Moral Encounters of the Artificial Kind: Towards a non-anthropocentric account of machine moral agency

by

Fabio Tollon

Thesis presented in partial fulfilment of the requirements for the degree of Master of Arts (Philosophy) in the Faculty of Arts and Social Sciences at Stellenbosch University

Supervisor: Dr Tanya De Villiers-Botha

December 2019

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2019

Abstract

The aim of this thesis is to advance a philosophically justifiable account of Artificial Moral Agency (AMA). Concerns about the moral status of Artificial Intelligence (AI) traditionally turn on questions of whether these systems are deserving of moral concern (i.e. if they are moral patients) or whether they can be sources of moral action (i.e. if they are moral agents). On the Organic View of Ethical Status, being a moral patient is a necessary condition for an entity to qualify as a moral agent. This view claims that because artificial agents (AAs) lack sentience, they cannot be proper subjects of moral concern and hence cannot be considered to be moral agents. I raise conceptual and epistemic issues with regards to the sense of sentience employed on this view, and I argue that the Organic View does not succeed in showing that machines cannot be moral patients. Nevertheless, irrespective of this failure, I also argue that the entire project is misdirected in that moral patiency need not be a necessary condition for moral agency. Moreover, I claim that whereas machines may conceivably be moral patients in the future, there is a strong case to be made that they are (or will very soon be) moral agents. Whereas it is often argued that machines cannot be agents *simpliciter*, let alone moral agents, I claim that this argument is predicated on a conception of agency that makes unwarranted metaphysical assumptions even in the case of human agents. Once I have established the shortcomings of this "standard account", I move to elaborate on other, more plausible, conceptions of agency, on which some machines clearly qualify as agents. Nevertheless, the argument is still often made that while some machines may be agents, they cannot be moral agents, given their ostensible lack of the requisite phenomenal states. Against this thesis, I argue that the requirement of internal states for moral agency is philosophically unsound, as it runs up against the problem of other minds. In place of such intentional accounts of moral agency, I provide a functionalist alternative, which makes conceptual room for the existence of AMAs. The implications of this thesis are that at some point in the future we may be faced with situations for which no human being is morally responsible, but a machine may be. Moreover, this responsibility holds, I claim, independently of whether the agent in question is "punishable" or not.

Abstrak

Hierdie tesis het ten doel om ŉ filosofies-geregverdigde beskrywing van Kunsmatige Morele Agentskap (KMA) te ontwikkel. Gewoonlik behels die vraagstuk na die morele status van Kunsmatige Intelligensie (KI) twee vrae: die morele belang waarop sulke stelsels geregtig is (dus, of hulle morele pasiënte is) en of sulke stelsels die bron van morele optrede kan wees (dus, of hulle morele agente is). Die Organiese Benadering tot Etiese Status hou voor dat om ŉ morele pasiënt te wees ŉ voorvereiste daarvoor is om ŉ morele agent te kan wees. Daar word dan verder aangevoer dat Kunsmatige Agente (KA) nie bewus is nie en gevolglik nie morele pasiënte kan wees nie. Uiteraard kan hulle dan ook nie morele agente wees nie. Die verstaan van "bewustheid" wat hier bearbei word, is egter konseptueel en epistemies verdag en ek voer gevolglik aan dat die Organiese Siening nie genoegsame bewys lewer dat masjiene nie morele pasiënte kan wees nie. Ongeag hierdie bevinding voer ek dan ook verder aan dat die aanname waarop die hele projek berus foutief is—om ŉ morele pasiënt te wees, is nie ŉ noodsaaklike voorvereiste daarvoor om ŉ morele agent te kan wees nie. Verder voer ek aan dat, terwyl masjiene in die toekoms morele pasiënte *mag* wees, hulle beslis morele agente *sal* wees (of selfs alreeds is). Daar word dikwels aangevoer dat masjiene nie eens *agente* kan wees nie, wat nog van morele agente. Ek voer egter aan dat hierdie siening ŉ verstaan van "agentskap" voorveronderstel wat op ongeregverdige metafisiese aannames berus, selfs in die geval van die mens se agentskap. Ek bespreek hierdie tekortkominge en stel dan ŉ meer geloofwaardige siening van agentskap voor, een wat terselfdertyd ook ruimte laat vir masjienagentskap. Terwyl sommige denkers toegee dat masjiene wel agente kan wees, hou hulle steeds vol dat masjiene te kort skiet as morele agente, siende dat hulle nie oor die nodige fenomenele vermoëns beskik nie. Hierdie vereiste word egter deur die "anderverstandsprobleem" ondermyn—ons kan doodeenvoudig nie vasstel of enigiemand anders (hetsy mens of masjien) oor sulke fenomenele vermoëns besit nie. Teenoor sulke intensionele verstane van morele agentskap stel ek dan ŉ funksionalistiese verstaan, wat terselfdertyd ook ruimte laat vir masjiene as morele agente. My bevindinge impliseer dat ons in die toekoms ons in situasies sal bevind waarvoor geen mens moreel verantwoordelik is nie, maar ŉ masjien wel. Hierdie verantwoordelikheid word nie beïnvloed deur die masjien se kapasiteit om gestraf te word nie.

Acknowledgments

There are several people who deserve to be thanked for putting up with me during the writing of this thesis.

Firstly, I would like to thank my supervisor, Dr Tanya De Villiers-Botha. The rigorous standard you maintained in your feedback, attention to detail, and, most notably, your willingness to let me find my philosophical voice, are things I am incredibly grateful for.

Secondly, to Deryck, Daniel, and Lize. Thank you all for listening to me drone on about machines, agents, patients, and fish. Deryck for the cultural scaffolds you afforded, Daniel for making me sometimes take life seriously, and Lize for showing me what genius looks like.

Thirdly, a thank you to my family, without whom none of this would have been possible. A special mention to my sister, Liana, who has had to put up with my nonsense more than most. Your belief in me matters more thank you think.

Lastly, I am grateful to both my examiners Dr Susan Hall and Dr Chris Wareham, who provided insightful comments and suggestions which aided me in producing a polished final product.

Any and all errors that remain are my own.

Contents

Introduction

What exactly is Artificial Intelligence (AI)? Can a machine think? Should we accord moral status to all entities capable of thought? Can and should machines be held responsible or accountable for any actions of theirs that affect human beings? These (and many other) questions are becoming increasingly pressing in the philosophy of AI. How exactly one goes about answering these questions depends on many prior philosophical commitments and, especially when it comes to AI, the potential need for revising these commitments.

Our *own* intelligence has been an object of inquiry for biological human beings for thousands of years, as we have been attempting to figure out how a collection of bits of matter in motion (ourselves) can, perceive, predict, manipulate and understand the world around us (Russell and Norvig, 2010: 1). *Artificial* Intelligence, on the other hand, is "a cross-disciplinary approach to understanding, modeling, and replicating intelligence and cognitive processes by invoking various computational, mathematical, logical, mechanical, and even biological principles and devices" (Frankish and Ramsey, 2014: 1). AI attempts not just to *understand* intelligent systems, but also to *build* and *design* them (Russell and Norvig, 2010: 1). There are various models on which this is done. In this thesis, I will adopt a *rational agent* approach to understanding AI, which claims that when constructing an AI, the goal should be to create a system capable of achieving the best outcome, or when there is some uncertainty, the best outcome to be expected, given the task that it is to perform (*ibid.*: 4). This "best outcome to be expected" should be evaluated from the perspective of *action*, and in this way it is concerned with intelligent *behaviour* in artefacts[1] (*ibid.*: 5).[2] The reason for adopting this approach is twofold: firstly, it is more generally applicable than the more formalistic and restrictive "Laws of Thought" (LoT) approach, for example, which attempts to codify all knowledge and represent it in logical notation. This becomes a major obstacle when one encounters informal information or situations where we do not have absolute certainty with respect to the variables involved. Secondly, this rational agent approach is more amenable to scientific investigation, as opposed to approaches which are reliant on human behavior or thought (*ibid.*: 5). The reason for this is that the standard of rationality that AIs can deploy is mathematically well-defined

[1] An artefact is an object made by a human being (Johnson and Noorman, 2014: 144).

[2] This type of approach can be contrasted with three other broadly defined approaches to AI. Firstly, there is the "cognitive modelling" approach, which attempts to create machines that think just like human beings. Secondly, there is the "laws of thought" approach, which is the study of mental faculties via the usage of computational models. Thirdly there is the "Turing Test" approach, which aims to design machines that perform functions that, if performed by humans, would require intelligence. For a detailed discussion of all of the aforementioned methodologies see Russell and Norvig (2010).

and completely general, which contrasts sharply with the type of "rationality" exhibited by human beings in our day-to-day interactions (see Russell and Norvig, 2010: 5 for more on this). To note this contrast is not to claim that we are "irrational" in the sense that we are insane or emotionally unstable, but rather that we are not perfect decision makers (see Kahneman, 2000).[3] In this way the rational agent approach is not hamstrung by human biases in decision making, and can instead focus on the general principles that might be used in the construction of AI.

But what exactly is an "agent", and can any suitably programmed AI ever truly be considered a "rational agent"? The rational agent approach presupposes this, but is this presupposition well-grounded or the result of a faulty intuition? One common sense understanding of "agent" might be that it is something that can act, and it is quite clear that many already existing instances of AI can be construed as being capable of action in this common-sense specification. A simple example is that of a Roomba vacuum cleaner: when it is zipping around over the floor of your house it is clearly *doing* something. The crucial question then becomes whether this doing is in fact an action, and whether this kind of action should qualify our cleaning companions as *agents*. While Roomba vacuums might be a trivial example, as nobody is arguing that these machines have anything like the complexity required to qualify as agents in the sense that human beings are agents, they do raise an important question. This question is whether the transition from non-agent to agent is simply a matter of complexity, and whether this complexity can be specified by given criteria in a non-anthropocentric way. In other words, would it be possible to have necessary and sufficient criteria for a conception of agency that can capture both artificial and biological entities? Recent trends in AI research have been geared toward the creation of artefacts that act in ways that are increasingly autonomous, adaptive and interactive, which may eventually lead to these entities performing actions which are entirely independent from human beings (Floridi and Sanders, 2004). The possibility of the creation of these types of machines raises a number of pertinent ethical issues, concerning both our obligations toward such entities and the type of responsibility ascribed to them for any actions they undertake independently of substantive human influence (Bostrom and Yudkowsky, 2011: 1).

In this introductory chapter, I will outline the various contours of the debate surrounding the moral status of machines. This will involve a brief outline of the metaphysics of agency, and

[3] See Bortolotti (2015) for a related discussion, in which she seeks to undermine the "rationality assumption" imbued in most of western philosophy.

how recent trends in technological development may come to disrupt standard conceptions of the relationship between agents, actions and events. Following from this I will put forward a series of provisional definitions with the hope that they will serve as a means of disambiguating any claims which are made throughout the paper.

The Metaphysics of Agency

This dissertation will investigate the nature of agency and then seek to apply this discussion to one of the most philosophically interesting moral questions to date: can an intelligent machine[4] be a moral agent? With this in mind, I will unpack some key distinctions in the metaphysics of agency, which deals specifically with the relationship between agents and actions (Schlosser, 2015). Within this framework I will address the multiple issues that arise when we retain an anthropocentric conception of agency. As will become clear, one of the most pernicious issues in this debate is the failure of standard accounts of agency to account for the emergence of increasingly independent artificial artefacts.

Justification

At first glance, this type of argument might seem preposterous. How can we ever hold *machines* responsible for their actions? Generally, most of us only allow that other human beings and (hopefully) animals are proper subjects of moral concern, viewing them as having a moral *stake*. Moreover, we tend to consider competent adult human beings to be responsible for their actions and thus morally accountable. To argue that a machine or suitably complex algorithm should be awarded the same (or similar) type of moral status is *surely* absurd. *We* are responsible in the sense that we can claim *authorship* for what we have done, by having *intentions* which can be guided by the use of reason (Wegner, 2002). Our *reason responsiveness* is crucial, as the process of reason-giving and -accepting has arguably been the key to our cultural evolution and the development of our moral frameworks (see Dennett, 2003). Furthermore, we view ourselves as autonomous and our reasons as our own, and it is in this way that we come to be responsible for what we do.

[4] A machine is a complex system composed of artificially constructed components which, when taken together, can perform certain tasks and/or operations.

Consider the approval some may experience when using all the new-fangled features of a modern gadget, such as a smartphone, or the anger and disapproval we might feel towards the same phone when it fails to perform all of its stipulated functions. In our more reflective moods, we appreciate the illogicality of these responses, and, moreover, that these reactions have no *moral* significance. After some reflection we recognise that such emotional responses to inanimate objects are unreasonable: there is a clear distinction (at least intuitively) between things that really are worthy of moral responses (and/or concern) and things that are not. While we might be proud of these artificial entities and get angry at them when they "misbehave", we do not think they are being disobedient or are "out to get us". In other words, we do not consider their failings to be the result of them misbehaving; rather, we might instead say that they *malfunction* (Johnson and Noorman, 2014: 154).

Moreover, when we evaluate moral situations, we tend to think in terms of giving moral stakeholders their due: giving them what they *deserve* based either on how they have behaved or whether they have been harmed. To go back to the example of the malfunctioning phone: there arises, firstly, the question of whether the phone is misbehaving *intentionally*, in the common-sense usage of the term ("on purpose"), and whether the phone could in some sense be *responsible* for its behaviour, and hence could possibly be held morally responsible. This is a question of moral *agency*. Conversely, the further question may arise of whether, if we wanted to punish the phone by, for example, beating it with a stick, would we be doing it a *moral harm*. In other words, do we owe it certain moral *obligations*? This is a question of moral *patiency*. These two questions can be viewed as fundamental to all moral philosophy: *who* or *what* is deserving of moral concern, and *who* or *what* can be said to be (morally) responsible for their actions (Gunkel, 2012)? On the one hand, the emergence of artificially intelligent systems, properly conceptualised as artificial agents[5] (AAs), may complicate many presuppositions of who or what can count as a source moral action. On the other hand, trends in contemporary macro-ethics have been geared towards expanding the boundaries of moral concern by focusing on the nature of who or what should count as a moral patient, independent of whether the entity in question is a moral agent or not (Floridi and Sanders, 2004).[6] Added to this is the prevailing

[5] An artificial agent is artificial in the sense that it has been manufactured by intentional agents out of pre-existing materials, which are external to the manufacturers themselves (Himma, 2009: 21). It is an "agent" in the minimal sense that it is capable of performing actions (Floridi and Sanders, 2004: 349). A simple example of such an artificial agent would be a cellphone, as it is manufactured by humans and can perform actions, such as basic arithmetic functions or responding to queries via online searches.

[6] A patient-orientated approach to ethics is not concerned with the perpetrator of a specific action, but rather attempts to zero in on the *victim* or receiver of the action (Floridi, 1999). This type of approach to ethics is considered non-standard and has been incredibly influential in both the "animal liberation" movement and "deep

assumption in the literature on artefactual agency that a necessary condition for being a moral agent is that one is also a moral patient (see Floridi and Sanders, 2004; Torrance, 2008). It is with this in mind that any investigation into moral agency must first address the question of moral patiency. Important for the purposes of this thesis is that the implications of adopting this framework for machines are clear: if a machine cannot be considered a moral patient, then it cannot be a moral agent either. I, however, will argue that moral patiency is not a necessary condition for moral agency.

A brief consideration of "the animal question" may serve as a useful illustration of how questions of agency and patiency have come to change how we view the moral boundaries dividing Us and Them. Descartes understood the animal and the machine as indistinguishable, referring to animals as mere *automata* (Gunkel, 2014: 119). In this way Descartes instantiated a dualism between the world of the animal and the world of human beings. According to Descartes, animals, unlike human beings, lack reason and by extension the capacity for rational thought. In this way, they operated like mindless entities, executing predetermined responses to external stimuli, which is perhaps a less scientific way of saying that the behavior of *all* animals is genetically predetermined. By likening animals to automata, Descartes was making a deep ontological point about both machines and animals: both are composed of a different *substance* when compared with humans, and so have a shared ontological identity, which marks them as both substantively distinct from and inferior to human beings (Gunkel, 2012: 3). While human beings, according to Descartes, can claim to be autonomous paragons of reason, machines and animals are simply following a deterministic set of instructions. This implies that neither animals nor machines could be agents. Due to this, animals and machines are not included in our moral universe, as it makes no sense to punish an action if the entity in question *could not have done otherwise*. Moreover, as neither are capable of the affect that Descartes thought was due to mind only, they were not moral patients, as they could not suffer harm. Recently, though, the question of animal affect has been revised and they are taken to be capable of suffering (see Singer, 1975, 2011). Hence, non-human animals are now broadly considered to be legitimate subjects of moral concern: not as agents, but as patients. The subject matter of the remainder of this thesis will be centered around whether we might be able to perform a similar expansion of our moral universe with respect to machines, but as potential moral *agents* rather than patients.

ecology" approaches to environmentalism (Leopold, 1948; Naess, 1973; Singer, 1975, 2011). The latter both place an emphasis on the *victims* of moral harms; the harm we do to animals and the environment respectively.

My goal in this thesis is therefore to challenge the seemingly innocuous intuition that machines can never be subject to moral assessment for the actions that they perform. I am not here claiming that currently existing artificial systems must be ascribed moral responsibility for their actions, but, rather, that we should seriously consider the possibility that in the (near) future we may have to once again broaden our moral boundaries, given current technological developments. With the above considerations in mind, my thesis statement can be summarized as follows: The emergence of complex artificial artefacts will in all likelihood force us to extend the boundaries of the concept of agency. In the near future, the key conditions we deem necessary and sufficient for agency may be met by these technological systems, and when that happens, we must be ready to admit that these agents will also come to be sources of moral action, making them moral agents that can be held responsible for their actions.

This thesis statement claims many things. The jump from "mere" agency to moral agency is perhaps the most controversial claim that I make. Then there is also my exclusive focus on the concept of agency and the seeming neglect of moral patiency. In later sections, I will provide arguments defending these claims and my approach. For now, however, I would like to orientate the reader with an overview of what exactly the metaphysics of agency entails. In what follows I will define some of the more technical terms I will be using in this thesis and attempt to address any ambiguities with which I can foresee the reader having problems.

Disambiguation

Before going any further, it would be helpful to introduce some tentative definitions in order to avoid any misunderstandings as my argument progresses. The first two concepts to be defined are *moral agency* and *moral patiency*:

(1) Moral Patients: A class of entities that can in principle qualify as receivers of moral action.

(2) Moral Agents: A class of entities that can in principle qualify as sources of moral action (Floridi and Sanders, 2004: 349-350).

In this paper I will focus most of my attention on the analysis of the concept of a moral *agent*—more specifically, in the case of machines, *artificial* moral agents. Thus, we should be clear on the difference between "natural" and "artificial" entities:

(3) Natural entities/systems: these entities are natural in the sense that their existence can be explained in terms of physical and biological processes that are not the result of human artifice.

(4) Artificial entities/artefacts/systems: these entities are artificial in the sense that they are manufactured by intentional agents (i.e. humans) out of pre-existing materials, which are external to the manufacturers themselves (Himma, 2009: 21). Machines and AIs are examples of artificial entities.

In this paper, therefore, I will be interested in and focus on the moral status of *artificial* entities. In the category of artificial entities there are two further "types":

(5) Telerobots: these are remotely controlled machines that make only minimally-autonomous decisions.

(6) Autonomous machines: machines that are "autonomous"[7] in the engineering sense of the term, which simply means that these robots must be capable of making at least some of their major decisions based on their own programming (Sullins, 2011: 26).

This thesis will have implications for our understanding of the class of artificial entities known as autonomous machines. When these robots make decisions, the programmers are to some extent responsible for their actions, but perhaps not wholly so (this is an idea that will be developed further in my final chapter) (*ibid.*). The type of "responsibility" attributable to these robots could vary between the "decision" of a robotic vacuum cleaner to ram itself into your foot, to the complex future scenario in which a robotic caregiver might have to interact with a person in need of urgent medical care. Regardless, the machines which will potentially pose the most interesting moral questions are those that are not yet in operation, but which, using current technological capacities, can be predicted to arise in the near future. Most of our present-day AI systems are in a strong sense tethered to the interests and intentions of their human operators: either through deterministic programming, or else via interactive control

[7] The history of "autonomy" is a nefarious philosophical problem on its own, but I will not concern myself with these issues at this point. In later parts of this thesis I will problematize the usage of "autonomy" in this sense in the case of machines, but for now this definition serves my purposes.

through some form of human supervision. These technological systems mostly function as an extension of ourselves, and it is clear that the moral responsibility for any actions performed by these machines lies in the hands of the human operator or programmer. However, the very real possibility of AAs that act in more autonomous/independent ways could disrupt this way of viewing our moral relationship to technology. Technological examples of this possibility come from self-driving cars and the military application of autonomous drones (see Sparrow, 2007; Müller, 2014; Nyholm, 2017). In both cases there is an explicit goal on the part of developers to make these systems act independently of human control, and the success of unpiloted drones and of self-driving cars attest to the efficaciousness of this design goal.

The question of what to do with an artificial system which is capable of having a causal influence on events in a given context, and is not tethered to human action, thus arises. More specifically, the question arises as to what type of moral status can be coherently attributed to such artificial systems in a way that is unencumbered by anthropocentric intuitions regarding moral agency.

The Road Ahead

Before getting into the details of agency, however, the question of moral patiency must be addressed. Animals are rightly considered to be moral patients and are therefore accorded a type of moral stake: we are not to unnecessarily harm animals as they have the capacity to *experience pain* or to *suffer*. This capacity for subjective states is rooted in them being *sentient* creatures. With the aforementioned in mind, I will therefore outline one of the most intuitive accounts of moral standing: the Organic View of Ethical Status (Torrance, 2008). The main justification for this strategic move is a presupposition that has served to inform most discourse in the literature surrounding the possibility of artificial moral agency, namely the purported ontological relationship between moral agency and moral patiency. As I outlined earlier, this presupposition claims that, in principle, only moral patients can be moral agents. In other words, moral patiency is a necessary condition for moral agency (Floridi and Sanders, 2004). In my next chapter, therefore, I provide an exposition and critique of Torrance's influential defense of this presupposition. In the following chapter I sketch the possible contours along which the concept of artificial agency may refer. As I will show, it is an illegitimately anthropocentric assumption that only entities with *moral autonomy* or *intentionality* (in a problematic sense that will be specified) can be moral agents. Moreover, these criteria incorporate metaphysically contested concepts into our understanding of moral agency, making it uncertain whether human beings even qualify as moral agents on this view. Following this, in my third chapter, I will provide a *functionalist* account of agency, and also of moral agency, that is not vulnerable to the same kind of objections as the intentional account detailed in Chapter Two. I then detail two possible ways to understand this functionalist account, in the form of conservative and progressive approaches to moral agency. I will claim that we should be progressive in our conceptualization of moral agency, with the implication that an agent can be morally responsible without necessarily being able to appreciate its “punishment”. In my conclusion I will show how a progressive account of functionalist moral agency is to be preferred on both philosophical and methodological grounds. Moreover, such an account will allow us to deal with the challenges posed by morally efficacious artificial agents in a way that is consistent with our dealings with human moral agents.

Chapter 1: Patiency Fails

In my introduction I touched on one of the basic assumptions of standard approaches to ethics, and its implications for moral agency: that moral agents need to be moral patients. In this chapter I will delineate one of the most intuitive accounts of who or what is deserving of moral consideration: The Organic View of Ethical Status (hereafter simply the "Organic View") (Torrance, 2008). On this view only moral patients can ever be moral agents. Moreover, this common-sense view claims that, in principle, only things which are biologically alive can ever be subjects of moral concern and, hence, by extension, sources of moral action. One justification for this approach is that in order for an entity to be assigned responsibility for an action it must have some kind of "moral sense", and only entities that are moral patients have this capacity. A good articulation of this view comes from Steve Torrance (2008), who claims that in order to even consider the question of moral agency, the entity in question must first answer to a prior judgement in which it is deemed to be a moral patient. The arguments that he presents in defence of this view contain many of the characteristics that make a regular appearance in the literature on machine moral agency/patiency, such as questions of sentience, intentionality, and the conceptual relationship between moral agents and moral patients (see Floridi and Sanders, 2004; Johnson and Miller, 2008; Himma, 2009; Sullins, 2011; Johnson and Noorman, 2014). Therefore, if the Organic View can be undermined then many "standard" assumptions in the literature can also be shown to be unsound.

One of the main tenets of the Organic View is that the ascription of moral *patiency* is a necessary condition for moral agency. In what follows, I will argue that this intuitive account of moral ascription relies on illegitimate anthropocentric presuppositions, as opposed to sound philosophical argumentation. After showing that the Organic View fails to provide a valid and coherent account of moral patiency in the first place, I go on to propose an alternative, more plausible, account, which I will argue also lends itself to philosophical investigations into the possibility of machine moral patiency. More importantly, I will show that on this more plausible account, moral patiency need *not* be a precondition for moral agency. I will go on to argue that the issue of moral *agency* is much more pressing than moral patiency in the case of machines, and so that will be the focus of the rest of my thesis. Before doing this, however, in order to contextualize current work on moral patiency and agency, I will provide a brief overview of "standard" versus "non-standard" approaches to ethics, and how there has been a general shift towards "non-standard" approaches in recent history.[8] It is due to this shift that the question of moral patiency has become paramount, as non-standard views focus on the *receivers* of moral actions, as opposed to the agent-orientated approach of standard accounts.

1.1 Moral Patiency

As per definition (1) (see introduction), a moral patient refers to the class of entities that can in principle qualify as receivers of moral action. In other words, if an entity is a moral patient, it would be of moral concern. It would be an entity towards which we would have certain moral duties/obligations and responsibilities on a given moral theory (Gunkel, 2012: 93). In recent years, trends in macro-ethics more generally have been geared towards expanding the boundaries of moral consideration by focusing on the nature of who or what should count as a moral patient, independently of questions relating to whether the entity in question is a moral agent or not (Floridi and Sanders, 2004). [9] Significantly, however, on this account all moral agents are moral patients. This type of approach has been termed "non-standard" and stands in contrast to standard approaches. Standard approaches claim that there is a one-to-one correspondence between moral agents and moral patients: all moral agents are moral patients and vice versa (*ibid.*: 350). Confusingly, the standard approach to ethics is currently relatively uncommon, with non-standard approaches dominating contemporary ethical discourse. Non-standard approaches to ethics have been motivated, in part, by human beings having a bad track record when it comes to extending our boundaries of moral concern. As an example, for a time, the scope of moral patiency in western countries and colonies was only extended to white Europeans, with slaves not being considered worthy of any moral consideration. Thankfully, over time, we have come to appreciate that all human beings, no matter what creed or colour, are rightly deserving of moral concern and are therefore moral patients. The underlying philosophical arguments in favour of the expansion of our moral universe has centered around, in the case of slaves, the correct understanding of personhood[10], and, secondly, in the case of

[8] "Standard" approaches to ethics focus on the perpetrator of the action, and instead of asking "can they suffer?" (as in non-standard approaches) asks "are they rational?".

[9] A patient-orientated approach to ethics is not concerned with the perpetrator of a specific action, but rather attempts to zero in on the *victim* or receiver of the action (Floridi, 1999). This type of approach to ethics is considered non-standard and has been incredibly influential in both the "animal liberation" movement and "deep ecology" approaches to environmentalism (Leopold, 1948; Naess, 1973; Singer, 1975, 2011). In both aforementioned approaches there is an emphasis on the *victims* of moral harms: on the one hand the harms inflicted upon animals, and on the other the environmental harm enacted upon our planet.

[10] The question of whether an artificial entity may come to be considered a "person" in the morally (or legally) relevant sense is both a pertinent and tricky one to answer without a detailed exposition. I will not address this

non-human animals, around whether they can suffer or not (Singer, 1975). It is with the above in mind that any investigation into machine moral agency needs to first address the question of moral patiency. The traditional defense of this position (which finds expression in the Organic View) is that in order to be morally responsible for an action (a moral agent) an entity must be capable of moral reasoning. This moral reasoning is taken to include the capacity for a type of "moral sense", which requires that the entity also be a moral patient. The implications of this non-standard approach to ethics for machines is clear: if a machine cannot be considered a moral patient, then it cannot be a moral agent either. In this chapter, I will seek to problematize this assumption by showing how both arguments *against* machine moral patiency and claims that moral patiency is a requirement for moral agency, are not only illegitimately anthropocentric but are predicated on an invalid account of (human) patiency.

Singer's (1975) arguments in support of the equal consideration of all sentient life served as the philosophical foundation of the animal rights movement, a movement focused on the rights of specifically *non-human* entities. Here, it is quite unproblematic to assume that animals can be moral patients. Moreover, questions of animal agency do not often arise. A machine can also be construed as a non-human entity, and so the question to now be considered is whether machines might also be deserving of some kind of moral concern, and if so, on what grounds? The question is further complicated by the possibility that machines may more plausibly be thought of as agents than animals are. To illustrate the issues that arise, I will focus on Torrance's articulation of the Organic View. It is perhaps the most intuitive view of moral ascription as it claims, straightforwardly enough, that only biological systems can ever, in principle, have moral agency. This intuition seems to have some purchase, as it makes little sense to hold an AA morally responsible for an action if it does not have some kind of psychological capacity or "moral sense" with which to reflect on the moral action, and it is claimed that machines, like animals, do not have this capacity (Floridi and Sanders, 2004: 367). In order to make his case, Torrance centres his discussion around two factors which feature prominently in the Organic View: firstly, he claims that *sentience* (or phenomenal consciousness) is a key factor in the type of rationality proper moral entities (humans) exhibit, and, secondly, that biological constitution is of fundamental moral significance for this capacity (2008: 505). In this chapter, I specifically focus on the first of these claims, and the reason for this explicit focus will become clear in my exposition of Torrance's account. While Torrance

question here, as it is not essential to my current argument. See Gunkel (2012: 42-54) for a discussion of this topic.

does not explicitly claim to endorse the Organic View, he does have a favourable disposition towards it. He claims to be willing to concede that it may be wrong (or at least in need of further qualification) (*ibid.*: 505). I will attempt to show that the Organic View is not just in need of further qualification, but rather in need of a complete revision of its presuppositions and philosophical methodology. In order to make my argument, I will first put forward the case made by Torrance (*ibid.*: 503) that AAs do not have "empathic rationality", with the implication that machines, unless they can be designated as "sentient", cannot be proper subjects of moral concern (i.e. moral patients). From this, I then show how the conception of sentience Torrance operationalises in his account is illegitimately anthropocentric and in need of revision due to conceptual and epistemic shortcomings.

1.2 The Organic View of Ethical Status

According to Torrance (*ibid*) there are five key components to the Organic View:

> 1. There is a crucial dichotomy between beings that possess organic or biological characteristics, on the one hand, and 'mere' machines on the other.
>
> 2. It is appropriate to consider only a genuine organism (whether human or animal; whether naturally occurring or artificially synthesized) as being a candidate for intrinsic moral status—so that nothing that is clearly on the machine side of the machine-organism divide can coherently be considered as having any intrinsic moral status.
>
> 3. Moral thinking, feeling and action arises organically out of the biological history of the human species and perhaps many more primitive species which may have certain forms of moral status, at least in prototypical or embryonic form.
>
> 4. Only beings, which are capable of sentient feeling or phenomenal awareness could be genuine subjects of either moral concern or moral appraisal.
>
> 5. Only biological organisms have the ability to be genuinely sentient or conscious (*ibid.*: 502-503).

Torrance believes that only moral patients are capable of being moral agents (*ibid.*: 509). This type of view is reflective of a broader intuition outlined earlier, which is captured in the non-standard approach to ethics. The intuition is that it is only appropriate to morally appraise the actions of a specific kind of being, one which is, in the first case, a proper subject of moral concern. Only entities that are subjects of moral concern (i.e. moral patients) can be held to certain moral requirements (i.e. moral agents), as only moral patients are capable of the appropriate kind of moral reasoning required for moral agency. In this way his arguments, as they will be presented below, revolve around the question of patiency as the key factor in determining the type of moral ascriptions we might give to machines (now or in the future). In other words, for Torrance, if we wish to assign the capacity of moral agency to an entity, this ascription must answer to a prior judgement of whether the entity in question is a moral patient.

1.2.1 Empathic Rationality

In this section I deal specifically with claim four of the Organic View: "Only beings, which are capable of sentient feeling or phenomenal awareness could be genuine subjects of either moral concern or moral appraisal" (*ibid.*: 503). The reason for focusing on this aspect of the Organic View is that, if it is found wanting, it undermines the entire argument. The criterion of sentience is what grounds Torrance's conception of agency, and so if this criterion fails specifically then so does Torrance's account of moral agency more generally. If this aspect of his argument is faulty, then, there is no room for moral agents, which would contradict our ordinary conceptions of ourselves as moral agents. This will become clear as my critique develops.

Torrance begins his argument by asking us to imagine an AA that has a certain minimum level of rationality and has the cognitive ability to recognise that certain beings have sentient states, and thus moral interests. Moreover, the AA can reason about the effects that different courses of action may have on these sentient creatures. Yet, this type of agent does not have the capacity to *feel* moral harms (i.e. is not a moral patient, on Torrance's construal). Such agents, due to their ability to *cognitively* apprehend and interpret the behavioural cues of other entities, and to infer from these that the entity in question could be undergoing a moral harm, etc., *might* be thought of as being fitting subjects of moral appraisal (*ibid.*: 510).

Nevertheless, the problem with this view, according to Torrance (*ibid*), is with assuming that the type of rationality required for moral agency is simply cognitive or intellectual, as this would provide us with an anaemic account of moral standing. Torrance suggests that the kind of rationality that is required for an entity to legitimately be given the status of moral agent may turn out to be different from the kind that could be achieved by an AI system. He argues that the type of rationality traditionally associated with *humanity's* moral responsibility is fundamentally tied to our sentient nature (in other words, our capacity for *affect*). Thus the claim is that being a moral agent requires (human) sentience (or affect) (*ibid*: 510). The argument goes as follows: our kind of rationality involves the capacity for a kind of affective or empathetic identification with the experiential states of others, where such identification is integrally available to the agent as an essential component in its moral decision-making procedures (*ibid.*: 510). Torrance (*ibid.*: 516) calls this kind of rationality *empathic rationality* and contrasts it with the purely *cognitive* or *intellectual rationality*, which might be attributable to intelligent, computationally-based AAs. While we expect information-processing systems to make decisions in a purely mechanistic way, Torrance claims that we have different standards when it comes to our moral decision-making procedures, as we expect human beings to factor the potential experiential consequences of their actions into their moral reasoning (*ibid.*: 511). Significantly, he claims that entities that are only capable of intellectual rationality would not have a "real" or "true" understanding of the experiential states of others. Such an entity could simply not understand how its actions might affect others. Hence, due to their lack of capacity for affect, not only can AAs not be considered to be moral patients as they cannot suffer or be harmed, but, more importantly for our purposes, AAs also fail to qualify as moral *agents* as they are necessarily incapable of moral reasoning.

Thus, Torrance's argument is that moral decision making requires the capacity for "engaged empathic rational reflection" (*ibid.*:511), which requires the ability to identify with the experiential states of others. Any rational agent that is not also sentient (in a manner equivalent to human sentience) would not have this empathic ability, since a precondition for a "true" understanding of experiential states is that one is able to have these states oneself. Since only entities capable of being "ethical consumers" can have this type of empathic rationality (*ibid.*: 499) other types of agents are precluded from being subject to moral evaluation, as without the ability to take a "moral point of view", it would be absurd to then evaluate actions undertaken by such agents using moral criteria.[11] On the Organic View, then, we are forced to

[11] The example of a psychopath is interesting in light of the present discussion as it is assumed that while having the ability to reason *practically*, psychopaths appear to lack the ability to reason *morally* (Litton, 2008: 350). This seems to map on to the argument presented by Torrance, as he claims that while machines can reason practically, they cannot reason morally. However, in the case of the psychopath it can be argued that there is, at a deeper level, still a cognitive deficit which leads to this moral inability, something Torrance is not willing to admit is at issue in the case of the machine (see Litton, 2008; Torrance, 2008).

conclude that entities lacking a specific type of sentience cannot be moral agents. I will claim that this way of viewing moral ascription is flawed, and that we ought to steer clear of a reliance on the supposed presence of internal, qualitative states as a justification for such ascriptions. For now, however, let us continue along the counters of the Organic View, as there is a deeper presupposition which must first be explicated before my critique can be put forward.

The question that now arises for Torrance, is what, at a deeper level, results in creatures with the capacity for sentience (i.e. moral patients), whatever their functional similarities to artificial systems, that are worthy of moral concern? In other words, what is it about the *constitution* of AAs that excludes them from having a moral stake and thus from being morally appraised? According to the Organic View the biological makeup of these creatures is causally important with respect to their moral standing (*ibid.*: 511). The next section will explore the justification that Torrance provides for this assumption.

1.2.2 Self-Maintenance

The central claim made by Torrance that will be addressed at this juncture is that there is an essential link between moral categories and categories of biological organism. What this implies, for Torrance's purposes, is that morality is the domain of (biological) creatures that have an internally organized existence (and by extension have the capacity for affect) rather than an externally organized one – that is, creatures that exist not simply as artefacts whose constituent parts have been cobbled together by external designers, but which exist in a "more autonomous" sense (*ibid.*: 512). In this way, no artificial entity (see definition (4) in the introduction) can be of any moral concern, and only certain natural entities (see definition (3) in the introduction) can have this standing. Moreover, the population of our moral universe should only contain entities which are *self-organizing* (*ibid.*: 512). These entities, by virtue of being self-organizing, are by extension *self-maintaining*, with an inherent drive to survive (*ibid.*: 512). Biological organisms, by actively engaging with their environments and maintaining a boundary between themselves and the world, perform tasks which are not accessible to electronically powered, computational mechanisms. These AAs have no *inherent* motivation to self-maintain (i.e. they do not "care" what becomes of them): it is their external makers who perform this task and who care (*ibid.*: 512). Any AA would thus fall outside of this specification, as they would all be the product of *human* research and development, meaning that any sense of "meaning", "understanding" or "valuing" they may exhibit would be derivative from the intentional design bestowed upon them by their (hopefully) benevolent carbon overlords (us). Only entities that are self-maintaining in this sense, then, have the capacity for affect, and by extension can be considered sentient. In this way, Torrance excludes the possibility of AIs being considered sentient as he believes that sentience/affect cannot be explicitly programmed.

This type of argument does not in principle exclude the possibility that a kind of artificial life may come to exhibit the type of self-maintenance outlined above. What Torrance is instead claiming is that no entity without this type of self-care and self-maintenance will be capable of having phenomenal/qualitative/affective aspects to its experience (*ibid*.: 500).[12] Torrance, therefore, should not be read as necessarily being in disagreement with "the Principle of Substrate Non-Discrimination" which 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" (Bostrom and Yudkowsky, 2011: 8) [emphasis mine].[13] On this principle, substrate, holding other variables constant, lacks fundamental moral significance. What Torrance (and the Organic View more generally) *is* claiming, however, is that the type of conscious experience that human beings have the capacity for is different from that of any artificially constructed agent, because of the different causal histories associated with each kind of entity. On this view, human beings (and other biological entities) are *autopoietic*, while artificial entities are not (Torrance, 2008: 513). Autopoiesis is a term of art borrowed from the philosophy of biology, and it essentially designates a class of creatures that are *self-creating* (*ibid*.: 513). Here, the notion of an autopoietic system is meant to serve the function of a scientific support structure which can buttress the empirical validity of the Organic View. The type of self-creation exhibited by biological systems is purportedly characterized by "the appropriate exchange of its internal components with its environment, and via the maintenance of a boundary with its environment" (*ibid*.: 513). By way of this continuous interaction with its environment "autopoietic entities are radically distinguished from 'mere' mechanisms, since, unlike the latter, they enact their own continued existence, and their own purpose or point of view" (*ibid.:* 513). As should be clear from the previous quote, an essential component of this account is the notion of "lived experience" or "sentience" (*ibid*.:

[12] Torrance does not believe that functionalist accounts of mind fully capture the qualitative aspects of experience. He thus believes in the metaphysical possibility of "philosophical zombies", humans which look and behave indistinguishably from us but lack phenomenal conscious states of experience (Torrance, 2008). This is a thorny philosophical issue in its own right, but I will not go into further detail here.

[13] Bostrom and Yudkowsky (2011: 8) claim that not upholding this principle would amount to endorsing a kind of racism: in the same way that skin colour lacks fundamental moral significance, so does substrate.

515). These are supposed to require a particular causal history—one exhibited by biological creatures but not by AI. In this way, Torrance believes that he has managed to justify the exclusion of AAs from the realm of moral consideration, at least until they are capable of exhibiting the type of autonomous self-organisation and self-maintenance outlined above (*ibid.*: 515). To put it rather crudely, on the Organic View, machines cannot be proper subjects of moral concern because they are not *biologically alive* (Gunkel, 2012: 129).

Nevertheless, it should be clear that Torrance's usage of autopoiesis to buttress the Organic View simply passes the buck: his claim that only systems that are internally organised or self-maintaining avoids the accusation of violating the principle of substance neutrality by focussing on causal history instead. He suggests that causal history, rather than substrate, determines sentience and thus claims to uphold the Organic View's claim that only biological entities can have phenomenal states of mind and can thus count as proper subjects of moral concern. Yet, Torrance simply claims that autopoietic systems have a "point of view" that equates to phenomenality without making an argument to this effect (Torrance, 2008: 513). He therefore equivocates between "internal [self]-organization" and sentience. While making the case that biological systems are self-organising, he assumes that they are also the only entities capable of being sentient, since he believes that this follows from having "a point of view" in the sense of having a locus of self-maintenance. The first thing to notice is that single-celled organisms and plants are also self-maintaining and therefore autopoietic in this sense, but it would be quite a stretch to argue that they have phenomenal mental states. Secondly, there seems to be no argument for why exactly this *particular* type of causal history is required for the emergence of phenomenal awareness: what excludes the possibility of it arising from programming, for example? There seems to be no *necessary* reason for why the emergence of something like phenomenal consciousness is precluded from occurring in artificial systems. Moreover, Torrance does not explain why we should not consider all biological entities as sentient, seeing as they would all be autopoietic systems by definition.

As should be clear from the above critique, the criterion of sentience is *the* key to moral ascription on the Organic View, notwithstanding whether or not we can describe the system as self-maintaining or not. If empathy is integral to the way in which moral reasoning operates and, furthermore, empathy is necessarily tethered to sentience (which in turn implies moral patiency), then according to the Organic View we are forced to conclude that entities lacking sentience cannot be moral agents. I claim that this way of viewing moral ascription is problematic in that we cannot even be sure that human beings necessarily have the requisite internal states. I will show that this conception of moral agency relies on a conception of sentience that is unwarranted. We ought to steer clear of a reliance on internal, qualitative states as the sole justification for our moral ascriptions. In my critique of the Organic View, therefore, I will specifically focus on the issues surrounding the usage of sentience as central to the construction of our moral landscape.

1.3 Problems with the Organic View of Ethical Status

The first ambiguity that needs to be addressed is the vague way in which internal, experiential states are operationalised in Torrance's articulation of the Organic View. Here, only organisms capable of having some kind of "qualitative experience" of pain (or any other such experiential state) will qualify as moral patients (and by extension as moral agents).[14] As we saw, through the mechanism of empathic rationality, entities capable of having experiential states can use these affective responses to guide their reasoning procedures and in this way come to adopt a "moral point of view". Anything which is incapable of this empathic form of reasoning, on the Organic View, cannot be a proper subject of moral concern, as such entities would be incapable of engaging in the type of moral decision-making required for this type of attribution. They would be incapable of factoring into their reasoning how their decisions may impact the experiential states of others, as, due to their inability to have these experiential states themselves, they would not have a "real" understanding of these states. Moreover, Torrance (2014) is a realist about mental states and claims that there is an *objective* answer to the question as to an entity's psychological state. This realism about mental states works to reinforce his views regarding our moral ascriptions to AAs: Torrance's specific form of realism claims that even if there were no functional or cognitive difference between an artificial and biological system, there would still be a *phenomenal*[15] difference (*ibid.*: 13). This phenomenal difference is of fundamental moral significance for Torrance, given his claim that some form of conscious experience is a prerequisite for moral patiency. While I do feel that this presupposition hamstrings his argument, I will not go into any specific detail in this regard.[16] My focus is more

[14] For the sake of argument, I focus here on the experience of pain, but logically it would be possible to subject any type of internal mental state to the same type of analysis. Any theory which posits an "experience of *X*" claim must eventually answer to the question of *who* or *what* (i.e. what *type of mind*) is *experiencing*, or capable of experiencing, *X* and how we can know that.

[15] Phenomenal in the sense of having the capacity for conscious awareness.

[16] My own view is that there is in fact no difference between what can be "functionally" known about the mind and "phenomenal" aspects of mind: the phenomenal is a just a special case of the functional, and in this way, there is no "hard problem" of consciousness. See Chalmers (1996) for a defense of the hard problem, and Cohen and Dennett (2011) for a subsequent critique.

general and is more concerned with the inherent ambiguity in the operationalisation of "phenomenal" aspects of experience as a justification for moral concern. In what follows I, firstly, bring to light conceptual ambiguities inherent to the Organic View, and, secondly, discuss how the distinction between the mere "appearance" of something and the "real thing" operationalised in the Organic View is a problematic one.

1.3.1 Conceptual Issues

To see the ambiguity more clearly, an example put forward by Daniel Dennett (1996) offers a wonderful (albeit grisly) illustration of this by using the case of an amputated limb. Dennett asks us to imagine that:

> A man's arm has been cut off in a terrible accident, but the surgeons think they can reattach it. While it is lying there, still soft and warm, on the operating table, does it feel pain? A silly suggestion you reply; it takes a mind to feel pain, and as long as the arm is not attached to a body with a mind, whatever you do to the arm can't cause suffering in any mind (1996: 16-17).

Our intuition is that, although it might be possible to argue that the detached arm on the table may be capable of adverse nerve stimulus (i.e. pain), without being attached to some kind of mind this pain can never constitute suffering. The *experience* of pain is equivalent to suffering, and without an *experiencer* pain in itself can be of no moral significance (Gunkel, 2012: 115). At this point a defender of the Organic View can agree, as this seems to be the exact point that they are arguing for, as only *genuinely* sentient creatures would be deserving of moral concern. Such sentient creatures are the equivalent of an "experiencer of pain" in the example above, in that they are the "experiencers of moral violation"; however, in what follows I will argue that this is a problematically anthropocentric stance to adopt.

While it might be reasonable to attribute the status of moral patient to certain classes of sentient animals, as we go further down the phylogenetic tree, and as creatures differ from us in their external appearance, we tend to be less likely to attribute the requisite kind of sentience to them. We are inclined to view other *hominids* as sentient, but most would not award this same ascription to other creatures which perhaps have more "basic" minds, such as molluscs. We tend to think of them as analogous to the arm on the table: capable, perhaps, of adverse nerve stimulus, but not *sentient* to the required degree, not capable of *experiencing* pain. Moreover, the Organic View itself does not give us a clear criterion for sentience (of the requisite kind), and so we have to rely on our intuitions to determine which kinds of creatures are moral patients, and these intuitions are geared towards including those entities that look like us and excluding those that look less like us. These intuitions do not necessarily track "actual" sentience, and so the criterion of sentience does not help us, in practice, to identify moral patients. To see this more clearly consider the example of fish, more specifically, fish cognition. Our *perception* of an animal's intelligence is often a key criterion (although not the only one) for whether we consider them to be sentient or not, and fish are rarely considered to be intelligent or phenomenally sentient in a manner akin to humans or even mammals. Moreover, fish are very rarely (if ever) accorded the same type of moral concern as are warm-blooded, non-human animals. Standard reasons given for such claims is that fish lack the requisite neural complexity in order to have the right kind of "experience". Such endothermism[17] (in the case of fish, specifically) stems from a disjunction between the public perception of fish intelligence and scientific reality (Brown, 2015). There is ample scientific evidence supporting the conclusion that "fish perception and cognitive abilities often match or exceed other vertebrates'" (*ibid.*). For example, fish are capable of tool use and display evidence of complex social organisation and interaction (such as signs of cooperation and reconciliation). The point here is not to outline all of the ways in which fish cognition may be measured. Rather, the key issue is that if we use our traditional metrics of intelligence when it comes to animals (such as tool use and social organisation), then we are forced to conclude that fish are on par with (and at times exceed) other "sentient" vertebrates in these criteria. The next question, then, would be whether, following from the fact that fish exhibit "intelligent" behaviour, they are also phenomenally sentient and hence capable of similar kinds of suffering? Our intuitions surrounding fish sentience and their capacity to feel and suffer seem to be biased away from accepting them as sentient "enough" to merit moral concern. It seems that we struggle to empathise with fish as

> We cannot hear them vocalise, and they lack recognisable facial expressions both of which are primary cues for human empathy. Because we are not familiar with them, we do not notice behavioural signs indicative of poor welfare (*ibid.*).

[17] That is, unfair moral discrimination based on the temperature of an entity's blood.

This implies that a proper, scientific, construal of fish behaviour would support the conclusion that fish have relatively complex cognitive capacities, are capable of suffering, and are therefore sentient in a manner similar to creatures that are accorded moral concern (*ibid.*). To bring this back to the Organic View, the issue that the example above was meant to highlight is that how we go about identifying moral patients should not be guided by scientifically illegitimate and anthropocentric conceptions of "sentience". By not giving us a clear definition of sentience, the Organic View relies on our intuitions, which, as the example above demonstrates, are not good guides to "real" or "genuine" sentience ascription.

Applying the discussion above to the question of whether an artificial system could, in principle, be the subject of moral concern highlights the potential for moral harm in the future. In the same way that we have biases that cause us to accord a lesser moral status to non-human entities that do not sufficiently look like us, we may be biased against machines based on their unfamiliar structures. This is not to claim that sentience can have no purchase whatsoever when it comes to moral ascription, but rather to assert that the vague description of sentience used in the Organic View provides an anthropocentric understanding of what *constitutes* sentience in the first place. While seemingly shifting the focus from substrate to "causal history" or historical development in how we evaluate whether entities are sentient or not, the Organic View still equates the appropriate history with a biological one. While claiming to be substrate neutral what we instead find is that this view has shifted the goalposts, while keeping the criteria the same: it is still only biological entities that can be genuinely sentient as only biological entities can have the requisite history. The Organic View, as I have shown above, fails to provide a convincing argument as to why exactly this should be the case. Moreover, even within biological species we still struggle to accurately discriminate between creatures that are "genuinely" capable of affect or not, making use of anthropocentric intuitions instead of argument. The continued usage of this conception of sentience would therefore exclude machines from moral consideration in principle.

1.3.2 Epistemic Issues

The second complication to be unpacked is the distinction between a mere ersatz phenomenon and its "true" instantiation. This is an idea which has a considerable amount of philosophical baggage, has been around since at least Plato, and which is a recurring theme throughout the Western philosophical canon (Gunkel, 2012: 138). By making use of sentience as the underlying capacity which qualifies/disqualifies an entity as having a moral stake, what the Organic View is in fact claiming is that only entities with the *real* capacity for phenomenal states qualify: the mere appearance of behavioural cues that point to phenomenal states (as may be the case with anthropomorphic robots) is not enough to ground our moral ascriptions, and as such only entities that are *genuinely* sentient can be accorded a moral stake. However, how exactly are we to go about "proving" that an organism is sentient, *really* sentient (i.e. "phenomenally conscious")? As Dennett points out, "everybody agrees that sentience requires sensitivity plus some further as yet unidentified 'factor *x*'" (1996: 66). Therefore, considering my discussion above regarding how we define sentience in non-human creatures, how are we to make an epistemically sound judgment as to what counts as, for example, "real" pain versus the mere "appearance" of pain? The fuzzy nature of the concept being employed (sentience), in conjunction with its subjective nature, renders it immune to such an analysis.

To see how this might be the case, consider the classic British television game show *Would I Lie to You?* In the show, contestants are split into two teams, competing against one another in attempts at deception. In each round, one contestant from each team is randomly selected and reads out loud a card with a note on it. The content of the note is unknown to the contestants until they read it, and the goal for the contestant who has read the card out loud (the speaker) is to convince their opponents that what has been read is in fact the truth. The content on the cards is of a personal nature, and so only the contestant who is reading the card will know whether it is the truth or not: the opponents have no idea and are allowed to ask probing questions, which the speaker must attempt to answer in a believable way. Once the questioning is over, the opposing team can decide to either claim that they believe the speaker to be telling the truth or claim that they believe them to have lied. After they have submitted their decision, the speaker reveals whether the note was in fact true or a lie, and if the opponents guessed correctly, they receive a point.

While British television can be as dry as academic philosophy, that is not the point I wish to make. In the case of the game show there is a type of deception at play: the speaker is attempting to convince the other team of the truth or falsity of their note. Similarly, when discussing questions of true sentience versus ersatz-sentience, we are attempting to figure who or what is on either side of the divide. We interpret the available evidence and then need to come to a sound judgment about the entity in question. However, and this point is crucial, in the case of the game, the deception is removed: we are shown the veracity of the matter when the speaker reveals whether they were telling the truth or not. In the case of our sentience ascriptions, we have no such epistemic security: we do not have the privileged access required in order to know whether we have made the correct judgment or not. There is no verifiable test we can perform in order to determine whether we have made the correct kind of ascription. The reason we have these issues is epistemic opacity—we do not have direct access to the qualitative states of others and are therefore not in a good position to judge whether an entity is "truly" sentient or not.[18] Consider the advent of advanced neuroimaging technology, such as functional magnetic resonance imaging (fMRI), which allows us to detect brain activity associated with blood flow. This type of technology allows us peer into the "moving parts" in the brain which may be correlated with sentience. However, talk of internal states and the talk of how we describe, scientifically, the information that an fMRI machine represents to us are two very different language games. We therefore cannot know whether two equivalent systems – one inorganic the other organic – are phenomenally different by merely putting them through a scanner. To attempt to explain what these internal states "feel like" in terms of neurophysiology and physics would be a category mistake (Powers, 2013: 233).

This issue precludes us from being able to use "true" sentience, as specified in the Organic View, as a qualification for moral status, whether biological or artificial. However, it is possible that in the future, once we have a more coherent conceptualisation of sentience and how it comes about, we may use this criterion to determine who or what counts as a moral patient. The argument I have put forward simply undermines the specific notion of "real" sentience put forward by the Organic View. What this implies, for my purposes, is that there is now conceptual space for the notion that some future artificial system may come to be sentient in such a way as to coherently be accorded a moral stake. We simply don't know if various entities—including other people—are only apparently or "truly" sentient. Hence, we could decide to treat all apparently sentient creatures as moral patients, which implies at some point AI may be worthy of this type of moral attribution. However, I am doubtful of whether such an overreliance on internal, qualitative, states can ever be the only factors that feature into our ascription of moral patiency. Therefore, in what follows I provide a tentative suggestion for future investigation into questions of moral patiency in general, and machine moral patiency in particular.

[18] One could, of course, bite the bullet and say that based on external cues we could work *as-if* the entity in question was sentient, but Torrance is not in a position to do this as his argument requires that these qualitative states of experience are *objective* and not merely pragmatic as-if inferences.

1.4 Towards a Coherent Account of Moral Patiency

From the failure of the Organic View, I would like to tentatively suggest a model for future research into moral patiency, a model which does not operate on the same anthropocentric biases as the Organic View. Consider how we come to infer the psychological states of others on a day-to-day basis (usually without the use of advanced neuroimaging equipment): we largely use external cues in order to make plausible predications about what might be going on in their craniums. However, this type of projection is not necessarily indicative of the "real" type of phenomenal ascription required for sentience as specified above, but it at the very least provides a predictive model that we can use to infer what might be going on in other people's heads. This methodological approach, formalized by Daniel Dennett, is known as the intentional stance (Dennett, 1989). This "intentional stance"[19] treats the agent in question as a rational one, and then attempts to figure out which beliefs and desires the agent ought to have in light of this capacity (*ibid.*: 17). Imbued in Dennett's exposition of the intentional stance is a willingness to let go of certain, outdated conceptual categories. He is willing to acknowledge that mental postulates such as "beliefs", "desires", etc. are useful for predicting behaviour, but are not good guides as to what is *really going on in the brain*. They are therefore not good theoretical entities, which is why the intentional stance must remain (and is) *non-committal* (or theory neutral) with regards to the internal structures that underlie the specific competencies that an investigator is explaining (Stich, 1981: 44; Yu and Fuller, 1986: 454; Dennett, 2009: 10).

Likewise, we might be able to use the intentional stance to try to determine whether an entity in question is indeed worthy of moral consideration, based on certain behavioural cues.[20] This approach should not, however, be seen as exhaustive: it is only a helpful heuristic as to whether an entity is in fact sentient. While relying *only* on behaviouristic cues would mean that we would accord a moral stake to anything capable of, for example, mimicking pain, this would be a mischaracterization of my proposed usage of Dennett's methodology. My suggestion is simply that we take these behavioural cues seriously, as opposed to only relying on the

[19] Adopting the intentional stance allows us to better predict human (or any other intentional system, such as a thermostat) behaviour (Dennett, 1989, 2009). By providing the intellectual tools for the adoption of the intentional stance Dennett provided a more scientific way in which to account for and explain certain aspects of our "folk psychology". In sum, Dennett's strategy provides a naturalistic basis for folk psychological categories. This strategic move decomposes the essentialist distinction between "original" and mere "as-if" (or "derived") intentionality and provides a Darwinian account of the concept.

[20] These could be signs that are indicative of suffering, for example vocalizations (sighing or moaning), facial expressions (grimacing, frowning, rapid blinking, etc.) or bodily movement (being hunched over, exterior rigidity, etc.).

presumed capacity to have "real" qualitative states of experience or having a particular causal history, as exemplified in the Organic View. In addition to behavioural cues, we might look to other cues indicative of an entity's internal constitution and what this tells us about the likelihood of this entity having the capacity for affect. This type of naturalistic approach is exemplified in the example of fish cognition above, in which philosophical intuition is consistent with and does not contradict our best science (Ritchie, 2008; Brown, 2015). We might find that we over-ascribe the capacity for affect on this approach, but it is surely better to err on the side of caution when it comes to moral concern.

Two more examples of behaviouristic and functional approaches to moral ascription are the Moral Turing Test (Gerdes and Øhrstrøm, 2015)[21] and Turing Triage Test (Sparrow, 2004). The first of these tests asks whether an artificial system "acts at least according to the ethical standards that are normally considered acceptable in human society" (Gerdes and Øhrstrøm, 2015: 99).[22] If the system can pass such a test, then it can be worthy of moral consideration. Secondly, there is the Turing Triage Test (Sparrow, 2004). This test proposes that in a "triage" situation (one in which a choice must be made as to which of two human lives to save), if one replaces one human person with an AA, and the moral character of the dilemma remains intact, then the AA would have achieved moral standing comparable to that of human beings (*ibid.*: 203). Both of the aforementioned propose novel ways in which we might come to understand the basic moral contours of our relationships with intelligent machines in the future. This type of approach may perhaps lead to a philosophically coherent account of machine moral patiency. For now, however, our concern will be with machine moral *agency*, as this issue is of more practical importance in light of contemporary trends in AI research and applications, as I will outline below.

1.5 Patiency as Speculative

Using the intentional stance as a stepping stone, we can perhaps come to better understand our moral ascriptions: instead of trying to identify whether an entity in question is *really* experiencing pain or any other internal state, we can adopt the logical-behavioristic tenets of this approach and use external behavioral cues in order to co-determine what the correct moral

[21] Also see Wallach and Allen (2009: 70) for an exposition of the comparative Moral Turing Test (cMMT), which asks "which of these agents is less moral than the other?" as opposed to the question of which entity is the artificial agent, posed in the MTT.

[22] For a critique of the Moral Turing Test, see Arnold and Scheutz (2016).

response might be, on a given moral theory. In this way an artificial entity *may come to be* considered a moral patient in the future, if it is capable of exhibiting these external cues convincingly (in conjunction with certain other, sufficiently non-anthropocentric criteria) (see, e.g., Johansson, 2010). However, the current technological state of artificial artefacts is not heavily geared towards creating entities with the explicit potential of being considered moral patients (Royakkers and van Est, 2015).[23] It is highly unlikely that any current form of robotic technology would exhibit the kind of behavioral responses which would convince us of its standing as a genuine moral patient. While I have not presented a positive account of what exactly it takes to qualify as a moral patient, suffice it to say that nobody seems to be claiming that any currently existing technology would qualify as a patient in this sense The more pressing moral concerns come from the development of robotic cars, police robots and military robots (*ibid.*). In all three of the aforementioned cases the question of patiency is, at the moment, secondary: it does not matter much whether we have a certain moral *obligation* towards a robotic self-driving car, as researchers are not currently implementing or attempting to develop cars that would be worthy of this type of consideration. What is presently of substantially more practical importance is what happens when a self-driving car is causally involved in the infliction of a moral harm. In such a scenario, we can see the seeds of a potential "responsibility gap"(Gunkel, 2017; Nyholm, 2017) emerging, in which it might not be clear whether the human user or the technology itself is "really" responsible for the moral harm. In such a situation, how ought we to go about assigning moral responsibility? The answer to this question will ultimately be one which is concerned with moral agency, which will necessitate a conception of "moral agency" adequate to these novel circumstances. I will argue that in scenarios such as these, our concern with moral agency should not include the question of whether the technology in question is a moral patient or not. Considering my discussion above it should be clear that we would be wise not to ground our ascriptions of moral agency on the shaky foundations of the Organic View, and by extension by any appeal to the presence of some kind of internal mental state. Moreover, it should be clear that a self-driving car, for example, can potentially perpetrate morally significant harms, even while it cannot plausibly be thought to be a moral patient. My proposed mobilization of the behavioristic tenets of the intentional stance to overcome the problem of moral patiency were not developed in detail, as

[23] Sex robots and care robots provide the most technologically plausible example of potential entities that may have the potential of being considered moral patients in the future (Royakkers and van Est, 2015; Danaher, 2017a). These examples, however, pose complications that will only arise in the future, whereas questions of moral agency are already confronting us in the form of military drones and self-driving cars.

they are merely suggestions for future research into this question. This is sufficient for my purposes, however, as I will be arguing that moral patiency and moral agency can come apart, and that it is both possible and necessary for us to attribute moral agency to an AI that is not a moral patient. This claim will be justified in the forthcoming chapter, in which I will discuss machine moral agency in substantially more depth.

1.6 Conclusion

In this chapter I have shown how an overreliance on internal mental states hamstrings philosophical investigation of moral patiency more generally, and machine moral patiency specifically. I have not explicitly provided a worked-out account of machine moral patiency. Therefore, while it *might* be possible to extend the scope of moral concern to machines, it is *speculative* whether this will actually happen (as it depends on whether a plausible theory of the requisite developmental history is put forward and whether we are swayed by external cues that are suggestive of an internal state of suffering, etc.). What is relatively *certain*, however, is that we will soon be faced with situations in which robots will be causally involved in actions for which some type of moral evaluation will be required, and where the role of the human designer and/or operator is more and more obscured (such as the case of self-driving cars mentioned above) (Johnson, 2015: 709). Furthermore, the actions of these machines will significantly affect the lives of moral patients (us!), notwithstanding the fact that they might not be included in this class of moral ascription. In the next chapter I will therefore outline the so-called "standard account" of agency. I will show how, much like the Organic View presented in this chapter, the standard account of agency is also problematically anthropocentric. I will provide a non-anthropocentric account of agency and use it to potentially ground a philosophically defensible notion of moral agency. This non-anthropocentric account of moral agency may allow us to morally evaluate the actions undertaken by machines, independently of whether they are moral patients or not (Powers, 2013).

Chapter 2: Conceptualizing Agency

2.1 Introduction

In the previous chapter, I highlighted the complications that arise when we attempt to ground our conception of moral patiency within an anthropocentric understanding of sentience. In addition to these complications, it might be reasonable to remain agnostic as to whether an artificial entity does or can have qualitative mental states ("phenomenal consciousness"), as we can i) perhaps attempt to avoid designing systems with the capacity for these internal states in the first place (Bryson, 2018: 23) or ii) posit that it might in fact be impossible for a machine to have these kinds of internal states in principle (Dennett, 1978).[24] In light of this, my discussion of machine moral standing will shift away from a focus on moral patiency and turn instead to moral agency. Consider an event which took place in 2006, in which a driverless magnetic levitation train crash in Germany lead to the death of twenty-three people when the train smashed into a maintenance truck (Harding, 2006). According to eye-witness accounts from survivors, they felt helpless, as they could see the maintenance truck on the track through the train's panorama front window but could do nothing as there was no driver to alert. In cases such as this, who should be held responsible? The system responsible for operating the train, the human designers who developed the program, or the company that owns the train (Boyles, 2017)? In this case, it was claimed that "human error" was the primary cause in the tragedy, and that if protocol had been correctly followed, the disaster could have been averted (Harding, 2006). An apparent breakdown in communications lead to certain regulations not being followed. Had these regulations been followed, it was claimed, the accident would not have occurred. However, as machines become increasingly autonomous, there could come a point at which it is no longer possible to discern whether or not any human error could in fact have been clearly causally efficacious in bringing about a certain moral outcome (Grodzinsky, Miller and Wolf, 2008: 121). The key issue that arises in such discussions is one of *attributability*, and, more specifically, whether we can attribute the capacity for *moral agency* to an artificial agent. The ability to make such an ascription would lead to the resolution of potential "responsibility-gaps" (Nyholm, 2017): cases in which warranted moral attributions are

[24] This could be quite a contested claim, as it might be suggested that an "unforeseen consequence" of the increasingly complex nature of artificial systems may lead to the development of systems capable of "suffering" or having some such internal state without us ever intentionally designing them as such. For now, however, I assume that we can foresee many of these potential consequences.

currently indeterminate. This idea will be explored in further detail in the next chapter. For now, I will focus specifically on the concept of moral agency.

Instead of asking the question of whether an entity is deserving of some kind of moral concern, moral agency is concerned with whether an entity is capable of moral action. As has been mentioned previously, an agent is simply a being with the capacity to *act* (Schlosser, 2015). A moral action would be a type of action that it would make sense to evaluate using moral criteria. Inevitably this type of discussion leads to further questions concerning responsibility, as it is traditionally supposed that a moral action is one that an entity can be morally responsible for by being accorded praise or blame for the action in question. This type of moral responsibility has historically been reserved for certain biological entities (generally, adult humans). However, the emergence of increasingly complex and autonomous artificial systems might call into question the assumption that human beings can consistently occupy this type of elevated ontological position while machines cannot.

Of course, it is the *capacity for agency* that makes someone eligible for moral praise or blame, and thus for any ascription of moral responsibility (Eshleman, 2016). It is with this in mind that, in this chapter, I will investigate how the conceptual framework provided by various accounts of agency can help us to better understand the potentially morally-laden roles that increasingly autonomous machines can come to fulfill in human society now and in the future. I will start off by sketching what is termed the "standard account" of agency and firstly show how this account conflates various conceptions of agency. The standard account requires that an entity be phenomenally conscious for it to be considered an agent more generally, and therefore a moral agent specifically. I claim that we need to move beyond whether an entity is "phenomenally conscious" or not when ascribing agency. What follows from this critique of the standard account is that a conceptual space is opened in which it becomes possible to attribute agency to artificial systems. Secondly, I will provide an exposition of three types of agency that might *prima facie* be accorded to machines (Johnson and Noorman, 2014). Two of these types of agency are seemingly uncontroversial, as they deal with artefacts that operate in functionally equivalent ways when compared with human actions. The third conception, however, is much contested, as it deals with the autonomy of the potential agent in question. It is also this sense of autonomy that grounds various notions of *moral* responsibility, and so, in order for an agent to be a moral agent it must, supposedly, meet this requirement. I will flesh out the details of this view, and then in my final chapter show how the specific senses of *intentionality* and *autonomy,* as emphasized by Johnson (2006), are necessarily anthropocentric, and hence cannot provide clarity on the issue of whether or not artefacts qualify for this kind of agency. They are anthropocentric in that, in a way similar to the standard account, they privilege the significance of the type of phenomenal consciousness that human beings have in how we determine who or what is a moral agent.

2.2 The Standard Account of Agency

According to Himma (2009: 19-20), the standard account of agency claims that in order for the entity in question to qualify as an agent it must be capable of performing an *action*. Actions are a kind of doing, but not all doings are actions. Take the example of digesting one's food: digestion is a doing, but it is not an action. Eating food, on the other hand, can be understood as an action, in that one can decide whether or not to eat food, and it is in virtue of having this type of active ability to make decisions regarding doings that an entity in question can be said to be an agent (*ibid*.: 19). You do not consciously decide whether to digest your food or not; subsequent to the consumption of copious amounts of carbohydrates (which are of course offset by vigorous physical exercise) there is an *automatic* somatic response (digestion). Your body responds to the intake of food by initiating the breakdown of large, insoluble food molecules, into smaller water-soluble molecules, so that they can be absorbed into the bloodstream. Gastroenterology aside, what I wish to emphasise here is that digestion is not an *intentional* act, and that eating food, on the other hand, is. Eating is not simply an automatic response by one's nervous system: it involves a coordinated effort by the mind/body to, for example, consciously decide what to eat, make the food, transport it from plate to mouth, etc. The mental state(s) associated with these actions are clearly about something other than the mental state itself. On the standard account, only entities capable of this type of *action* can be considered agents. There is an important distinction here between our everyday meaning of the term "intentionality" and the more technical way in which it is operationalised in the philosophy of mind. The everyday meaning has to do with whether one "intends" to perform an action; in other words, whether one does it "on purpose". This type of intentionality seems to presuppose some kind of linguistic competence on the part of the agent in question, and so Himma uses the term in its technical sense, as developed by Brentano (*ibid*.: 20). This technical meaning has to do with the "directedness" or "aboutness" of our mental states. What this ordinarily implies is that mental states are intentional when they are about or directed at something other than the mental state itself.[25] The advantage of this view is that it captures the fact that dogs, cats, etc. are seen as intentional beings, notwithstanding the fact that they are incapable of linguistic communication and are not considered *rational* agents. Intentional acts, as Himma (*ibid*.: 20) describes them (and as they are used in the standard account), therefore refer to the restricted, technical understanding of the term "intentional".

Consider once again the example of eating and digestion that was introduced earlier: eating, or perhaps more specifically, the decision about *what to eat* can be understood as being intentional in the technical sense outlined above. It is intentional in that the mental state accompanying the action is not about the action itself but directed towards or about some other object (what to eat). This is not the case for digestion: there is no accompanying mental state which would make an automatic somatic response, such as digestion, intentional in the sense specified above.[26] These intentional acts are accompanied by a certain type of *phenomenal mental state* (or certain neurophysiological correlates). Exactly what type of mental state this should be is a subject of great philosophical controversy, but I will not attempt to resolve the issue here.[27] Suffice it to say that, on the standard account, there are certain intentional mental states that separate mere doings from actions (*ibid*.: 20).

2.2.1 The Complex Carbonist Account of Agency

The view presented above supports what I will term "the Complex Carbonist Account of Agency" (hereafter "the Carbonist Account"). The reason I have operationalised this terminology is due to an unfortunate (in my opinion, at least) implication of the standard account – it suggests that only sufficiently complex carbon-based creatures can come to count as agents.[28] The reason for this is that the capacity for phenomenal consciousness seems to be presupposed by the notion of agency as presented above (*ibid*.: 27). My goal, therefore, is to

[25] For example, I may desire an almond toffee milkshake, and I can act on this desire by purchasing the aforementioned delight. This act, in the sense specified, is intentional in that my mental state (*desire for an almond toffee milkshake*) is *directed towards* something other than the mental state itself.

[26] It might be argued that some lower-level cognitive state which regulates digestion could be intentionally directed. However, on the standard account, it seems that when we traffic in the language of intentionality, we presuppose some phenomenologically loaded understanding of the concept.

[27] One could argue that this mental state should be a belief/desire pairing (see Davidson, 1980), or perhaps a type of "volition" or "willing". All the aforementioned are contentious philosophical theories with no standardized or generally accepted definitions, and each comes with its own set of problems.

[28] In the future it may be that the standard account is not Carbonist in this way. Consider an artificial system designed in a functionally equivalent way to a human being (such as a humanoid). In this case it might be reasonable to confer moral agency to such a system on the standard account. However, for the purposes of this investigation, I am more concerned with artificial systems which do not explicitly seek to functionally represent human decision-making procedures or aesthetics.

expand the definition of agency by illuminating the shortcomings of the standard account. It may in fact be the case that the standard account does not *necessarily* mandate this type of carbon chauvinism, but at present its restrictive conception of agency means it cannot accommodate non-Carbonist views of the concept. Consider again the requirement that actions require intentional mental states in order to be classified as such. These "mental states" are by definition conscious mental events, and according to the standard account only biological (carbon-based) entities are conscious in this appropriate sense (*ibid.*: 27). The reason these mental states are considered to be conscious is that if, from a first-person perspective, one does not have access to the aforementioned states, then the "reasons" for performing an action would be arbitrary. In other words, on this view, "real" agents are not compelled to perform acts; they have *reasons* for doing so. Such "reasons" for action, on the Carbonist Account, must be accessible via introspection from a first-person point of view, and for these reasons to be "grasped", they must in some way be conscious (*ibid.*: 24). This "grasping" of a reason does not necessarily imply the ability to articulate it: even a cat has something like conscious access to the fact that she is eating because she is hungry. In order for these reasons to "count" they should not be *directly* caused by some external stimulus: for example, the action should not be undertaken due to some abnormal electrical discharge within the brain or something clearly outside of the control and awareness of even the most basic agent.

Humans are not agents in exactly the same way that cats are – while cats may have conscious access to their mental states, human beings have the further capacity to give communicable reasons for their actions (and this links to us also being both *moral* and *rational* agents, who can be held accountable, which will be discussed shortly). Notwithstanding this, it is clear that human beings also have access to their mental states and can therefore perform intentional acts. However, an "intentional act" seems to be a Carbonist term of art with the exclusive focus on acts that can be performed by creatures with a certain kind of *mind*—the kind that can have certain phenomenal states of consciousness.[29] Hence, the kind of mind presupposed in this account is of the type that we humans (and perhaps certain animals) possess. This conception of agency itself is thus dependant on the prior ascription of a certain kind of consciousness: in this way only conscious beings can be agents, and since only carbon-based beings are conscious

[29] For evidence of this, see both the *Stanford Encyclopedia of Philosophy* (Jacob, 2019) and *Routledge Encyclopedia of Philosophy* (Crane, 2011) entries on "Intentionality". For example, the *Stanford* entry states that "Intentionality is the power of minds to be about, to represent, or to stand for, things, properties and states of affairs" (Jacob, 2019), and the *Routledge* entry states that "intentionality is the mind's capacity to direct itself on things" (Crane, 2011).While citing philosophical encyclopedias is not an argument, it is suggestive of the general view of practicing philosophers.

(according to the standard view), only carbon-based beings have the capacity for agency. While the standard account does not mandate the need for carbon-based existence for consciousness to arise, this is an implication of the view. On this account, there is nothing else differentiating "pseudo" agents from "real" agents besides their different substrates. The standard account of agency thus precludes the concept of machine agency from successfully referring. And, as moral agency is a special class of agency, and thus dependant on a prior capacity for agency, machines, on the view above, can never be moral agents. Over and above simply assuming that only carbon-based entities can be conscious, the standard account is not sensitive to the various types of agency that might be possible and thus, potentially, attributable to machines. Its focus on "intentional acts" blurs various conceptual lines that can be drawn between different kinds of agency. In what follows I will outline three conceptions of agency that are all distinct and valid accounts of the concept and that are not distinguishable on the Carbonist account. Moreover, they are all potentially applicable to artificial systems. These are: causally efficacious agency, "acting for" agency, and autonomous agency (Johnson and Noorman, 2014). I will then go on to assess the implications of these different conceptions of agency for the kind of moral responsibility attributable to machines. Before doing this, however, I will briefly show how the standard account above also fails on its own terms: it fails, in some cases, to accurately describe how human beings are agents.

2.3 Types of Agency

As stated in my introduction, the metaphysics of agency is concerned with the relationship between *actions* and *events*. The most widely accepted metaphysical view of agency is event-causal, where it is claimed that agency should be explained in terms of agent-involving states and events (Schlosser, 2015). In other words, agency should be understood in terms of *causation*, and, more specifically, in terms of the causal role the agent plays in the production of a certain event. Agents, therefore, are entities capable of having a certain effect on the world, where this effect usually corresponds to certain goals (in the form of desires, beliefs, intentions, etc.) that the agent has. On the Carbonist account above, this would mean understanding agency in terms of causation by the relevant (phenomenal) mental states of the agent in question, as these relate to the agent's goal(s). It does not seem clear, however, why talk of agency, in this sense, *necessitates* any phenomenologically loaded sense of intentionality. Firstly, it seems many non-human forms of agency are possible, ones which are not necessarily tethered to any phenomenologically loaded mental states. Take the example of a bee when it pollinates a flower. The bee, as an agent of pollination, is clearly playing a causal role in the production of some event. On this event-causal construal, therefore, we can understand the bee as an agent. The Carbonist account, therefore, seems to be too narrow in its construal of agency.

A second issue has to do with what the addition of phenomenologically loaded mental states adds to our understanding of agency. It seems that such an addition would serve to complicate matters substantially, in a case where a more economical answer is available. There are many instances of human behaviour that can be explained without any appeal to representational mental states. These are cases in which humans showcase the ability to interact with their environment seamlessly, seemingly without any conscious deliberation or planning (Schlosser, 2015). On the Carbonist account we are forced to concede that such examples of "skilled coping" (*ibid.*) would not qualify as instances of human agency. This would have the counterintuitive result of suggesting that actions such as driving a car or riding a bicycle are not cases of human agency. Moreover, the explanation of such actions in terms of representational mental states places a high burden on the agent's information processing systems. This leads to highly uneconomical descriptions of agents who perform such actions, as the suggested explanation is that there are set of highly specific mental representations that can account for the behaviour. Better accounts of such action are possible by appeal to, for example, certain behavioural dispositions of the agent (*ibid.*).

When we talk about human beings (who are paradigmatic agents) performing actions, there is therefore no necessary need to appeal to their phenomenology in order to explain their actions, as there may be cases where human beings perform actions that do not require any such explanation. Moreover, how are we to go about identifying whether the person does in fact have the correct mental representations? A plausible answer here would be that we have inferred the correct mental state when our identification leads to *predictive success.* In other words, the ascription of mental states can be said to have been successful if such ascriptions support the successful prediction of behaviour. However, this response does not require the further postulate of phenomenologically loaded mental states. In fact, as noted above, such a suggestion would make our ascriptions of agency incredibly costly. To avoid this, we can simply claim that there are cases of human agency that can be explained without necessarily appealing to any internal mental states. In other words, instead of using the technical sense of intentionality above, we can make use of the broader conception of the term, which seems to better track our intuitions regarding the types of agents that are possible. This opens up room for alternative understandings of agency, and suggests, at the very least, the plausibility of machine agency.

The aforementioned "standard account" of agency, therefore, seems to have little purchase in allowing us to better understand both the agential roles human beings are capable of and the potential role machines could play. In what follows, I will therefore paint with broader strokes in order to account for the various ways in which technological artefacts may come to be "agents" that affect both the capacity for human action and the range of possible events that may arise from such action. A key distinction between the standard account of agency and the account which will be presented below is their different usage of the term "intentionality". In the previous account, as was pointed out, intentionality referred to the "aboutness" of our mental states specifically. In the exposition that follows, however, the term intentionality will be used much more broadly. It will be used to denote an action that has been performed "on purpose" or by an entity that has goals which serve as reasons for its actions. The exclusion of "phenomenal" intentionality is justified in this sense as it captures the fact both human beings and artificial systems, while not necessarily exhibiting "phenomenal" intentionality, still come to play agential roles. In this sense, then, referring to the intentionality of technology would denote the fact that technological artefacts are designed in certain ways to achieve certain outcomes. This will become clear as the argument develops.

2.3.1 Causally Efficacious Agency

In the context of potential artificial agency perhaps the most comprehensible and comprehensive conception of "agency" is put forward by Johnson and Noorman (2014), namely the causally efficacious entity. This conception of agency simply refers to the ability of some entities—specifically technological artefacts—to have a causal influence on various states of affairs, as extensions of the agency of the humans that produce and use them (*ibid.*: 148). This includes both artefacts that may be separated from humans in both time and space (for example, attitude control[30] in spacecraft in nominal orbit around the earth) or those deployed directly by a human being (for example, someone firing the Stream Machine DB-1500 Double Barrel water cannon[31]). A fair question to raise at this junction is whether in fact it makes sense to consider these types of artefacts agents at all. Perhaps we should

[30] Attitude control is the controlling of the orientation of an object with reference to an inertial frame, or another entity (e.g. a nearby object, the celestial sphere, etc.).

[31] This magnificent hand-pumped artifact can fire a stream of water as far as 20 meters away.

conceptualize them as *tools* instead. The reason for preferring the terminology of agent as opposed to tool is that these artefacts have human intentions programmed/encoded *into* them (Johnson, 2006: 202). This is in contrast to a tool, such as a hammer, which may be used by a person to perform a specific task but does not have the specifications of this task programmed/encoded into its very makeup. It cannot in any way perform or represent the task independently of human manipulation. The key distinction, then, between a tool and a technological artefact, is that the latter has a form of intentionality as a key feature of its makeup, while the former does not (*ibid.*: 202).[32] Consider the simple example of a search engine: keys are pressed in a specific order in an appropriate box and then a button is pressed. The search engine then goes through a set of processes that have been programmed into it by a human being. The "reasons" for the program doing what it does are therefore necessarily tethered to the intentions of the human being that created it.

It makes sense to think of such artefacts as possessing "agency" to the extent that the ubiquity and specific design of these types of artefacts make a difference to the effective outcomes available to us. For example they make possible novel means with which to achieve our ends by increasing the amount of potential action schemes at our disposal (Illies and Meijers, 2009: 422). These artefacts can therefore be seen as enlarging the possible range of actions available to a particular agent in a given situation. Yet, while it is clear that artefacts can thus have causal efficacy in the sense that they may *contribute* to the creation of certain novel states of affairs, this causal contribution is only efficacious in *conjunction* with the actions of human beings (Johnson, 2011: 197; Johnson and Noorman, 2014: 149). The reason we can think of these causally efficacious artefacts as agents is the fact that they make substantial causal contributions to certain outcomes. In this way the causal efficaciousness of an entity leads, in the form of a non-trivial action performed by that entity, to a specific event.

Moreover, according to Johnson and Noorman, we can legitimately think of these artefacts as agents due to the fact that their manufacturers have certain intentions (aims) when designing and creating these artefacts, and so these artefacts have significance in relation to humans (2014: 149). The type of agency that we can extend to artefacts under this conception would thus not be one that involves any meaningful sense of responsibility on the part of the artefact, and, by extension, would not entail a distinctly *moral* type of agency (as for an entity to be a

[32] The intentionality of the program should be understood in functional terms, according to Johnson (2002: 202). What this means is that the functionality of these systems has been intentionally created by human designers, and so is necessarily tethered to and wholly determined by human intentions. The human intentions, in this sense, give the "reasons" for why the technological artefact acts in a particular way.

moral agent it should be morally responsible for the action it has performed). While Johnson and Noorman concede that artefacts can be causally efficacious in the production of various states of affairs, their contribution in this regard is *always* in combination with that of human beings (*ibid.*: 149). On this conception of agency, therefore, we can only consider entities that act in "causal conjunction" with human beings. The next conception of "agency" that will be unpacked can be employed for machines that perform tasks on behalf of, but independently from, human operators and so can be seen as a special case of causally efficacious agency.

2.3.2 "Acting for" Agency

This conception of agency focusses on artefacts that act on the behalf of human operators in a type of "surrogate" role (Johnson and Powers, 2008; Johnson and Noorman, 2014). In an analogous way, when it comes to human beings, surrogate agency occurs when one person acts on behalf of another. In these cases, the surrogate agent is meant to represent a client, and therefore is constrained by certain rules and has certain responsibilities imposed upon them.[33] This type of agency involves a type of representation: the surrogate agent is meant to use his or her expertise to perform tasks and provide assistance to, and act as a representative of, the client, but does not act out of his or her own accord in that capacity (Johnson and Noorman, 2014: 149). When it comes to artificial systems, this "acting for" type of agency occurs in those artefacts that replace or act on behalf of humans in certain domains. Take the example of a stockbroker: in the past, in order to have a trade executed one would have to phone a stockbroker and request the purchase/sale of a specific share. The stockbroker, acting on your behalf, would then find a willing buyer/seller in the market and execute the trade. The reality today is much different: individuals can now create accounts on trading platforms and buy shares online without the need of a stockbroker. Furthermore, the exchanges on which these trades are made are also run by computers: inputting a "sell order" places your request in an order book, but this order book is not a literal one, as it might have been in the past, and so there is no need to leave the comfort of your home to perform these tasks. Current online order books are fluid, competitive spaces in which high frequency trading occurs, without the need for humans to keep record: this job is taken care of by the computer powering the system.[34]

[33] For example, lawyers are not allowed to represent clients whose interests may conflict with that of another client.

[34] Another interesting development in automated trading has been the explosion of this technology as it applies to cryptocurrency markets. These markets have been heavily impacted by the emergence of "trading bots" which

Technical details aside, what the aforementioned example brings to light is how tasks that were once the exclusive domain of human beings are now being performed by artificial systems without too much "hands-on" human involvement. The function of the tasks performed by these systems, however, is still the same: the purpose of an automated trade is still the same as a trade executed by a human, as in both cases the end being pursued is the purchase/sale of some stock at the behest of a given client. What has changed, however, is the means by which that specific end is obtained—the artefact acts within given parameters but does not have each action specifically stipulated by a human operator. Some authors (Johnson, 2006; Johnson and Powers, 2008; Johnson and Noorman, 2014) go on to claim that because of this, these technological artefacts have a greater degree of intentionality than do causally efficacious agents, in the sense that these artefacts have human intentions modelled into the way that they operate. The causally efficacious agent was simply one that had an influence on outcomes, in conjunction with human beings. The "acting for" agent, on Johnson and Noorman's construal, should be understood in terms of analogy: it can be useful to think of artificial systems as if they act on our behalf (in an analogous way to how a lawyer represents her client), but the decisions they make are not the same as ones made by human beings. The range of actions available to them is still a direct function of the intentions of their programmers/designers, and is in this sense "determined", whereas human action, according to Johnson and Noorman at least, is not. These agents differ from causally efficacious agents in that they have a greater degree of independence from direct human intervention, and thus have, to a greater degree than the causally efficacious agent, human intentionality modelled into their potential range of action.

Johnson claims that when we evaluate the behaviour of computer systems "there is a triad of intentionality at work, the intentionality of the computer system designer, the intentionality of the system, and the intentionality of the user" (2006: 202; Johnson and Powers, 2005).[35] Nevertheless, these artefacts, in order to function as desired, are fundamentally anchored to their human designers and users (Johnson, 2006: 202). This is true of systems whose proximate behaviour is independent of human operators, as even in such cases the way the system functions is tethered to its design and use, both aspects which involve human agents. These human agents have internal states, such as beliefs, desires, etc., and, according to Johnson

replace the individual as the executor of a buy/sell instruction. The human operator simply inputs certain key parameters and the bot does the rest.

[35] Johnson and Powers (2005: 100) refer to this as "Technological Moral Action" (TMA).

(*ibid.*: 198), these states are by necessity mental states. In the next chapter I will problematise this argument, as it seems to commit the same kind of error as the standard account of agency presented above, in that it claims that phenomenal consciousness is a necessary condition for moral agency.

If the tasks that are delegated to these kinds of artificial agents have moral consequences, this would provide another way in which to conceptualize the role such artefacts could perhaps play in our moral lives (Johnson and Noorman, 2014: 155). Consider, for example, automatic emergency braking (AEB) technology, which automatically applies the brakes when it detects an object near the front of the vehicle. This simple system has been enormously successful, and research indicates that it could lead to reductions in "pedestrian crashes, right turn crashes, head on crashes, rear end crashes and hit fixed object crashes" (Doecke *et al.*, 2012). We can usefully think of AEB as assisting us in being better and safer drivers, leading to decreased road fatalities and injuries. These artificial systems, of which AEB is an example, can therefore be seen as performing delegated tasks which can have moral significance. We can therefore meaningfully think of them as being morally relevant *entities*. However, according to Johnson (2006), because of the type of intentionality these entities have, they cannot be considered to be moral *agents*. Johnson claims that the intentionality that we can accord to technological artefacts is only a product of the intentionality of a designer and a user, and so this intentionality is moot without some human input (*ibid.*: 201). When designers engage in the process of producing an artefact, they create them to act in a specific way, and these artefacts remain determined to behave in this way. While human users can introduce novel inputs, the conjunction of designer- and user-intentionality wholly determines the type of intentionality exhibited by these types of computer systems (*ibid.:* 201). So, while it is reasonable to assess the significance of the delegated tasks performed by these artefacts as potentially giving rise to moral consequences, it would be a category mistake "to claim that humans and artefacts are interchangeable components in moral action" in such instances (Johnson and Noorman, 2014: 153). For example, consider the type of moral appraisal we might accord to a traffic light versus a traffic officer directing traffic: while these two entities are, in a functional sense, performing the same task, they are not morally the same (Johnson and Miller, 2008: 129; Noorman and Johnson, 2014: 153).

In order to press this point further, Johnson and Miller carve a distinction between "scientific" and "operational" models and how we evaluate each one respectively (2008: 129). According to the authors, scientific models are tested against the real world, and in this way these types of models are constrained by the natural world (*ibid.*: 129). For example, we can be sure we have a good model of a physical system when our model of this system accurately represents what actually occurs in the natural world. Operational models, on the other hand, have no such constraints (besides, of course, their physical/programmed constraints). These models are aimed at achieving maximum utility: they are designed to realise specific outcomes, without the need to model or represent what is actually going on in the natural world *(ibid.*: 129). For example, a trading bot (as discussed above) need not in any way model human thinking before executing a trade. All that is important for such a bot, for example, is that it generates the maximum amount of profit, given certain constraints. Moreover, the efficacy of such systems is often exactly that they exceed the utility provided by human decision making, usually in cases where complex mathematical relationships between numerous variables need to be calculated. In light of this, Johnson and Miller argue that because only the *function* of the tasks is the same (when comparing human action to operational models), we should not think of such systems as moral agents, as this would reduce morality to functionality, an idea to which they are directly opposed to (*ibid.*: 129). I think this line of reasoning is faulty and relies on an anthropocentric intuition regarding responsibility ascription, but a detailed critique of this will have to wait until the next chapter. For now, all that should be noted is that artefacts can be agents that, when acting on behalf of human beings, participate in acts that have moral consequences. This, however, does not necessarily mean they are morally responsible for the actions they participate in bringing about: once again, in the current literature, this responsibility is reserved for human beings.

2.3.3 Autonomous Agency

The third and final conceptualization of agency to be dealt with in this chapter is that of autonomous agency. On the face of it there are two ways in which we might come to understand the "autonomous" aspect of this account. Firstly, there is the type of autonomy that we usually ascribe to human beings. This type of autonomy has a distinctly moral dimension and, according to Johnson and Noorman (2014: 151), it is due to our autonomy in this sense that we have the capacity for moral agency. The moral dimension has to do with how this sense of "true" autonomy (in the history of moral philosophy at least) has come to be taken as a distinctive feature of humans which distinguishes us from other types of entities. "True" autonomy is often used in discussions of moral agency as the key ingredient which lends credence to the idea that only human beings qualify as moral agents, as we are the only entities with the capacity for this kind of autonomy (see Johnson, 2006, 2015; Johnson and Miller, 2008; Johnson and Powers, 2008; Johnson and Noorman, 2014). Hence, it is due to the fact that individual human beings (apparently?) act for reasons that they can claim "authorship" for, that they can be said to be truly autonomous and this is what allows us to hold one another morally responsible (also see Wegner, 2002: 2). According to Johnson and Noorman (2014: 151) if a being does not have the capacity to *freely* choose how to act, then it makes no sense to have a set of rules specifying how such an entity *ought* to behave. In other words, the type of autonomy requisite for moral agency here can be stated as the capacity to *freely choose how one acts* (*ibid.*: 151).

However, there is a second understanding of "autonomous agency" that has to do with how we might define it in a non-moral, engineering sense. This sense of autonomy simply refers to artefacts that are capable of operating independently of human control (Johnson and Noorman, 2014: 151; Johnson, 2015: 708). Computer scientists commonly refer to "autonomous" programs in order to highlight their ability to carry out tasks on behalf of humans and, furthermore, to do so in a highly independent manner (Alonso, 2014). A simplistic example of such a system might be a machine-learning algorithm which is better equipped at operating in novel environments than a simple, preprogramed algorithm. Nevertheless, this capacity for operational or functional independence is, according to Johnson and Noorman (2008: 152), not sufficient to ground a coherent account of *moral* agency, since, they argue, such agents do not freely choose how to act in any meaningful sense. So, while the authors do not suggest we eliminate the standard convention of speaking about "autonomous" machines, they insist on carefully articulating which sense of autonomy is being used. "True" autonomy, on their view, should be reserved for human beings. We should be sensitive to the specific sense of autonomy we mobilise, as confusing the two senses specified here can lead to misunderstandings that may have moral consequences (*ibid.*: 152).

To see how this might play out, it will be helpful to consider how the conception of "truly" autonomous agency not only grounds morality as such, but also confers a particular kind of moral *status* on its holder (*ibid.*: 155, my emphasis). As stated above, this conception of agency has historically served the purpose of separating humans from other entities. The traditional means by which this has been achieved is by postulating that human beings exercise a distinct type of *freedom* in their decision making, which is what grounds a coherent sense of moral responsibility. Freedom in this sense is about having meaningful control over one's actions, a type of control which makes a decision or action *up to the agent* and *not* other external circumstances. It is possible for agents of this kind to have done otherwise—they deliberately and freely choose their actions. Moreover, the sense of freedom described above has a sense of autonomy embedded into its definition: if this free decision is not the product of the specific agent in question, and is rather due to external pressures, then we cannot meaningfully consider the action to be free, and hence we would be hard-pressed to hold the agent in question morally responsible for such an act. An example of such a decision would be if an agent was coerced into performing some action (perhaps by physical force or psychological manipulation), in which case we would not consider the act to have been performed "freely".

These apparent differences in capacity for autonomous action also influences the types of rights we can coherently accord to various entities. On the basis of being autonomous moral agents, humans are accorded several clusters of positive and negative rights, and differences in the type of moral standing we possess can alter the kinds of rights we are extended (*ibid.*: 155). For example, in democratic states there is a minimum legal voting age. One justification for this type of law is the claim that one should only be allowed to vote when one reaches an age of political maturity: an age at which one can exercise the necessary capacities to *consciously* cast a *well-informed* vote. In this instance, one's capacity to make informed—and hence, ostensibly *free*—political decisions, captured in a minimum voting age, comes to inform the type of rights one is conferred (i.e. the right to vote). It is against this background that it is argued that we should be careful to distinguish between the two conceptions of autonomous agency identified here and realise that artefacts should not be understood as having the morally relevant kind of autonomy, as we cannot reasonably consider them to be *freely* choosing how they act. *Their* actions are always tethered to the intentions of their designers and end users. While I agree that there are various ways in which "autonomy" refers, I will argue that the two senses of "autonomy" specified above provide a problematic account of the concept, especially in the case of "true autonomy". The reason for this is the reliance on a prescriptive, psychologically-loaded, and ultimately unfounded analysis of "responsibility". I will take up this critique in more detail in the next chapter.

2.4 Towards a Non-Anthropocentric Account of Moral Agency

While I believe that the three conceptualizations of agency introduced above capture many of the ways in which we can meaningfully consider the roles that artificial artefacts play, I think that the methodological approach adopted by the aforementioned authors leads them to a misguided analysis. The two key issues I have with the account presented above have to do with, firstly, the sense of "intentionality" mobilised and, secondly, the inadequate account of "true autonomy". Underlying both defences of these concepts is an anthropocentric presupposition by Johnson: that there is something "mysterious" and unique about human behaviour, and that this mysterious, non-deterministic aspect of human decision making makes us "free", and therefore morally responsible for our actions (2006: 200). Exposure to the philosophical debate surrounding free will is not something I would wish on my worst enemy, but my sentiment toward this debate is not an argument. The real issue with gesturing towards human freedom as a way of grounding our moral autonomy is that one then brings metaphysically contested claims from the free will literature into a debate about moral agency. Johnson's claim rests on the fact that she presupposes some form of incompatibilism[36], more specifically libertarianism[37] (about free will, not politics). This is a controversial position to hold and is in no way the generally accepted view in philosophical debates on free will (O'Connor and Franklin, 2019). In this way her argument that the "freedom" of human decision making is what grounds the special type of autonomy that we apparently have generates far more problems than it does solutions. It is against this background that my next chapter will instead focus specifically on her usage of the concepts of "intentionality" and "autonomy" and show how her account misses some key nuances which have significant implications for our understanding of the potential for machine moral agency. I will claim that while machines may not come to exhibit the criteria required for *intentional* moral agency this is not the key issue at the heart of discussions surrounding machine morality. Instead, I will suggest that machines may come to exhibit *functional* moral agency. I will then go on to outline what conditions need to be met in order for this type of moral agency to successfully refer.

[36] This view claims that the truth of determinism is incompatible with freely-willed human action.

[37] Libertarians claim that determinism rules out free will but make the further point that our world is in fact indeterministic. It is in these indeterminacies that human decision making occurs, with the implication that these decisions are free, as they are not necessarily bound to any antecedent causal events and laws that would make them perfectly predictable.

Chapter 3: Functional Moral Agency

In the previous chapter I illustrated how the Carbonist Account of agency is flawed, on its own terms. I argued that this account of agency is too narrow in its over-reliance on internal, phenomenal mental states as necessary for agency. I then went on to discuss more plausible, broadly functionalist, accounts of agency. To do this I made use of Johnson and Noorman's (2014) three different conceptions of artefactual agency. While providing a better conceptualisation of the possibility of AAs than the Carbonist Account, I claimed that their account still has an unfortunate overreliance on internal mental states as the primary criterion in our attributions of *moral* agency. Therefore, while Johnson and Noorman's account provides us with an understanding of how we may have *agents* without mental states, and also offers a more nuanced account of the moral *roles* that AAs may perform (by identifying them as moral *entities*), this account insufficiently addresses the potential of artificial agents to also be artificial moral agents (AMAs). In order for an entity to be considered a "real" moral agent, on Johnson's account, it must still have *intentionality*, and it is this specific sense of intentionality (2006; Johnson and Noorman, 2014) that I claim is at issue. Johnson understandably uses human beings as the paradigmatic case of moral agents, as we consider most human beings to be moral agents, and so any account of moral agency that contradicts this assumption would seriously undermine our conventional understanding of "moral agency" as well as our conventional ways of holding each other responsible. With this in mind I will show how her exposition fails to properly account for how *human beings* are in fact moral agents. Following from this shortcoming, I develop an alternative, functionalist, account of moral agency, which preserves human beings as moral agents. I then go on to argue that this alternative account can also make provision for AMAs.

As I hope to show, a more plausible account of moral agency will also serve us well in a future in which AAs will become increasingly independent, as it will allow for the possibility of considering them as genuine AMAs. The value of creating a space in our moral landscape for AMAs is that by doing so we can potentially resolve the so-called "responsibility-" and "retribution-gaps" that may arise from independent AAs (Champagne and Tonkens, 2013; Müller, 2014; Gunkel, 2017; Nyholm, 2017). The former refers to cases in which it is unclear whether a human being was responsible for a given morally significant action while it is clear that an AA was causally involved in the action. The latter refers to the problem of what to do when it becomes clear that AAs *were* responsible for producing moral harms. In such scenarios, people may feel the strong urge to punish somebody for the moral harm, but there may be no appropriate (human) target for this punishment (Nyholm, 2017). The functional account of moral agency that I develop here may help us to solve the aforementioned responsibility-gap but may not be of much assistance when it comes to issues of retribution. This is not to claim that this issue is not one worthy of serious philosophical consideration, but simply to state that it will not be my explicit focus. It will be my claim, however, that the possibility of identifying an AMA will allow us to address such moral wrongs in a way that is more just than would be possible if we fail to address the responsibility gap.

My argument in this chapter will proceed in four main steps: Firstly, I will evaluate Johnson's (2006) argument that because machines do not have "intendings to act" they should not be considered as moral agents. Secondly, I will introduce the notion of *functional* moral agency, as an alternative to the *intentional* moral agency argued for by Johnson and show how the former criterion for moral attributability does not have the same shortcomings as the latter. Thirdly, I will provide a positive account of functional moral agency. While this account will draw on the work of Floridi and Sanders (2004), I will also propose a terminological amendment to their account. I will then, following Floridi and Sanders once again, flesh out a distinction between the identification of a moral agent versus the evaluation of a moral agent and show how these two processes can come apart. Briefly, whereas the identification of a moral agent simply refers to the fact that the entity in question can be said to be the source of some moral action (i.e. is causally efficacious in the production of some action with moral consequences), the evaluation of a moral agent would require that the entity be subject to moral assessment (i.e. be held morally *responsible*). If these two aspects can indeed come apart then it opens the possibility for AMAs to be identified as sources of moral action without necessarily being morally responsible in the traditional sense as a result (*ibid.*). Lastly, then, I will outline two possible options available to us following from this discussion: the conservative and the progressive approach to AMAs.

3.1 "Intentional" Agency

Johnson (2006: 198) enumerates five key conditions human beings need to meet in order for their actions to be considered properly moral actions, i.e. morally-evaluable actions. In addition, she claims that *moral* agents are the only entities that are capable of performing such actions, so if the entity is not capable of performing the type of action specified below then it cannot be considered to be a moral agent.

1. The agent [that performs the action] must have an internal mental state, which consists of its own beliefs, desires, intentions, etc. Together, these intentional states give rise to reasons for action.
2. The action must be an outward, embodied event.
3. The internal state must be the direct cause of the outward event
4. The outward behaviour has an outward effect
5. This effect is on a moral patient (i.e. is of moral significance).

Conditions (2), (4), and (5) establish that a moral action is an event that has physical consequences that affect a moral patient. An entity is a *moral agent* by virtue of the fact that it can perform such an action. Johnson wants to use the argument above to argue that computers cannot be considered moral agents, but, as should be clear, her argument centres on intentional states, as she assumes that computers do not meet criterion (1). Her conclusion is that computers cannot be moral agents because they cannot perform moral actions. However, she does not argue for why only actions that are accompanied by the requisite internal mental states named in (1) qualify as properly moral. Implicit is simply the assumption that moral actions must be performed by agents with mental states that are sufficiently like the mental states supposedly possessed by human agents in such circumstances, because humans are considered to be properly moral agents. Hence, her argument comes down to the circular claim that computers cannot be moral agents because they cannot perform moral actions and they cannot perform moral actions because they are not moral agents. Despite the circularity identified above, I will proceed to scrutinize condition (1) more closely, as it forms the crux of many arguments against machine moral agency. As already hinted at, this argument rests on illegitimate and unwarranted anthropocentric intuitions. In what follows I will refer to Johnson's argument as the *intentional account* of moral agency.

While Johnson concedes that computers may meet conditions (2) through (5) (*ibid.*: 199), her main objection to the idea of AMAs turns on condition (1), in that she denies that artificial entities can be moral agents as they do not have the requisite mental states that could have been the cause of the events that we would understand as their actions. Moreover, although entities may act in certain ways which, from an external standpoint, may mimic human behaviours, because these behaviours are not the result of some internal *mental* state (as they ostensibly are in the case of humans), according to Johnson we cannot infer moral capabilities on the part of these entities. In humans, such intentional states combine to form reasons for acting (in

Johnson's terminology "intendings to act"), and these reasons are internal to them (2006). For Johnson, this *independence* is essential, and, since AAs have the essential feature of having been designed by paradigmatic moral agents (human beings), this ontogenetic fact undermines their ability to be considered "real" moral agents, as she claims,

> [n]o matter how independently, automatically, and interactively computer systems of the future behave, they will be the products (direct or indirect) of human behaviour, human social institutions, and human decision (*ibid*.: 197).

In this way, when Johnson claims that AAs have "intentionality", this intentionality is to be understood in derivative terms, i.e. as derived from the intentionality of their human designers (*ibid*.: 201). What this implies for my purposes is that "real" intentionality is reserved for human beings and AAs can only have a derived type of intentionality (Grodzinsky, Miller and Wolf, 2008; Powers, 2013). This "derived" kind of intentionality[38] is insufficient as a criterion for grounding moral agency. Key to this distinction is that human beings have *internal* mental states, which AAs lack.

3.1.1 The Problem of Other Minds

Key to Johnson's (2006) account (as specified in condition (1) above) is that in order for an entity to be a moral agent it must have the "right" kind of internal mental states (in the form of "intendings to act"). The question that arises from this stipulation is just how we should go about figuring out just exactly what the content of these mental states is (or whether they are "really" there in the first place). We cannot cut open the cranium of a conspecific and infer the content of their mental states by peering at the wetware that is their physical brain. The usual means by which we go about attributing mental states to one another is through a third-person, publicly accessible means of evaluation (Torrance, 2014: 20). In what follows, I will therefore provide a critique of Johnson's assumption that specifically *internal* mental states can serve as a necessary requirement for moral agency.

As mentioned above, the way in which we go about attributing internal mental states to one another rests on external behaviouristic cues: the only evidence we have to go on when

[38] See Dennett (2009) for a critique of the distinction between "derived" and "original" intentionality. I will not pursue this line of argument any further, as my main concern will be with the inherent epistemic issues surrounding the sense of intentionality Johnson makes use of.

attempting to figure out what caused another human being to behave in a specific way is evidence which is expressed in some external way, in order for us to be able to evaluate it (Powers, 2013: 232). Therefore, having this requirement (the requirement of having verifiable internal mental states) in place when it comes to AMAs suggests a standard which cannot even be upheld when it comes to the attribution of moral agency to human beings. In practice we simply do not have privileged access to such internal mental states (except maybe our own) (Floridi and Sanders, 2004: 365). While it may be possible in principle (perhaps with the future development of an ever finer-grained neuroscience) to have some kind of access to an agent's internal mental states, in practice this possibility is not guaranteed (*ibid.*: 365). For now, if we have the choice between abstract metaphysical speculation (inferring the presence and nature of others' internal mental states) and using clearly identifiable, observable cues, it seems obvious that we should prefer the latter method of investigation. The latter provides us with greater epistemic security than the former, in the form of verifiable data points which we can use to make relatively sound judgements. If all that we reliably have to go on when judging whether human beings are moral agents are external cues that they provide, then positing the verifiable existence of internal mental states as a necessary requirement for moral behaviour is a condition that cannot even be met by humans. We have no guarantee that these states *exist at all*. We cannot prove, with certainty, that they do: We simply make the inference that they exist based on our observations of the behaviour of others and by inferring from our own case. This is not to claim that internal mental states should play no role in the determination of moral agency, but rather that they cannot play a role *simpliciter* in this determination, given that we are precluded from definitively knowing whether they exist or not. Johnson's stipulation that agents *must* have the right kind of internal mental states, therefore, precludes us even from having a plausible account of *human* moral agency. This is surely not a desirable outcome, as any theory of moral agency that we endorse should at the very least preserve the status of human beings as moral agents (*ibid.*: 357). Hence, any account that wants to preserve humans as moral agents must concede that the presence of *apparent* internal states of the requisite kind is a necessary precondition for moral agency. In what follows I will outline a functionalist account of moral agency. Such an account does not posit that the existence of internal mental states are necessary for moral agency and so is not subject to the same types of issues outlined above.

3.2 Functionalism

A functionalist account of mind claims that it is not *internal constitution* that makes something a mental state of a particular kind, but rather the *role* it plays, or the *function* it serves in the particular system of which it is part (Johansson, 2011; Levin, 2018). Applied to the study of moral agency, this approach claims that we need only look at certain behaviours and reactions that are functionally equivalent to similar ones in human beings to ground our moral ascriptions (Wallach and Allen, 2009). Thus the basic assumption regarding *attributability*, when answered in functionalist terms, is that "intentional understanding is simply irrelevant" (*ibid.*: 58). On this account, human beings qualify as moral agents when fulfilling certain functionalist criteria (which will be specified below). Such an account excludes the prerequisite that human beings need to have particular kinds of internal mental states. A general functionalist account of moral agency, then, by preserving our current understanding of (some) human beings as moral agents, can therefore provide a more plausible theoretical foundation for moral agency in general than intentional accounts.[39]

With this in mind, I will discuss a functionalist account of moral agency, as it is developed by Floridi and Sanders (2004). However, I will amend their account, arguing that their appeal to "autonomy" in questions relating to moral agency creates more problems than it solves. I will also respond to the objection that a functional account of moral agency gives us a less than optimal account of moral assessment. I will show that a functional account of moral agency can allow for meaningful moral assessment, which implies that we can have AMAs in the full sense. Before doing any of this, however, I will introduce the "Method of Abstraction", a key framework for functional accounts of moral agency.

3.2.1 Levels of Abstraction

The "Method of Abstraction" is an approach borrowed from "modelling in science where the variables in the model correspond to observables in reality, all others being abstracted" (*ibid.*: 354). This type of approach therefore allows for us to "abstract away" certain properties in a

[39] See List and Pettit (2011) for an account of group agency, as it is applied to the potential for corporate agents. The type of framework developed in their book is complementary to the one I present here, as they also argue for a functionalist conception of agency. In their book, however, they seek to provide a coherent account of how a multitude of individual agents can act together as a group, and how this group can itself be considered an agent. My focus, however, is more specifically geared towards individual agents. Wallach and Allen (2009) also provide a wonderful functional account of moral agency, as it relates specifically to the *design* of AMAs, and so they do not devote much time in their book to a conceptual analysis of agency per se.

given environment that may be deemed non-essential to the entity under analysis. Applied to moral agency, then, what this approach suggests is that when determining whether an agent is indeed a moral agent or not would depend on the level of abstraction that is specified. For the purposes of my argument I will adopt the same level of abstraction proposed by Floridi and Sanders (*ibid.*). Their proposal is that the correct level of abstraction for evaluating whether an entity is a moral agent or not comes down to what we might observe through a video camera over a period of 30 seconds.[40] A key reason for adopting this level of abstraction is that it helps to alleviate some of the issues inherent to intentional accounts of moral agency by focussing on easily verifiable criteria, such as what is amenable to observation, instead of internal states.[41]

The "lower" the level of abstraction, the closer it is to reality, and the more concrete and detailed it is. The more we abstract (i.e. the "higher" we go) the more observables we can "abstract away", and so the particular level of abstraction we are at refers to a given, circumscribed collection of observables (*ibid.*: 354). For practical purposes, this implies that when analysing a particular entity (such as a moral agent), at a particular level of abstraction (for example, observations made through a video camera over a period of 30 seconds), each definition that is employed refers to a given set of criteria that are observable at that level of abstraction. To abstract is not to simply overlook a particular aspect of an entity that is not included at that level, but rather to acknowledge that it does not have immediate relevance to the discourse in question (*ibid.*: 353). For example, intentional accounts of AMA set the level of abstraction quite low: the criteria embraced by these accounts tend to correspond to that of an adult human being and all that we take that to entail. The basic functionalist claim, however, is that the level of abstraction proposed by intentional accounts of moral agency is *too low*, and that by raising it we can come to appreciate less anthropocentric perspectives on which other entities may qualify as moral agents. In other words, by abstracting away certain criteria (such as responsibility, free will, or mental states) we can have more philosophically coherent accounts of moral agency, even if these accounts turn out to be "mind-less"[42] (Floridi and Sanders, 2004; Powers, 2013). The benefit of adopting this methodology in this instance is that we will be able to preserve the philosophical coherency of the view that human beings are in

[40] One might object that the proposed level of abstraction does not do much work in the argument I am presenting. However, I am merely reporting on Floridi and Sanders' usage, not advocating for it explicitly. It may be that this proposed level of abstraction is inappropriate for discussions on moral agency.

[41] One could argue against this and claim that this is the incorrect level of abstraction to adopt for evaluations of moral agency (see Grodzinsky, Miller and Wolf, 2008). However, I believe that this is the correct level to adopt as it best captures our intuitions with regards to how we determine whether an entity is an agent or not: whether we can see it have an effect, in the form of an action, on the world.

[42] "mind-less" in the sense of not necessarily requiring mental states.

fact moral agents without importing unjustifiable assumptions as is the case at the "intentional" level. The method of abstraction proposed above allows us to disregard redundant features that we assume exist in adult human beings but cannot reliably prove exist. They are redundant in that disregarding them does not change the way in which we identify adult human beings and assess their moral actions in practice. Moreover, the functionalist account has the advantage of allowing, at least in principle, the potential inclusion of machines into our moral universe. Intentional accounts prohibit this by an appeal to unjustified intuitions regarding human exceptionalism from both other biological and artificial entities. In what follows I will therefore outline three key functionalist criteria of moral agency, as proposed by Floridi and Sanders (2004).[43]

3.2.2 Functionalist "Moral" Agency

The three functionalist criteria of agency that Floridi and Sanders (*ibid.*) propose are:

1. Interactivity: the entity has the capacity to interact with its environment.
2. Autonomy: the entity in question has the ability to change itself without direct intervention from its environment. In other words, the entity can change states without direct external influences.
3. Adaptability: the entity is capable of updating the transition *rules*[44] by which it changes states (*ibid.*: 357-358).

If an entity satisfies the three aforementioned criteria, it is an agent. A few examples (both natural and artificial) may help to clarify the conditions stipulated above. The obvious example is that of an adult human being: we can interact with our environment, we can change states without direct external intervention, and we are capable of adapting to novel environments and can thereby update the "transition rules" by which we operate: We can learn new habits or change our preferences over time, with these changes resulting in new choice architectures becoming available to us.

[43] These criteria are part of a much larger project known as "information ethics" (Floridi, 1999) which seeks to develop a minimalist theory of deserts not tethered to concepts such as "life", "pain", etc. Also see van den Hoven and Weckert (2008) for an overview of the field.

[44] This ensures that the agent in question has the ability to learn from its interactions with its environment and then use this information to change the way it learns.

Next, consider an entity which meets none of the criteria: a rock. Recall that the proposed level of abstraction is what would be possible to observe through a video camera over a time period of 30 seconds. With this stipulation in place, a rock would not be capable of interacting with its environment (perhaps over a longer period of time erosion may occur, potentially satisfying the interactivity condition) (*ibid.*: 358). The rock is also straightforwardly not autonomous, as it cannot do anything unless acted upon, and is therefore not adaptive either. Let us contemplate a more complex system next: an email filtering bot. The purpose of such bots is to monitor incoming emails, and, by learning the preferences of the user, filter unwanted mail out of one's inbox. The bot is interactive in that it receives inputs (in the form of incoming mail) and has outputs (in the form of filtered mail in the user's inbox). It is autonomous in that it can change states in response to these inputs without the intervention of the user. The entire point of such systems is that they have the ability to discriminate between junk and important mail without explicit input from the user. Moreover, it is adaptable in that it can learn the preferences of the user and use this information to update the way its filtration mechanism functions, changing the internal rules by which it operates. This type of bot, therefore, can be considered an agent on the functionalist definition. The next question would be whether it could be considered a *moral* agent. A moral agent would have to meet an additional stipulation:

> 4. An agent is a moral agent "if it is capable of morally quantifiable action" (*ibid.*: 364).

While the specification of "morally quantifiable action" is certainly one that requires some fleshing out, I will only do so in a later section of this chapter. For now, however, I will simply outline some positive upshots of this account more generally. The first feature to note is that this type of specification is neutral about whether the agent has internal mental states and is therefore immune against the problem of other minds outlined earlier. Secondly, this account also preserves the view that human beings count as standard moral agents, as our actions are clearly morally quantifiable in certain cases (such as when we transgress moral norms like committing murder), meaning that certain of our actions can cause morally beneficial or harmful outcomes to occur. Moreover, our moral sensibilities can (and do) change over time, as can be seen by the (depressingly recent and restricted) acknowledgment that gay marriage is not immoral, which was denied in the past. If an AA were to similarly meet the four criteria enumerated above, we would have to concede that they would qualify as AMAs.[45]

3.3 Problematising Autonomy

While I agree with the interactivity and adaptability conditions stipulated above, the autonomy condition (2) is one that I have terminological reservations about. The main reason for this is that the term "autonomous" comes with heavy philosophical baggage. To see why, one need look no further than my previous chapter. There I showed how Johnson argues that we should be cognisant of the distinction between "autonomy" as it is used in the engineering sense and "autonomy" as it is used when applied to human beings, especially in the context of moral theorising. The engineering sense refers to how an entity may be able to operate outside human control; the moral sense refers to a "special" capacity that human beings have, elevating us above the natural world and making us morally responsible for our actions. I have already presented an argument to the effect that this presupposition is both metaphysically contested and unjustifiably anthropocentric (in that it simply presupposes that human beings have free will and that humans are thus the only moral agents). It is this "metaphysical baggage" that I believe makes the criterion of autonomy problematic. However, the claim that the concept comes with unfortunate connotations is no substitute for an argument. Therefore, using the aforementioned as a stepping stone, I will show why we have good reasons to "rethink" the criterion of autonomy (also see Alterman, 2000).

3.3.1 Losing the Baggage

By "autonomous", what is usually meant is the ability of some entity to change states without being directly caused to do so by some external influence (see (2) above).[46] This is a very weak sense of the term (in contrast with the way it is traditionally understood in moral and/or political

[45] This means that animals, in given moral contexts, could also qualify as moral agents. An example of this might be a trained police dog who misidentifies a suspect and attacks an innocent person. In such a situation we can rightly consider the dog to be a moral agent as it is causally efficacious in the production of a moral harm. This will become clear as the chapter develops. Children could similarly be considered moral agents, although we treat animals and children differently than we do adult human beings. Once again, as my chapter develops the relationship between being a moral agent and moral responsibility will become more salient.

[46] Changing states simply refers to an entity's ability to update its internal model of the world in light of new information from its environment. This can be as simple as a thermostat keeping the temperature at a set level despite the temperature dropping in the environment.

philosophy) but the basic idea can be grasped with this definition.[47] It captures the major difference between how the term is used in the design of AI systems and how it refers when applied to human beings: the engineering sense and the so-called "moral" sense. A potential reason for stipulating autonomy as a condition for functional moral agency is that there is a pervasive assumption in AI research that one of the main goals of creating machine intelligences is to create machines that can act autonomously in the engineering sense: reasoning, thinking and acting *on their own, without human intervention* (Alterman, 2000: 15; Van de Voort, Pieters and Consoli, 2015: 45). This is a design specification that has almost reached the level of ideology in AI research and development (Etzioni and Etzioni, 2016). However, as is argued by Alterman (2000: 19), identifying machine autonomy, is already problematic as the distinction between the non-autonomous "getting ready" stage and the autonomous "running" stage of a specific AI is a spurious one at best. In the first "getting ready" stage a system is prepared for deployment in some task environment. In the second stage the systems "runs" according to its design (*ibid.*: 19). Traditionally it was supposed that these two stages are what separate the "autonomous" from the "non-autonomous" states of the machine. However, consider a case where a system has completed the "getting ready" stage and is ready to "run", and suppose that while entering its "running" state in its given task environment the system encounters an error. In such a situation, it would be necessary to take the system back into the "getting ready" stage in the hope of fixing the bug. In this way there is a cycling between the "getting ready" and "running" stages, which entails cycling between stages of "autonomous" and "non-autonomous" learning (*ibid.*: 19). This means that the system's "intelligence" is a function of both of these stages, and so it becomes unclear where we should be drawing the line between what counts as autonomous or non-autonomous in terms of the states of the machine. According to a Alterman , "if the system is intelligent, credit largely goes to how it was developed which is a joint person-machine practice" (*ibid.*: 20). In other words, if the system is considered intelligent, this is already largely a carbon-sillicon collaborative effort. Instead of asking whether the system is autonomous or not, then, we should perhaps intead inquire as to how its behaviour might be independent (in a matter of degree, not type) from its human designers'.

My claim is that this very weak sense of autonomy, as it is used in the design of AI, invites confusion. For example, the most common usage of the term "autonomous" in discussions on machine ethics usually revolves around military applications (see Sparrow, 2007; Müller,

[47] For example, see Christman (2018) for an exposition on how autonomy refers in the moral and political arenas.

2014). A key issue here, however, can be noted in the metaphysical baggage that comes with the ascription of autonomy to a system. Consider this remark by Sparrow where he claims that "autonomy and moral responsibility go hand in hand" (2007: 65). What this would imply is that if an AA were to be considered autonomous in this weak sense, it would be also have the capacity to be held morally responsible for certain actions. This line of reasoning would force us to concede that every autonomous system is also capable of being held morally responsible. However, it is possible that some machines may indeed be autonomous without being morally responsible. For example, a military drone might be sent to execute a strike on a certain, predetermined location. This location is programmed (by a human) into the drone before it takes off, but from the moment of take-off the drone acts autonomously in executing the strike. Let us assume that the strike is unsuccessful, as civilians instead of terrorists were at the strike location. In this case, while the drone is autonomous, we would not hold the drone morally responsible for this outcome, as it is clear that the moral harm was due to human error. In this weak sense the criterion of autonomy would provide an implausible account of agency more generally, as it would never allow for minimally "autonomous" machines that are not morally responsible for their actions. In other words, by relying on "autonomy" as a criterion, Floridi and Sanders' account overextends the concept of moral agency. Therefore, instead of trafficking in the language of autonomy, I propose that we instead utilise independence from human control as a criterion for agency.

What this would entail is that instead of asking whether this or that AA is autonomous or not, we should instead investigate to what extent it operates independently of human control. It seems as though these questions would have similar (if not the same!) answers, but the main point is that in the latter case we know with more precision what we are talking about. For example, consider an example of a military drone which is deemed to be autonomous. If such a system were to perform some moral harm, under Sparrow's analysis, we would say that the drone is morally responsible for this harm (*ibid.*). As I outlined above, this would miss key steps in the analysis, as in such a situation we skip from autonomy to responsibility without, for example, asking whether the entity is also adaptable. What seems to happen when we use this type of language is that the surface grammar of the term "autonomy" does not track well with its underlying logical form. With this in mind, it should be clear that my proposal is a semantic one: by changing our terminology from "autonomy" to "independence", the hope is that we will avoid the aforementioned metaphysical disputes. This will not necessarily change the *content* of the "autonomy" condition above, and so my amendment to Floridi and Sanders' account should not be read as my being in disagreement with them on that score. Rather I propose a renaming of the criterion in order to avoid unnecessary metaphysical disputes and baggage.

In the case above, therefore, asking whether a system is autonomous or not requires us to first disambiguate the term, specify the sense in which we are using it, and even then, we might still not answer the question in the right way because of some metaphysically contested baggage which may hinder our investigation. On the other hand, asking whether an entity is independent or not is a much more specific and clear question, a question with far less baggage, and one which would prompt more coherent and to-the-point answers, aiding us in our investigation of functional moral agency. This independence can be understood by observing how many of its decisions the AI is capable of making based on its own programming, and outside the direct influence of human control/supervision.

3.4 (A)moral Responsibility?

Perhaps the most vexing question that remains unanswered (or unasked?) at this point is, in this functional account of moral agency, what the relationship between a functional moral agent and moral responsibility might be. On traditional accounts it is presupposed that being a moral agent is a necessary condition for being held morally responsible for one’s actions. The functionalist account accepts this but is not as restrictive as to who *qualifies* as a moral agent, and by extension who can be considered to be morally responsible. A key reason for this lies in dismantling an anthropocentric intuition which claims that in order for an agent to be a moral agent it must be capable of being subject to moral evaluation. This need not be the case, though, as according to Floridi and Sanders it is possible to have moral agents that are not morally responsible/evaluable (2004: 367). In the next section I will present their argument to this effect. Following from this I will go on to show how their account ultimately falls short of providing a fully satisfactory account of moral assessment. They do not go far enough in their own analysis, as while showing how we can have moral agents without mental states, they do not take this argument to its logical conclusion. They propose that being punishable is necessary in order for a moral agent to be morally responsible, and that this is necessarily tethered to the agent in question having the “correct” psychological responses to such punishment. I will show how this need not necessarily be the case and claim we can also have morally responsible agents who lack the “correct” mental/psychological states.

3.4.1 Identification and Evaluation

In this section I will briefly trace the contours of the argument provided by Floridi and Sanders as to what constitutes the "morally quantifiable action" criterion identified above. They propose that identification and evaluation as a moral agent can come apart. In order for an AA to be a moral agent, on this account, it would only have to be the *source* of a particular moral outcome, and not necessarily morally responsible for that outcome. They claim that talk of moral responsibility is unnecessarily "soaked with anthropocentrism" (*ibid.*: 366). This anthropocentrism pumps the intuition that in order for an entity to be considered a moral agent, it must necessarily be capable of being subject to moral evaluation and sanction. Floridi and Sanders suggest that this intuition confuses *identification* of an entity as a moral agent with the *evaluation* of that entity as a morally responsible agent (*ibid.*: 367).

Consider the example of children, as described by Floridi and Sanders (*ibid.*). It is conventionally assumed that a good parent will identify their children as sources of moral action (and therefore as moral agents). However, up until a certain age (eighteen[48] usually being the upper limit on this front) parents do not hold their children to be fully morally responsible, i.e. subject them to moral evaluation equivalent to that of an adult, despite the fact that they are clearly moral agents (in a limited sense). An eight-year-old who misbehaves is rightfully reprimanded, but the parent does not consider the child to be capable of being subject to full-blown moral evaluation and punishment in the way an adult would be. In this case, according to Floridi and Sanders, we can assert that a child has been causally efficacious in the production of some moral outcome (and is therefore a moral agent), without it *necessarily* being the case that the child is also morally responsible and thus fully liable to sanction.

To bring the discussion back to AA, an implication of this view is that once identified as the source of a morally quantifiable action, an AA can be attributed the status of AMA. Importantly, however, the further claim is made by Floridi and Sanders that mental/psychological states are a *necessary* requirement for entities to be subject to full-blown moral evaluation. Nevertheless, this is not a requirement for identification as a moral agent. Identification of whether an entity is a moral agent or not is only the first step in figuring out whether the entity is also morally responsible for some specific event (*ibid.*: 367). The further determination of moral responsibility depends on evaluative, backward-looking moral considerations and is importantly tethered to the agent in question being capable of having the

[48] Our best neuroscience, however, suggests that our neocortex is only fully developed by age 25, so perhaps eighteen is too soon for us to be morally responsible for our actions (see Sapolsky, 2018).

correct psychological responses to its “punishment”. In this way, it is argued, an AMA can be causally efficacious in the production of some moral harm, without the further claim that the AMA is also morally *responsible* and hence liable to *punishment* (*ibid.*: 367).[49] In other words, causal efficaciousness in the production of some moral event is a necessary but insufficient condition for moral responsibility (*ibid.*: 371).

Floridi and Sanders, therefore, implicitly maintain that moral responsibility is reserved for human beings and preclude AMAs from such assessment, as we have the correct psychological responses when accused of performing a moral harm while AMAs do not. In the sections that follow I will argue against this claim and present the case that moral responsibility need not be tethered to an agent’s capacity to be punished. Before doing this, however, I will provide an outline for how “punishability” features in moral assessment more generally.

3.4.2 Moral Agency, Moral Responsibility, and Causal Efficaciousness

An important feature of the functionalist account presented thus far is that it is not at all necessary for an AMA to have anything like mental/psychological states for it to be a moral agent. This is important for present purposes as it is traditionally supposed that when we praise and blame one another we do so in order to ensure that we and those around us get what we deserve. Traditionally tied to this basic sense of desert is that those who do wrong should be punished (or made to feel bad) and those who do good should be rewarded (or made to feel good). It should be clear that on this conception “desert”, at least for human beings, turns on us having the capacity for subjective states of experience. Even the pragmatic turn initiated by Strawson (1993), which grounds moral assessment in terms of *reactive attitudes* still traffics in the language of subjective states (e.g. resentment, gratitude, guilt). On this account, praising and blaming an agent for some behaviour is a way to signal that the behaviour is either to be encouraged or discouraged.[50] With this information the immoral agent, at the very least, should update her moral sensibilities and become a better moral agent in the future. In this aforementioned sense, then, there is a pedagogical element at play when it comes to morally

[49] Floridi and Sanders (2004: 367) make the same point using different (and confusing) terminology. They claim that identification as the source of a moral harm means an AMA is morally accountable (descriptively speaking), and that moral responsibility (normatively speaking) is tethered to evaluative, backward looking moral considerations.

[50] This is in line with a broadly *consequentialist* view of moral responsibility, which claims that praise and blame are appropriate when they can lead to an agent changing their behaviour in some morally desirable way. This is in contrast to *merit*-based approaches which claim that praise and blame are only appropriate if an agent *deserves* such responses (Eshleman, 2016).

evaluating an agent. When applied to human beings, it is clear that these reactive attitudes stem from or are aimed at various psychological states (such as a feeling of resentment at being harmed, or guilt at harming another). Being morally assessable in this sense then entails being subject to desert—liable to praise or blame (Vargas, 2015: 16), and being liable to praise and blame entails being amenable to these. This implies that in order to be morally evaluable, an agent needs to have psychological states. Floridi and Sanders seem to implicitly accept this common-sense description. For what is the use of apportioning praise or blame if the agent cannot appreciate these? I will address this issue below.

What should be emphasised is that the sense of desert I have in mind is not one that is metaphysically demanding, and so does not turn in any significant way on questions of whether our world is deterministic/indeterministic or whether we have free will or not (see Strawson, 1993; Roskies and Malle, 2013). This is what I take from Strawson: ascriptions of moral responsibility do not turn on major metaphysical theories but are instead grounded in social and/or relational terms. We do not evaluate people morally on the basis of whether we think they do or do not have free will (or some other grand metaphysical thesis). Instead, we do so when the following four conditions are met: Firstly, when they have performed an action that has moral consequences. Secondly, when we can determine that they are responsible for their actions, in that the action was a choice made by the agent and was not performed under duress or some direct external compulsion. Thirdly, when they are morally evaluable agents, such as being adult human beings of sound mind and therefore capable of sound judgment. And finally, if they have the capacity to appreciate the wrong that they have done and are amenable to either praise or blame. It is this last criterion that I wish to question, as the ability to appreciate whether one has done a moral wrong need not be necessary in order for an agent to be morally evaluable, as is claimed by Floridi and Sanders. In other words my claim here is that moral amenability, and more specifically moral responsibility, is a function of the social role an agent performs and *not* the product of his or her internal states *simpliciter* (see Coeckelbergh, 2010; cf. Torrance, 2014). What I further aim to show, contra Floridi and Sanders, is that this grounding of moral assessment need not be tethered to the existence of psychological states in the moral agent. I will use the example of psychopaths to make this point clear.

3.4.3 Psychopaths and "Punishability"

There are clear cases in which psychopaths[51] are causally efficacious in the production of moral harms, and we do in fact hold them morally responsible for these actions. We also punish psychopaths when we find them guilty of performing morally bankrupt actions: they are placed in prison. This punishment is independent of whether they could have in fact "done otherwise" (given their constitution). Moreover, it is unclear whether punishment would even lead to an improvement in their behaviour (due to their lacking an emotional component to their decision-making) (Litton, 2008). However, there are clear pragmatic benefits to society with this kind of system in place, as it ensures that psychopaths are kept in effective quarantine and are not capable of harming others. In this way, then, they may be considered similar to AMAs in that they lack the capacity for emotion but do in fact possess the ability to reason practically. We accept both to have the capacity to be causally efficacious in the production of moral harms, and "punish" both even when it is unclear that this punishment has any effect on them at all. In the case of psychopaths, this punishment consists in their removal from society, in the case of an AMA it might refer to reprogramming the system or simply turning it off.[52] Although it seems counterintuitive, this account suggests that an agent may be morally responsible without the agent having the capacity to appreciate why it is being punished. So, even though psychopaths are not "punishable" in the sense that they are reactive (or care about) praise or blame, we punish them in any case for the pragmatic, societal, benefits this confers. Their lack of feeling, therefore, does not preclude us from finding them to be responsible moral agents. In much the same way, an AMA may not "feel bad" when punished, but this need not stop us from potentially holding them morally responsible. They are still moral agents "deserving" of "punishment", even if that means we just turn them off or reprogram them for the benefit of society. The example above, therefore, introduces the possibility that there is a case to be made for a sense of moral responsibility that is independent of whether the agent in question is "punishable" or not. Floridi and Sanders resist this claim, while I will argue below that we

[51] The usage of psychopaths as an example may be objectionable, as there is an argument to be made that they are perhaps not morally responsible for their actions precisely because they lack the correct psychological responses. Moreover, one could maintain that those who hold psychopaths morally responsible are in fact incorrect for doing so. For the purposes of this thesis, however, I will assume that there is at least a plausible case to be made that psychopaths are indeed morally responsible for their actions. It is clear that this argument would need to be developed in a number of ways, but that will not be my concern here.

[52] Recall that my discussion of moral agency does not presuppose a prior ascription of moral patiency, and so worries about personhood or arguments regarding "robot rights" are of no significance here. This is a significant and constructive aspect of my thesis, and I will investigate it in more detail in my conclusion

should embrace it. There are therefore two possible paths one could follow, what I will term the *conservative* and the *progressive* accounts of functional moral agency.[53]

3.5 Conservative and Progressive Moral Agency

In the account presented thus far identification as a moral agent implies that the entity in question was causally efficacious in the production of some moral event. This is still short of being held morally *responsible*, as being causally efficacious in the production of a specific event is simply to claim that the agent was the *source* of the morally relevant outcome. Positing that the agent is morally responsible, on the other hand, would be a much stronger claim. But what exactly does it mean to be held morally responsible, and can an AA ever be held morally responsible in a way analogous to how human beings are held morally responsible? While Floridi and Sanders' account creates the conceptual space for AMAs, they do not go on to consider what the implications might be for these entities to also be considered morally responsible for their actions. According to Floridi and Sanders, to hold an agent morally responsible necessarily entails that the agent can be praised or blamed for its actions, as they state "that it would be ridiculous to praise or blame an AA for its behaviour or charge it with a moral accusation" (2004: 366). It thus seems as though they would resist the conceptual leap to claiming that AMAs can be morally responsible for their actions. Their argument is more concerned with establishing that AAs can coherently be thought of as sources of moral action (and therefore as moral agents). I will therefore take their argument further and consider the implications of this more radical extension of their account. I will introduce two potential routes we may take in our approach to understanding moral responsibility as it relates specifically to AAs. The first view I will term *conservative* moral agency (the view that AMAs can only be causally efficacious in moral outcomes, not morally responsible) and contrast it with *progressive* moral agency (the view that AMAs can indeed be morally responsible). What is important to keep in mind, however, is that in both cases one should still accept the thesis as I have defended it up until this point: that identification as moral agent and evaluation as moral agent can indeed come apart in our practices of moral assessment. In what follows I will outline both the conservative and progressive accounts of AMA and flesh out the implications of each view.

[53] This terminology is not my own but is borrowed from Davies (2007). Davies uses the distinction to show the difference between "deeply" naturalistic conceptual orientations and ones that are ostensibly naturalistic but are insufficiently so (2007: 39).

3.5.1 Conservative Moral Agency

What I have termed the conservative account of moral agency claims that in principle only entities with psychological/subjective states of experience can be appropriately evaluated as being morally responsible for their actions. This does not mean that AAs are not moral agents, only that they cannot be prescriptively assessed (i.e. be subject to full blown moral evaluation like an adult human being might be). This type of argument is essentially Strawsonian as it claims that having certain intentional/dispositional/psychological states of mind are essential to the attribution of praise and blame for certain actions. Tied to this are basic questions of desert and how these are linked to psychological states more generally, in that only entities with the "right" kind of phenomenology (so far as we know we are the only creatures with the correct hardware/software combination for this) have the ability to have their actions morally sanctioned. On this account the fact that we have the capacity to feel emotions such as guilt or resentment as a result of our own behaviour (or the behaviour of others) is essential to whether or not we *deserve* to be held morally responsible. However, one could present a counterargument to the effect that we cannot ever really "know" whether an entity in fact has the "right" phenomenology (this would be a variation of the problem of other minds introduced earlier). In response to this, a defender of the conservative view could claim that yes, consciousness and its associated phenomenology are a core requirement of being held morally responsible. However, how we *identify* whether an entity is conscious or not should be understood in behaviouristic terms: if an entity behaves like it has mental states, then we should concede that it does in fact have those mental states (Johansson, 2010). Yet, a defender of the conservative view would still need to add a kind of "discontinuity" principle, positing some essential difference between human moral agents and AMAs, which need not necessarily turn on any mentalistic criteria (Fossa, 2018: 123). The reason for this is that on the tenets of the conservative view only human moral agents can be morally responsible, while AMAs cannot. An example of this type of principle could be that AMAs are always *sensible tools*, in that they "do not have the possibility to play a conclusive role in determining the final ends and the moral values of humanity" (*ibid.*: 123). I will not go into detail as to what the correct distinguishing factor may be, as I believe attempting to posit such a discontinuity rests on an anthropocentric intuition: it relies on the fact that in the past the only entities capable of moral evaluation and action have been human beings, but this is no guarantee that it will remain this way into the future. Furthermore, how are we to say in advance what tasks an AMA may be able to solve in the future, or whether the ubiquity of AMAs come to dominate our moral discourse and, in this way, set the standard for what counts as good moral reasoning or not?

The implications of this conservative account are therefore that a significant burden is placed on the programmers and developers who fund and design AI systems. If AMAs cannot be considered morally responsible, then what seems to emerge are potential "responsibility-gaps" (Nyholm, 2017). As outlined earlier, these are cases in which we are uncertain whether a human agent (programmer/developer, etc.) or an AMA was more causally efficacious in bringing about some moral outcome. Consider this example of a near future banking algorithm from Bostrom and Yudkowsky (2011: 1):

> A rejected applicant brings a lawsuit against the bank, alleging that the algorithm is discriminating racially against mortgage applicants. The bank replies that this is impossible, since the algorithm is deliberately blinded to the race of the applicants. Indeed, that was part of the bank's rationale for implementing the system. Even so, statistics show that the bank's approval rate for black applicants has been steadily dropping. Submitting ten apparently equally qualified genuine applicants (as determined by a separate panel of human judges) shows that the algorithm accepts white applicants and rejects black applicants. What could possibly be happening?

Getting to the correct answer in such a situation may prove to be impossible in practice, as it may not be possible to understand exactly why or how the algorithm is coming to its racially skewed conclusions.[54] In such a state of affairs, who is to be held responsible for the actions of the algorithm? According to the conservative view above, we should look towards the executives at the bank and the programmers of the algorithm in our evaluative responsibility analysis. On this account, to take seriously the claim that the algorithm itself is morally responsible is a closed-off possibility. For a more progressive account of moral agency, however, such a move is available.

[54] This would in large part depend on the type of machine learning algorithm in question: neural networks and genetic algorithms are much less transparent than systems based on Bayesian networks or decision trees (Bostrom and Yudkowsky, 2011: 1). In the case of more transparent systems, it may be possible to identify the cause of the problem, which may be that the algorithm, for example, uses the addresses of applicants from historically poverty-stricken areas and based on this denies their applications.

3.5.2 Progressive Moral Agency

The progressive account goes further than the thesis presented thus far and in fact claims that AMAs can indeed be subject to moral evaluation in the same way that human beings are. The reason for this is that there is indeed *continuity* or *homogeneity* between human moral agents and AMAs and that, moreover, there is no qualitative difference between human moral agents and AMAs (Fossa, 2018: 115). This argument resists the temptation to presuppose that only entities with psychological internal/subjective states can be morally assessable in principle. Instead it asserts that there is a homogeneity between human moral agents and AMAs, and that, in principle, it should be possible for an AMA to be morally responsible in the same way that human beings are morally responsible. This is captured in the four criteria stipulated above that must be met in order for an entity to be considered a moral agent. As it was shown earlier, both human beings and machines can qualify as moral agents on this account. As we saw, this view showed why and how psychological states are spurious postulates in our ascriptions of moral *agency*. The progressive account claims that these postulates are also spurious when it comes to moral evaluation, i.e. in attributing moral *responsibility*.

One of the main reasons for this is that, on the conservative account, the buck of moral responsibility still stops and starts with human beings and their associated psychology. This account then still has no cogent answer to the critique of other minds above, as it still postulates the existence of particular internal mental states as a discontinuity principle between human moral agents and AMAs. The progressive account incorporates the view that an entity can be morally responsible for an action whether or not it is "punishable" or not. As I have shown, it is possible to hold an agent morally responsible without the further requirement that the agent also needs to be "amenable" to "punishment" (as in the case of psychopath).

Consider once again the example from above of the opaque banking algorithm. To reiterate, the conservative response to such a situation would be to look towards the programmers of the algorithm or the owners of the bank who decided to implement the algorithm and to hold them morally responsible for the harm that occurred. On such a view, the AI that executed the decision would be thought of as an AMA in that is the *source* of the moral harm, but it would not be considered *morally responsible*. The progressive response, however, allows for the possibility that the AI itself is morally responsible. Now this would of course depend on certain conditions being met, such that, at the given level of abstraction, we determine that the AMA *could have behaved otherwise than it did*. The key point of failure for the conservative account, as we saw, is the potential emergence of a responsibility-gap, a situation in which our responsibility ascriptions are indeterminate: these would be states of affairs in which we cannot clearly identify who or what is to be held morally responsible for an action and where we end up holding no one accountable. In the case of the banking algorithm above where it may be impossible to know whether a human being was morally responsible or not, how are we to assign responsibility for the moral harm that occurred? It might emerge after a detailed investigation of the role played by the software developers that through no fault of their own the AI started discriminating based on race. Using the progressive account just discussed allows for a greater number of possible moral responses, which could aid in reducing their occurrence and in compensating victims for the harms that AMAs may cause.

The progressive account provides a solution to this problem by proposing that there is no need to posit substantive differences between human moral agents and AMAs: both are part of a homogenous grouping—a moral type—and should be evaluated as such. This approach is functionalist "all the way down", in that it assumes that human modes of existence (understood in formally specified functional criteria) can be reproduced technologically in the form of an AI meeting the functionalist criteria outlined above.

3.6 Concluding Remarks

In this chapter I have shown how "intentional" accounts of moral agency fall short of providing a philosophically coherent account of the concept. From this conceptual failure, I introduced a functionalist account of moral agency, which provides a much more secure foundation for our investigations into moral assessment. Using Floridi and Sanders' criteria as a starting point, I developed a fleshed-out functionalist account of moral agency. Central to this account was a disambiguation between identification as moral agent and evaluation as a moral agent. It was shown how these two factors can come apart, and that it is therefore possible for an entity to be considered a moral agent without it necessarily being morally responsible. Following from this I outlined two potential routes one could take: conservative or progressive accounts of moral agency. The conservative account claims that there is a qualitative difference between human moral agents and AMAs, while the progressive account claims that there is no such difference. I argued that we should be progressive in our conceptualization of moral agency. Doing this resists the illegitimate temptation to assume that in order to be morally responsible an agent should also be punishable. Moreover, this progressive orientation also allows us to solve the emergence of potential responsibility-gaps. In my conclusion, I will consider these implications in further detail and claim that this progressive approach is preferable on both philosophical and methodological grounds.

Conclusion

A central theme throughout this thesis has been that we should keep our conceptions of moral agency and moral patiency distinct, and that talk of the one does not *necessitate* talk of the other. To assume that one cannot be a moral agent without being a moral patient is to operate with an illegitimately anthropocentric account of moral agency where moral agents must necessarily have particular phenomenal mental states. Once we relieve ourselves of this problematic intuition, we can come to appreciate a functionalist approach and the better warranted, and nuanced, understanding it gives us of moral agency. Being progressive in our understanding of moral agency not only provides a more plausible account of the concept in general, but also allows for the new kinds of moral agency made possible by developments in AI. In what follows, I will first provide a positive account of this approach and then show how it can aid us in better understanding the moral role(s) that AI can and will come to play in our lives. Equally important is the way in which this account opens new possibilities for dealing with the moral questions that will arise from the actions of AI. In what follows, I will outline both the philosophical and methodological implications of this construal.

Moral Encounters of the Artificial Kind

“Philosophical benefits” here refers to what might be gained from this approach to moral agency apart from the practical benefits of the approach in how it allows us to resolve moral problems that may arise from the actions of artificial agents specifically. In other words, the focus here is on the theoretical benefits to be gained from understanding moral agency as described here, independently of such practical goals. With methodological benefits, on the other hand, the focus is on how this novel understanding of moral agency aids us in achieving some specific practical purposes when it comes to the possibility of artificial moral agents, such as the effective design of machines that can behave morally.

Philosophical Benefits

The philosophical benefits of the approach defended in this thesis should be clear: a functionalist understanding of moral agency does not suffer the same theoretical vulnerabilities as its primary alternative, the intentional account. The functional account does not fall prey to the same anthropocentric biases that were shown to be prevalent in the intentional account. Importantly, it avoids the problem of other minds, where it is assumed that moral agents have specific kinds of phenomenal mental states that enable them to make truly moral decisions. We saw that it is not even possible to prove that human beings—the quintessential moral agents—have such states. The functionalist account allows us to preserve the moral status of human beings, without needing to account for metaphysically problematic states. It also allows us to determine that given actions are sufficiently independent to be assessed morally, without making metaphysical presuppositions regarding free will and indeterminate actions and without needing to delve into the fraught metaphysical terrain surrounding these. It also allows us to account for moral agents without presupposing the capacity for moral patiency, which, as we saw in Chapter 1, is itself a problematic concept. It introduces a more coherent philosophical framework for dealing with moral assessment more generally, which may perhaps lead to revisions of longstanding assumptions in the literature on responsibility (such as the novel idea that identification as moral agent and evaluation as morally accountable can come apart). It becomes possible to hold moral agents morally responsible, irrespective of whether or not they are liable to punishment. Perhaps most importantly, it allows us to avoid the anthropocentric biases that permeate our thinking around morality and moral responsibility. It allows for the possibility of recognising other kinds of moral entities and also to conceive of new ways of dealing with these not predicated on the way in which we treat human moral entities. In what follows I will outline the methodological benefits of my approach which have not been explicitly dealt with thus far.

Methodological Benefits

As stated earlier, the methodological benefits here refer to those advantages that serve a specific practical purpose when it comes to dealing with moral questions. As my discussion thus far has shown, using the functionalist framework more generally, and its application to moral agency specifically, allows us to better understand and deal with emerging technologies and the novel moral encounters they will give rise to. Moreover, it gives us a coherent way of thinking through the implications of a capacity for moral agency for regulating the emergence of increasingly complex artificial systems. In order to make the above points salient it may be useful to briefly touch on two examples which have become increasingly popular in discussions of artificial moral agency: military drones and self-driving cars (see Sparrow, 2007; Wallach and Allen, 2009; Müller, 2014; Royakkers and van Est, 2015; Nyholm and Smids, 2016; Nyholm, 2017; Himmelreich, 2018; Wolkenstein, 2018).

Military Robots

Let us start with an examination of "Killer robots"—weapons systems capable of performing lethal military operations that were once the domain of human beings. An example of this type of system is the "Predator" drone, an unpiloted combat aerial vehicle capable of remotely performing military operations such as air-to-ground missile launches (Sparrow, 2007: 63; Royakkers and van Est, 2015: 560).[55] Talk of drone technology has recently become part of our common lexicon, former US president Barrack Obama's controversial use of drones to wage war in Iraq being a key trigger point for this debate. My point here, however, is not to take a stand on whether we are justified in using drones in war or whether their use has a net benefit. Rather, I wish to show how the effective characterisation of drones as AMAs can help us navigate potentially treacherous moral terrain when they are used. To do this, however, we must first understand the goals and moral reach of this technology. The express goal of drones is to operate in situations in which in which it is considered too dangerous for human pilots to operate.

Consider an example from the Kosovo war, in which NATO aircraft were forced to fly above 15,000 feet to avoid enemy fire. In this case, any bombs deployed would have had to be dropped from this height. In one instance, this tragically resulted in NATO aircraft mistaking a convoy of busses transporting refugees for Serbian tanks, and subsequently bombing them (Royakkers and van Est, 2015: 560). In a state of affairs such as this, an unpiloted drone would be preferred, as it could fly at a lower altitude, taking greater care in target selection and the subsequent use of lethal force. Such drones also reduce the need for human lives to be put in danger in military operations, creating a new class of 'cubicle warriors' (*ibid.*: 560). They may also be cheaper than human soldiers in the long run (a military drone does not need a pension scheme or a hospital plan), and outperform human soldiers in specific domains (human soldiers tend to require sleep to function optimally) (Müller, 2014: 4). There is therefore a strong *prima facie* case for driving the project to create ever more complex drone technology, and this is indeed reflected in the US government having funded research into the construction of autonomous

[55] Drones are just one example of many remotely operated vehicles (ROVs) that are being deployed in warzones. Another example is a remotely operated machine gun which makes use of the special weapons observation remote direct-action system (SWORDS), manufactured by Foster-Miller Inc. (Wallach and Allen, 2009: 20).

robots since the early 2000’s via the Defence Advanced Research Projects Agency (DARPA) (Wallach and Allen, 2009: 49).[56] One could even argue that it would be morally impermissible to place a solider in a life-threatening situation if that same task could be carried out by a military robot, in which case the use of such robots could be ethically defensible, and even encouraged.

Armed with this understanding of military drones more generally, we can consider a situation in which drones take lethal action and civilian casualties are incurred. This is not mere speculation: it is estimated that since 2004 between 769 and 1,725 civilians have been killed in drone strikes in Pakistan, Yemen, Somalia and Afghanistan (*Drone Warfare*, 2019). Moreover, drones are not infallible, and we can foresee a scenario in which a decision is made to launch a strike, but the target is misidentified (as in the Kosovo example above). “Standard” approaches to moral agency would dictate that when evaluating such civilian deaths, we should exclusively look towards the human beings that can be held morally responsible for these deaths, since holding the drone responsible would be conceptually inappropriate. This is generally because i) human operators are taken to be ultimately responsible for the actions of such drones, and ii) because moral responsibility is taken to entail punishment, and drones cannot be punished (e.g. Sparrow, 2007: 74).

This view may have made sense in the past, when it was clear that generally human judgment was the ultimate cause of the moral harm. The situation today, however, is much changed. We will soon be faced with technology capable of performing strikes quite independently of human control. Indeed, the explicit goal of these weapons is that they “would be able to select and attack targets without intervention by a human operator” (Müller, 2014: 1). In cases like this, responsibility-gaps emerge[57], with standard approaches lacking the conceptual tools to plug these gaps. Using my progressive approach outlined earlier, however, we can indeed speak of both accountable and potentially morally responsible drones without the requirement that they need to have internal mental states, free will, etc., or that they need to be punishable for their actions (cf. Bigman *et al.*, 2019). All that the claim of moral responsibility means in such instances is that upon review of the drone’s actions, it was determined that the drone *should have behaved otherwise*. This is not to deny the importance of holding human beings morally

[56] The US Department of Defence spends $5 billion per year on ‘unmanned systems’ (their sexist terminology, not mine), while DARPA has an annual budget of $3 billion (Müller, 2014: 4).

[57] Cases in which it is unclear whether a human being was responsible for a given moral action while an AA was causally efficacious in the production of the action involved. In situations such as this the assignment of moral responsibility is indeterminate.

responsible nor does it serve to deflect responsibility away from genuinely bad actors, as some authors worry (see Johnson and Miller, 2008). Instead, this approach allows us to acknowledge that with the emergence of these types of technologies, in some instances, there may be no human being who can coherently be held morally responsible for a moral wrong; the cause of, and the responsible party for, the moral harm might be the same: an AMA. Having a philosophically coherent framework of drone responsibility in place we will be in a better position to not only identify that a moral harm has occurred, but also, in some cases to take corrective actions and redress that harm. In the "no responsibility" case, because our responsibility ascriptions are indeterminate, our ability to respond appropriately and correct the harm will not be as effective.

Self-Driving Cars

A second interesting example of potential AMAs comes from the emergence of self-driving cars (see Nyholm and Smids, 2016; Himmelreich, 2018; Wolkenstein, 2018). There are already plenty of Advanced Driver Assistance Systems (ADAS) in place in automobiles, which support the driver, but do not automate the entire process of driving (Royakkers and van Est, 2015: 555). Examples of this are systems which alert the driver when they are drifting out of their lane or when they are unintentionally speeding. Moreover, these systems have progressed from simply alerting drivers to actually correcting the behaviour: by intervening and getting the car back in its lane or reducing the car's speed (*ibid.*: 555). The logical extension of this kind of technology is the fully independent self-driving car, capable of operating without any human intervention.

It is estimated that these vehicles will massively reduce road accidents, as around 90% of current accidents occur due to human error (*ibid.*: 557). The benefits of having self-driving cars can be easily appreciated: they do not get drunk, they will refrain from aggressive behaviour, they will always give their undivided attention to the road, and they will always stick to the speed limit. However, what is to be done in cases where these systems come to cause moral harms and no human being can be found responsible? The account of AMA presented in this thesis can help to solve the emergence of such responsibility-gaps in the specific case of self-driving cars. Recent incidents involving a Google test-drive vehicle and a Tesla model S suggest that the emergence of such a responsibility-gap is not far off and that they can be potentially pernicious.

In the Tesla case, a model S collided with a truck that its sensors had failed to detect, which resulted in the man in the Tesla being killed instantly. Tesla, while expressing condolences to the dead man's family, accepted no responsibility for the accident. Instead, they claimed that the customer is always in control and therefore responsible for what happens, even when the car is in "Autopilot"[58] mode (Nyholm, 2017: 2). Tesla did, however, promise to update their equipment in light of the incident.[59] Tesla's response to the situation, therefore, was to admit that their car was causally efficacious in the production of the moral harm that occurred (the car could therefore be usefully construed as an AMA, although this could also be questioned, as will be discussed below), but that the car itself was not morally responsible for the action, the driver was. Moreover, Tesla also claimed that they themselves were not morally (or legally) responsible for the incident. It should be clear that the responsibility gap allows for the possibility that no-one need take moral responsibility for the harm that results from the actions of an artificial system.

In the former case a Google test-drive vehicle crashed into the side of a bus (*ibid.*, 2017: 1). The Google car assumed the bus would yield to allow it (the Google car) to merge into an adjacent lane. The bus did no such thing, and the Google car attempted to merge anyway, resulting in the accident. Google claimed responsibility as the bus had no obligation to yield, and they admitted that the car *should have allowed* the bus to pass before initiating the movement. They cited the fact that driving is often a series of negotiations, in which drivers try to predict other drivers' movements; and in this case the Google car got it wrong (Ziegler, 2016) Thankfully no human being was harmed in this case, but it differs from the Tesla case in that Google claimed (at least partial) responsibility for what happened (*ibid.*, 2016).

These two cases highlight the current landscape of responses available to companies with respect to the contemporary state of self-driving car technology (if it even merits being called this in the case of Tesla specifically). In the Tesla case, it might be unreasonable to claim that the car was morally responsible for what happened, as Tesla explicitly requires that whenever Autopilot is enabled the user should keep their hands on the steering wheel, entrenching the idea that this technology is still very much in the realm of being a driving *assistance* system

[58] While Tesla refer to the system as Autopilot, this seems a strange choice, as they also claim that this system is, strictly speaking, a vehicle *assistance* feature (Lambert, 2019).

[59] What exactly "updating their equipment" might mean is worth noting. Tesla's Autopilot system makes use of a variety of sensors, and in the incident described it was claimed that the ride height of the trucks trailer confused the radar system on the model S. This resulted in the radar system mistakenly thinking that the trailer was in fact a road sign.

(Lambert, 2019). The Tesla vehicle may not be an AMA, as the independence of the car from human control is questionable, given the requirement that drivers keep their hands on the steering wheel. Nevertheless, if it becomes apparent that the human "driver" is entirely superfluous to the actions taken by the car, the claim that the car is not an agent (and hence not morally responsible) becomes questionable. If we were to find that the car is sufficiently autonomous to be deemed an agent, moral implications will follow both for it and for its designers. Google, for example, accepted that it was their fault that the accident occurred, and promised to update their systems in light of the incident.[60] In the Google case, the car really was operating independently of human control, and so it is reasonable to designate it as an AA (and therefore a potential AMA).

Nevertheless, based on the facts in both cases as they stand, there seems to be no good case for proceeding with an investigation into whether the cars *themselves* should be held morally responsible for the harms that occurred. In the Tesla case, drivers who make use of the Autopilot system are still (provisionally) responsible for any harms that occur when this system is operational. This is because Autopilot always requires them to keep their hands on the steering wheel and remain alert (they can presumably override decisions made by the car). However, we can imagine future scenarios where this might not be the case. An example of this could be an update to Autopilot which enables the car to change lanes independently, without requiring permission from the driver first, for example, or where it takes actions that cannot be overridden. In these situations, at the very least, more work is required in investigating the moral contours of our relationships to technologies that may operate independently of human beings, but in which human beings still exercise a significant degree of *supervisor* control (see Nyholm, 2017 for a discussion). The questions that such an investigation may raise are therefore related to the role of *human supervision* in self-driving cars, as Tesla's current software requires that the driver is in a constant state of vigilance, even when Autopilot is engaged. In the Google case there was no morally quantifiable action that occurred, and so their "claiming responsibility" simply amounts to the car being designated an AA rather than an AMA. Nevertheless, were moral harm to occur, various issues would need to be addressed: "To what extent was the action due to faulty design or programming?" "To what extent was the action due to the use made of the AMA by the user?" "To what extent was the action due to independent decisions made by the AMA?" "If the AMA acted wrongly, what

[60] Thankfully, no moral harm occurred, as this would have massively complicated matters for Google's PR team, as they would then have had to read this thesis.

should be done with it?" "Should it be decommissioned, discontinued, or reprogrammed, for example?" What we gain by using the framework provided in this thesis is that we are not forced to simply claim either "it's the user" or "it's the programmer/developer", which opens up conceptual space for self-driving cars to be causally efficacious *and* responsible in the production of moral harms. This allows for a greater number of possible responses to such moral harms, which would ideally help us reduce their occurrence in the future as well as to recompense for them.

Spooling Back the Reel

The argument in this thesis has proceeded in three distinct stages, each stage corresponding to a particular chapter. In my opening chapter on moral patiency, I argued against a prevailing assumption in the literature on moral agency: that a necessary condition for moral agency is that the entity in question is also a moral patient. I showed that the sense of sentience underpinning standard assumptions regarding moral patiency are vague and often illegitimately anthropocentric. If the underlying assumptions around moral patiency are problematic, then it makes no sense to use the resultant (problematic) conceptions of moral patiency to ground our ascriptions of moral agency, as is often done in work on the possibility of artificial agency. Toward the end of Chapter 1, I proposed an alternative framework for addressing questions of moral patiency, but nonetheless went on to claim that questions relating to moral *agency* are far more pressing in light of recent technological developments. It is with this in mind that my second chapter tackled the concept of moral agency, independently of considerations pertaining to moral patiency. To this end, I investigated the "standard account" of agency and exposed some problematic assumptions underlying this account. Using Johnson and Noorman's account of agency, I then proceeded to show how a coherent account of agency in general also has the conceptual room for AAs. I showed that the standard account precluded talk of AAs, while Johnson and Noorman's exposition created the conceptual room for such investigations. Notwithstanding this, however, their account of AA was also found wanting with regards to its use of original intentionality as a grounding principle in when it comes to determining whether an agent is a *moral* agent. Accounts that rely of original intentionality presuppose that there is something inherently "mysterious" about decision making in human beings that in principle renders it closed to artificial replication. This is a metaphysically contested position, especially as it cannot even be coherently shown that *human* agents possess such intentionality. In my final chapter, therefore, I proposed that we adopt a *functional* as opposed to an *intentional* framework in our philosophising of moral agency in general. This framework is to be favoured both in and of itself (as it provides a more plausible account of the status of human beings as moral agents) and due to the fact that it allows for the possibility of AMAs, which is a pressing moral topic. Within this functionalist framework I argued that it may indeed be possible for AMAs to be morally responsible for certain actions that they perform, in much the same way that human beings can be said to be responsible. This understanding of responsibility involves distinguishing between causal efficaciousness, moral responsibility and "punishability", and claims that each of these can come apart in our practices of moral assessment. This allows us to hold AMAs morally responsible for their actions and to act against them accordingly, despite their not being able to subjectively "appreciate" their punishment as a human moral agent might.

Avenues for Future Research

In this last section, I would like to make salient aspects of this thesis that I think are of potential philosophical significance and could therefore be further developed by other scholars. The first avenue for future research would be providing non-anthropocentric accounts of moral patiency. Good attempts at this are provided by Sparrow (2004), Wareham (2011), Coeckelbergh (2014), and Danaher (2017b). A key issue faced by any account of moral patiency is how such frameworks ought to deal with cases where the AI in question does not necessarily have humanoid features but nonetheless exhibits certain external cues that lead us to believe that it should be accorded some kind of moral concern. Moreover, exactly what "machine consciousness" entails need not necessarily be anything like human consciousness, making the solution to the question of machine moral patiency even more complicated.

Secondly, there are implications for the use of "entertainment" robots, a key exemplar of which would be the sex robot (Royakkers and van Est, 2015). The use and distribution of such robots raise questions concerning the role of consent and ownership, and how (if it all) these concepts refer in this case. If we concede that such robots are AMAs, can they give meaningful consent? Moreover, can we legitimately speak of acts such as "robotic rape", and punish those performing such acts (see Danaher, 2017)? More work needs to be done at both the philosophical and regulatory levels to unpack solutions to these and other pertinent questions.

Lastly, I believe considerable work needs to be done on outlining and guarding against the widespread hubris on the part of humanity with respect to our artificial companions. There is a growing sense of unease among scholars working in AI related fields about the potential of a "superintelligence" emerging (Bostrom, 2014). The worry is that such an intelligence would very quickly surpass human levels of intelligence and may come to pose an existential risk to the future of humanity, in that it might not necessarily have anthropomorphic goals (see Bostrom, 2014: 116 for an argument to this effect).[61] Moreover, embedded in this line of reasoning is the fact that current and future technologies will be able to outperform human beings at certain (narrowly specified, for the moment) tasks, and thus could reach levels of perfection not even imaginable by flesh and blood moral agents. A perhaps counterintuitive implication of this view might be that human beings become archaic vestiges of what the ideal moral agent may have been but are unable to keep pace with the moral development of AMAs. In light of this, we might actually be better off creating a species of robots not incumbered with our evolutionary baggage (see Hintze *et al.*, 2015; Duffy, 2018) who would constitute a massive moral improvement compared to us (Dietrich, 2001: 5). We should, however, be sensitive to the fact that benevolence and intelligence need not go hand in hand[62]: just because an AMA is superintelligent does not necessarily imply that it will adopt a view of morality that is congruous with human survival (Bostrom, 2014).

The interaction of the biological and the artificial is a feature of human existence that seems to be here to stay. The ubiquity of AI both in our day-to-day interactions and in its role in the maintenance of social cohesion more generally are at a point where attempts to spool back the reel of progress are impossible. What we need to do, in light of this fact, is to be sensitive to both the philosophical and methodological aspects of our relationship to technology. Philosophically, we must be open to the idea that our traditional understanding of various concepts may come under stress and need to be revised or replaced. Methodologically, we must be careful in our design and implementation of AMAs, as we might be creating artificial gods capable of eliminating our species.

[61] See Danaher (2015) for an exposition and critique of Bostrom's argument in this regard.

[62] Also known as the orthogonality thesis (see Bostrom, 2012).

Reference List

Alonso, E. (2014) ‘Actions and Agents’, in Frankish, K. and Ramsey, W. M. (eds) *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, pp. 232–246.

Alterman, R. (2000) ‘Rethinking autonomy’, *Minds and Machines*, 10(1), pp. 15–30. doi: 10.1023/A:1008351215377.

Arnold, T. and Scheutz, M. (2016) ‘Against the moral Turing test: accountable design and the moral reasoning of autonomous systems’, *Ethics and Information Technology*. Springer Netherlands, 18(2), pp. 103–115. doi: 10.1007/s10676-016-9389-x.

Bigman, Y. E. *et al.* (2019) ‘Holding Robots Responsible: The Elements of Machine Morality’, *Trends in Cognitive Sciences*. Elsevier Ltd, 23(5), pp. 365–368. doi: 10.1016/j.tics.2019.02.008.

Bortolotti, L. (2015) *Irrationality*. Cambridge: Polity Press. doi: 10.1002/ejoc.201200111.

Bostrom, N. (2012) ‘The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents’, *Minds and Machines*, 22(2), pp. 71–85.

Bostrom, N. (2014) *Superintelligence*. Oxford: Oxford University Press.

Bostrom, N. and Yudkowsky, E. (2011) ‘The Ethics of Artificial Intelligence’, in *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press.

Boyles, R. J. M. (2017) ‘Philosophical signposts for artificial moral agent frameworks’, *Suri*, 6(2), pp. 92–109.

Brown, C. (2015) ‘Fish intelligence, sentience and ethics’, *Animal Cognition*, 18(1), pp. 1–17. doi: 10.1007/s10071-014-0761-0.

Bryson, J. J. (2018) ‘Patiency is not a virtue: the design of intelligent systems and systems of ethics’, *Ethics and Information Technology*. Springer Netherlands, 20(1), pp. 15–26. doi: 10.1007/s10676-018-9448-6.

Chalmers, D. J. (1996) *The Conscious Mind*. Oxford: Oxford University Press.

Champagne, M. and Tonkens, R. (2013) ‘Bridging the Responsibility Gap’, *Philosophy and Technology*, 28(1), pp. 125–137.

Christman, J. (2018) *Autonomy in Moral and Political Philosophy*. Spring 201. Edited by E. N. Zalta.

The Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/.

Coeckelbergh, M. (2010) 'Robot rights? Towards a social-relational justification of moral consideration', *Ethics and Information Technology*, 12(3), pp. 209–221. doi: 10.1007/s10676-010-9235-5.

Coeckelbergh, M. (2014) 'The Moral Standing of Machines : Towards a Relational and Non-Cartesian Moral Hermeneutics', pp. 61–77. doi: 10.1007/s13347-013-0133-8.

Cohen, M. A. and Dennett, D. C. (2011) 'Consciousness cannot be separated from function', *Trends in Cognitive Sciences*, 15(8), pp. 358–364. doi: 10.1016/j.tics.2011.06.008.

Crane, T. (2011) *Intentionality*. Routledge Encyclopedia of Philosophy, Taylor and Francis. doi: doi:10.4324/9780415249126-V019-2.

Danaher, J. (2015) 'Why AI Doomsayers are Like Sceptical Theists and Why it Matters', *Minds and Machines*, 25(3), pp. 231–246. doi: 10.1007/s11023-015-9365-y.

Danaher, J. (2017a) 'Robotic Rape and Robotic Child Sexual Abuse: Should They be Criminalised?', *Criminal Law and Philosophy*, 11(1), pp. 71–95. doi: 10.1007/s11572-014-9362-x.

Danaher, J. (2017b) 'The rise of the robots and the crisis of moral patiency', *AI and Society*. Springer London, 0(0), pp. 1–8. doi: 10.1007/s00146-017-0773-9.

Davidson, D. (1980) *Essays on Actions and Events*. Oxford: Oxford University Press.

Davies, P. S. (2007) *Distributed Cognition and the Will: Individual Volition and Social Context*. Edited by D. Ross et al. London, England: The MIT Press. Available at: http://books.google.co.uk/books?id=sX6g_vw4yKcC.

Dennett, D. C. (2009) 'Intentional Systems Theory', *The Oxford Handbook of Philosophy of Mind*, pp. 1–22. doi: 10.1093/oxfordhb/9780199262618.003.0020.

Dennett, D. C. (1978) 'Why you can't make a computer that feels pain', *Synthese*, 38(3), pp. 415–456. doi: 10.1007/BF00486638.

Dennett, D. C. (1989) *The Intentional Stance*. Cambridge, Massachusetts: MIT Press. doi: 10.1017/S0140525X00058611.

Dennett, D. C. (1996) *Kinds of Minds: Toward an Understanding of Consciousness*. New York: Basic Books.

Dennett, D. C. (2003) *Freedom Evolves*. New York: Viking.

Dietrich, E. (2001) ‘Homo sapiens 2.0: Why we should build the better robots of our nature’, *Journal of Experimental and Theoretical Artificial Intelligence*, 13(4), pp. 323–328. doi: 10.1080/09528130110100289.

Doecke, S. D. *et al.* (2012) ‘The potential of autonomous emergency braking systems to mitigate passenger vehicle crashes’, *Australasian Road Safety Research, Policing and Education Conference*. Wellington, New Zealand.

Drone Warfare (2019) *The Bureau of Investigative Journalism*. Available at: https://www.thebureauinvestigates.com/projects/drone-war (Accessed: 11 July 2019).

Duffy, B. (2018) *The Perils of Perception*. London: Atlantic Books.

Eshleman, A. (2016) *Moral Responsibility*. Winter 201, *Stanford Encyclopedia of Philosophy*. Winter 201. Edited by E. N. Zalta. Available at: https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/.

Etzioni, A. and Etzioni, O. (2016) ‘AI assisted ethics’, *Ethics and Information Technology*, 18(2). doi: 10.1007/s10676-016-9400-6.

Floridi, L. (1999) ‘Information ethics : On the philosophical foundation of computer ethics’, *Ethics and Information Technology*, 1, pp. 37–56. doi: DOI\t10.1023/A:1010018611096.

Floridi, L. and Sanders, J. W. (2004) ‘On the Morality of Artificial Agents’, *Minds and Machine*, 14, pp. 349–379. doi: 10.2139/ssrn.1124296.

Fossa, F. (2018) ‘Artificial moral agents: moral mentors or sensible tools?’, *Ethics and Information Technology*. Springer Netherlands, 20(2), pp. 115–126. doi: 10.1007/s10676-018-9451-y.

Frankish, K. and Ramsey, M., W. (2014) ‘Introduction’, in Frankish, K. and Ramsey, M., W. (eds) *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, pp. 1–11.

Gerdes, A. and Øhrstrøm, P. (2015) ‘Issues in robot ethics seen through the lens of a moral turing test’, *Journal of Information, Communication and Ethics in Society*, 13(2), pp. 98–109. doi: 10.1108/JICES-09-2014-0038.

Grodzinsky, F. S., Miller, K. W. and Wolf, M. J. (2008) ‘The ethics of designing artificial agents’, *Ethics and Information Technology*, 10(2–3), pp. 115–121. doi: 10.1007/s10676-008-9163-9.

Gunkel, D. J. (2012) *The Machine Question*. London: MIT Press.

Gunkel, D. J. (2014) ‘A vindication of the rights of machines’, *Philosophy and Technology*, 27(1), pp. 113–132. doi: 10.1007/s13347-013-0121-z.

Gunkel, D. J. (2017) ‘Mind the gap: responsible robotics and the problem of responsibility’, *Ethics and Information Technology*. doi: 10.1007/s10676-017-9428-2.

Harding, L. (2006) ‘At least 23 die as driverless train crashes into maintenance truck’, *The Guardian*, 23 September. Available at: https://www.theguardian.com/world/2006/sep/23/germany.topstories3.

Himma, K. E. (2009) ‘Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?’, *Ethics and Information Technology*, 11(1), pp. 19–29. doi: 10.1007/s10676-008-9167-5.

Himmelreich, J. (2018) ‘Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations’, *Ethical Theory and Moral Practice*. Ethical Theory and Moral Practice, 21(3), pp. 669–684. doi: 10.1007/s10677-018-9896-4.

Hintze, A. *et al.* (2015) ‘Risk sensitivity as an evolutionary adaptation’, *Scientific Reports*, 5, pp. 1–7. doi: 10.1038/srep08242.

van den Hoven, J. and Weckert, J. (2008) *Information technology and moral philosophy*, *Information Technology and Moral Philosophy*. doi: 10.1017/CBO9780511498725.

Illies, C. and Meijers, A. (2009) ‘Artefacts Without Agency’, *The Monist*, 92(3), pp. 420–440. doi: 10.2174/138920312803582960.

Jacob, P. (2019) *Intentionality*. Spring 201. Edited by E. N. Zalta. Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/archives/spr2019/entries/intentionality/.

Johansson, L. (2010) ‘The Functional Morality of Robots’, *International Journal of Technoethics*, 1(4), pp. 65–73.

Johansson, L. (2011) ‘Robots and moral agency’, *Journal of Technoethics*. Available at: http://www.diva-portal.org/smash/get/diva2:410512/FULLTEXT02.pdf.

Johnson, D. G. (2006) ‘Computer systems: Moral entities but not moral agents’, *Ethics and Information Technology*, 8, pp. 195–204. doi: 10.1017/CBO9780511978036.012.

Johnson, D. G. (2015) ‘Technology with No Human Responsibility?’, *Journal of Business Ethics*, 127(4), pp. 707–715. doi: 10.1007/s.

Johnson, D. G. and Miller, K. W. (2008) 'Un-making artificial moral agents', *Ethics and Information Technology*, 10(2–3), pp. 123–133. doi: 10.1007/s10676-008-9174-6.

Johnson, D. G. and Noorman, M. (2014) 'Artefactual Agency and Artefactual Moral Agency', in Kroes, P. and Verbeek, P.-P. (eds) *The Moral Status of Technical Artefacts*. New York: Springer, pp. 143–158. doi: 10.1007/978-94-007-7914-3.

Johnson, D. G. and Powers, T. M. (2005) 'Computer systems and responsibility: A normative look at technological complexity', *Ethics and Information Technology*, 7(2), pp. 99–107. doi: 10.1007/s10676-005-4585-0.

Johnson, D. G. and Powers, T. M. (2008) 'Computers as surrogate agents', in Van Den Hoven, J. and Weckert, J. (eds) *Information Technology and Moral Philosophy*. Cambridge: Cambridge University Press.

Kahneman, D. (2000) 'Experienced Utility and Objective Happiness: A Moment-Based Approach', in Tversky, A. (ed.) *Choices, Values and Frames*. New York: Cambridge University Press.

Lambert, F. (2019) *Tesla Autopilot, electrek*. Available at: https://electrek.co/guides/tesla-autopilot/ (Accessed: 10 August 2019).

Leopold, A. (1948) 'A Land Ethic', in *A sand county almanac with essays on conservation from Round River*. New York: Oxford University Press.

Levin, J. (2018) *Functionalism*. Fall 2018. Edited by E. N. Zalta. The Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/archives/fall2018/entries/functionalism/.

List, C. and Pettit, P. (2011) *Group Agency: The Possibility, Design, and Status of Corporate Agents*. New York: Oxford University Press.

Litton, P. (2008) 'Responsibility status of the psychopath: On moral reasoning and rational self-governance', *Rutgers Law Journal*, 39(349), pp. 350–392.

Müller, V. C. (2014) 'Autonomous killer robots are probably good news', *Frontiers in Artificial Intelligence and Applications*, 273, pp. 297–305. doi: 10.3233/978-1-61499-480-0-297.

Naess, A. (1973) 'The shallow and the deep long-range ecology movements', *Inquiry*, (16), pp. 95–100.

Noorman, M. and Johnson, D. G. (2014) 'Negotiating autonomy and responsibility in military robots', *Ethics and Information Technology*, 16(1). doi: 10.1007/s10676-013-9335-0.

Nyholm, S. (2017) ‘Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci’, *Science and Engineering Ethics*. Springer Netherlands, pp. 1–19. doi: 10.1007/s11948-017-9943-x.

Nyholm, S. and Smids, J. (2016) ‘The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?’, *Ethical Theory and Moral Practice*. Ethical Theory and Moral Practice, 19(5), pp. 1275–1289. doi: 10.1007/s10677-016-9745-2.

O’Connor, T. and Franklin, C. (2019) *Free Will*. Summer 201. Edited by Edward N. Zalta. The Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/freewill/.

Powers, T. M. (2013) ‘On the Moral Agency of Computers’, *Topoi*, 32(2), pp. 227–236. doi: 10.1007/s11245-012-9149-4.

Ritchie, J. (2008) *Understanding Naturalism*. Stocksfield: Acumen.

Roskies, A. L. and Malle, B. F. (2013) ‘A Strawsonian look at desert’, *Philosophical Explorations*, 16(2), pp. 133–152. doi: 10.1080/13869795.2013.787439.

Royakkers, L. and van Est, R. (2015) ‘A Literature Review on New Robotics: Automation from Love to War’, *International Journal of Social Robotics*. Springer Netherlands, 7(5), pp. 549–570. doi: 10.1007/s12369-015-0295-x.

Russell, S. and Norvig, P. (2010) *Artificial Intelligence: A Modern Approach*. 3rd edn. Edited by S. Russell and P. Norvig. Boston: Prentice Hall.

Sapolsky, R. (2018) *Behave*. London: Vintage.

Schlosser, M. (2015) *Agency*. Fall 2015. Edited by E. N. Zalta. Available at: https://plato.stanford.edu/archives/fall2015/entries/agency/.

Singer, P. (1975) *Animal liberation: a new ethics for our treatment of animals*. New York: New York Review of Books.

Singer, P. (2011) *The Expanding Circle: ethics, evolution and moral progress*. New Jersey: Princetown University Press.

Sparrow, R. (2004) ‘The Turing Triage Test’, *Ethics and Information Technology*, 6(4), pp. 203–213. doi: 10.1007/s10676-004-6491-2.

Sparrow, R. (2007) ‘Killer Robots’, *Journal of Applied Philosophy*, 24(1), pp. 62–78. doi: 10.1111/j.1468-5930.2007.00346.x.

Stich, S. P. (1981) ‘Dennett on Intentional Systems’, *Functionalism and the Philosophy of Mind*, 12(1), pp. 39–62.

Strawson, P. (1993) ‘Freedom and Resentment’, in Fischer, J. M. and Ravizza, M. (eds) *Perspectives on Moral Responsibility*. London: Cornell University Press, pp. 45–66. Available at: http://books.google.com/books?id=0ncN3TuDQ7cC&pg=PA34&dq=intitle:Perspectives+on+Moral+Responsibility&hl=&cd=1&source=gbs_api.

Sullins, J. P. (2011) ‘When Is a Robot a Moral Agent?’, *Machine Ethics*, 6(2001), pp. 151–161. doi: 10.1017/CBO9780511978036.021.

Torrance, S. (2008) ‘Ethics and consciousness in artificial agents’, *AI and Society*, 22(4), pp. 495–521. doi: 10.1007/s00146-007-0091-8.

Torrance, S. (2014) ‘Artificial consciousness and artificial ethics: Between realism and social relationism’, *Philosophy and Technology*, 27(1), pp. 9–29. doi: 10.1007/s13347-013-0136-5.

Vargas, M. R. (2015) ‘Moral Responsibility & Desert: Social, Scaffolded, & Revisionist’, *The University of San Francisco School of Law*, 12, pp. 1–38.

Van de Voort, M., Pieters, W. and Consoli, L. (2015) ‘Refining the ethics of computer-made decisions: a classification of moral mediation by ubiquitous machines’, *Ethics and Information Technology*, 17(1), pp. 41–56. doi: 10.1007/s10676-015-9360-2.

Wallach, W. and Allen, C. (2009) *Moral Machines*. New York: Oxford University Press.

Wareham, C. (2011) ‘On the Moral Equality of Artificial Agents’, *International Journal of Technoethics*, 2(1), pp. 35–42. doi: 10.4018/jte.2011010103.

Wegner, D. M. (2002) *Illusion of Conscious Will*. London, England: Bradford Books. doi: 10.1073/pnas.0703993104.

Wolkenstein, A. (2018) ‘What has the Trolley Dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars’, *Ethics and Information Technology*. Springer Netherlands, 20(3), pp. 163–173. doi: 10.1007/s10676-018-9456-6.

Yu, P. and Fuller, G. (1986) ‘A Critique of Dennett’, *Synthese*, 66(3), pp. 453–476.

Ziegler, C. (2016) *A Google self-driving car caused a crash for the first time, The Verge*. Available at:

https://www.theverge.com/2016/2/29/11134344/google-self-driving-car-crash-report (Accessed: 10 July 2019).