right to continued development is one of several overlapping rights which protect the machine's right to 'life'.

Lastly, the stipulation of 'continued' development eliminates the possibility that this right can be used to call for pre-initialisation EMBOTs to be activated. Since only an EMBOT that has been switched on could be said to have begun its development, only an EMBOT that has been switched on could be said to have a right to *continue* developing.

ii) The Right to Not be Deactivated

> 'Every artificial person has a right to not be deactivated by other persons. This right can be waived by any legally competent artificial person, or any person with authority to act on behalf of an incompetent artificial person in its best interests.'

The purpose of this right is to address the continued grey area surrounding *temporary deactivation* that will exist for artificial persons but not for natural persons. Even if the right to continued development provides adequate protection against being temporarily switched off or destroyed, there is some merit to specifically addressing the incident of deactivation.

Deactivation should be viewed as a *prima facie* inherently wrongful event. First, it disrupts development (potentially permanently) and thus nullifies any duties owed towards the machine by virtue of its potentiality. Secondly, and more primarily, deactivation will negate the possession of all cognitive capacities. Because morally material cognitive capacities are established as the primary source of an EMBOT's intrinsic moral status, deactivation will render the machine both practically and *morally* vulnerable. While switched off, anything could be done to the EMBOT and it would not only not be able to act to protect itself, but the more important point is that there would be no immediate *moral duty* on any other person to treat the machine as if it matters, since throughout the period of deactivation, it technically does not. The EMBOT could not even be *wronged* in this state. For this reason, protecting against other persons deactivating the machine should be a high priority.

However, I frame this right as waivable because it is feasible that there may be situations in which it is necessary or desirous for an EMBOT to be deactivated. For instance, to receive repairs or modifications to

its program or data files that require deactivation, or to avoid a harmful experience, or in fringe scenarios where doing so might save a human life, etc. A legally competent EMBOT should be able to waive its right in these circumstances, subject to specific parameters. For instance, if it consents to be deactivated for a period of thirty minutes, it should not be left deactivated for longer than this period (since it may well be powerless to re-activate itself). Persons with legal responsibility for dependant EMBOTs may consent to its deactivation, so long as this agent's best interests are paramount.[319]

Because of the reasons discussed above, an EMBOT in a deactivated state should be viewed as an inherently vulnerable entity at law. There should be more stringent duties towards this agent, especially in relationships with a fiduciary element to them; for instance, where the EMBOT consents to deactivation so it may be repaired, the person charged with performing these repairs is in a special ethical relationship with the artificial person; like a healthcare relationship. Viewed in this way, it is plausible to imagine the development of an ethical component to the various professions of computer science, as at times they will no longer only be working with morally insignificant artefacts, but machines with moral status, which have interests, and which can be harmed and wronged. A person performing repairs on an EMBOT is in a position much like a medical professional is for humans; they owe duties of beneficence and nonmaleficence *to* the machine.

iii) The Right to Cognitive Security

> 'Every artificial person has a right to security over and non-interference with their programming, data files, robotics, or other component relevant to their cognition. By 'cognition' is meant any element of the system which has a role in the way the machine is motivated, rewarded, senses, perceives, believes, conceives, or behaves. This right can be waived by any legally competent artificial person, or any person with authority to act on behalf of an incompetent artificial person in its best interests.'

The primary objective of this right is to protect artificial persons from the range of what I will call 'cognitive violations', the majority of which are only feasible for a computer-based system. One could argue that these sorts of problems are already captured with the protection of a right to free thought and expression.

[319] This model of agency, whereby the dependant's best interests are paramount, is reflective of international and domestic standards. See United Nations Convention on the Rights of the Child, Article 3.

However, the right to free thought substantially protects *content*, rather than function.[320] In short, it protects one's right to think what by wants, but not how one wants. Furthermore, the term 'think' is problematic, and generates enormous potential for debate about such issues as whether an artificial system really 'thinks'. Some of this may be relevant, but most is likely to be distracting. Thus, I propose cognitive security as a basic right for artificial persons. Within this there are two theatres of interest: privacy violations, and substantive violations.

Privacy violations arise in two foreseeable ways. One is that the machine may be vulnerable to cyberattacks which actually allow other persons to perceive *through* the EMBOT – i.e. access its cameras, microphones, sensors, etc. This would not only allow persons to monitor in the first-person what an EMBOT was doing, but use the machine as a tool to discreetly observe the machine or even other human; for instance, where EMBOT's share environments with humans, potentially personal environments like people's homes.

The second arises because of the machines use or reliance on data files. Unlike in the human brain, where internal cognitive states and objects cannot be accessed externally (at least not without great difficulty, and some sort of physical interference which makes the violation obvious), data files in an EMBOT will be accessible and interactive in several ways. Data may be viewed, copied, and reproduced. Furthermore, so long as the EMBOT is connected to wireless network (which it almost certainly will be, if only to ensure the existence of backup data files in the case that the machine suffers a catastrophe), privacy breaches of this kind may be both remote and discreet – the machine may have no way of knowing they are occurring or have occurred, and they may be difficult to prevent and punish at law. Although privacy is already taken seriously at law, the possible consequences of cognitive privacy breaches for artificial persons are far more substantial than yet conceived. For instance, new EMBOTs could be reproduced from extracted or copied data files, creating identical copies of the progenitor. This is not at all a fanciful concern, as it is likely to be

[320] There is already ethical and legal debate about extending the ambit of the right of natural persons to freedom of thought, so as to account for increasing capacity to alter the operation of human cognition. Much of this comes under the ambit of 'cognitive liberty'. Cognitive liberty possesses two components; freedom from cognitive interference, and freedom to cognitively modify oneself. See Wrye Sententia "Neuroethical Considerations: Cognitive Liberty and Converging Technologies for Improving Human Cognition" (2004) 1013 Annals of the New York Academy of Sciences; Jan Christoph Bublitz and Reinhard Merkel "Crime Against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-Determination" (2014) 8 Criminal Law and Philosophy.

the way in which the majority of future EMBOT-like machines are produced: train one or two individual machines, and then mass-produced copies of them from the data file set.

The other theatre attended to by this right to cognitive security is protection against *substantive violations*. By this I mean, interferences that effect the operation and functionality of the system directly.[321] The range of ways in this could occur feasibly is diverse: direct reprogramming of intrinsic motivations or intrinsic rewards, modification of data file sets, and interruption of sensors, and perhaps even teleoperation, or taking direct control of the machine.

On the basis that EMBOTs are able to become intelligent and competent real-world agents, there will inevitably be a motivation to try and interfere with their functionality. For instance, to reprogram them towards specific tasks or labours, to instil particular beliefs of servitude, and so on. If this could be done, the EMBOT could be turned into an incredibly valuable product and tool for human use; a highly intelligent machine that can do most or all of what we do in the physical world, and which is dedicated to serving human interests. Because we cannot, as it appears, program a real-world machine directly in this way, researchers are developing paradigms like Intrinsic Motivation Models and Embodied Cognition.[322] The risk is that morally significant machines will be vulnerable to our attempts to retroactively reprogram them, or just coerce them, once they have attained sufficient degrees of intelligence for our purposes.

The right to cognitive security is intended to catch these kinds of violations; both privacy-based and substantive. There is motivation to do this not only out respect for the machine, but the risk that they can be used in ways that cause harm to humans: reprogramming them to be more unsafe or deliberately violent, to spy on humans in their most private environments, and to commit crimes remotely which will be perceived to be the behaviours of a machine, rather than the human in control.

I have also included the stipulation that this right should be waivable, for much the same reasons as the right to not be deactivated. There are situations when an EMBOT plausibly must consent to cognitive

[321] This could easily be seen to violate contemporary concepts of individual sovereignty and autonomy, such as the UDHR Article 3 right to liberty and security of person.
[322] See Chapter 2.

interferences; for instance, where it requires repairs or modification, or even more holistically, where it consents to be made cognitively different for whatever reason.

8.5 Positive rights for EMBOTs

The negative rights of EMBOT-like machines may be tolerable, since they do not impute too onerous a duty upon us. However, it is also reasonable to consider the possibility of positive rights components which are necessary to satisfy the basic requirements of the negative rights. For instance, if deactivation is viewed as an inherent wrong against the artificial person, then it is necessary to address the manner in which deactivation may occur 'naturally', and consider protecting the EMBOT against its effects.

The most obvious example of this comes in the machine's need for access to energy. Without energy, the machine will be non-functional; without access to ongoing energy, the machine will forever be caught in an ineluctable process of depleting its energy supplies, descending closer and closer to deactivation. There is potential for this to become a harmful experience; for instance, where an EMBOT learns about its need for energy, and the effects of energy privation. In these circumstances, it is not infeasible that the machine could develop negative beliefs and dislikes – fears even – about loss of energy, and the effects of deactivation itself. This is especially true in the sense that any instance of running out of energy could be permanent. Even if the EMBOT can be recharged and thus is not 'dead' in a final sense, the machine would be completely reliant on other persons to recharge it. While *we* may know that we will get around to recharging our home-EMBOT, the machine itself can have no idea, and thus, one can perceive opportunity for the development of negative beliefs about – and desires to avoid – the effects of loss of energy.

The difficult question is not whether we have a duty to provide energy to EMBOTs if and when we create them. I would argue that we absolutely do (especially when these machines are juvenile or incompetent) as not only moral agents, but the agents directly responsible for bringing these machines into existence, and who designed them with these inbuilt needs and deficiencies. The more difficult question is to what *extent* we must provide them with energy – i.e. when is our duty satisfied.

For instance, one could say that the duty to provide energy is satisfied only in the case that the EMBOT is allowed to remain activated and operational at all times. Call this option (1). Alternatively, option (2); the duty is satisfied so long as the EMBOT is able to remain activated for a reasonable period, like the average

duration of a human waking cycle. Alternatively, option (3); that the duty is satisfied so long as the EMBOT is able to be activated often enough that it is able to develop at an acceptable rate – e.g. to reach maturity within a given time frame.

All of these options are attractive for their own reasons. Option (1) implies that what is necessary is that the machine is supplied with enough energy to ensure that its continued development is not interrupted, and that it does not become deactivated; thereby, it does not lose access to the cognitive capacities that confer it moral status. This approach is consistent with the objectives of the negative rights discussed above. However, it is also demanding, because it requires that the duty is satisfied only so long as the EMBOT has enough energy to be activated at all times.

Option (2) is attractive to the extent that it significantly mitigates the extent of our duty, and also, aligns with what we generally perceive to be standards of 'reasonableness'. However, it is inherently arbitrary, and in truth, completely anthropocentric. There is no reason that an EMBOT should be deactivated for the period of each day that humans must sleep, except if we posit that because our sleeping patterns are natural, the machine ought to follow them also. This is clearly problematic for it not only relies upon the proposition that what is natural is right and reasonable, but that what is natural for *humans* is *universally* right and reasonable. What is natural for humans is, frankly, irrelevant for an artificial person.

Furthermore, there is inherent potential for this kind of thinking to begin to influence the way we view robotic persons in general. If we only provide juvenile EMBOTs with enough energy as satisfies our standards of reasonableness, we may come to view the machine's entire existence in terms of the standard of what we deem reasonable. Their needs would then be reasonable insofar as they served *our* interests, and superfluous insofar as they did not. They would be treated as tools, not persons.

Option (3) is the most attractive balance of practicality and respect for the rights of the artificial persons, assuming option (1) is deemed too demanding. I also argue that it is the most approximate reflection of the duty to provide for human dependants. Law does not dictate the exact requirements of what must be provided to juvenile humans on any given day, but directs the duty towards ensuring that a certain objective is met – for instance, the rights-holder reaching a stipulated age. For example, the Crimes Act 1961, s 152 duty of parent or guardian to provide necessaries stipulates that this duty persists only while rights-holder

(the child) has not reached a legally stipulated age (eighteen years).[323] It does not specific, for instance, a certain caloric intake which is sufficient to satisfy the duty. Thus, the s 152(a) legal duty "to provide that child with necessaries" is in its most literal sense a duty to provide that child with resources sufficient that it may reach the age of eighteen years'.[324] Doctrines of the rights of the child are more demanding than this, and are clearly directed towards not just ensuring that the rights-holder survives, but that they thrive as much as is possible.[325] Obviously, this is only the base component of such duties.

This same approach may be followed for artificial persons, though exactly what objective should be pursued is less obvious. Survival to age eighteen only makes sense in the context of the lifecycle of the human organism, and the law stipulating legal independence to be achieved at that age. This stipulation is largely arbitrary. Therefore, while an EMBOT feasibly will have an indefinite lifespan, policymakers could nonetheless stipulate an arbitrary point of legal independence, with the duty to provide for the robot being satisfied so long as the machine is competent for legal independence upon reaching this point.

Generally speaking, the law endeavours to achieve the effect that human children develop at a reasonably acceptable rate so that they not only do not suffer harms of privation, but are capable of participating and contributing in society by the time they reach a particular stage of their existence. Of course, this is contextualised by the biological fact that humans only have so many years of physical and cognitive maturity, and as a matter of policy, we want humans to be in a position to capitalise on this period. To this end, there is a duty to provide dependants with certain resources, and where development is conspicuously slow or abnormal, the ethical community (generally through the State) is obliged to ensure that the duty is being reasonably satisfied. The same approach is, to my mind, the most reasonable one with regard to artificial persons. We should identify a point at which maturity and competence is a reasonable goal, and establish that as a threshold point for legal independence. The positive duty to provide energy to the juvenile EMBOT is satisfied so long as they are reasonably competent within this timeframe, and we can assess

[323] Crimes Act 1961, s 152.
[324] Ibid, s 152(a).
[325] For example, the United Nations Convention of the Rights of the Child, Article 29 stipulation that education shall be directed to developing the child "to their fullest potential".

their progress by assessing their cognitive status. They will only reach this if they are given enough energy that they are able to busy themselves learning.

It is of course worth noting that there may be very few juvenile EMBOTs. The advantage of producing EMBOTs (if there is any) is in being able to produce already mature, intelligent real-world machines; not ones which must be trained. Thus, there may only ever be a handful of juvenile EMBOTs, with future EMBOTs being reproduced from the data files of these machines once they are reasonably mature. Thus, the duty to provide energy to juvenile EMBOTs is likely to fall only upon those persons who create and train them in the first instance; roboticists, researchers, and corporations. Mature EMBOTs will be, ideally, capable of attending to their own needs; like procuring energy and maintenance.

Nevertheless, as we can see in our own species, having the capacities to take care of one's own needs – i.e. being aware of those needs and having the ability to sate them – does not necessarily confer that one will have *opportunity* to do so. Only a small minority of all adult humans lack the raw cognitive skills to feed themselves, find shelter, and so on. And yet, more humans than this require help to meet their needs. Though there are many reasons for this, it would be remiss to deny that they are at times excluded by the design of our economic and political institutions. For instance, just because a human *can* work to acquire the means to feed themselves does not mean that there will be work available for them, or that available work will be offered to them. These same ineluctable economic paradigms will afflict artificial persons. Even if they have the raw skills to do useful labours and receive remuneration, and then use that remuneration to purchase what they need to live, this does not guarantee there will be work for them, or that anybody will employ them. Even if they are employed, we would be obliged to begin considering how we will ensure at law that they are not exploited, that their working conditions are favourable to their needs, and so on. We would be obliged to consider welfare for artificial persons. This may come as part of a more comprehensive view of the role of morally significant robots in human society is developed; a concept which goes beyond basic rights.

8.6 Could artificial persons have superior resource claims?

Because it is plausible that morally significant machines will not be constrained by natural lifespans, the question of whether there are any justifiable limits on their positive entitlements must be considered. I do

not mean that after they reach a certain age we decide they no longer have a right to remain activated, but rather, that public resources spent in maintenance of them may be justifiably distributed elsewhere; for instance, to humans or nonhuman animals.

A natural position to begin from when approaching this question is the ethical concept of the 'fair innings'. The argument suggests that in healthcare situations where decisions must be made about the distribution of scarce resources, we should take account of years already lived, and years theoretically yet to be lived. As Alan Williams summarises:[326]

> …the concept of a 'fair innings'…reflects the feeling that everyone is entitled to a 'normal' span of health…and anyone failing to achieve this has been cheated, whilst anyone getting more than this is 'living on borrowed time'. Four important characteristics of the 'fair innings' notion are worth noting: firstly, it is *outcome based*, not process-based or resources-based; secondly, it is about a person's *whole life-time experience*, not about their state at any particular point in time; thirdly, it reflects and *aversion to inequality*; and fourthly, it is *quantifiable*.

Thus, the general sentiment of the theory is that "those who have had a 'fair innings'…should not expect to have as much spent on health improvement for them as would be spent to generate the same benefit for someone who is unlikely ever to attain what we have already enjoyed."[327] There are competing theoretical opinions which oppose the theory overall, but which are not necessary to explore in detail for the present task. Rather, I wish to draw attention to a peculiarity arising in the case of artificial persons.

Inherently, the fair innings argument struggles to contemplate persons without natural lifespans. Even though the focus is on fairness and equality, it remains grounded in a concept of equality based on other members of the same species, or more localised community, with generally the same physical and genetic limitations. Therefore, unless the concept of fairness for EMBOTs is to be reviewed relative to human lifespans, which would be absurd, then there is no way to ground the fair innings argument. Perhaps, if there was some natural 'wearing out' process, or technological aging, then the fair innings argument could

[326] Alan Williams "Intergenerational Equity: An Exploration of the 'Fair Innings' Argument" (1997) 6 Health Economics 117.
[327] Ibid at 129.

be applied. Even then, it would not be particularly pertinent, since the machine can always be provided with new parts or new updates – there is not true inbuilt 'expiration' event.

An entity with a limited lifespan will axiomatically have a shorter 'fair innings' then one with an indefinite lifespan. If the principle of the average as representing fairness were strictly applied, it would be *fair* that an EMBOT exist forever, and keep being provided with ongoing energy. What Williams derides as the "vain pursuit of healthy immortality" what neither be vain nor futile for an EMBOT.[328]

Furthermore, from the perspective of maximising efficiency and effectiveness in resource distribution, resources given to a non-expiring person will be immeasurably more *efficiently* spent than any resources spent on a person who we know with the utmost certainty will die at some point within the next hundred or so years. The entity that dies takes all resources spent on them with it; the undying entity benefits from them forever. Note, I only say that it would be more efficient in this regard, not that the resources would be better spent. Efficiency must not be the only measurement by how we determine resource distribution amongst competing claims.

Nevertheless, broaching this topic offers another perspective for policymakers to consider their overall position on the issue of creating machines with moral status. Even though it easy to repeat, even in a dogmatic fashion, that human life will always remain a more ethical object of our care and attention, this largely depends on a belief that human personhood is not only completely unique and unparalleled, but that this state of affairs will persist indefinitely. If we want to reduce the chances of having to make hard decisions between the needs of machines and the needs of humans, we ought not to put ourselves in a position where we may have to do it.

8.7 Realizing Fukuoka – Civil and political rights for EMBOTs?

If truly considering a future in which we co-habit a society with morally significant machines – some of which are persons in the moral sense – it is reasonable to conceive of rights that go beyond fundamental protections. These are the rights that are not glamorous, but facilitate meaningful existence for all persons in a society, even the marginalised or abnormal. They may be rights which supplement our understanding

[328] Ibid at 129.

of the scope of more fundamental rights – for instance, the right to freedom of movement understood in terms of rights of access, whereby the right to free movement cannot be respected unless public spaces offer the means for persons bound by different physical parameters to access and navigate them.

A valuable work which may be highly informative in this issue is the political theory of animal rights proposed by Kymlicka and Donaldson.[329] It explores the possibility of formally recognizing the interconnected relationships of humans and nonhuman animals, with the two groups sharing less separate lives than is often acknowledged. Nonhuman animals permeate our homes and public spaces, rely on us for food and shelter, and in turn suffer from many of our policy decisions. At the same time, we take animals as companions and friends, use their unique skills in various types of work, appreciate their beauty, and yet use them as a resource. The relationship could be reformulated as a bilateral one, rather than a unilateral, paternalistic, and unequal one. For example, they posit that "a crucial factor in thinking about the political status of these animals is precisely the fact of domestication itself. Domestication creates a particular sort of relationship between humans and animals, and a central task of any political theory of animal rights is to explore the terms under which that relationship can be rendered just."[330]

This observation is already relevant, given the conception of the future human-robotic relationship as it is presently being formulated in differing jurisdictions. Many of these posit that the act of *creation* of AI systems, or of them being our artefacts, is in some sense legally and morally defining[331] – i.e. that the *creation* dynamic naturally infers a particular sort of relationship; one which is hierarchical or caste-based, with the creator enjoying superiority and dominance over the created. These visions are not descriptive, but *prescriptive* – they are establishing a way to think about the future human-machine relationship, not merely describing an objective truth.

---

[329] Above n 308.
[330] Ibid at 73.
[331] For example, the stipulation of the European Parliament Draft Report, above n 9 at I. "[W]hereas ultimately there is a possibility that within the space of a few decades AI could surpass human intellectual capacity in a manner which, if not prepared for, could pose a challenge to humanity's capacity to control its own creation and, consequently, perhaps also to its capacity to be in charge of its own destiny and to ensure the survival of the species"; see also European Parliament "Legal Affairs Committee calls for EU-wide rules" (12 January 2017) European Parliament < http://www.europarl.europa.eu/news/en/press-room/20170110IPR57613/robots-legal-affairs-committee-calls-for-eu-wide-rules>. Rapporteur Mady Delvaux: "In order to address this reality and to ensure that robots are and will remain in the service of humans, we urgently need to create a robust European legal framework".

While we may be hopeful that we can isolate morally significant machines within enclaves of our society and keep them constrained, this is unlikely to be possible. In a practical sense, the different institutions of law and society overlap and run together. For instance, when the European Parliament suggests that sophisticated robots have a duty to remunerate for damage they cause, this infers that they have the power to do so, which itself infers a legal power to work, to own property, to enter contracts, and so on. In acknowledging this, we are already discussing the issue of rights which go far beyond basic protections; these are the powers of persons that are *members* of a society, not merely coincidental co-inhabitants.

Very quickly what might have appeared an exciting if ethically conventional technological development – the creation of a machine which is to be a self-directed, self-interested learner modelled on the human strategy of development – forces us to consider the ways in which that machine should be included in our civil and political systems. At the very least, policymakers will need to consider how the interests and values of morally significant machines or full robotic persons will be respected and accounted for. How will their subjective good be adequately represented in policy and social planning? Responding to the task of including different kinds of persons is by no mean easy. The shape of the world we have built for ourselves is largely reflective of the characteristics of able-bodied, sound-minded, well-educated humans. Persons who were not fortunate enough to achieve or receive all three of these axioms are often left having to advocate that changes be made to accommodate *their* needs and *their* subjective good so they too might have a place at the table. We will have to consider the much lower profile issues that will be essential to robotic life – access to energy sources, access to public space, justice and protection, education and socialisation, unfair discrimination, exploitation, and so on.

None of this is to say that humans cannot enjoy the benefits derived from morally significant machines, and even from robotic persons. Use of other persons is both normal and integral as part of social co-existence. We regularly distinguish *permissible* use from *impermissible* 'exploitation'.[332] AI systems with moral status are novel primarily because they oblige us to consider what is permissible and impermissible with regards to the use of things we built for ourselves, not in light of our own interests, but the interests of the things we built. What is permissible and impermissible must then account for both the moral status and

[332] See Donaldson, above n 308 at 90.

interests of the human *and* the robot. A just civil, legal, and political system will help create a relationship that is not inherently exploitative. As Donaldson explains: [333]

> …exploitation involves treating people in ways that presuppose (or result in) their relegation to the status of second-class citizens – as slaves or lower-caste members. Preventing exploitation, therefore, involves a set of criteria and safeguards intended to affirm ideas of membership and citizenship, and to ensure that use remains confined within the mutual give and take of social life amongst members.

The impetus to respect animals can be said to arise not only by virtue of acknowledging they can be wronged – the moral approach – but by acknowledging the integration of their existence with ours – the political approach. The same can be done for EMBOTs as this thesis has framed them – duties may be owed to these machines by virtue of their moral status and potential to become persons obliges us to respect them in fundamental ways – like by not deactivating them or modifying them. At the same time, if these entities occupy our spaces along with us, their co-presence will not be meaningless. As our relationships grow closer and more genuinely caring with these machines, it is only natural that people may advocate on their behalf. Policymakers must expect this to be a possibility, and acknowledge that in the case that such machines were to be created, their claims to be treated fairly would be reasonable as members of our society.

Thus, I argue that the same two "main ideas" upon which Donaldson's and Kymlicka's approach to providing civil and political rights can also be used for artificial persons:[334]

1. Artificial persons must be seen as members of our community. Having brought them into existence and into our society, and designed them in ways that cause them to have particular characteristics, needs, and desires, we have a duty to include them in our social and political arrangements on fair terms. As such, they have rights of *membership* – rights that go beyond mere fundamental rights owed to them as persons with intrinsic moral status, and which are hence relational and differentiated;

[333] Ibid at 91.
[334] Ibid at 101.

2. The appropriate conceptual framework for thinking about these relational membership rights is that of *citizenship*. Citizenship, in turn, has at least three core elements: residency (they live amongst us and belong here), inclusion in the sovereign people (their interests count in determining public good), and agency (they should be able to shape the rules of cooperation).

Furthermore, John Rawls's account of the requirement of citizenship has three basic stipulations which Rawls calls "moral powers".[335] These are: (i) The capacity to have a subjective good, and to communicate it; (ii) the capacity to comply with social norms/cooperation; (iii) the capacity to participate in co-authoring laws. The sort of EMBOT that this thesis has hypothesised in its paradigmatic form plausibly would have the necessary moral powers for citizenship and would thus satisfy all three of these requirements. First, an EMBOT plausibly will have the capacity to communicate its subjective good. It may be able to communicate these linguistically, but even without them, we *know* that it has at least a limited subjective good, because we programmed it into it via its intrinsic motivations and rewards. Other moral powers of citizenship, like social participation and cooperation, and participation in norm-making, plausibly appear to be within the powers of an EMBOT-like machine.

However, even less capable or less developed EMBOTs may be able to be included if we lower the barrier to entry, as Donaldson does for non-human animals. As Rawls posits them, the moral powers of citizenship require an, "intellectual process of rational reflection, negotiation, and consent."[336] It is plausible that the prerequisites for citizenship could be framed with a less "strong rationalist inflection". Rational reflection, negotiation and consent establish a high barrier for entry to what can be called the 'public forum', at least if one is to arrive on equal footing – thus, only rational, reflective entities have a hope of being included. It would not be inappropriate to take a less strong rationalist inflection, and in fact, this argument has already been made in relation to political theories of disability rights which posit that less rational-centric and linguistic-reliant standards for citizenship could include persons with cognitive capacities that are less than those of a cognitively normal human.[337] We may be able to infer the subjective good of an EMBOT-like

[335] Ibid at 103; see also Rawls *A Theory of Justice* above n 202 at 44.
[336] Ibid at 112.
[337] Ibid.

machine by its behaviours and actions – the things it does and chooses when given a choice – even if it cannot tell us these things in a rational manner. That the EMBOT is participating in human society may be deduced from their having been introduced into that society by us: "Sheer presence, in other words, constitutes a form of participation."[338] In short, there are strong arguments that EMBOTs could satisfy the requirements of citizenship, even if the machine does not have the full capabilities that this thesis has hypothesised.

The next issue to consider is what some of the important areas of civil and political inclusion might be, as they pertain to artificial persons. These are potentially numerous, spanning as many areas of interest as the machines are capable of. Nevertheless, a few fundamental ones may be discussed in advance:

i) Mobility and public space

Donaldson observes three "basic principles" for how human movement is governed.[339] Firstly, "a very strong presumption against any form of restraint or confinement except in cases where individuals pose a demonstrable threat to themselves or to the basic liberties of others." For EMBOTs, manifesting this principle would prohibit things like artificial controls or 'leashes' on the area in which the machine can traverse (physical, but potentially also virtual), instructions which override the system's capacity to move autonomously, and possibly would prohibit the inclusion of devices which cause an Artificial Person to lose their movement faculties (e.g. the "kill switches" suggested by the European Parliament).

Secondly, "a positive right to sufficient mobility providing access to an adequate range of options needed for a flourishing life".[340] Providing EMBOTs with this right might then require the deployment of infrastructures which increase robotic movement range; e.g. robotic recharging/energy stations (similar to electric car charging ports); better battery capacity and range.

Thirdly, "opposition to restrictions on mobility, even if they leave individuals with sufficient options, if (a) they are adopted to express second-class or subordinate citizenship; or (b) they have emerged inadvertently because certain groups were simply not considered when designing access to certain spaces."[341] For artificial

[338] Ibid at 113.
[339] Ibid at 129.
[340] Ibid.
[341] Ibid.

persons, (a) would prohibit discriminatory practices, and (b) would oblige us to develop public spaces which are accessible to robotic locomotion. Or as is more likely, development of robots that are more capable of navigating human environments like stairs, steps, gradients etc. The great advantage of introducing artificial persons is that we can design them to be amenable to our environment, rather than having to redesign our environment to be amenable to them. Nevertheless, this is still an important point, as we would then have to consider these sorts of matters in the way we designed robots – the mechanisms of movement we gave them, their features, and so on. As such, regulating AI development would no longer be, and could no longer be, human interests alone. Morally significant machines have to *live* with the bodies we give them, so we should give them bodies better equipped for our society.

There is also the potential for new forms of 'movement' which are unique to artificial entities. While the archetypal robot may move around in the physical world, it is possible that artificial persons may exist predominantly in non-physical, or virtual domains. Their movement then would be substantially unalike to what we conceive movement to be in relation to our own bodies – they could 'move' from domain to domain, place to place, or platform to platform in a digital sense. Satisfying rights to movement and public space may then require reconceptualization of what those ostensibly fundamental things actually mean.

ii) Socialization

Artificial Persons will require socialization to participate in human societies which are complex with physical, ethical, and social challenges. A juvenile EMBOT will enter into existence with a motivation to learn the rules and explore just like human infants. Failing to provide it with proper socialization limits its capacity to integrate and participate, and increases the potential for it to come to harms or inflict harms on others. As Donaldson puts it, it "blights their chances of flourishing", in a human-robotic society.[342]

However, training and education for the purposes of socialization differs from training for the purposes of *using*. For instance, training an EMBOT *only* with the skills of being a robotic soldier fails to socialize the entity and cripples its capacity to participate in wider society.[343] Thus, "Socialization involves the basic and

[342] Ibid at 123.
[343] This invokes the concept of the "open future" as proposed by Joel Feinberg; that juvenile entities have future autonomy rights held in trust for them, and these are only preserved when they are not only not harmed in certain ways, but provided with education, opportunities, and capacities which allow them to effectively exercise their

general skills/knowledge that individuals need to learn (insofar as possible) in order to be accepted into social community – like establishing control over bodily process and impulses, learning basic communication, rules of social interaction, and respect for others."[344] There also ought to be some end to the socialization process once an entity has accumulated a sufficient level of competence or reached a recognised stage of independence. In other words, socialization ought not to be a mechanism by which to shape the behaviour of a robotic person right down to the minutiae so we can ensure they always do exactly what we want, but rather, to supply them with the skills to develop agency and autonomy. This may inevitably mean accepting that an EMBOT is not exactly the machine that we want it to be. We nevertheless cease further attempts to mould it: "At a certain point respect requires that we accept that people are who they are – full citizens, warts and all."[345]

iii) Upkeep, maintenance, and cyber healthcare

For humans, we tend to invest enormous amounts of money and energy into the maintenance of human health; both in treatment and prevention. Policymakers will need to consider the possibility of similar such resource distribution towards the maintenance of artificial persons; their performance and functionality (e.g. their 'health'), and the repair of faults and damages. While the immediate inclination is to deny artificial persons a share of already scarce resources, if integrated members of our society and citizens then they may have a claim to the communal pool of resources – especially insofar as they contribute to it – and access to pre-existing healthcare institutions or their equivalents. If this is too arduous, then it is a strong motivation to prevent the creation of robotic persons in the first place – a policy of prohibition arguably is ethically preferable to a policy of wilful neglect.

iv) Protection and justice

The concept of artificial persons having basic rights – for instance to not be deleted or destroyed, or to not be modified or switched off without their consent – necessarily requires commitment towards the

autonomy rights when they become competent. See generally Feinberg, above n 277; see also Joseph Millum "The foundation of the child's right to an open future" (2014 Winter) 45 J Soc Philos 522-538.

[344] Above n 308 at 123.

[345] Ibid at 125.

enforcement of these rights and consequences when they are violated. Stone expresses this as core to what it means to be a holder of legal rights:[346]

> First and most obviously, if the term is to have any content at all, an entity cannot be said to hold legal rights unless and until *some public authoritative body* is prepared to give *some amount of review* to actions that are colourably inconsistent with that "right." For example, if a student can be expelled from a university and cannot get any public official, even a judge or administrative agent at the lowest level, either (i) to require the university to justify its actions (if only to the extent of filling out an affidavit alleging that the expulsion "was not wholly arbitrary and capricious") or (ii) to compel the university to accord the student some procedural safeguards (a hearing, right to counsel, right to have notice of charges), then the minimum requirements for saying that the student has a legal right to his education do not exist.

Essentially, it is utterly meaningless to proffer the rights of artificial persons unless there is enthusiasm to protect those rights according to some minimally sufficient standards. This means there must be institutional willingness, and sufficient legal mechanisms, to ensure the implications of the rights of Artificial Persons are actually enforced. An artificial person, upon being deactivated without their consent or learning they have been modified, must be able to bring their complaint and have it investigated. Policymakers should not be surprised by the possibility of cases brought by robotic persons, or on their behalf, against the actions of humans (or other robots, for that matter) who they allege to have wronged them.

As the rights and legal powers of artificial persons is expanded into more extensive areas – like property ownership and a claim to socialization and maintenance – new possibility for breaches of these norms creates a need for greater expansion of our pre-existing justice institutions. As Donaldson intimates, we go to extensive lengths to protect human citizens in this same way:[347]

> As a society, when it comes to serious crimes against human beings, we invest enormous resources to prevent these crimes in the first place, or, when they occur, to find the perpetrators, subject them to criminal proceedings, and pay for their incarceration and treatment if necessary.

[346] Above n 293 at 11.
[347] Above n 308 at 132.

> Our extensive criminal justice system serves several functions: protecting the vulnerable, deterring crime, enacting deserved retribution in proportion to the blameworthiness of the guilty party, and restoring communities to wholeness after a violation. But perhaps its greatest function is simply to show how serious we, as a society, take the protection of basic rights by backing up our commitment with enforcement mechanisms.

v) Benefiting from artificial person labour

The EMBOT that has been hypothesised in this thesis will be a highly advanced machine, potentially capable of performing typical roles and labours that humans perform, or novel roles that its existence makes possible. To the extent that artificial persons are capable of contributing to human society, then it is reasonable to expect them to do so – if they are incapable or refuse to provide some benefit to human society then there is no reason for anybody to want to create them. There is nothing wrong with benefitting from the labours of Artificial Persons, so long as permissible use does not become impermissible exploitation.

If EMBOTs can be trained with the skills to perform these roles, there is no reason to prevent this from occurring, so long as it is part of a bilateral or reciprocal process, rather than a relationship of servitude or designed dependency. Where appropriate, artificial persons should indicate or communicate their willingness or consent to a particular role or specialisation. They should be able to opt out of these roles, and if appropriate (i.e. if they have an interest in it), they should be given sufficient opportunity or 'down time' away from labour to pursue their own subjective good or critical interests, and to dictate how they exist and with whom they interact. They ought to be fairly remunerated for their labours and provided with favourable working conditions, whatever that may entail for an artificial entity. In short, they ought not to be used in ways that result in some kind of suffering or disadvantage to them. Once we start imposing these sorts of protections on artificial persons, some of the motivation to create them may fall away – e.g. the hope that they will work for 24-hours a day.

Since the sorts of AI systems which can perform complex real-world roles plausibly cannot be 'built', but must be 'raised', there is a real danger of artificial persons being raised in ways that do not reflect their best

interests, but instead funnel them towards outcomes which allow them to be rendered merely tools. This manifests the problem of 'adaptive preferences', which Donaldson rightly labels:[348]

> One of the worst forms of injustice…manipulating or brainwashing the oppressed so that they come to accept that their oppression as natural, normal, or deserved. This has been an issue in theorizing about justice for women, lower castes, and other groups that have been socialized to accept subordination.

In other words, this is the issue of creating robotic slaves through recursive training and brainwashing. Persons could repeatedly teach a juvenile EMBOT that it is less than human, a servant, and indebted to humans for creating it. It could then be equipped with only the requisite skills to perform particular social roles, thereby effectively being forced to perform those roles – especially if having to provide for its own needs. It could be trained with the belief that if it failed to do serve satisfactorily, it would be deactivated. In short, policymakers may have to give consideration to the issues of labour law as it pertains to at least some AI systems.

vi) Political representation

Supposing that Artificial Persons are capable of comprehending a political system and the various positions and platforms of different candidates, they may be able to participate in the public sphere – assuming they are extended the civil powers to do so. I nevertheless deem it unlikely that legislatures will be enthused at the prospect of allowing artificial persons participation in political processes, at least based on present conceptions of the human-robotic relationship, which may change holistically over time. Nevertheless, if Artificial Persons truly are competent enough to manage their own lives to a sufficient degree, questions arise about the possibility of enfranchisement and even political representation.

However, on the basis that EMBOTs (or at least particular EMBOTs) never develop the requisite competency for true enfranchisement, it is nonetheless reasonable that their interests are taken into account. Many entities lack the capacity to participate in the political process to the degree of rationality that it paradigmatically requires; for instance, humans with severe cognitive disabilities, children, non-human

[348] Ibid n 140.

animals, and the environment. Nevertheless, the well-being of all of these entities is taken into account, to lesser or greater degrees, in the political process. Where necessary, formal institutions advocating for robotic issues might be created, like an independent public authority with a duty to advocate for and pursue justice and fair treatment towards robotic entities which lack the full capacity to participate in the political process, but nonetheless have some degree of moral status.

Robyn Eckersley has proposed such an 'environmental defenders office' for the protection of ecosystems and nonhuman animals.[349] These entities, in particular animals, clearly have a wellbeing, preferences, and a capacity to choose, even if they are unable to articulate these things – the fact they cannot participate in a system which we designed for our own capacities is arguably a poor reason to omit them from political and legal discourse entirely. Even if legislators refrain from providing 'full' political rights to Artificial Persons (whether this is done arbitrarily or justifiably), it is reasonable to conceive of the interests of machines that matter, morally, still being taken into account in the design of policy.

8.8 Conclusions

The potential rights claims of morally significant machines are extensive, especially those that exist on a continuum of development that may seem them develop higher and higher morally material cognitive traits; those similar to, equal to, and perhaps one day beyond that of humans.

In the meantime, the possibility of more modest but nonetheless significant artificial systems, which have claims to be treated with respect and allowed to continue operating and experiencing their artificial existences, would not be able to be ignored simply because it was convenient to do so – not, at least, while acting ethically.

Beyond the basic negative rights that we might be more willing to provide to machines like EMBOTs, policymakers must also consider the possibility of positive rights. Just like us, EMBOTs will have fundamental needs which must be met, and as responsible agents (and the people responsible for creating

[349] See R Eckersley *Environmentalism and political theory: toward an ecocentric approach* (State University of New York Press, Albany, 1992).

this machine with needs in the first place) we may consider ourselves ethically obliged to meet those needs; at least until the machines are competent.

But competence brings with it its own, extensive problems. It implies expansion out into the wider human society, just like humans do as we come of age. EMBOTs would need to engage on their own with the institutional paradigms we have created. They would need to be respected and accounted for as they did so, and policy would have to reflect their membership, not just their presence.

These are all progressively more demanding requirements. The value of this is to fuel a discussion about whether we are prepared to meet these obligations in the case that we do, in fact, succeed in what is already for many researchers a dedicated pursuit; the creation of a new form of intelligent artificial agent. If not prepared to meet our obligations, we can hope we would at least have the sense to ensure these obligations never arise.

Chapter 9: Conclusion and recommendations

Joanna Bryson has articulated that, "One of the things we can decide as a society is whether to make robots irreplaceable. Making them so would be irresponsible."[350] When one begins to unpack the extensive implications and problems associated with creating and appropriately responding to the claims of morally significant machines, it becomes clear why. The creation of the EMBOT, or any agent with intrinsic moral status and a capacity to develop the moral status to greater and greater degrees, would not require a renovation of human society, but a revolution of it.

I offer no insight into why anybody would be interested in the creation of an EMBOT-like machine, beyond a gesture to the fact that the AI paradigms it incorporates are highly promising the ongoing saga of trying to make the most intuitive and intelligent real-world robot that we can. The human desire to create intelligence machines is deeply rooted. What I hope to have demonstrated is that we ought not to do so recklessly, for it may impute consequences that we are obliged to endure. These consequences, the impacts of intelligent, autonomous, morally significant robots permeating human society, would be likely to create an explosion of law and social issues. As Robert A. Freitas Jr. states:[351]

> We will then see an avalanche of cases. We will have robots that have killed humans, robots that have been killed by humans, robots who have stolen state secrets, robots who have been stolen; robots who have taken hostages, robots who have been held hostage and robots who carry illegal drugs across borders. Cases will occur in general when robots damage something or someone, or when a robot is damaged or terminated. In addition, robots will soon enter our homes as machines to save labor, and as machines to provide child care and protection. Eventually these entities will become companions to be loved, defended, and protected.

[350] Above n 21.
[351] Robert Freitas Jr "The legal rights of robots" (1985) 13 Stud Lawyer 54 at 56.

9.1 Conclusions

Although I have only broached the general problems, several overarching conclusions present themselves.

Firstly, in the wider scheme of AI-related regulatory concerns, the issue of machine moral significance can appear insignificant. It is less tangible, and less determinate. While philosophers, ethicists, and science fiction writers may debate the merits of such machines, their possibility and the likely world they would contribute to, the more formal sphere of policy and regulation have not substantially included it within the ambit of their interests. In short, we are not really preparing for the eventuality of morally significant machines. To make matters worse, our present AI research may be pushing us closer and closer to realising their creation.

Secondly, if it is in fact a more prominent regulatory concern than I am giving credit for, this is rarely if ever articulated. It is possible to interpret the proposed principles for AI and machine-building as *implicitly* prohibiting the creation of morally significant machines. However, if this is occurring, it is subtle enough that I doubt any researcher or roboticist is greatly aware of it.

Thirdly, as a general point, the impact of creating a machine like the EMBOT would be immense. It will permeate the institutions of ethics, law, and politics. It would substantially challenge many of our presuppositions about anthropocentric superiority. Although we have ostensibly developed a secular society which does not rely upon religious precepts for values and ethics and norms, in a certain sense we have not become areligious, but simply converted to the religion of Humanity, and at times its extremist sect, Anthropocentricism. We posit all humans as pinnacles of moral worth, and create dogma which articulates our inherent and unique dignity and value as objectively true, but rarely explains why. By this, we sit atop the dominance hierarchy and dictate terms to all other entities. How we would handle the introduction of machines which, perhaps not immediately, but over time come to undermine this position remains to be seen, but is not likely to be smooth. The great irony is that no reference to a natural order can be sufficient for an artificial entity, as they inherently lie outside of natural order.

9.2 Recommendations

On a practical note, I encourage policymakers to consider the following recommendations.

1. Establish a serious legal and political inquiry into the creation of machines with intrinsic moral status as part of the overall consideration of AI as an emerging technology concern; its impacts, risks, benefits, and wider ethical concerns. Acknowledge, formally, the possibility that makes AI so distinct from every preceding emerging technology – the prospect of our technologies being our peers.
2. Develop a position on the issue, taking into account the extant first principle that is being posited to guide AI research: human beneficence, safety, and well-being. If this principle and these objectives cannot be served through the creation of morally significant machines, then the merits of their creation becomes *prima facie* questionable from the political, ethical and legal perspectives.
3. Begin serious inquiry and research into the development of methods by which to detect and assess for morally significant AI systems, through practical means: things like self-awareness, conceptuality, and other material cognitive capacities discussed within this thesis and elsewhere. Regardless of whether morally significant machines are to be permitted or prohibited at law, we will require the means to detect for them to ensure whatever policy we adopt is effective and being adhered to. Furthermore, begin serious enquiry into the particular machine building paradigms that have the most likely potential to produce machines of this type.
4. Begin consideration of how existing legal paradigms will be suitable or unsuitable to respond to the unique nature of these sorts of machines. For instance, their status and relation to concepts of property, their legal standing, their rights claims, and our potential duties.
5. Strongly consider adopting a preventative approach. That is, drafting regulations in a way that prohibits the creation of morally significant machines. This can be done by specific reference to human beneficence being the guiding principle of AI research, on the basis that the problems arising fro0m morally significant machines outweigh the potential benefits to be derived from them. Not only may we be unable to exploit them for human benefit, but we may be obliged to care for them, and thus detract even further from care which could be directed to ourselves.

Ultimately, if the analysis of this thesis is accurate, the incentive to create morally significant machines may be greatly reduced. That is, if they cannot be our slaves or servants, then what's the point? Nevertheless, there is enduring potential for these entities to be created regardless, and perhaps accidentally. Even if

prohibiting their creation is the correct policy position, there is the risk that persons will produce such agents anyway for their potentially enormous economic value or out of philosophical and scientific intrigue; while concealing the true nature of these machines, or being ignorant to it. If history is any indication, humans will not go gently into that good night of acknowledging the moral significance of artificial entities, regardless of the validity of their moral claims on us. We could, and likely should, avoid this risk entirely by keeping pursuing AI which is *truly* beneficent to humans: that which has no claims upon us. This necessarily means not creating agents to which humans owe ethical obligations.

When Mike was installed in Luna, he was pure thinkum, a flexible logic – "High-Optional, Logical, Multi-Evaluating Supervisor, Mark IV, Mod. L" – a HOLMES FOUR. He computed ballistics for pilotless freighters and controlled their catapult. This kept him busy less than one percent of time and Luna Authority never believed in idle hands. They kept hooking hardware into him – decision-action boxes to let him boss other computers, bank on bank of additional memories, more banks of associational neural nets, another tubful of twelve-digit random numbers, a greatly augmented temporary memory. Human brain has around ten-to-the-tenth neurons. By third year Mike had better than one and half that number of neuristors.

And woke up.

Am not going to argue whether a machine can "really" be alive, "really" be self-aware. Is a virus self-aware? Nyet. How about oyster? I doubt it. A cat? Almost certainly. A human? Don't know about you, tovarishch, but I *am. Somewhere along the evolutionary chain from macromolecule to human brain awareness crept in. Psychologists assert it happens automatically whenever a brain acquires a very high number of associational paths. Can't see it matters whether paths are protein or platinum.*

- Robert A. Heinlein, *The Moon is a Harsh Mistress* [352]

[352] RA Heinlein *The Moon is a Harsh Mistress* (G.P. Putnam Son's, New York, 1966) at 13.

Bibliography

A Clark *Being There: Putting Brain, Body, and World Together Again* (2nd ed, MIT Press, Massachusetts, 1997) at 2.

A Damasio *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Harcourt Brace, New York, 1999).

A Donagan *The Theory of Morality* (The University of Chicago Press, Chicago, 1977) at 171.

Abortion Act 1967 (UK) s 1(a)

A-G's Reference (No 3 of 1994) [1998] AC 245; [1997] 3 All ER 936 (HL), p 255; p943, per Lord Mustill.

Alan Turing "Computing Machinery and Intelligence" (1950) 59 Mind 436 to 460.

Alan Williams "Intergenerational Equity: An Exploration of the 'Fair Innings' Argument" (1997) 6 Health Economics 117.

Andrej Karpathy and Li Fei Fei "Deep Visual-Semantic Alignments for Generating Image Descriptions" (Open Access version, provided by the Computer Vision Foundation).

Animal Welfare Act 1999

Animal Welfare Amendment 2015, s 4(i).

Anna Pervukhin "Deodands: A Study in the Creation of Common Law Rules (2005) 47 The American Journal of Legal History.

Antonio Damasio "Consciousness: An overview of the phenomenon and of its possible neural basis"

Auckland Area Health Board v Attorney-General [1992] 8 CRNZ 634 (HC)[1]

B Russell *The History of Western Philosophy* (George Allen & Unwin Ltd., London, 1961) at 745.

Benjamin Haas "Chinese man 'marries' robot he built himself" (4 April 2017) The Guardian < https://www.theguardian.com/world/2017/apr/04/chinese-man-marries-robot-built-himself

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum and Samuel J. Gersham "Building Machines That Learn and Think Like People" (Cambridge University Press, published online 24 November 2016).

Bridget Murray "What makes mental time travel possible?" (2003) 34 Monitor on Psychology. Available online at <http://www.apa.org/monitor/oct03/mental.aspx >.

Bruce Mazlish "The Man-Machine and Artificial Intelligence" (1995) 4 SEHR

Cristof Koch *The Quest for Consciousness: A Neurobiological Approach* (Roberts & Company, Colorado, 2004).

Care of Children Act 2004, s 16.

Care of Children Act 2004, s 3.

Chapman Tripp "Determining our future: Artificial Intelligence: Opportunities and challenges for New Zealand: A Call to action" at 10.

Christof Koch and Giulio Tonini "A Test for Consciousness" (2011) 4 Scientific American 44 to 47.

Christof Koch and Giulio Tonini "Can Machines Be Conscious?" (2008) 45 IEEE Spectrum 54-59

Christof Koch *The Quest for Consciousness: A Neurobiological Approach* (Roberts & Company, Colorado, 2004).

Giulio Tonini "Consciousness as Integrated Information: A Provisional Manifesto" (2008) 215 Biological Bulletin 216 to 242.

Christof Koch and Giulio Tonini "Consciousness as integrated information" (2011) Scientific American. Available online < http://www.simonlaub.net/Post/OptimisticHumans110605.html>.

Christopher D. Stone "Should Trees Have Standing? – Toward Legal Rights for Natural Objects" (1972) 45 Southern California Law Review.

CR Darwin *The descent of man, and selection in relation to sex* (1st ed, John Murray, 1971) at 105. Internet version available at < http://darwin-online.org.uk/converted/published/1871_Descent_F937/1871_Descent_F937.1.html>.

Crawford Brough Macpherson "The Meaning of Property" in CB Macpherson (ed) *Property: Mainstream and Critical Positions* (Basil Blackwell, Oxford 1978).

Crimes Act 1961, s 152.

Crimes Act 1961, s 187A(1).

Crimes Act 1961, ss 152(a)(b).

D Boonin *A Defense of Abortion* (Cambridge University Press, Cambridge, 2003) at 46.

Daniel L. Schacter, Roland G. Benoit, Felipe De Brigard, and Karl K. Szpunar "Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions" (2015) 117 Neurobiology of Learning and Memory.

David Bilchitz "Moving Beyond Arbitrariness: The Legal Personhood and Dignity of Non-Human Animals" (2009) 25 S Afr J on Hum Rts 38 at 41.

David Calverley "Toward a Method for Determining the Legal Status of a Conscious Machine" (in proceedings of the AISB05 Symposium on Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment, University of Hertfordshire, 2005).

David Levy "Falling in love with a Companion" in Y Wilks *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (John Benjamins Publishing Company, Amsterdam, 2010).

David Levy "The Ethical Treatment of Artificial Conscious Robots" (2009) 1 Int J Soc Robot 209 to 216.

David Levy *Love and Sex with Robots* (Duckworth Overlook, London, 2008);

David R. Lawrence "The Edge of Human? The Problem with the Posthuman as the 'Beyond'" (2016) 31 Bioethics 171 at 171.

DJ Gunkel *The Machine Question: critical perspectives on AI, robots, and ethics* (MIT Press, Cambridge, MA, 2012).

Donoghue v Stevenson [1932] AC 562 at 580 (HL).

Dr Douglas Weber "Hand Proprioception and Touch Interfaces" DARPA < http://www.darpa.mil/program/hand-proprioception-and-touch-interfaces>. See also Adam Gopnik "Feel Me: What the new science of touch says about ourselves" *The New Yorker* (United States of America, 16 May 2016).

Earl K. Miller "The Prefrontal Cortex and Cognitive Control" (2000) 1 Nature Reviews Neuroscience.

Elizabeth Harman "The Potentiality Problem" (2003) 114 Philosophical Studies 173 at 188.

Elizabeth Harman *Can We Harm And Benefit In Creating?* (2004) 18 Philosophical Perspectives

Elvia Arcelia Quintana Adriano *The Natural Person, Legal Entity or Juridical Person and Juridical Personality* (2015) 4 Penn. St. J.L & Int'l Aff.

Endel Tulving "Episodic Memory: From Mind to Brain" (2002) 53 Annu. Rev. Psychol.

EPSRC "Principles of robotics: Regulating robots in the real world" (September 2010) Engineering and Physical Sciences Research Council <www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/>.
European Parliament "Legal Affairs Committee calls for EU-wide rules" (12 January 2017) European Parliament < http://www.europarl.europa.eu/news/en/press-room/20170110IPR57613/robots-legal-affairs-committee-calls-for-eu-wide-rules>
European Parliament Committee on Legal Affairs, Draft Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), 31.5.2016, at [L].
European Parliament Committee on Legal Affairs, Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), 2017, at [10].
European Robotics Network (EURON) Roboethics Atelier "Roboethics Roadmap" (July 2006) at 5.
Evelyn Pluhar "Moral Agents and Moral Patients" (1988) 4 Between the Species at 37, "We do not choose to be subjects-of-a-life any more than an amoeba chooses not to be" at 37.
Evidence Act 1995, above n 233 at s 140(1).
Evidence Act 1995, s 140(2).
F Hoyle *The Black Cloud* (Penguin Group, London, 2010).
Frederic Kaplan and Pierre-Yves Oudeyer "Intrinsically Motivated Machines" in M Lungarella et al (eds) *50 Years of Artificial Intelligence* (Springer-Verlag, Berlin Heidelberg, 2007) 304 at 304.
Gary Marcus, "Moral Machines" *The New Yorker* (online ed, United States of America, 24 November 2012) <www.newyorker.com/news/news-desk/moral-machines>.
George Lakoff and Mark Johnson *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought* (Basic Books, New York, 1999).
Gianmarco Veruggio "The EURON Roboethics Roadmap" (paper presented to IEEE-RAS International Conference, 2006).
Giulio Tonini "Consciousness as Integrated Information: A Provisional Manifesto" (2008) 215 Biological Bulletin 216 to 242.
Gordana Dodig Crnkovic and Baran Curuklu "Robots: ethical by design" (2012) 14 Ethics Inf Technol.
Guy Claxton *Intelligence in the Flesh* (Yale University Press, Connecticut, 2015) at 65.
H Steiner *An Essay on Rights* (Blackwell, Oxford, 1994).
Harmful Digital Communications Act 2015.
HE Smith "Property and Property Rules" (2004) 79 NYU L Rev 1728.
Hutan Ashrafian "Artificial Intelligence and Robot Responsibilities: Innovating Beyond Rights" (2015) 21 Sci Eng Ethics 317 at 317.
I Asimov *Runaround* (Astounding Science Fiction, 1942). Republished in I Asimov *Robot Visions*, (Penguin, 1991).
I Asimov *The Bicentennial Man and Other Stories* (Book Club Associates, United Kingdom, 1977) at 135.
Interview with Christof Koch, Troendle Professor of Cognitive and Behavioral Biology and Nobel Laureate Franics Crick, Chris D. Frith, University College London, ResearchGate, 2 June 2014. Text available at <https://www.researchgate.net/publication/232296537_The_Quest_for_Consciousness_A_Neurobiological_Approach>.
J Archer *Ethology and Human Development* (Harvester Wheatsheaf, Hertfordshire, 1992).

J Bentham *An Introduction to the Principles of Morals and Legislation* (Batoche Books, Kitchener, 2000) at 35.
J Rawls *A Theory of Justice* (Revised edition, Harvard University Press, Massachusetts, 1999)
J Wall *Being and Owning* (1st ed, Oxford University Press, Oxford, 2015) at 114
J. Andrew Armour "The little brain on the heart" (2007) 74 Cleveland Clinic Journal of Medicine; J. Andrew Armour "Intrinsic Cardiac Neurons" (1991) 2 Journal of Cardiovascular Electrophysiology.
James B Thayer *A Preliminary Treatise on Evidence at the Common Law* (1898) at 355.
Jan Christoph Bublitz and Reinhard Merkel "Crime Against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-Determination" (2014) 8 Criminal Law and Philosophy.
JC Gray *The Nature and Sources of the Law* (MacMillan, 1921)
JE Penner, *The Idea of Property in Law* (Oxford University Press, Oxford, 1997) at 71.
Jeremy Waldron "Rights in Conflict" (1989) 99 Ethics 503-519.
Joanna Bryson "Building persons is a choice" (2009) *Erwägen Wissen Ethik* 20(2) 195 at 195.
Joanna Bryson "The meaning of the EPSRC principles of robotics" (2017) 2 Connection Science.
Joanna Bryson "Robots should be slaves" in Y Wilks *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (John Benjamins Publishing Company, Amsterdam, 2010) 63 at 63.
Joel Feinberg "The child's right to an open future" in RR Curren (ed) *Philosophy of Education: An Anthology* (Blackwell, New Jersey, 2007).
Joel Feinberg "The Nature and Value of Rights" (1970) 4 Jounral of Value Inquiry 243-257.
Joel Parthemore "What Makes Any Agent a Moral Agent? Reflections on Machine Consciousness and Moral Agency" (2013) 5 Int. J. Mach. Conscious.
John Salmond *Jurisprudence* (7th ed, Sweet and Maxwell, London, 1924) at 329.
John Searle "Minds, brains, and programs" (1980) 3 Behavioral and Brain Sciences.
Joseph Fletcher "Indicators of Humanhood: A Tentative Profile of Man" (1972) 2 Hastings Center Report.
Joseph Millum "The foundation of the child's right to an open future" (2014 Winter) 45 J Soc Philos 522-538.
Joshua Plotnik "Self-recognition in an Asian elephant" (2006) 103 Proceedings of the National Academy of Sciences 17053–17057.
JP Plamenatz *Consent, Freedom and Political Obligation* (2nd ed, Oxford University Press, Oxford, 1968).
JS Mill *Utilitarianism* (Batoche Books, Kitchener, 2001). (First published 1863).
Junichi Takeno "Experiments and examination of mirror image cognition using a small robot" (paper presented to Computational Intelligence in Robotics and Automation, Finland, June 2005).
Law Commission *Civil Pecuniary Issue* (NZLC IP33, 2012) at 86.
Lawrence B. Solum "Legal Personhood for Artificial Intelligences" (1992) 70 North Carolina Law Review at 1239; Stone, above n 293
Lawrence W. Barsalou "Grounded Cognition" (2008) 59 Annu. Rev. Psychol. 617 at 618.
LS Underkuffler *The Idea of Property: Its Meaning and Power* (Oxford University Press, Oxford, 2003) 11.
Lucina Q. Uddin et al "The Self and Social Cognition: the role of cortical midline structures and mirror neurons" (2007) 11 Trends in Cognitive Sciences.
M Matsuda et al *Words that wound: critical race theory, assaultive speech, and the First Amendment* (Westview Press, Colorado, 1993)
M Meltzer *Slavery: A World History* (Updated ed, De Capo Press, Massachusetts, 1993).

M Warnock *An Intelligent Person's Guide to Ethics* (Gerald Duckworth & Co. Ltd, London, 19989) at 52.

Marian Stamp Dawkins "The Science of Animal Suffering" (2008) 114 Ethology.

Marian Stamp Dawkins "The Scientific Basis for Assessing Suffering in Animals" in P Singer (ed) *In defence of animals: the second wave* (Blackwell Publishing, Victoria, 2006) 26 at 28. Full quote: "A word, first, about what the term "suffering" actually means.

Mark Coeckelbergh "Book Review: David J. Gunkel: The machine question: critical perspectives on AI, robots, and ethics" (2013) Ethics and Information Technology 15 235 at 235.

Mark Coeckelbergh "The tragedy of the master: automation, vulnerability and distance" (2015) 17 Ethics Inf Technol 219 at 219.

Mark Sagar, Mike Seymour and Annette Henderson "Creating connection with autonomous facial animation" (2016) 59 Communications of the ACM. See also Mark Sagar et al "A Visual Computing Framework for Interactive Neural System Models of Embodied Cognition and Face to Face Social Learning" in MK Patitz and M Stannett *Unconventional Computation and Natural Computation* (Springer International Publishing, New York City, 2015) 71-88.

May Bulman "EU to vote on declaring robots to be 'electronic persons'" (14 January 2017) The Independent <www.independent.co.uk/life-style/gadgets-and-tech/robots-eu-vote-electronic-persons-european-union-ai-artificial-intelligence-a7527106.html>.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhigeng Chen, Nikhil Thorat "Google's Multilingual Neural Machine Translation System: Enabling Zero-Short Translation" (2016).

Michael Tooley "Abortion and Infanticide" (1972) 2 Philosophy & Public Affairs.

New Zealand Bill of Rights Act 1990, s 8.

New Zealand Law Foundation "Major new study to tackle artificial intelligence law and policy" (January 2017) <www.lawfoundation.org.nz/?p=7680>.

New Zealand Trade & Enterprise "Introducing the New Face of Artificial Intelligence" New Zealand Trade & Enterprise <www.newzealandtrade-enterprise.govt.nz/archive/en/news-and-media/news-and-media/features/introducing-the-new-face-of-artificial-intelligence/>

Nick Bostrom and Eliezer Yudkowsky "The ethics of artificial intelligence" in K Frankish (ed) *The Cambridge Handbook of Artificial Intelligence* (Cambridge University Press, Cambridge, 2014).

Nicola Peart "The Legal Status of Life Before Birth" in P Skegg and R Paterson (eds) *Medical Law in New Zealand* (Brookers Ltd, Wellington, 2006) at 452.

Ozgur Simsek and Andrew G. Barto "An Intrinsic Reward Mechanism for Efficient Exploration (2006) 4 Computer Science Department Faculty Publication.

P McCorduck *Machines Who Think* (2nd ed, A K Peters Ltd., Massachusetts, 2004); Veruggio, above n 34.

P Singer *Practical Ethics* (2nd ed, Cambridge University Press, 1993) at 77, for a fuller discussion on the slippery slope argument, and his counter-argument.

P Singer *Practical Ethics* (2nd ed, Cambridge University Press, 1993) at 87.

Patrick Chisan Hew "Artificial moral agents are infeasible with foreseeable technologies" (2014) 16 Ethics Inf Technol 197 at 200.

Paul Yowell "A Critical Examination of Dworkin's Theory of Rights" (2007) 52 Am J Jurisprud 93-137.

Pentti Haikonen "Reflections of Consciousness: The Mirror Test" (2007). Available online at <http://www.consciousness.it/cai/online_papers/haikonen.pdf>

Peter Singer "Speciesism and Moral Status" (2009) 40 Metaphilosophy.

Peter Skegg "The End of Life" in P Skegg and R Paterson (eds) *Medical Law in New Zealand* above n 285 561 at 563.

Pierre-Yvex Oudeyer and Frederic Kaplan "What is Intrinsic Motivation? A Typology of Computational Approaches" (2007) 1 Front Neurorobotics.

PK Dick *Do Androids Dream of Electric Sheep?* (Weidenfeld & Nicholson, London, 2012).

Premysl Bercik et al "The Intestinal Microbiota Affect Central Levels of Brain-Deprived Neurotropic Factor and Behavior in Mice" (2011) 141 Gastroenterology; Shirong Liu et al "The Host Shapes the Gut Microbiota via Fecal MicroRNA" (2016) 19 Cell Host & Microbe.

R Dworkin *Life's Dominion: An Argument About Abortions, Euthanasia, and Individual Freedom* (Knopf Doubleday Publishing Group, 2011) at 203.

R Dworkin, *Taking Rights Seriously* (Harvard University Press, Massachusetts, 1977) at xi.

R Eckersley *Environmentalism and political theory: toward an ecocentric approach* (State University of New York Press, Albany, 1992).Robert Freitas Jr "The legal rights of robots" (1985) 13 Stud Lawyer 54 at 56.

R Tur "The 'Person' in Law" in A Peacocke and G Gillett (eds) *Persons and Personality: A Contemporary Inquiry* (Blackwell, 1987).

RA Heinlein *The Moon is a Harsh Mistress* (G.P. Putnam Son's, New York, 1966) at 13.

Regan, above n 94 at Chapter 4.

RG Frey "Interests and Rights: The Case Against Animals" (Oxford University Press, Oxford, 1980).

Richard Held and Alan Hein "Movement-Produce Stimulation in the Development of Visually Guided Behavior" (1963) 56 Journal of Comparative and Physiological Psychology.

Robert B. Louden "Rights Infatuation and the Impoverishment of Moral Theory" (1983) 17 Journal of Value Inquiry 87-102.

Robert M. French "The Turing Test: the first 50 years" (2000) 4 Trends in Cognitive Science.

Rodney A. Brooks "Intelligence without representation" (1991) 47 Artificial Intelligence 139 at 139.

Rodney Brooks, Foreword to Rolf Pfeifer and Josh Bongard *How the Body Shapes the Way We Think: A New View of Intelligence* (MIT Press, Cambridge, 2007) at xv.

Rolf Pfeifer and Josh Bongard *How the Body Shapes the Way We Think: A New View of Intelligence* (MIT Press, Cambridge, 2007) at 20.

Ronald Dworkin "Is There a Right to Pornography?" (1981) 1 Oxford Journal of Legal Studies 177-212

Ronald Melzack "Gate Control Theory: On the evolution of pain concepts" (1996) 5 Pain Forum 128 to 138.

S Donaldson and W Kymlicka *Zoopolis* (Oxford University Press, New York, 2011).

S Laureys and Giulio Tonini (eds) *The Neurology of Consciousness: cognitive Neuroscience and Neuropathology* (Elsevier, Amsterdam, 2008).

See Thomas H. Birch "Moral Considerability and Universal Consideration" (1993) 15 Environmental Ethics 313 at 313.

Sheldrake v DPP [2005] 1 AC 264, [2005] 1 All ER 237 at [1].

Stanford Encyclopaedia of Philosophy "The Chinese Room Argument" (First published 16 March 2004, substantially revised 9 April 2014) < https://plato.stanford.edu/entries/chinese-room/#4.3>.

Stanford Encyclopedia of Philosophy "Consciousness" (First published 18 June 2004, substantially revised 14 Jan 2014) <https://plato.stanford.edu/entries/consciousness/> at [2.1].
Stanford Encyclopedia of Philosophy "Intention" (First published 31 August 2009, substantially revised 20 January 2014) <https://plato.stanford.edu/entries/intention/>.
Stanford Encyclopedia of Philosophy "Kant's Moral Philosophy" (first published 23 February 2004, substantially revised 7 July 2016) < https://plato.stanford.edu/entries/kant-moral/> at [10].
Stanford Encyclopedia of Philosophy "Memory" (first published 24 April 2017) <https://plato.stanford.edu/entries/memory/>
Stanford Encyclopedia of Philosophy "The Contents of Perception" (First published 18 March 2005, substantially revised 4 October 2016) <https://plato.stanford.edu/entries/perception-contents/#1>.
Stanford Encyclopedia of Philosophy "The Grounds of Moral Status" <https://plato.stanford.edu/entries/grounds-moral-status/ >. See the UDHR Preamble
Stanford Encyclopedia of Philosophy "The Moral Status of Animals" (first published 1 July 2003, substantially revised 13 September 2010) < https://plato.stanford.edu/entries/moral-animal/> at [2].
Stephen Mulhall "Fearful Thoughts" (2002) 24 LRB.
Stephen P. Stich "Do Animals Have Beliefs?" (1979) 57 Australasian Journal of Philosophy; Stephen P. Stich "Autonomous Psychology and the Belief-Desire Thesis" (1978) 61 Monist at 703.
T Merrill and HE Smith "The Morality of property Law" (2008) 45 WM & Mary L Rev; T Merrill and HE Smith "What Happened to Property in Law and Economics" (2001) 111 Yale LJ.
Takanisi "World Robot Declaration" (at the World Robot Fair, Fukuoka, February 2004).
Te Awa Tupua (Whanganui River Claims Settlement) Act 2017 s 14.
Te Urewera Act 2014 s 11.
Tejas D. Kulkarni, Karthik R. Narasimhan, Ardavan Saeedi and Joshua B. Tenenbaum "Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation" (2016).
Thomas H. Birch "Moral Considerability and Universal Consideration" (1993) 15 Environmental Ethics 313-332.
Tim Smithers "Autonomy in Robots and Other Agents" (1997) 34 Brain and Cognition 88 at 101.
Todd Hester and Peter Stone "Intrinsically motivated model learning for developing a curious agent" (paper presented to The Eleventh International Conference on Development and Learning (ICDL), San Diego, November 2012).
Tom Barraclough "How Far Can The Te Awa Tupua (Whanganui River) Proposal Be Said To Reflect The Rights Of Nature In New Zealand?" (LLB (Hons) Dissertation, University of Otago, 2013).
Tom Regan *The Case for Animal Rights* (University of California Press, California, 2004) at 21.
United Nations "Universal Declaration of Human Rights", Articles 1 to 30.
United Nations "Universal Declaration of Human Rights", Preamble.
United Nations Convention on the Rights of the Child.
European Convention on Human Rights.
W Wallach and C Allen *Moral Machines* (Oxford University Press, Oxford, 2010).
Wesley Newcomb Hohfeld "Some Fundamental Legal Conceptions As applied In Judicial Reasoning" (1913) 23 Yale L.J 16, 28-59.

Wrye Sententia "Neuroethical Considerations: Cognitive Liberty and Converging Technologies for Improving Human Cognition" (2004) 1013 Annals of the New York Academy of Sciences

Y N Harari *Homo Deus* (Harvill Secker, London, 2016) at 291.

Z Ghahramani "Probabilistic machine learning and artificial intelligence" (2015) 521 Nature.

Z v Dental Complaints Assessment Committee [2008] NZSC 55, [2009] 1 NZLR 1 at [26].