CAN MORALITY BE COMPUTED? AN EXPLORATION OF WHETHER MACHINES CAN BE MORAL AGENTS

Gillian Eileen Cross

(Person Number: 8805370J)

A Research Report submitted to the Faculty of Humanities, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Arts, Applied Ethics for Professionals

31 May, 2015, Johannesburg

ABSTRACT

Technology is an integral part of our daily lives and continues to advance rapidly, impacting our physical and social worlds. We increasingly rely on advanced machines to act in more autonomous and sophisticated ways. What would happen if artificial forms of intelligence developed to the point where machines behaved more like us and we started treating them as people? Ethics should anticipate and account for such possibilities so that science does not move faster than our moral understanding.

My thesis states that when we are able to feel gratitude or resentment towards the actions of artificially intelligent machines we can be said to see them as morally responsible agents. I argue that standard ethics frames morality developmentally – only when we reach adulthood are we deemed able to enter into the type of relationships where we can hold one another morally responsible for our actions. I apply a more abstract notion of moral development to future versions of technology and couple this with a definition of morality as a relational or social construct. This allows me to argue that machines could develop to a point in the future where we react to them morally as we would to humans. Questions on whether we ought to react in this way are muted as relationally we quite simply would be unable to feel otherwise. Objections from definitions of moral agency based on innate qualities, specifically those associated with the concept of intelligence, are dispelled in favour of a relational definition.

DECLARATION

I declare that this research report is my own unaided work. It is submitted for the degree of Master of Arts, Applied Ethics for Professionals, in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any other degree or examination in any other university.

_____

GILLIAN EILEEN CROSS

31 day of May, 2015

ACKNOWLEDGMENTS

I would like to express deep appreciation to my supervisor, Dr Robert Kowalenko, for his rigour, depth of commentary, patience and interest in progressing my learning. My sincere thanks to the ongoing support received from Dr Brian Penrose, a Programme Director, par excellence. To faculty, the passion for what you do and your deep levels of expertise in your fields is evident and inspirational and I valued each lecture and interchange. Finally, to my fellow classmates, thanks for fine debates and active engagement, you made the experience that bit richer and rewarding.

TABLE OF CONTENTS

Prologue

Popular culture is peppered with stories of man's quest to create artificial forms of intelligence. From Mary Shelley's novel *Frankenstein* published in 1818 about an eccentric scientist who creates a grotesque creature in an unorthodox experiment; to the Technicolor *The Wizard of Oz (*1939), which sees the heroine Dorothy befriend a number of characters in a strange land, who want more human-like qualities such as Tin Man who desires a heart. Contemporary science-fiction (sci-fi) provides a fertile platform for exploring the ramifications of worlds populated by increasingly sophisticated forms of non-human intelligence. Many movie scripts explored the very nature of being, and in some cases, touched upon the ethical considerations that these ontic journeys raise.

Growing up in the 1970's and 80's my imagination was captured by these visual glimpses of future universes inhabited by humans, robots and other forms of Artificial Intelligence (AI), many of which behaved and even looked a bit like us. Stanley Kubrick's classic film, *2001: A Space Odyssey* (1968), reached cult status via its main character a machine called HAL 9000. HAL the Heuristically programmed ALgorithmic computer was a sentient computer that controlled the systems of the Discovery One spacecraft and interacted with the ship's astronaut crew. Being a computer he was visually represented as a red camera eye on instrumentation panels on the ship. Despite having no distinct, physical form, audiences were captivated by his human-like behaviour as he learned, reflected upon his actions and conversed with the crew. HAL spawned a host of robot-actors featured in films that followed. The likeable C-3PO an android programmed to interact with people on etiquette and protocol, entertained audiences with his camp, British accent and dry-wit in George Lucas's *Star Wars* (1977). A dystopia in Ridley Scott's *Blade Runner* (1982) featured a world where bioengineered *replicants* looked and behaved like us and could only be identified apart from humans by an expert conducting a series of extensive tests.

Admittedly my selection of topic stems from the fascination sci-fi held for me in my youth. My delight was evident upon discovering a branch of ethics that explored whether machines[1] as artificial forms of intelligence could merit moral consideration.

[1] The terms machine intelligence and artificial intelligence are used interchangeably.

1. Introduction

A recent issue of *The Economist* reads like a sci-fi novel with its special report entitled 'Rise of the Robots' (Anon, 2014)[2]. The collection of articles explores a range of topics from concerns relating to the autonomy of lethal military robots; to the advent of self-driving cars which open up a myriad of legal, regulatory and even ethical issues. Most relevant is mention of Paro, a robot used to provide a source of calm and reassurance to people suffering from Alzheimer's in nursing homes in Japan, parts of Europe and America. Denmark's council of ethics is reviewing the likes of Paro and whether it is acceptable for robots to be designed to fool mentally impaired people into thinking they have feelings. Increasing sophistication and autonomy of such machines give further poignancy to Dennett's question 'when HAL kills, who's to blame' (Dennett, 1997, p353)? Consider the ethical implications if Sony's robotic dog AIBO (released in 1999); or Google's self-driving car (prototyped 2013) harmed or even killed people by accident. Such rapid advancements in biotechnology, nanotechnology and information technology, particularly in the last 10 years, have increased the relevance of moral issues surrounding the evolving human-technology relationship. As humans become ever more reliant on technology we increasingly live in what Wiener, the founding father of computer ethics, called the 'automatic age' (Wiener in Bynum, 2006, p167). In this age, Wiener believes machines are integrated not only into the physical environment, but also into the social fabric of society.

Some contemporary ethicists refer to the information age as the 'fourth revolution' where 'today we are slowly accepting the idea that we are not standalone and unique entities, but rather informationally embodied organisms (inforgs), mutually connected and embedded in an informational environment (infosphere), which we share both with natural and artificial agents, similar to us in many respects' (Floridi, 2010, p10). Others view various forms of software such as social media, the internet and even decision support software as forms of 'cognitive enhancement' for human beings (Bostrom and Sandberg, 2009, p320). Cyborg technology offers 'even tighter integration of information technology with human minds' (Allen in Floridi, 2010, p226). For example, brain-implanted electrodes have been used to control prosthetic limbs in humans (Lebedev and Nicolelis in Allen, 2010, p226). What was

[2] Where no page numbers are referenced in-text, it is due to the fact that the text is an online publication without page numbers.

thought of as sci-fi is fast becoming a reality.

What happens 'when science moves faster than moral understanding' (Sandel, 2004, p1)? To Sandel's point, ethics cannot lag behind such developments and needs to account for reality. The reality is that we are becoming more dependent on machines, which in turn are acting more autonomously. What would happen at some point in the future, if robots developed to the point where they behaved more like we did and we started treating them as people? Ethics should anticipate and account for such possibilities.

In exploring my report question, I use the terms morality, moral responsibility and moral agency. Although different concepts, they are deeply linked. Morality, our sense of ethical right and wrong, is enforced in the concept of moral responsibility, our obligation and duty to treat each other in a morally responsible manner, generally described as exhibiting a sense of goodwill towards each other. Moral agency, our ability to act with moral consequence, is the manifestation in action of our responsibility to be moral. You can be morally responsible without being a moral agent. For example, if I am immobilised in a wheelchair and unable to physically intervene to save a drowning child, I am in fact morally responsible to want to save the child, although I do not have the agency to do so. Conversely, the drowning child can be saved by an agent who is not deemed to be a moral agent. For example, a dog can save the child. The dog has agency as it saved the child, and in so doing, it has led to a moral result. However it is not a moral agent, as we do not hold dogs morally responsible because we believe they lack the ability to appreciate the moral consequences of their actions. There is a conceptual possibility that a mind could have morality, but no moral agency, simply because it is not capable of being an agent, for example, a brain-in-a-vat.

I argue for a developmental view of machine moral agency as follows:

Sub-Argument:

> (P1) Universally, infants and small children are not socially developed enough to be held morally responsible for their actions.
>
> (P2) When we reach adulthood, we are deemed socially mature and able to enter

into the type of inter-personal relationships with each other, which include holding one another morally responsible via moral praise or blame.

(P3) Morally responsible adults, capable of moral action, are considered moral agents.

Moral responsibility and moral agency are developmental constructs.

Main-argument:

(P1) Moral agency is a developmental construct.

(P2) Artificially intelligent agents may reach a point of development in the future, where we exhibit feelings towards their actions that approximate praise (gratitude) or blame (resentment).

When we can be said to feel gratitude or resentment towards the actions of artificially intelligent agents, we can be said to deem them morally responsible and thus moral agents.

An outline of my report follows. Section 1.1 motivates the relevance of the topic reviewing commentary on the moral implications of the evolving relationship between man and machine. Section 1.2 reviews the question of machine agency and section 1.3 of machine moral agency. This review highlights a common feature - most theorists apply standard definitions of intelligence to machines using the degree of fit of the *definiens* to the *definiendum* to argue either for or against machine moral agency. The discussion on intelligence illustrates borderline cases in human moral agency, for example, teenagers which implies that moral agency is a developmental construct, a point pivotal to my argument. I start building a working definition of machine autonomy in section 1.2 as autonomy relates to the capacity for moral agency and add to it in section 1.3. Section 1 ends with the claim that human moral agency has very little to do with intelligence, which allows me to review alternate theories of moral agency for machines.

Section 2 and section 3 review theories which accord moral agency on aspects other than

intelligence in the hope that something suitable may surface to help me extend moral agency to machines. Section 2 looks at behavioural claims that if and only if we see machines looking and/or behaving like us, do we accord them moral agency. This falls short of being normatively adequate as the rightness or wrongness thereof is not covered. My argument finally finds support in section 3 with a theory by Strawson that moral agency is based upon the type of relationships we are able to enter into. This definition is attractive as it avoids the pitfalls of associating moral agency with innate qualities linked to notions personhood, in favour of external, social constructs. It allows me to argue that machines could develop to a point where we react to them as we do to each other, including on moral matters. Section 4 references the enhancement debate, specifically moral agency as a threshold concept, which aligns to my claim that it is a developmental construct. Section 5 concludes by asking whether our definition of machine autonomy and machine moral agency has been augmented post-Strawson and reviews the implications thereof for my thesis.

1.1 The evolving man-machine interface

A phenomenological approach towards information ethics has in common the belief in a 'co-constitutive relationship' between humans and technology (Introna, 2011). This approach paves the way for my paper as it starts us thinking differently about the relationship between humans and machines, particularly sophisticated forms of artificial intelligence. Broadly, such theorists believe that technology and society continually draw on each other for their on-going sense and meaning. Intuitively this idea has appeal. Technology does shape the way we live our lives and interact with each other and is something most of us use daily. Technology is not an artefact alone. It is also the technological attitude or disposition that made the artefact appear as meaningful and necessary to us in the first place. On this view technology is not just something external to us or 'out there', it is also immediately something 'in here', at the very source of our humanity (Stiegler in Introna, 2011). This means that when we design new AI we are also designing the type of humans we are or want to become. Stiegler takes a more radical approach that 'the human and the technical are co-original' (Stiegler in Introna, 2011). He believes that human evolution is intrinsically linked to the use of technology. 'The freeing of hands when in the upright position' allowed us to use tools and *exteriorize* technology, which changed the way we were able to master our world and in turn the way we viewed our world and our place in it. This initiated a

process of *interiorization* which he calls *epiphylogenesis* – the 'extra-genetic, co-evolution' of the human and the technical (Stiegler in Introna 2011). Moreover, he believes that without this technicity, society and culture would not be possible as humans would be unable to exist in time. Simply put, there would be no way to mark the different eras of humanity. Heidegger believes that we view the world technologically. We have what he calls a 'technological mood'. Such 'enframing' means that we already understand technology and that we are pre-disposed towards finding technical solutions to problems, including ethical ones (Heidegger in Introna, 2011). Furthermore, some believe that 'moral responsibility cannot be properly understood without recognising the active role of technology in shaping human action' (Jonas 1984; Verbeek 2006; Johnson and Powers 2005; Waelbers 2009; all in Noorman, 2014).

What era would follow from our current, technological era? Information ethicist Floridi would say that we are moving into the era of 'A-LIVE' or Artificially Live where our understanding of the world is informational (Floridi, 2010, p9). In essence Floridi's theory is ontological, his state of being worthy of moral consideration is an information object, broadly anything that comprises packets of information. This is everything in the universe including thoughts and experiences. Thus the way we make sense of reality in the A-LIVE era shifts from the material to the informational. Reality is de-physicalized and premium is placed on usage and interactivity. Such an era would certainly re-define the way we even see a distinction between man and machine. The question is no longer whether it is AI or human but whether it is interactive and usable. Definitions of moral agency will shift to anything that helps facilitate the flow of information (Floridi, 2010, p9). It seems the locus of moral consideration is entities that process information, rather than whether they are sentient or not. Phenomenologists suggest a powerful relationship between man and machine, which makes it difficult to separate out their actions, including the consideration of moral matters. Traditional ethical theories are unable to account for this phenomenon adequately and a revised approach is needed to frame how our relationship is evolving with technology.

1.2 Can machines be agents?

Traditionally we ascribe responsibility for moral acts to adult, human agents performing

actions that have moral consequences. We hold each other morally responsible when our voluntary actions have some form of morally significant outcome that is either worthy of blame or praise. Obviously, not everything in the world is morally relevant. Some actions, such as a man tying his shoelaces, are morally irrelevant. Similarly a machine installing a part on a car assembly line is neither a morally relevant action nor one which leads to a morally relevant outcome. Some actions, even if performed by those we deem to be moral agents, have outcomes which are morally irrelevant. Conversely, other actions performed by agents who are not deemed to be morally responsible, can nevertheless have moral outcomes or consequences. For example, a car assembly robot could malfunction and kill a factory worker, or a lion could eat a child. When such actions have morally relevant outcomes we are tempted to treat the entity that performed the action (i.e., the assembly robot and lion) as moral agents. Neither a car assembly robot, nor a lion is a moral agent as both lack any appreciation of the moral consequences of their actions. The act itself, even if it has a moral consequence, needs to be considered separately from the status of the agent performing it. Thus, something more than an act with a moral consequence is needed for agency. Our very idea of morality is intertwined with our ability to act but also with our ability to appreciate the moral consequences of our actions. We hold agents accountable for their acts because we see them as having control over their actions. The ability of agents to act of their own accord, to have free-will, is a key feature that makes them morally responsible in traditional ethics. When free-will is compromised, for example, when agents are coerced into performing acts, we mitigate the agent's moral responsibility for the action. To illustrate, killing your boss in cold blood would rightly receive a severe moral and legal sanction of murder; but being instructed to shoot your boss by an irate college, who is holding a gun to your head, could morally mitigate such judgement. It is worth reviewing this example from a legal standpoint, as the law separates out action from intent. Most crimes require the presence of both *actus reus* (the guilty act) and *mens rea* (the guilty mind). Using our boss example, shooting the boss is a guilty act, however, if coerced into performing it by a colleague holding a gun to your head, your mental state was not one of an intention to kill, i.e., you did not have a guilty mind. As a result you did not use free-will in performing the act. Conversely, if you shot your boss in cold blood, you have performed a guilty act with a corresponding guilty mind. This illustrates that a morally responsible agent is not merely a person who is able to do moral right or wrong. Beyond this, the agent is accountable for

his/her morally significant conduct by choice (or in our example, accountability is waived where no choice exists). Moral accountability[3] can only happen in the absence of coercion and is thus deeply connected to the concept of the agent's ability to have free will. It is also connected to the agent's ability to have a *mens rea* (guilty mind). When conditions such as consciousness are placed on what is required to have a *mens rea*, it follows that current-day machines would be excluded from moral agency as 'we cannot yet (and may never be able to) build conscious machines' (Himma, 2009, p24).

Many would contest that while machines can perform functional actions, for example, they can perform the acts they were designed for such as installing parts on a car on an assembly line; they cannot be said to be moral agents because they do not act freely and are unable to appreciate the consequences of their actions. 'It makes no sense to praise or censure something that lacks conscious mental states no matter how sophisticated a technology it may be' (Himma, 2009, p24). 'You do not scold your webbot, that is obvious' (Floridi and Sanders in Himma, 2009, p25). What happens when sophisticated machines appear to perform autonomous actions, the outcomes of which have moral consequence? Who would be responsible if the self-driving Googlecar malfunctioned and crashed into a school bus killing children? When only humans are accorded moral agency, individual software developers could be held morally accountable if they had not considered this aspect when they designed the software to drive the car. They could undergo censorship for violating a code of ethics for software design, if such code existed. Google itself could only be judged morally responsible if one does not accord any moral agency to the car, and locates all agency (and responsibility) in the designers, and by extension, in the company for which they work. At face value, Googlecar could be considered a moral agent. It appears to act autonomously (e.g., it drives itself) and seems responsive to identifying and reacting to things in its environment (e.g., brake when people in front). However, upon a closer diagnostic the car is not a moral agent. It cannot change its own software, which would equate to freedom of choice or real autonomy for a robot, and is still deterministically following a programme of 'drive-self' created by a human programmer. Real autonomy involves a process of learning, for example, bump corner adjust process, which the car does

[3] I use the terms moral responsibility and moral accountability interchangeably in this report and take them to have the same meaning.

do; however it learns according to the programmer's notion of what learning is, not its own. The programmer is still the agent, not the car. Superficially, the more machines appear to behave like us and do the things we do, we may be prone to ascribe moral-agent-like qualities to them. Regardless of whether Googlecar can be said to have free will in terms of its actions, we may perceive it as an agent due to its independent behaviour. We can see no human agents driving the car. Humans may have programmed the software but they are absent from the actual action of driving. I am aware of the pitfalls of associating such perceived autonomy with moral agency (and of the extrapolation that just because we could do something, it is ethically right to do so). Certainly a form of autonomy is necessary for moral agency; however, it is not sufficient for full moral agency. The ability to act independently and carry out an action, does not quite embrace the capacity to reflect on a situation and to form intentions about how to act. Googlecar drives itself, it is autonomous in this sense, but that is not enough to make it a full moral agent. It does not have free-will because it is bound by the prompts of its programme. It must be noted however, that both notions of autonomy and free-will are ethically controversial in traditional ethics which lacks agreement on definitions thereof; a feature that can leave the definition of human moral agency open to a degree of interpretation. I first review these definitional difficulties and from here, will provide a working definition of autonomy which I believe could be applicable to machines.

To illustrate some definitional difficulties, Schmidt and Kraemer entertain an interesting dialogue on notions of machine (or what they call artefact) autonomy. They believe that all machines start off non-autonomous because of their origin; they were created by another; but that they can achieve a type of autonomy via a process brought about by developing learning skills and adaptive behaviour. In this sense they could be said to have the capacity for independent action programmed into them by their creators (Schmidt and Kraemer, 2006, p75). They use a more classical definition of autonomy, taken from the Greek words *autos* (self) and *nomos* (law), which means the ability to impose a formal law on oneself. While etymologically, 'autonomous' was used in ancient Greece to describe politically independent states it is useful for our discussion on moral agents. Schmidt et al's argument is interesting insofar as it believes neither machines nor people are ultimately autonomous. 'For conceptual reasons it is a terminological mishap to apply the notion of autonomy to

artefacts [machines] because it would be a mistake to apply it to humans' (Schmidt et al, 2006, p74). They believe the field of AI has diluted the definition of autonomy to a weak definition, replacing it by 'the merely functional use of the term' to enable the reference to machines as 'autonomous agents...in a narrow and derivative sense' (Schmidt et al, 2006, p74). Their article primarily focuses on why humans cannot be said to be fully autonomous, and they are almost Wittgensteinian in their concern about how AI has changed the meaning of the term semantically over time. Although they do not specify what this weak definition comprises, they do say that AI uses the term to mean 'human-like' which is 'totally erroneous as humans are, at best, only semi-autonomous' (Schmidt et al, 2006, p79). Machines would fail the moral agency test on a human-like definition of autonomy. Humans would fail for different reasons, as their development of self is a social construct, hence they can never be fully autonomous (alone) as what is means to be human is developed in the presence of others (Schmidt et al, 2006, p78-79). Unlike humans however, machines could not be semi-autonomous because this would require them to have beliefs about their beliefs or a moral will of their own. Googlecar follows a programme it does not have an inherent belief that avoiding harming people by not running them over is a moral good.

Definitional difficulties continue in the concept of free-will (the capacity to freely choose one's acts) which 'is not well understood…and remains deeply contested among certain schools of philosophy in general' (Himma, 2009, p22). Himma alludes to the free-will debate in classical ethics and takes a deterministic stance that everything that happens in the world is subject to causal laws, including moral thought and action. Non-moral facts, such as our desires and emotions, fully determine moral facts by pre-conditioning our moral views and thus negate free will and moral responsibility. For Himma we are 'not clear in what sense our choices are free…and if we don't know what free-will is, we are not going to be able to model it technologically' (Himma, 2009, p28). Agents could not have done otherwise when acting as their actions are the result of antecedent, casual conditions and they are thus neither free nor morally responsible. On this basis, moral responsibility is impossible for Himma as no plausible interpretation of free will can be provided. The problem with such a view, as McKenna and Russell point out in their analysis of what they refer to as 'the dilemma of determinism', is that 'its conclusions seem both intellectually incredible and humanly impossible for us to accept or live with' (McKenna and Russell, 2008, p3). On

Himma's reading, we may as well open up all the prisons, release prisoners regardless of their crime and discontinue all retributive forms of justice, as we are incapable of being morally responsible and are therefore not punishable for morally bad acts.

Consciousness is often seen as part of free-will, and therefore of moral agency. Himma believes it reasonable to assume that on standard accounts of moral agency, 'each of the necessary capacities presuppose the capacity for consciousness'. For example, 'accountability… is sensibly attributed only to conscious beings…it is hard to make sense of a non-conscious thing freely choosing anything' (Himma, 2009, p24-25). For Himma the type of consciousness needed for moral agency, is an 'intentional mental state' on the part of the agent (Himma, 2009, p24). He goes on to say that the very nature of the way we react to agents who behave in morally good (praise) or morally bad (censure) ways only makes sense if the agent is conscious. Van Gulick, in referring to the subject philosophically says that 'perhaps no aspect of mind is more familiar or more puzzling than consciousness and our conscious experience of self and world' (Van Gulick, 2014). He refers to the problem of consciousness as one of a lack of 'any agreed upon theory of consciousness' (Van Gulick, 2014). Related, but not identical, is the 'problem of other minds', which holds that we can never truly know the mind of another (Himma, 2009, p28). We can make assumptions on what other people are thinking and how they think but this is indirect, only each individual person or entity can truly know what goes on in their own mind. It is conceivable that we might achieve a better understanding of consciousness without solving the problem of other minds, and vice versa. These examples highlight some of the difficulties in traditional definitions of moral agency. Many of the definitions used to confirm or deny machines moral agency come from standard ethical definitions based on the concepts of autonomy, free will and associated concepts such as consciousness, which lack universal consensus in traditional ethics. Apart from definitional difficulties, the task of proving whether machines can exhibit free-will or consciousness is extremely difficult, if not insurmountable, as our concept of both is inextricably linked to our concept of being human. Bringsjord et al say it would be difficult to build a machine that would have to 'think for itself' and have 'free-will' (Bringsjord, Bello and Ferrucci, 2001, p4). Moor refers to 'bright-line arguments' which propose a line between people and machines which machines cannot cross, that prevent machines from becoming full ethical agents, specifically on the basis 'that no machine can

have consciousness, intentionality and free-will' (Moor, 2006, p20).

I will use as my departure a working definition of machine autonomy as anything that is self-propelling, in order to gradually add additional conditions. Being self-propelling is probably necessary for autonomy as otherwise we cannot have agency. I realise that self-propelling on its own is grossly inadequate: a little plane with a rubber propeller is self-propelling, but it is not autonomous. I take self-propulsion to mean something that is capable of performing actions by itself, in the sense that Googlecar can drive without humans being present. I am under no illusion that Googelcar is only driving because it has been programmed to do so. When such autonomous machine action results in moral outcomes, the working definition extends to something similar to that proposed by Floridi that machine (or what he calls information) agents are 'any interactive, autonomous, and adaptable system that can perform morally quantifiable acts' (Floridi and Sanders in Tavani, 2008, p159). Granted Floridi's definition needs tightening and is weak as many systems have internal change mechanisms, such as the weather or indeed any organism, starting with microbes, which does not make them moral agents. I will however, use it as a starting point, to roughly mean that machines performing actions by themselves that are capable of adjusting these actions, for example, via a learning algorithm, can be said to be autonomous. For example, Googlecar drives per its programme without direct human intervention. It has a learning algorithm which enables it to brake when people are in front, reverse when obstacles are in the way and so forth, which allows it to adapt its actions (or system) to external changes in the environment. In this way, specific instances augment its programming rules which tell it to avoid harm or injury to people. Clearly autonomy means far more than this on standard, philosophical readings. I now review and comment on what others have taken its meaning to be, in order to contextualise my working definition and identify those aspects that align most closely to the literature.

Feinberg says there are at least four different meanings of autonomy in moral and political philosophy: the capacity to govern oneself, a set of rights expressive of one's sovereignty over oneself, a personal ideal and the actual condition of self-government (Feinberg in Christman, 2015). I will focus specifically on the first two, as they encapsulate the notion of moral autonomy, defined as 'the capacity to impose a moral law on oneself and following Kant [autonomy] is the organising principle of all morality' (Hill in Christman, 2015). Personal

autonomy is seen as a trait that individuals can exhibit relative to any aspect of their lives, both moral and non-moral, and as such is not limited to questions of moral obligation (Dworkin in Christman, 2015). I favour definitions of autonomy as 'a condition or state of functioning' as they suit the utility of machines. Accordingly, I explore Kant in more detail. Conversely, I typically avoid definitions of autonomy 'as a capacity' as most such capacities are deeply linked to 'definitions of personhood' which are ill suited to machines (Franklin-Hall, 2013, p227; Schmidt et al, 2006, p75). Furthermore, autonomy links to agency in the sense that an agent 'possesses a right to autonomy if [it] is (in some sense) either acting autonomously or at least capable of doing so (Feinberg in Franklin-Hall, 2013, p227).

The question of autonomy has played a central role in the development of modern Western moral philosophy. It could be argued that the advent of Christianity, theologically, required us to be autonomous agents in the sense that an act of faith was a freely chosen one not imposed by God. People, who chose religion, chose not to do evil. This crystallised into a view, reflected in the work of Montaigne (1595), that morality needed to go beyond just sectarian principles and that it might arise from resources within human nature itself, understood as human self-governance or autonomy which culminated in the works of Kant (1785), Reid (1788) and Bentham (1789) (Schneewind, 2011, p147). Contemporary definitions of autonomy, as the ability of individuals to self-govern, are influenced by theorists from this era (Schneewind, 2011, p147). Although a predecessor of the abovementioned theorists, ancient Greek philosopher Aristotle, saw the key to human flourishing, which he believed was the moral purpose of our lives, as autonomy, defined as the ability to be the best one could be in society in a number of self-chosen ways (Bynum, 2006, p160). This self-chosen drive is echoed in Kant as the ability to impose laws upon or govern oneself. Kant believed the purpose of morality was to impose absolute duties on us in order to guide what we ought to do in moral situations. For him, this 'special kind of moral necessity could only arise from a law we impose on ourselves' (Schneewind, 2011, p151). The key to successful application of this autonomy, or self-governing action, was freedom in the sense of something derived from within us instead of that externally imposed upon us. Our practical rationality or reason was our internal-self-guiding-law. This 'moral law…is not a requirement to do good to others. It tells us rather to act only in ways which we could rationally agree to have everyone act'. We can answer whether our actions are

moral, by asking 'can I without self-contradiction will this [action] to be a law according to which everyone always acts' (Schneewind paraphrasing Kant, 2011, p151)? We are fully autonomous in this sense, as it is always possible for us to know what the right thing to do is by appealing to our rationality, without depending upon external sources for motivation or guidance. For Kant, such self-imposition of the moral law (i.e., rationality) is autonomy. As this law has 'no content provided by sense or desire or any contingent aspect of our situation, it must be universal' and is thus normative (Christman, 2015). Furthermore, Kant's law is the source of all moral value, a value which is objective, 'not contingent upon psychological facts about us'. Rather 'it is the unavoidable implication of the exercise of practical reason' (Taylor in Christman, 2015). In summary, autonomy for Kant is both a model of practical reason in determining our moral obligation and a feature of other persons deserving moral respect from us (Christman, 2015).

Could we view Kant's rules that guide our moral behaviour, as analogous to a computer programme? Specifically, for Kant, the machine would only be autonomous if, like us, it could decide on its own which among a number of alternative programmes, it is best to follow. Jeremy Bentham (1781), who like Kant believed that a basic principle of morality had to be one which could actually be used by everyone alike, pointed out that his theory of act utilitarianism (doing the greatest good for the greatest number of people) involved performing 'moral arithmetic'. Machine ethicists Michael and Susan Anderson suggest that Bentham's theory can be programmed into machines using 'an algorithm to compute the best action…with the highest net pleasure being the right action' (Anderson, 2007, p18). Granted, the machine is running a programme developed by people; however, its programme may enable it to 'learn' from different situations and in so doing, go beyond its original programme. Similarly, Schmidt et al in their discussion of artificial autonomy refer to a machine as 'being an autonomous agent…abiding by self-imposed moral laws that prove its independence from the original intentions of its creator' (Schmidt et al, 2006, p 77). Many would however argue that no outcome of a learning programme is really unintended by the programmer, since it came about in a series of computing steps all of which have been intended by the programmer. It may just be that increasing complexity makes it appear to us that the outcome is somehow 'new', or 'unintended'. Perhaps however, increasing complexity is all you need for definitions of autonomous moral agency based on the type of

relationship we are able to independently have with an entity (see my discussion on Strawson in section 3 for relational definitions of moral agency).

So where does Kant leave us in our working definition of machine autonomy? The premium Kant places on rationality, devoid of psychological values and desires, is an attractive fit with our notion of machines and the software programmes that drive them. The universality of rational no matter what, makes it easier to conceptualise a sense of moral agency being accorded to something machine-like. Kant's ethics do without moral sense, and any (moral) input from emotions, consciousness, etc. All it requires is rationality, which is the closest thing machines can deliver. My working definition of machine autonomy can now extend to include the ability to process information rationally. Thus we now have a working definition of machine autonomy as anything that is self-propelling (i.e., able to perform actions by itself which are driven by software programmes); that is able to process information rationally (i.e., using rules of logic and reason); that can adapt its actions relative to stimuli received from its environment (e.g., using a learning programme) which in turn, allow it to alter its responses, and in so doing change its internal system (i.e., add to its learning programme).

I now review critiques of Kant and of moral agency as rationality, to identify whether there are aspects which could augment my working definition of machine autonomy. Self-described non-Kantians (Benson 1994; Grovier 1993; Lehrer 1997; Westland 2014, all in Christman, 2015), believe that being autonomous implies a measure of self-worth, and thus subjectivity, in that we must trust our decision-making capabilities enough to put ourselves in a position of responsibility in the first place. Korsgaard takes this further with her 'practical identity', an individual-specific 'conception of the right and the good bound up with one's self-conception and self-worth' (Franklin-Hall, 2013, p228-229). Traditional critiques of autonomy-based moral views, and Kant's in particular, concern the way in which the theory of autonomy is grounded in our 'cognitive abilities rather than in our emotions and affective connections' (Williams 1985; Stocker 1976; both in Christman, 2015). Critics believe that when we make moral decisions, 'we bring both our ability to reflect using practical reason [Kant] but also our passions [non-Kantian] emotions, desires, felt commitments, senses of attraction and aversion, alienation and comfort' (Christman, 2015). For example, parents behave paternalistically towards their children. This is actually the opposite of viewing children autonomously, as they have not yet attained a 'capacity for

reasoning' (Franklin-Hall, 2013, p224) and as such 'cannot constitute themselves as moral agents all at once' (Schapiro in Franklin-Hall, 2013, p229). Parents perform such obligation from the emotional basis of love, care and moral and educational guardianship. In such cases, we use both our ability to rationalise and our ability to 'engage in the right passions', where emotions are part of our decision-making. Thus we value ourselves and others as 'passionate reasoners not merely reasoners per se' (Christman, 2015). This introduces a notion of individual subjectivity in our moral decision making as autonomous agents. Further critiques of Kant say that if the source of normativity involved in autonomy need only be hypothetical, i.e., by asking whether a decision would be made by all rational people in the same situation, this is insufficient motivation to ground a sense of obligation to the decision. Resultantly, morality is imposed by 'an idealised more rational self' rather than the 'actual self' working with an 'actual set of judgements made by the agent in question' (Christman, 2015). Hegel's critique of Kant encapsulates such notions, namely, whether autonomy can be the crux of moral obligation and respect if it is thought of in a purely procedural manner (as a by-product of rational thought, per Kant) if no substantive commitments or value orientations are included? Thus, can autonomy be objective? My working definition of machine autonomy would steer away from the abovementioned critique of Kant, as it implies that machines would need to be able to conceptualise concepts of value and use a variety of emotions in decision making, which they are unable to do.

Various theorists believe personal autonomy has intrinsic value independent of a fully worked out view of practical reason. John Stuart Mill viewed it as being 'one of the elements of well-being' (Mill in Christman, 2015). Such views see autonomy as the capacity to reflect on one's values, character and commitments and tie the concept to a notion of personhood. Dennett in advising on which entities should be called persons, sees personal autonomy as involving rational beings; with states of consciousness or at least states to which psychological or mental or intentional predicates are ascribed; our treatment of which is related to the entity being a person; capable of reciprocation in some way; able to verbally communicate and that is conscious in some special way, often identified as self-conscious (Dennett, in Schmidt et al, 2006, p75-76). This special sort of consciousness as a pre-condition for moral agency is iterated by Anscombe, Sartre and Frankfurt. Frankfurt takes it further, saying consciousness (and thus autonomy) requires that we have 'higher-order

volitions' or the ability to reflect on our basic, first-order desires, such as being hungry or wanting to take drugs. Our reflection on these desires and our responsible choosing of which ones to accept, reflects our autonomous personhood (Dennett, 1978, p285; quoting Frankfurt, 1971, p14; both in Schmidt et al, 2006, p76). If we built a working definition of machine autonomy to embrace this rich version of personhood, we would have a definition that few machines could conceivably satisfy and as such it would not be very conducive to my report. However, I am wary that this exclusion exposes me the possibility of 'defining my way through' a philosophical problem, i.e., finding a solution to a question by adopting a definition that fits that solution. As such, my definition has to be one that people would recognize as a definition of 'autonomy', even if in a minimal sense.

I briefly review critiques of autonomy as personhood to see whether these could help inform my working definition of autonomy. Definitions of autonomy as personhood struggle with the concept of why someone's autonomy is intrinsically valuable when he/she uses it to say, harm him/herself or to make morally poor choices. Conly believes that such definitions fail to account for the fact that when people make moral decisions, they are prone to systematic biases and distortions (Conly in Christman, 2015). As people have different capacities moral autonomy could obtain in degrees. Why then should personal autonomy be seen to be equally valuable in people who display different levels of it? Autonomy is often cited as the reason for treating all individuals equally from a moral perspective (Christman, 2015), but if it is not an all-or-nothing characteristic, this commitment to moral equality becomes problematic (Arneson in Christman, 2015). Thus, is it difficult to maintain that all autonomous beings, so defined, have equal moral status or that their interests deserve the same weight. Such difficulties make Kant's normative and universal notion of autonomy as rationality even more compelling[4]. What is interesting is the idea of autonomy as a threshold concept. This is the idea that autonomy and thus agency could be manifested in degrees. Put another way, it asks whether the abilities or capacities that constitute personal autonomy obtain all at once, or progressively (Christman, 2015)? Franklin-Hall entertains an interesting discussion, using a 'stage of life' theory to justify temporarily removing autonomy (and thereby asserting paternalism) with children, until they have developed the capacities

[4] I must note that I will endorse an idea of gradualism later, when I claim that morality is a developmental construct which can apply to future versions of machines.

necessary to be autonomous. Locke phrases this as 'we are born free, as we are born rational; not that [at first] we have actually the exercise of either; age that brings one, brings with it the other too' (Locke in Franklin-Hall, 2013, p223-224). Schapiro says 'children are as yet lacking in reason and therefore unfit to govern themselves; and Mill says that 'children lack imperfect reasoning and their lack of experience of the world and of themselves makes them not yet suitably proficient judges of where their own well-being lies' (Franklyn-Hall, 2013, p225). Similar notions are echoed in contemporary theorists, such as Archard (2004) and Freeman (1997). If we take a developmental view of technology, we know that current day machines are not yet autonomous on traditional readings which include notions of free-will and consciousness; but as new versions of technology evolve, machines may at some point develop capacities that more closely align to such readings. My working definition of machines learning from their environment is developmental to a specific machine's programme. With human morality the developmental view looks at an individual child developing. However with machines, a more abstract, developmental notion is also possible - future and different versions of the machine would develop over time and thus different machines could evolve the definition of machine autonomy further.

To summarise, the following is useful to my working definition from the critique of autonomy as personhood. The idea that autonomy could occur in degrees means my working definition of machine autonomy can be a weak one, namely, machines need to meet a minimum set of standards for autonomy. For our purposes the minimum standards are the ability to perform self-propelling actions; the rational processing of information (per Kant) and the alteration of the way information is processed, and thus actions are performed, relative to stimuli the machine receives from the environment (per Floridi). My working definition excludes notions of personhood such as free will and consciousness with the rider that the definition is meeting minimum standards for autonomy and a further qualification that as machines advance more standards could be added.

Many theorists believe that as we become more reliant on technology, it causes a blurring of moral accountability. A sophisticated technology like Googlecar falls prey to the 'problem of many hands' (Noorman, 2014). It is designed by a myriad of programmers, individuals and often companies, making the ascription of agency (and in turn moral responsibility) complicated. More autonomous forms of technology create a temporal and physical

distance between a person and the consequences of his / her actions. For example, remotely controlled military Drones, blur the causal connection between actions and events. The remoteness of long-range missiles further disassociates action and consequence morally. This disassociation starts changing our moral mind-set. The convenience of modern technology and its time/labour saving machine devices, lead to us framing our world with a 'device' mind-set which hides the full context of the world and disconnects us from it (Borgmann in Introna, 2011). For example, clicking a switch on a thermostat to make our room temperature warm, replaces the process of chopping wood, building the fire and maintaining it. Accordingly, we start to undervalue the natural (i.e., non-artificial) world, as it becomes increasingly hidden from us. This opacity of systems means that people often have no idea how they work and as such they have a tendency to rely either too much or too little on the accuracy of automated information (Cummings 2004; Parasuraman and Riley 1997; both in Noorman, 2014). Leveson and Turner (1993, in Noorman, 2014) discuss as an example, the malfunctioning radiation treatment machine Therac-25. Designed to treat cancer patients and for X-rays during a 2 year period in the 1980's the machine massively overdosed six patients, contributing to the death of three of them. In analysing the events, Leveson et al conclude that it was difficult to place the blame on a single person and that the event was a combination of human error, software errors, bad interface design and inadequate testing and quality assurance. Similarly the 1988 missile cruiser USS Vincennes shot down a civilian jet plane, killing all 290 passengers on board after its Aegis defensive system mistakenly identified the plane as an attacking military aircraft (Gray 1997, in Noorman, 2014). Two other warships in the area had correctly identified the aircraft as civilian however the crew were so overconfident in the ability of Aegis that they did not dispute the Vincennes identification as they believed Aegis had information about the craft that they did not have. A crew member later said 'we called her Robocruiser...she always seemed to have a picture…she always seemed to be telling everybody to get on or off the link as though her picture was better (as quoted in Gray 1997, p34; in Noorman, 2014). Nissenbaum believes the use of such technology has led to an erosion of moral accountability in current society. Such accountability is reduced by what she calls the problem of many hands; the acceptance of computer bugs as an inherent part of software development; the tendency for people to use computers as a scapegoat for responsibility (per the Aegis example) and the tendency of software companies to own code without

considering or taking responsibility for the outcome of its usage (Nissenbaum 1997, in Noorman, 2014). Such advanced machines disassociate us from our actions and substitute our agency.

Consider more practically, the question of machine agency and the blurring effect, in the following scenarios:

1. Scenario 1 - a crude forklift. Manually operated by someone physically pulling down on a heavy lever, which in turn activates a series of interconnecting cogs that lift the forks and the container thereon. In this scenario it is clear that the person pulling the lever is the agent. The outcome (the lifting of the container) is a direct result of the person's action (pulling the lever).

2. Scenario 2 - a button operated lifter. A person merely presses a button to activate the machine, which may be driven by a series of cogs and software to lift the container. It is still clear that the person's action lifts the container, but the physical effort of the agent is minimal.

3. Scenario 3 - a remote controlled lifter. A person has a remote control and activates the lifter from the comfort of his/her office a few kilometres away. The lifter comprises motion sensors driven by software which make it easy to identify and locate the container and lift it. This is similar to Googlecar. An onlooker would have no idea who the agent was and whether it was a machine or person operating the lifter. In turn, all the operator need do is activate the machine and resume work on other, unrelated tasks.

In scenario 3 there is a physical separation between the agent and the act of lifting. In such instances we may start applying agent language to something that is not an agent (i.e., to the lifting mechanism instead of the person operating the remote control). I am aware that just because we may be inclined to take this agent-surrogacy position, it does not necessarily follow that ethically we should take it. The fact however remains that there is a blurring of moral accountability between ourselves and machines in these instances.

Some theorists believe that the advent of domestic service robots of the Googlecar type will make the question of ethical accountability even more pressing as machines move from the

factory floor, with its relatively controlled environment, into our everyday lives which are often a lot less predictable (Allen, Wallach and Smit, 2006, p14-15). These scenarios question 'whether trolley cases could be one of the first frontiers for machine ethics' (Allen et al, 2006, p12). Trolley cases were first introduced by Philosopher Phillipa Foot in 1967 as a thought experiment. Variations are many, however the problem is framed as to what a person (and now machine) would do in a number of scenarios involving a runaway trolley on a railway track. For example, there are five people tied up and unable to move on the track, the trolley is headed straight for them. You are standing some distance off next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is a person on the side track. You do not have the ability to operate the lever in a way that would cause the trolley to derail without loss of life. What do you do? What ethical insights could a trolley case for a self-driving train teach us? Driverless train systems are being used in the London Underground and Paris metro systems today. What would we want such systems to compute where to steer an out of control train? At the very least, consideration of such scenarios from a moral perspective may change the way programmers develop the system. It may stimulate them to develop a programme that enables the train to calculate which scenario is more ethical. On act utilitarianism the option that loses the least life or does the least harm is selected. The lever would be flipped to divert the train to save the lives of the five at the expense of the one. According machines a type of moral agency sounds counter-intuitive, but upon closer reflection of the train scenario, we are unable to dispel the notion that it's worth considering the ethical ramifications associated with designing more sophisticated forms of AI. Thinking about AI as having such agency, at the very least, helps guide system design. The increased autonomy of technology in the train, Googlecar and forklift examples means we need to be less present at the point of moral action. Agency switches in a real-sense from us to machines as we outsource actions to them. In the train case, people are no longer there to flip the lever, the driverless train is. Allen et al point out that today we already have software programmes that violate ethical standards of privacy, such as search engines designed to collect data, without the knowledge of the user who initiated the query (Allen et al, 2006, p12). We are unaware that a simple search engine has been programmed with the agency to pre-decide how it uses our information.

Where does this discussion leave my working definition of machine autonomy? To recap, my working definition is tied to machines that are self-propelled. Drawing from Kant, they are able to process information rationally. Further they are able to respond to stimuli in their environment, which combined with relevant programming, enables them to adapt their actions accordingly. The definition is a weak one as the actions are guided by software programmes, which are ultimately created by human programmers, a feature which negates free-will and consciousness. The concept of a weak definition is borrowed from a critique of personal autonomy in traditional ethics, which holds that as people have different capacities, personal autonomy manifests in degrees. Thus, my working definition encapsulates a minimum set of criteria for machine autonomy, which should be reviewed as technology evolves.

To conclude, whether it is morally correct to do so or not, increased machine autonomy allows us to surrogate agency to machines and in instances where this results in moral outcomes, moral responsibility and thus moral agency can become blurred. Discussion on how other theorists have argued for and against machine moral agency could supply some answers in terms of how others see the rightness or wrongness thereof. In the next section, I conduct a review of theorists who associate concepts of intelligence with moral agency as most machine ethicists use human intelligence as a comparator in trying to define machine moral agency.

1.3 Can machines be moral agents?

The question of machine moral agency is deemed important enough to devote an entire school of ethics, called machine ethics, towards answering 'whether you can put ethics into a machine' (Moor, 2006, p19). This inter-disciplinary enterprise sees computer scientists and ethicists collaborating to try to build AI that exhibits a sense of morality. Researching whether ethics is programmable, underlines the founding paradigm of AI, which is to model the human brain as an intelligent agent. Accordingly, much literature on machine moral agency starts by defining particular types of intelligence as a pre-requisite to moral consideration. Opponents and proponents vary as to what constitutes intelligence and whether moral agency can be extended to machines using such definitions.

On traditional ethical views, what I have termed a weak definition of moral agency broadly

sees rationality as an essential quality. This simple-rationalist position believes that things are explicable via a system of theoretical, objective principles, similar to the theoretical truths of mathematics and that rational argument can explain all of nature and justify its actions (Dreyfus, 1988, p2). For example, Kant believed morality was based on principles of duty or obligation regarding what we ought to do morally. This duty was guided by our ability to act with self-restraint, which Kant viewed as a type of autonomy (which is included in my working definition of machine autonomy). Descartes believed that our moral behaviour could be guided by 'asking how a rational and good god would organise the world and how he would want man to act' (Dreyfus, 1988, p3). Kant's concept of a person is inextricably linked to *vernunft* (reason) and he also introduced the notion that adherence to sets of universal rules was needed for morality. Such beliefs formed the basis of Western philosophical thought and still underlie traditional views as to how we should evaluate human moral action today. Applying this traditional view to machines, at the extreme, theorists like Dietrich believe that because they are not emotional, machines may actually be able to behave more ethically than people. Machines do not favour their own interests, nor do they compete with each other and they do not behave unpredictably (Dietrich in Anderson, 2007, p17). Machine ethicists Michael and Susan Anderson concur that 'emotions can interfere with a being's ability to determine and perform the right action in an ethical dilemma…humans are prone to getting carried away by their emotions to the point where they are incapable of following moral principles' (Anderson, 2007, p19). Dreyfus in writing about what machines cannot do makes some interesting comments on how they process information as 'logic or inference machines' (Dreyfus, 1988, p53). In the 1950's the early computers were built by logicians such as Alan Turing (who will feature later in this report). They built computers that could 'manipulate symbols according to exact rules' (Dreyfus, 1988, p53). Computer programmers used symbols to represent elementary facts about the world and rules to represent relationships between them. 'Computers could then follow such rules to deduce how those facts affect each other and what happens when the facts change' (Dreyfus, 1988, p53). In this way computers simulate logical thinking (Dreyfus, 1988, p53). Dreyfus notes that humans suffering from a neurological disorder called *agnosia* process their world in much the same way as logic machines with total dependence upon analysis and rational explanation (Dreyfus, 1988, p64). For Dreyfus the simple rationalist view of intelligence applies to machines and to people with disorders that make them

behave like machines. On Kant's reading, adult humans are bearers of moral agency because of their ability to reason and rationalise, rather than their human-DNA or any other species-specific criterion. This leads modern computer ethicists like Sullins to claim that moral agency does not require personhood (Sullins, 2006, p23-29). Similarly, Floridi and Sanders want to extend the class of entities that can be involved in moral situations to include 'artificial agents' or what they call 'mindless-morality' (Floridi and Sanders, 2004, p349). My working definition of machine autonomy, a pre-cursor to machine moral agency, encapsulates a weak definition of intelligence on standard ethics.

To summarise the weak definition or simple-rationalist view of intelligence on standard ethics celebrates qualities of rationality, reason and logic and is guided by objective rules and principles of right action. This embraces what Allen, Wallach and Smit (2006, p16) have referred to as the Stoic view of ethics which sees emotions as irrelevant and dangerous to making ethically correct decisions. Ironically, this definition almost calls for humans to behave like machines when making moral decisions and taking moral action, devoid of emotions. Whilst normatively attractive, simple rationalists disregard a practical reality that people naturally use emotion when making moral decisions. They are more aligned to supporting the way computers process information (a machine's version of thinking rationally), rather than the way humans actually make moral decisions. Conversely, theorists such as Strawson believe that our emotions, or what he calls 'reactive attitudes', are our natural recourse when making ethical decisions and that we can only suspend these in favour of more objective attitudes, in exceptional cases like when dealing with people of unsound mind and so forth (Strawson, 1972, p6). On a weak definition of intelligence, would car assembly robots used in factories qualify as artificial moral agents? They have the capacity to recognise symbols (a car) and particular facts about a car in order to perform actions based on rules. For example, if car A is in front of me, install part Y; if car B is in front of me install part Z. This kind of primitive rule-following as exhibited by assembly robots may make them agents, but does not necessarily make them moral agents. For one, morality is more than recognising symbols and for another to be a moral agent, even in a minimal sense, the types of rules followed must be moral rules, which car assembly rules are not. Moral agency is thus somewhat more complex than the example given. Moreover, people do not behave like car assembly robots when making moral decisions, we do more

than recognise symbols and follow rules. Conflating morality-as-rationality is flawed, as in practise we do not behave in the way in which the definition suggests we do when making moral decisions. On this basis I argue that we must be cautious not to conflate weak or simple rationalist defintions of intelligence with moral agency too hastily.

Stronger definitions of intelligence are all based largely on qualities we associate with being human, or at least with being sentient. Many are similar to the qualities used to define personal autonomy, namely, the capacity for free-will, the ability to create, our intuition and emotions are used in varying degrees by theorists to argue against moral agency for machines. These definitions provide insurmountable barriers to machines ever being accorded moral agency. The criteria are too human and thus too onerous. Strong definitions are opposed to simple-rationalist views and such opposition has a long tradition. Pascal, a French philosopher who lived in the 17th century opposed Descartes' view that man should ask what a rational god would do to guide his actions. He very practically believed that in everyday life, all the average person had to rely on when making moral decisions was their background of custom and experience. Practically one had no choice but to trust one's emotions and intuitions (Dreyfus, 1988, p3). This led him to claim 'the heart has its reasons that reason does not know' (Pascal in Dreyfus, 1988, p3). Hume the great empiricist believed that knowledge was not grounded in theories or principles but rather based on habits formed by successful coping. Theorists like Heidegger and Merleau-Ponty, that human skill and experience was the greatest asset in decision making (Dreyfus, 1988, p3-5). Merleau-Ponty claimed people develop perception and understanding, based on our propensity for flexible styles of associative behaviour. For example, someone who knows how to drive a car can easily adapt the rules on how to change gears in a standard car, in order to drive a car with a gearshift on the steering column. Many opponents of machine moral agency use strong definitions of intelligence to demonstrate that machines cannot exhibit intelligence and therefore cannot be moral agents. Johnson and Kuflik argue that we have unique capacities which machines will never have such as mental states, intentionality, common sense, emotions and free will, which are key determinants in attributing moral agency (Johnson, 2006; Kuflik, 1999; both in Noorman, 2014). Johnson makes the case that computer technologies remain connected to the intentionality of their creators and users. Stahl says computers are not able to morally reason because they do not have the capacity

to understand the meaning of the information that they process (Stahl in Noorman, 2014).

In a retributive theory of punishment, Sparrow and Asaro point out that it makes no sense to treat computer systems as moral agents that can be held responsible, for they cannot suffer and thus cannot be punished (Sparrow, 2007; Asaro 2001; both in Noorman, 2014). For Sparrow et al the concept of morality is connected to punishment, if we cannot punish an entity for moral wrong, it is not a moral agent. Children fall into this category, the thought of punishing a young child for something it did that led to a morally poor outcome is considered unheard of universally. Young children, whilst able to perform acts that lead to moral outcomes, do so unintentionally, they are unable to fully understand the consequences of their actions in terms of moral right and wrong. On this basis, they are not yet moral agents and we would therefore frown upon punishing their actions accordingly. It would be difficult to ascribe retributive intuitions related to morality to machines. We do not regard machines as being punishable and therefore as being morally responsible (on Sparrow and Asaro's reading). The view of punishment implying morality could be a huge stumbling block in according morality to machines. When we are punished we are incarcerated in jail, a type of psychological torture. How would you psychologically torture a machine? How could mental torture be programmed into a machine? In the sci-fi movie 2001 Space Odyssey, HAL the ships computer malfunctions inexplicably. The crew decides that in order not to endanger the mission, they have to switch HAL off. HAL mentions that he fears being shut down and when the crew do flip the switch he starts regressing with his behaviour and speech from adult to more child-like, until eventually he shuts down. If machines do evolve to the point where they are similar to HAL, it could be said that we may be able to punish such a machine by switching it off, or destroying its software and memory. The thought of punishing current day machines in ways similar to which we punish ourselves, sounds absurd. Maybe it means that a retributive notion of punishment is inappropriate to defining machine morality? There are alternate theories of punishment, which involve rehabilitating offenders. Going further, consequentialist theories would favour reconciliation over retribution, particularly if the consequences of reconciliation retained harmony within the wider community. Similar notions are echoed in *Ubuntu,* roughly translated as our sense of communal humanity. Alternatives to punishment mean we do not need to stick solely to a definition of morality that involves notions of retributive justice. A machine could be rehabilitated by

reprogramming aspects of its software. For example, little robot, Paro who looks after the elderly, could be programmed in such a way that the machine may illicit a reactive attitude of gratitude in the elderly person it is taking care of. However, this programming does not mean that the machine is able to have feelings, but rather that the type of programme it runs, enhances certain feelings in us. Unfortunately the possibility of programming a machine with actual feelings is as remote today, as it is to build one that is conscious, like HAL appears to be in the movie. If it was possible to program genuine feelings of goodwill into a machine, we would have a machine that satisfies the 'strong' definition of intelligence, and most of our problems would be solved.

Dreyfus proposes a non-mechanical model of human intelligence to avoid definitions that equate humans to machines (Dreyfus, 1988, p15). He believes machines will never replicate human intelligence largely due to their ability to have intuition, often demonstrated by those with experienced skills (experts) who use hunches and intuition as their decision making core. In building his case he paints a picture of two different types of intelligence. Humans have intuitive intelligence, not rules-based, not logical (in this sense he disagrees with the traditional version of human moral intelligence per Kant). For him machine intelligence is about 'knowing-that' which is a type of knowledge based on facts and rules. Conversely, human intelligence is about 'know-how' based on experience and skills (Dreyfus, 1988, p2-8). This common-sense knowing results because humans process holistically rather than mechanically. Some of what we are metaphysically is innate, for example having a body with certain capacities. These are things we understand by virtue of being human, i.e., that insults make us angry (Dreyfus, 1988, p2-8). Despite attempts by MIT's AI laboratory to develop a humanoid robot called Cog, designed to go through 'an embodied infancy and childhood', Dennett believes that such understandings are not programmable (Dennett, 1997, p358). Dreyfus would concur that the main point of giving fictional character HAL a humanoid past was to give him the world knowledge required to understand the 'human condition' and hence be a moral agent (Dennett, 1997, p360). Dreyfus calls this the ability in humans to 'know without knowing' a state that is only acquired by having a human context, i.e., the experience of what it is to be human (Dreyfus, 1988, p2-8). These views relate a strong definition of intelligence to human context and experience. On such readings moral consideration seems relative to what we are innately and thus it cannot be accorded to

machines by definition.

Both weak and strong definitions show that there is a flaw in our understanding of intelligence itself in traditional ethics. Many of our concepts of intelligence, particularly on a simple rationalist reading, are machine like. As a person, I would not wish to live in a world where I was judged morally on mere rationality. In reality we use many of the qualities we think make us uniquely human like experiences, common sense and emotions when we make moral judgements. Simple rationalist theories fail to account for this. In fact they are more akin to computational processes that use logic, rationality and are devoid of emotion. Therefore simple-rationalist theories are actually robot-friendly, in that we understand ourselves as robots morally. In the field of AI, the mind is the brain. When you take away the brain (the matter), you take away the entire mind. For AI the brain, and hence the mind, is a machine. This is what Descartes precisely did not believe. He understood the mind as being separate from the body, a dualist view. If you take away the body, the mind (or the soul) may survive; a Christian doctrine with which Descartes stayed compatible. He understood animals as machines, because he believed they had no soul. So for AI, we are rather like Descartes' animals. People who work in AI try to create machines that function like a brain. If many concepts of human morality are machine like, why not blur the lines further and extend agency to machines? After all, we are dealing with new territory in information ethics and many theorists like Floridi call for a new ethic to account for the man-machine relationship. Much of the way we process our world outside of the ethical arena is also framed by rationalism. For example, the practise of Western medicine places a premium on logic, science and a mechanistic and symptomatic treatment process. In summary, the traditional (or weak, or simple-rationalist) definitions of intelligence are too narrow to apply to people insofar as they disregard the role emotions play in moral decision making in favour of rationality trumping all. Such definitions may be able to apply to certain types of advanced machines for this very reason. The strong definition however, is too people-centric to have relevance for machines as it appeals to innately human qualities. Both definitions disregard a wider question on whether intelligence is in fact a condition for morality. To analyse this question, the following set of claims is worth considering:

1. Intelligence implies moral responsibility/moral agency and moral responsibility/moral agency implies intelligence

2. Intelligence implies moral responsibility/moral agency but moral responsibility/moral agency does not imply intelligence

3. Intelligence and moral responsibility/moral agency are unrelated

4. Moral responsibility/moral agency implies intelligence

I have used both moral responsibility and moral agency here. Although they are separate terms, there is a conceptual possibility that a mind could have moral responsibility, but no moral agency, simply because it is not capable of being an agent. For example, a computer, might perhaps, if sufficiently sophisticated in a sci-fi type scenario, have the capacity to understand codes of conduct and moral norms and values, but no capacity for any sort of agency whatsoever, because it is not connected to any robotic machine, and hence cannot 'act' in any sort of way. While we do need to distinguish between moral responsibility and moral agency, I have retained both; however my research question focusses specifically on moral agency.

Claim 1 - intelligence implies moral responsibility/moral agency and moral responsibility/moral agency implies intelligence

This claim would assert a biconditional relation between intelligence and moral responsibility/moral agency. It means you are only truly intelligent, if and only if you are morally responsible/a moral agent; and if and only if you are morally responsible/a moral agent, are you are intelligent. This claim implies that intelligence can only be defined using moral terms. But it constitutes a straw-man argument - it is easy to show this cannot be the case. Human intelligence is far richer than simply moral forms thereof. We know that human intelligence is a multi-faceted and vast concept which extends way beyond the ethical domain. For example, people can have musical, creative and artistic forms of intelligence. Toddlers would fail an IQ test, but we would not call them unintelligent. Thus not all forms of intelligence are necessarily moral and not all forms of morality encompass all forms of intelligence. Computer engineers would of course support claim 1, as using a weak definition of intelligence, they could build a machine that is rational, logical and autonomous. Therefore on this definition machines could be built to be intelligent; and intelligent machines so built would be morally responsible/moral agents. On the bi-

conditional claim, both artificial intelligence and artificial morality could fairly easily be proved, which supports my straw-man claim.

Claim 2 - intelligence implies moral responsibility/moral agency but moral responsibility/moral agency does not imply intelligence

This claim would assert that if you are intelligent you are necessarily (capable of being) morally responsible/a moral agent, but if you are (capable of being) morally responsible/a moral agent you are not necessarily intelligent. If we say intelligence implies moral responsibility/moral agency, we would be committed to saying that if we are not morally responsible/a moral agent, we are not intelligent. This is problematic. Many psychopaths, like serial killers, have above average intelligence but are not morally responsible/moral agents. They epitomise a simple-rationalist definition of intelligence appearing devoid of emotion, calculating and rational. Yet some would argue that they are not morally responsible by virtue of their pathology. Would Oscar Pistorius have been sentenced differently if the court had found him to be a psychopath, favouring institutionalisation in a mental asylum rather than prison? Traditionally, we do not accord babies, children and severely mentally impaired adults' moral agency, but this does not make them immoral in our minds. It would mean that a mentally handicapped person cannot be a good person because he/she was not intelligent, or that children are not yet intelligent as they are not yet regarded as moral agents. Some may say that this is absurd but some may not. After all, one could argue that children are, in quite a real sense, 'stupid': they cannot read or write (well), they cannot operate machinery, orientate themselves in a city or a forest, judge risk, or, for that matter, pass an IQ test. Their IQ is in fact lower than that of adults as measured by standard IQ tests. The reason for which they require supervision at all times is, precisely, that they do not have the wherewithal (some may say, not yet the intelligence) necessary to avoid harm to themselves and others when left alone. On this basis some may say that children are not yet intelligent. They have the potential for intelligence, but that potential is not actualized as they have not yet matured into adults. Similar arguments could be made, *mutatis mutandis*, for handicapped people. The respect we treat a child with is because of its potential. Babies and children are not yet intelligent but we know that they develop into intelligent human beings (all things being equal). Thus it is not because of children's current properties that we treat them with respect, but rather because of their potential or

dispositional ones. We know that under normal developmental conditions, children develop into adults and are accorded moral agency. This analogy however is imperfect when applied to AI. AI may well develop dispositionally, as children do; however, the development is not of the same machine, but rather future versions of different machines.

Claim 3 - intelligence and moral responsibility/moral agency are unrelated. Another possibility is that intelligence and moral responsibility/moral agency is unrelated. I will argue in this report that intelligence has very little to do with moral responsibility/moral agency. Rather, moral responsibility/moral agency is based on the type of relationships we enter into with others and the way we react to these relationships (see section 3 for this discussion).

Claim 4 – morality responsibility/moral agency implies intelligence

On a simple, rational reading of intelligence, it could be said that moral judgement is cognitive and that being morally responsible/a moral agent is being rational. This is Kantian in the sense that moral responsibility/moral agency is rationality. Thus every moral mistake is a mistake of reasoning. For example, if you commit a crime, you violate the Categorical Imperative and have failed to see it as a universal moral principle. The Categorical Imperative guides your moral actions by asking whether you could expect all people to rationally act in the same way. Thus if you commit a crime you have made a mistake of reasoning. Therefore intelligence (defined as the capacity to be rational) is necessary for moral responsibility/moral agency. I do not endorse this claim as I argue and will show that intelligence has very little to do with moral responsibility/moral agency.

To conclude, associating moral responsibility/moral agency with intelligence is problematic in standard ethics and may only be part of what is needed for human (let alone machine) moral agency. Accordingly, alternative definitions of morality are explored, which could better suit machines, specifically, behavioural and relational definitions. What claim two inadvertently highlights is that regardless of how morality is defined, babies and young children, and sometimes teenagers, are not accorded moral agency, because they have not yet reached a sufficient level of development. Thus children are moral, but do no currently manifest moral agency, although they do have the potential to one day manifest it when they reach maturity. This ascribes a dispositional morality to children. Thus morality in traditional ethics is thought of as a developmental construct, and as such, something that

entities can develop as they evolve. This developmental view fits well with my argument that as technology advances it may advance to the point at which we are able to consider machines moral agents. The case of babies and small children shows that morality can be a dispositional property. Moral agency can therefore be a developmental construct and a dispositional one. We could see sophisticated robots today analogously to babies in terms of development. As artificial intelligence develops, robots could advance and become the equivalent of teenagers in the not to distance future and further into the future, even reach adulthood. On a developmental basis is it worth considering the ethical implications of robots advancing to the point where they are similar to adult humans. I do not consider whether robots are ready, or will ever be ready, to be accorded moral agency. This is what many machine ethicists try to do. Rather, I focus on the circumstances in which they could be accorded moral agency.

Moral grey areas such as whether we sentence teenagers in our legal system as adults or not, illustrate that often our judgement on ascription of moral agency in cases where entities are more or less developed, is guided by a matter of perspective. What makes some courts and legal systems, try a teenager for a crime as an adult, while others not? In certain states in America 16 year olds get tried as adults and sentenced to adult correctional facilities for punishment. Extending this thought, Floridi and Sanders call for a change of perspective on the way we should view machine agency. They suggest a mental sense-making tool to give us new perspective which they call a 'level of abstraction' (Floridi and Sanders in Noorman, 2014). When viewed at the appropriate 'level of abstraction' artificial agents can be described as 'being interactive, autonomous and adaptive' and thus sources of moral action and accountability (Floridi and Sanders in Noorman, 2014). To illustrate this concept Floridi et al use the following example. Suppose we join a conversation amongst 3 people half way through without knowing the topic of discussion. Anne observes that an anti-theft device is installed and it has had only one owner; Ben that its engine is not the original one and that the body has recently been repainted and Carole that it has a good resale value but that its spare parts are expensive. Each views the object under discussion relative to their own level of abstraction, i.e., according to their own interests and contexts. We guess they are talking about a car (or a motorbike or a plane) and that Anne's level of abstraction on what she has observed fits the mind-set of the owner, Ben's that of a mechanic and Carole's a car

insurance broker. Each view (level) is a collection of observables which makes it possible for us to analyse the system and develop a model of it or a series of models. It suggests that agency is a matter of perspective (or a number of different observables that form various perspectives). Floridi et al's example can be criticised in that it illustrates a change of context rather than a genuine level of abstraction. Anne, Ben and Carol practically identify features of the car important to each of them given their specific and differing interest in the car. There are no different levels of abstraction as the level of complexity is identical and hence it is the same abstraction. Rather the three people have different interests and therefore focus on different contexts for the car. Understanding a level of abstraction is akin to deciding how many details you put into a map. Do we just show the African continent, or do we include lines of latitude and longitude, or even contour lines to indicate topography and legends to indicate average temperatures and rainfall? A car engine viewed at one level of abstraction could simply indicate where the engine fits in relation to the overall chassis. Or viewed at another level, it could show petrol and oil mixing and powering the pistons of the engine. If we use levels of abstraction in this sense, rather than the car example suggested by Floridi et al, on certain perspectives, we are more likely to believe that machines are artificially intelligent and can have a type of agency, than on others. For example, we could have a telephone conversation with a machine and not realise that is it a machine because we can only hear its voice which sounds human. If we had a face to face chat with the machine, it would be obvious that we were having a discussion with a computer as physically we would see that a computer or audio-device was talking to us rather than a person. If technology evolved to the point where machines (e.g., Androids) looked and acted like us, would we accord them agency based on this level of abstraction? At this level of abstraction we would be unable to see the inner workings of the Android and externally it would appear human. But if we saw the Android at a different level, under the false skin, we would see a host of hardware, circuitry and a computer. On the first level of abstraction we may think the Android human and accord it agency. On the second we would know it was a machine and therefore not human and as such not a moral agent.

To conclude, definitions of agency based on intelligence are problematic and it may well be that machine moral agency calls for a different type of definition altogether. The definition would need to encompass a notion of agency as a developmental construct and should steer

away from qualities that are innately human. I now review the way things behave, instead of what things are, and corresponding behavioural arguments for machine agency as a possible alternative to definitions based on intelligence.

2. Behavioral arguments for moral agency

Machine ethics is a discipline which focuses on building machines that behave enough like us to be said to be moral agents. Moor and other machine ethicists like Michael and Susan Anderson focus on creating computer systems that can behave *as if* they are moral agents. While the challenge in some respects shifts from an ethical to an engineering one, the underlying assumption of machine ethics is that if machines can be said to behave like us, we could morally treat them the way we treat ourselves. Much of this research involves deconstructing how a machine processes information and equating its programming and processing ability to analogous human processes used in moral decision making. These ethicists refer to machines as 'Artificial Moral Agents' (AMA's) which they aim to engineer to make decisions which 'honour privacy, uphold shared ethical standards, protect civil rights, individual liberty and the welfare of others' (Allen et al, 2006, p13). This is a tall order for most people to live up to and certainly an ambitious task for programmers who will need to make such beneficent, artificial moral agents. Many theorists believe that AMA's exist today that can be programmed to have learning algorithms and ethical sub-routines which enable them to make a type of moral judgement of their own (Allen et al, 2006, p14; van den Hoven and Lokhorst in Moor, 2006, p20; Andersons in Moor, 2006, p20). A glance at the latest conference schedule for the Association for the Advancement of Artificial Intelligence (AAI)[5] held in Québec, Canada from 27 to 31 July 2014 provides interesting insights into current topics of research in the AI field, associating machines with human-like qualities:

The Importance of Cognition and Affect for Artificially Intelligent Decision Makers

Common and Common-Sense Knowledge Base for Cognition

Driven Sentiment Analysis

An Action Language for Belief-Based Cognitive Robotics in Continuous Domains

How Long Will It Take? Accurate Prediction of Ontology Reasoning Performance

[5] http://www.aaai.org/Conferences/conferences.php [online] Accessed 25 July 2014.

Moor (2006, p20) suggests a continuum concept of moral agency, where machines can qualify for limited moral agency at lower rungs of the scale. He identifies different types of moral responsibility which roughly relate to the three criteria for agency on a standard definition, namely:

- Implicit responsibility (a causal connection between the agent and its action),
- Explicit responsibility (agent ability to consider the consequences of its actions), and
- Full responsibility (agent freedom to choose to act).

On Moor, implicit moral agents are programmed to behave ethically or at least, avoid unethical behaviour. For example, auto-pilots in an aeroplane are programmed to ensure passengers arrive at their destination on time and safely. Explicit moral agents go further and use ethical principles to calculate the best course of action when confronted with an ethical dilemma. Examples of explicit moral agents are mostly in prototype phase, such as WD (named after the moral realist WD Ross), created by the Anderson's. WD makes decisions on *prima facie* duties and selects the relevance of the duty to the context using its learning algorithm which allows it to compare similar and dissimilar cases involving usage of those duties (Moor, 2006, p20). The resulting action is analogous to ethical behaviour. Defined on such actions machines can be said to have a type of moral agency. To summarise, implicit moral agents are programmed to behave ethically; whereas explicit moral agents can make ethical judgements using certain ethical principles. Neither is able to reasonably justify the judgements made; a quality which adult humans possess and which Moor calls full ethical agency. Dreyfus calls this the ability to judge one's judgements or have an opinion on one's opinion (Dreyfus, 1988, p2-8). Moor concludes that full moral agents can reasonably justify their ethical judgements and that machines are not quite there yet, and may never be. He believes however that machines can be limited, explicit moral agents and at least help prevent unethical outcomes. Machines do not have human consciousness, intentionality and free will and cannot have full ethical responsibility, but they can be partially morally assessable on explicit definitions (Moor, 2006, p20). Moor's full moral responsibility is similar to strong definitions of intelligence, again relying on qualities that are innately human, which would be difficult to extend to machines. His concept of a

moral continuum is far more useful for my purpose. I agree with Moor that is it useful to think of sophisticated machines as having a limited type of moral agency (in the same way that my working definition of machine autonomy indicates a degree of autonomy). In this sense, I advocate treating machines in the way we treat morally grey entities such as teenagers. Like teenagers, machines are not quite ready to be held fully accountable for their actions but they can be held accountable in a limited sense. Moor's continuum aligns well with my argument of moral agency as a developmental construct. Machines may well advance to the point of full moral agency with future improvements in technology (or they may not) but the continuum allows us to accord full moral agency to these entities if they someday do advance to this point.

Computer science and philosophy professors, Michael and Susan Anderson, try to prove Moor, using a range of experiments that show 'that it might be possible to incorporate an explicit ethical component into a machine' (Anderson, 2007, p25). They run a series of experiments to demonstrate that artificial intelligence can at least achieve Moor's minimal definition of agency. Their creation of EthEl, a system which looks after elderly patients, reminds them when to take their medication and judges when their refusal to do so is serious enough (potentially damaging enough to the patient) to override the need for patient autonomy by notifying a human overseer. EthEl is their real-world example of an explicit artificial agent. It is able to manage the ethical dilemma of when to sacrifice patient autonomy in favour of patient health. The Andersons' conclude that EthEl is programmed to follow certain types of ethical decision making, particularly act utilitarianism 'at least as well as human beings and, perhaps better' given the data processing and computational skills needed to weigh up the greatest overall harm or overall benefit in determining the best ethical course of action. In such instances, it may be easier to define 'harm' versus 'benefit' for a machine than the concepts of 'personhood' or 'virtue' (Anderson, 2007, p18). On this reading, machines are better suited to making, or processing, ethical decisions because of their data processing and computational abilities. This supports my previous claim that simple-rational definitions of human moral agency are very machine like. A more intuitive problem with most machines developed by machine ethicists is the nature of the machine itself. None of them are very compelling, even Paro who is designed to respond to affection and touch, can be considered patronising in comforting patients with dementia. Certainly

few would relish living in a nursing home for the elderly with EthEl as a caregiver. As a result these machines invite fairly obvious and harsh responses from their critics. After all, who would want a machine to tell on us in our old age if we do not take our medicine? If we look beyond the feelings of indignation at EthEl however, the Anderson's have created a machine that is an explicit moral agent on Moor's definition.

Machine ethicists like the Anderson's believe that the type of approach towards ethics in general guides the type of programming of ethics into machines. For example, generalism involves principle based approaches towards moral reasoning and EthEl is programmed using this construct. Programmes, such as Truth teller compare the results of actual ethical cases to figure out a case-based approach to moral consideration using the ethics of particularism (Anderson, 2007, p21). A criticism of machine ethicists like the Anderson's is levied by Dreyfus, who says that early work in AI suffers from the 'continuum problem' which is the belief that if initial research shows that machines can process simple forms they should be able to process complex ones (Dreyfus, 1988, p7). Bar-Hillel calls this the 'fallacy of the first step' which plagued the early, headily optimistic days of AI when researchers experienced initial breakthroughs which led them to believe that progress would be incremental (Dreyfus, 1988, p7-8). As a result, many machine ethicists were overambitious on what technology would be able to do in the future. In a similar way, perhaps EthEl and other such machines inject a false sense of optimism on what the future holds? Objections to AMA's are acknowledged by some machine ethicists as practically insurmountable currently. Moor's rebut is that just because machines cannot currently meet requirements it does not follow that they may never be able to (Moor, 2006, p20-21). It however remains doubtful whether ethics is as computable as machine ethicists believe. There is the possibility that key ethical principles could be 'computationally intractable, putting them beyond the limits of effective computation because of the essentially limitless consequence of any action' (Allen et al, 2006, p16). However in standard ethics there is often disagreement and hence 'limited understanding of what a proper ethical theory is' (Moor, 2006, p21). It would be unfair to judge machines on theories that we ourselves cannot agree upon in human moral decision making. Interesting objections come from Moor (2006, p21) and Allen et al (2007, p20) that the biggest challenge to machine ethics may be the challenge of ethics itself. 'The implementation of ethics can't be more complete than accepted ethical theory' (Allen et al,

2007, p20). Debate still ensues between ethicists on what the correct ethical action is on issues such as abortion, euthanasia and suicide, to give a few examples. This is similar to my claim that it is problematic to associate machine moral agency with hard to define capabilities on strong definitions of intelligence (e.g., free-will and consciousness). Ethics itself lacks agreement on the types of actions said to be ethical. Why then do we use such stringent standards in objecting to its applicability to machines?

I agree with Floridi that a degree of latitude and a different perspective is needed in extending agency to machines. Consider the following scenario. How would you respond morally if you discovered your boyfriend was a *Replicant* (Scott, 1982)? Replicants are examples of wet-AI, genetically engineered, organic robots. They are grown in a laboratory, but they are made of the same material that we are. They are virtually identical to an adult human, but have superior strength, agility and variable intelligence depending upon the model. Replicants vary from humans in terms of their emotional responses and empathy, which can only be detected by answering questions posed in the fictional Voight-Kampff test. If your boyfriend was a replicant, but neither you nor your boyfriend knew this and you found out that he was via a Voight-Kampff test, would you treat him differently? Up until the point of the test, you had both operated as normal adults and held each other accountable for your moral actions. Taking the test has changed nothing about your boyfriend, he is still the person you fell in love with, but he is not human. Like you, he fulfils standard definitions of moral responsibility, namely he is able to consider the consequences of his actions and he is able to freely choose to act (Eshelman, 2009; Jonas 1984; both in Noorman, 2014). He meets standard definitions of moral responsibility, and as such, should be treated as fully morally accountable. A different consideration asks whether with technological enhancements in bioengineering a 'theseus's ship' scenario is possible in the future. The riddle of Theseus's ship asks us to imagine a wooden ship restored by gradually replacing all its planks and beams (and all other parts) with new ones. The old parts are reassembled to create another ship identical to the original. The paradox is whether the ships remain the same as the original? There are now two ships, one remodelled and one reassembled, yet the resulting ships are clearly not the same (Deutsch, 2008). What would happen if over time various parts of us were replaced by organic, wet-AI parts until no original human parts remained? We still have the same physical appearance and the same

psychological and physical functioning, but are we the original? Practically, we have been replaced over time, from our human-parts, to organic parts grown in a laboratory. Would we deny ourselves the moral accord we had previously taken as a given when human because we have replaced our original human parts?

2.1 The Turing test

In the 1950's computer scientist, Alan Turing provided one of the first working definitions for artificial intelligence. His thought experiment, called the imitation game or Turing Test, is designed to see whether a machine can fool a person into thinking it is a person. If yes, the machine passes the test and as such, demonstrates a sufficient condition for intelligence (French, 2000, p116). Turing's test however is not a moral test; it merely tests for machine intelligence based on behaviour. I am assuming that Turing's test could be used as a moral test by replacing conversational questions used in the original with questions that indicate moral awareness. Some have proposed such a Moral Turing Test (MTT) which would be less dependent on conversational skills than the original Turing test and would use pairs of actual, morally significant actions as a test for moral intelligence (Allen, Varner, Zinser in Allen et al, 2006, p16). In Turing's test a man, woman and interrogator sit in separate rooms. They cannot see each other and communicate via teletype (today's version of email). The man pretends to be a women and the interrogator asks both a series of questions to try to ascertain who the real woman is. Unbeknown to the interrogator and the woman, the man is replaced by a machine half-way through. If the interrogator remains unable to notice the difference and cannot distinguish the real woman from the machine, the machine has passed the test. Turing believed 'there would be no reason to deny intelligence to a machine that could flawlessly imitate a human's unrestricted conversation' and that his test provided a sufficient condition for intelligence (French, 2000, p116). Thus the test of machine intelligence was in the eye of the beholder, it was up to us to judge. If we believe a machine converses like we do, it was in this sense intelligent like us. By extension, machines that pass a similar test with questions that indicate moral awareness could be said to be moral agents. Turing believed that machines are also capable of learning and makes reference to child machines that are able to learn and become more advanced, adult machines (Turing, 1964, p29). In this sense, he sees intelligence, and by extension moral intelligence as a developmental construct. Using Turing's learning construct perhaps machines as they

advance could go through a similar developmental process to children, developing gradually to the point of sophistication where, like adult humans, they could be said to have the capacities necessary to be morally aware. The idea that a machine can develop morally is used metaphysically, as this is not yet the case. The notion of machine morality as a developmental construct is abstract and is different to that of human morality. A significant difference is that in the case of the human child, all the development potential is in that individual child: it is not yet a moral agent, but*, ceteris paribus*, it soon will be upon maturity. In the case of a machine, a current, individual prototype is not yet a moral agent and never will be. However, future versions of the technology used in the machine may have the potential to lead, to gradually better, different prototypes, which may approach moral agency one day. The entire theory of AI can thus be seen developmentally. The precise locus of development potential is not in the individual machine, but rather in future, advanced prototypes of the original machine. Thus, it is difficult to locate the precise locus of the moral development potential of AI.

Artificial intelligence experiments with machines programming them to learn and develop. For example, Cog, a humanoid robot developed by MIT's AI team, is designed to go through 'an embodied infancy and childhood' (Dennett, 1997, p358). Cog can see with its video eyes and react to people accordingly. It is programmed to make friends and learn by playing with real things with its hands. This is all intended to enable Cog to acquire a memory and the team even plan to give it a virtual neuroendocrine system with virtual hormones (Dennett, 1997, p359). A search on more contemporary literature does not reveal whether the Cog team were successful with their plans. However, the team's web site[6], lists current projects as project Meso, a biochemical sub-system for a humanoid robot, a project on the Theory of Mind for a humanoid robot and Lazlo the humanoid face project. This is a good illustration of the problem of attributing a dispositional concept of morality to machines. Cog did not develop into anything in particular, and neither did the program of which the machine was a part. Nevertheless, perhaps some knowledge gained through Cog made its way into the other projects that the team is now involved in. Thus, it is in fact our AI knowledge that develops. As we work on more AI projects we gain experience to channel into the next

[6] http://www.ai.mit.edu/projects/humanoid-robotics-group/cog/current-projects.html [online] Accessed: 15 February, 2015.

project. This is the continuity and abstract concept of machine moral development. As our knowledge of technology develops, we in turn develop new and more advanced machines that may one day be able to be accorded moral agency. Further afield, examples of humanoid robots, outside of MIT include Aiko, a female Android, built by Canadian-Vietnamese designer Le Trung in 2007. Aiko is clothed in a silicone body, weighs 30 kg, measures 152 cm, and metaphorically 'speaks' fluent English and Japanese (in the sense that Deep Blue could be said to play chess). She is also skilled in cleaning, washing windows and vacuuming. According to her creator, she 'reads' books and distinguishes colors, knows how to learn and remember new things. This illustrates the hyperbole surrounding the description of AI by those who create it. No machine 'speaks' any human language yet, nor does it 'read' books and so forth. Aiko can respond to predetermined queries with predetermined answers. To be fluent in English, Aiko would have to pass the Turing Test. As such the described competencies are really tantamount to Searle's Chinese room (see an explanation of Searle in section 2.2). The following institutes are also attempting to build humanoid robots:

- Humanoid Interaction Lab, Tsukuba, Japan
- University of Waseda's Humanoid Robotics Institute
- Honda's Humanoid Robot

In addition to the MTT proposed by Allen, Varner, and Zinser (in Allen et al, 2006, p16); a myriad of variations of Turing's original test have been put forward. In the Triple Turing test (TTT), proposed by Harnad (in French, 2000, p119), the screen is removed between the rooms, so that both the physical appearance of the machine and its behaviour (conversation) are observed. Obviously Harnad wanted to remove the screen to minimise the ability to fool people into thinking a machine is a person. Certainly we are many, many years away from producing AI that both talks and looks indistinguishable to us, but it is worth considering Harnad's test, in such a future. If Harnad's test included moral questions, by extension we could say that we accord machines that passed it moral agency as they would be indistinguishable from the average moral agent both visually and behaviourally in their moral judgements. In the future we may not be able to distinguish man from machine. In such instances, Floridi would say that our perception of machines is no longer machine-like.

People would naturally start treating machines like people because they believe them to be like us. It does not matter whether they are merely following software programmes (or manipulating symbols like Searle, who objects to Turing) suggests. We affiliate with them regardless. Strawson would say we are unable to do otherwise due to our emotional natures which determine such reactions. This negates concerns on whether we morally ought to behave in this way.

2.2 Objections to Turing: Searle's Chinese room experiment

Searle agrees with Turing that you can fool people into thinking a machine is a person, but unlike Turing he says when the screen is removed, we would be unable to see the machine as anything but a machine, regardless of its levels of intelligence, or on MTT, regardless of its levels of moral awareness. By way of objection, Searle uses his own thought experiment, the *Chinese-Room* to illustrate his rebuttal (French, 2000, p119). Instead of a woman in a room and a man and then computer in the other, a native speaking Chinese person sits in one room and an English speaking person in the other. Only the English speaker knows that he/she cannot read, write or speak a word of Chinese. The Chinese person writes a question in Chinese characters on a piece of paper and slips it under the door. The English speaker has a code-book of sorts which enables him/her to match the Chinese characters to the English equivalents. Using this method the person translates the question and is able to answer it in written, Chinese characters. Resulting in a situation where the Chinese person thinks the English speaker is Chinese (or at least able to understand Chinese). Searle's test shows that while people can behave like machines, i.e., they can process symbols, the converse cannot be said to be true. Turing's machines are really just processing code, like the Chinese symbols, without thinking or being human. 'These agents are designed by their creators to mindlessly manipulate symbols that are perceived by naïve observers to be indicative of an underlying mind – but underneath there is little more than Searle's infamous rulebook' (Bringsjord et al, 2001, p3-4). So on Searle's reading machines cannot have moral agency, they merely execute programmes defined by binary code. There is no deliberation from the machine. Rather the deliberation on what code to write to allow the machine to execute, takes place in the mind of the person who writes the programme (the programmer). Hence the morality of the system is derivative of that of the programmer, a defining feature of a value sensitive design approach towards systems development (Brey,

2010, p47). A value sensitive design approach involves designing systems with ethical considerations in mind upfront. I do not cover this approach in my report as my concern is with the behavioural aspect of technology and whether it can have moral agency, rather than on the design of the technology itself.

Analogously, we can ask whether a computer really plays chess, or Googlecar really drives, or whether both are just executing their programmes. Is this still a case of Searle's Chinese room syndrome? Moor says no, even if ethics are derived from developers' programming, we can still evaluate machines as limited (explicit) moral agents. Chess programmes receive their chess know-how from humans but we still regard them as chess players (Moor, 2006, p20). When IBM's Deep Blue chess computer beat world champion Gary Kasparov in their first game in 1996 'the designers of Deep Blue didn't beat Kasparov, Deep blue did...it discovered the winning sequence of moves...and recognised and exploited a subtle flaw in Kasparov's game' (Dennett, 1997, p352). Even if software programmes operate machines, when we see them performing such actions, we identify those actions as playing chess or driving a car. We see their actions as their own behaviours. As the acts machines perform become increasingly disassociated from the software developers who built the drive car/play chess programmes, at some point we no longer care about the origin of the internal mechanism and we no longer understand it. It is easier to think of machines driving cars and playing chess by themselves. When such behaviour becomes complicated enough and has consequences that are moral, we are prone to accord moral agency to the entities performing the action. In so doing, we accord machines what I term a robot-friendly, moral agency. However, just because we may be inclined to accord machines moral agency on this basis, does not mean that we should accord it to them. There is no appeal to an ethical principle or concept to justify the moral rightness or wrongness of doing so. As such, we would be unable to say whether we were right to do so or whether this was a moral mistake.

To illustrate, suppose you had a cat. Would you react differently to the cat if you removed its skin and saw that underneath it was made of hardware, software and nuts and bolts? Searle would say when you saw that the inside of your cat was mechanical; you would no longer see your cat as a cat but as a robot, despite it having a biomass on-top of it. Taking this further, what if your father had surgery and you realised like the cat, he was a robot? Searle would again say that once you knew your father had mechanical innards, you would

be unable to see him as your father and would start thinking of him as a machine. The robot was built by humans, and as such, we know what their internals are, and with this knowing, we would be unable to forget. Searle would say of TTT, that it does not matter that the entity looks like and behaves like your father, once you knew it was a machine you would see it as such. Conversely, Turing and proponents of the TTT as a viable test of intelligence would say that there may be an adjustment period, but at some point you would forget that your father (and the cat) had mechanical innards and you would be unable to see them as anything but your father and your cat because they behaved like them (Turing) and behaved and looked like them (TTT). This example is fairly raw; most people would be unable to look past the image of mechanical innards when they looked at their father and their cat. Using a different example, what would happen if your father was cloned? The clone is made of human tissue, in fact it is made from cells cloned from your father, but it was grown in a petri-dish. Despite being an exact replica it is not your actual father. In such instance I am unsure what Searle's positon would be. Cloning could be seen as relevantly different from a machine, after all, the clone does have consciousness, understanding and all the other mental states that the original has. The point of the Chinese room is that the room, contrary to a person, does not 'understand'. If the Clone understands, then it passes the test. Turning and TTT would say we would forget the clone was a replica over time and think of it as our father. What would happen if your real father was rebuilt, he started off human, but over time each piece of him was replaced by wet-AI and genetically engineered tissue? He is no longer your original father. Do you still think or him as your father or not? My theory, similar to Turing and TTT, is that at some stage, we would forget that they were robots, clones or not the original if they behaved and acted the same. My intuition says that it would be easier to accept the rebuild scenario as this happens more slowly over time and starts off as your real father, but there is no logic to this intuition. Additionally, to say that TTT is an MTT is to assume that all that matters is that the robot, clone or non-original morally behaved like your father. I am unsure on whether I agree with this. Would I be able to view a robot that looked like a robot, but behaved like my father, as my father despite physical appearances? Intuitively, I would have to say no. I would need a combination of an MTT and TTT to continue to see the robot, clone or non-original as my father.

I am aware of the pitfalls of appealing to intuition in philosophical debate. Different

theorists have different positions on whether and to what degree we should appeal to our intuitions when making moral decisions. Just because something is intuitive does not mean it is true. I appeal to intuitions to capture what I call core intuitions, those which are universally shared, for example, babies are not morally responsible. Although I try to avoid appealing to specific intuitions, which are less likely to be shared universally, my research topic is fairly specialised. Where I do appeal to intuition I make the assumption that these may be universally shared. Much of my report does ultimately appeal to descriptive accounts of the way we behave, as do theorists I reference to support my views, such as Strawson's relational account of morality. While I try to find reflective equilibrium between descriptively adequate appeals and normative, ethical concepts; I acknowledge that my report is more heavily loaded on empirical accounts.

2.3 The Lovelace objection to Turing

A key objection in ascribing moral agency to machines is called the Lovelace objection. Lady Lovelace (Ada Byron, daughter of Lord Byron) was a 19th century mathematician and scientist. She contended that 'only when computers originate things should they be believed to have minds' (Bringsjord et al, 2001, p3). Origination is underpinned by concepts of free-will and creativity. In computing terms, machines consistently achieving outputs they were not programmed to achieve, which the programmer could not explain, could be said to have originated their own outputs and thus created something. Bringsford et al (2001, p5) agree with Lovelace that computers cannot create (or originate) things because they are merely following lines of code programmed into them. They do what we have programmed them to do. Essentially the Lovelace test boils down to saying computers will remain as Searle described them in his Chinese room experiment, capable of manipulating symbols, not actually creating. Even if a car assembly robot accidentally installed the incorrect part (a spare tire) to the bumper of a Toyota Camry and the position of the tire inspired designers to start designing a hybrid sedan-sports-utility-vehicle, we could still not credit the robot with originating the idea (Bringsford et al, 2001, p5). In other words, creativity is not simply random; it has to be non-random and intentional in order to be creative. Since the car assembly robot accidentally installs the incorrect part, the act is a random event and as such not creative. Turing counters that machines can create and that they 'take me by surprise with great frequency' which Moravec, in support, describes as being attributable to software

and coding bugs that 'are never fully tamed' (Turing and Moravec in Bringsford et al, 2001, p5). This is a weak re-joinder as computer bugs do not indicate machine creativity; rather they indicate a mistake made by programmers. Mistakes are random accidents and therefore not creative acts. The surprise Turing refers to is related to the fact that he did not know about the bug. This is an appeal to the ignorance of not knowing, rather than an indication of creativity on the part of the computer. The existence of computer bugs is therefore not a creative act and thus insufficient reason to say that machines are like people. Even if the unexpected behaviour is not due to a bug, it would still not amount to creative behaviour: often the complexity of software is such that we are unable to foresee a particular kind of behaviour, even though it is there in the code. In other words, complexity, and the epistemic problems it creates, is still not tantamount to creativity. It could be argued that because people build machines it is possible for us to know everything about them, thus machines cannot really surprise us. Post-hoc, we could trace why the machine installs the incorrect part on the Camry or how the software bug originated. Programmers programmed a set of rules into computer software which pre-determines what computers can do and therefore we can trace why software bugs occur. I argue that this does not matter. The point is that the perception of the act (creativity) overrides the internal mechanism that caused it (the software bugs). Due to the problem of many hands, we accord agency on the merit / demerit of our perception of the action, which we associate with the entity that performed it, regardless of whether the entity performing the action originated it. In reality, what people think they see they generally tend to believe? For example, much of the televised opening ceremony for the 2008 Beijing Olympics were rendered animations of fireworks displays and other activities which in reality did not happen at the physical event. TV audiences had no idea this was the case and believed the version they saw. Dennett says that it is our perception of the intentionality of something that gives it intentionality in our eyes (Dennett in Schmidt et al, 2006, p73-80). Standard ethical theories may well fall by the wayside in favour of popular opinions when we see AI behaving, and increasingly starting to look, like us. Although my claim is descriptive, and does not use normative principles to determine whether we ought to do this, Strawson says that appealing to such concepts would be irrelevant. People are emotional creatures and would be unable to suspend their reactive attitudes towards robots, clones or non-originals that looked and behaved like us (Strawson, 1972, p14). We would be unable to help our

feelings. If we felt the entity was our father, we would see it that way regardless of what his inner-workings comprised.

Bringsjord et al entertain a technical discussion on new technologies such as those described as 'creative' systems (BRUTUS, LETTER SPIRIT and COPYCAT); and an 'oracle' system (O-System) which may provide a machine that meets Lovelace's test. They however de-bunk any perceptions that these technologies actually exhibit creativity and free will, by explaining how their programmes work. Thus the creative system is merely a programme rather than the machine performing a genuinely creative act. Accordingly, they conclude that these technologies are not yet able to meet the test's requirements (Bringsjord et al, 2001, p21). Now that I am more expertly informed about how the technology above was programmed to behave, the rationale for these sophisticated forms of AI being unable to create or originate makes sense. However, had I not read these technical explanations (which I take as valid, particularly as some of the authors built the machines in question); I could quite easily (like the Chinese person), be fooled into believing that these machines were actually able to generate stories (BRUTUS), answer creative riddles (COPYCAT) and produce their own letter fonts (LETTER SPIRIT). I did not understand the internal mechanism causing the action and initially gave the machines moral agency. It dawned on me that I was guilty of taking over the appeal to ignorance, like Turing. Bringsjord et al built their machines, they know that their programmes are not really generating stories, nor answering creative riddles, nor originating fonts. Thus if we build something, we can always find out why it performs the way it does. Machines we build cannot be said to have agency as we built them and thus their actions are pre-determined. Treating something as moral pre-supposes choice and that the entity could act differently. Machines are acting on the programmes we give them, they are pre-determined and therefore lack agency on a definition of agency that rests on free will. However, Strawson and broadly the determinist school would say that everything is predetermined, including human moral action. On this reading, both humans and machines are pre-determined and therefore cannot be said to be moral agents. The problem with determinist, or incompatibilist schools of thought, is that they move in the direction of having to say human (and all) 'moral responsibility is impossible – since no plausible interpretation of free will can be provided. The problem…is that it seems both intellectually incredible and humanly impossible for us to accept or live with this conclusion' (McKenna et

al, 2008, p3). We do not stop holding each other morally responsible just because we may all be machines; or be unable to say what goes on in each other's thoughts and conscious minds; or because free will is complicated to define and may not exist. Something more is needed to bind us to sources of moral agency.

It is a (descriptive) fact about ourselves that the more things behave like us, the more likely we are to treat them like us, including matters moral. Doing this however, does not make it right. Even if we could treat the replicant like us, should we and why should we if we should? Animals, such as dogs, seem to have emotions, some of which are a bit like the ones we have. When we shout at them for doing something wrong, they cower, drop their eyes to the floor and behave as if they know something is wrong. Conversely, when we praise them they wag their tails and react favourably. However, given how incredibly difficult it is to correctly identify the emotion of say 'love' in a fellow human being; it would be very difficult to identify emotions across the species barrier with some degree of certainty. Contrary to popular myth, it is not any clearer that your dog 'loves' you, than your cat; or your canary; or your goldfish. Machines are far from being remotely like us or animals currently. The latest BMW is neither like a person nor an animal. Replicants are science-fiction and may only be a possibility in the very distant future (if at all). Extrapolating Strawson, whom I will discuss in detail in the section that follows, the rightness or wrongness of us treating machines who behave like normal adult humans is a moot point. On his view we would be unable to help ourselves from treating machines that look and behave like normal adult humans otherwise, because we are unable to suspend our feelings, what he calls our reactive attitudes, or range of emotions that we bring to bear on our interactions with them. 'In general, there is no question of us choosing or needing to justify the fact that we are liable to reactive attitudes and feelings. This is simply a given of our human nature as being fully human. Even if we had some theoretical reason to abandon or suspend these reactive attitudes it would be psychologically impossible for us to do this' (Strawson paraphrased in McKenna et al, 2008, p24e). Although Strawson applied his theory to 'ordinary inter-personal relationships' with 'other human beings', I extrapolate the applicability of his reactive attitudes to machines, specifically in future cases where machines could behave and look like us (Strawson, 1972, p5). Thus, if a machine behaved like us, we would be able to resent (or praise) it in instances where its actions demonstrated ill (or

good) will towards us. In such instances we would be unable to suspend our reactive attitudes in favour of an objective stance - a subdued, emotional stance which we are able to take in instances where entities do not qualify for moral responsibility, such as dogs, babies and severely mentally handicapped people. 'If your attitude towards something is wholly objective…you see it as an object and view it intellectually rather than with a sense of humanity' (Strawson, 1972, p9-10). I am presupposing that at some stage a machine's actions could clearly demonstrate ill will towards us, i.e., a strong definition of intelligence. If this is the case, then in Strawson's world, with machines that behaved like us we would be unable to suspend our reactive attitudes, or conversely, we would be unable to adopt an objective attitude towards them, which negates the question of what we ought to do. We inevitably would be unable to do otherwise. Presuming the Kantian principle that ought-implies-can, whether we ought to adopt an objective stance to machines in this instance is thus irrelevant. Strawson's theory is attractive for my argument as it says that morality is based on the feelings we have towards something. There are no human-centric, innate capabilities such as intelligence or creativity that we need to have to be morally responsible. Lovelace's objection is overcome. If I can feasibly demonstrate that we could resent (or praise) a machine, we can be said to accord it moral agency if we accept Strawson's approach to ethics. I now review Strawson and objections to his theory in detail.

3. Relational arguments for moral agency

Strawson's paper Freedom and Resentment attempts to resolve the classical debate on whether free will exists. The debate is important as many believe free-will is an essential element in according agents moral responsibility. Strawson tries to reconcile the concept of free-will with the various schools of determinism by arguing that their focus is too conceptual and claims they have ignored what really matters, namely, looking at what actually goes on in everyday life when we make moral decisions and hold a person morally responsible. He believes that the way we react to each other, both morally and non-morally, is deeply influenced by our perception of 'the attitudes and intentions that other human beings have towards us' (Strawson, 1972, p5). He names these 'reactive attitudes' which apply to all those involved in 'ordinary inter-personal relationships, ranging from the most intimate to the most casual' (Strawson, 1972, p6-7). Strawson presents a continuum of attitudes; the scope is as wide as the 'different kinds of relationships which we can have with

other people – as sharers of a common interest; as members of the same family; as colleagues; as friends; as lovers; as chance parties to an enormous range of transactions and encounters' (Strawson, 1972, p6). In moral terms we generally expect a degree of goodwill from all fellow human beings, both towards us directly as individuals and towards others. We in turn react with gratitude, or conversely if goodwill is not displayed, with resentment. Strawson distinguishes between 'personal' reactive attitudes to denote the way people react to us directly which he describes as 'others wills towards us', from that of 'impersonal or vicarious' reactive attitudes that relate to the way people react towards others which he explains as 'others wills towards others' (Strawson, 1972, p7). In this context, moral agents are those who are bound by 'self-reactive' attitudes such as 'feeling bound or obliged; feeling compunction; feeling guilty or remorseful or at least responsible; and the more complicated phenomenon of shame' (Strawson, 1972, p16). All three types of attitudes (personal, impersonal and self) are humanly connected.

McKenna and Russell refer to Strawson's descriptions of how we react to each other relationally as his 'naturalistic-way' (McKenna et al, 2008, p9). Strawson himself calls this his 'commonplaces' – descriptive accounts of the human emotional world. Contrary to rationalist views, he sees emotions as pivotal in determining the way we react morally. Being moral is about common human emotions, particularly our reactions and attitudes towards how we think others see us and how we, in turn, see people treating each other (Strawson, 1972, p24). In situations with a moral outcome, the importance we attach to someone's motivation for acting, is paramount and impacts the way we morally punish, reward or even excuse their actions. To illustrate, he uses an example of someone treading on your hand accidentally while trying to help you, as opposed to treading on it with a malevolent wish to injure. The physical pain felt is the same in both scenarios, but you resent the malevolent act and not the accidental one (Strawson, 1972, p6). People who intend us goodwill receive our gratitude even if their moral action has a negative consequence for us. Thus the way we feel about the intention of the agent, determines our moral reaction to its actions. Strawson makes clear that his interest in reactive attitudes constitutes a return to the moral sense tradition in ethical theory (McKenna et al, 2008, p8). He is derivative of 18th century predecessors such as David Hume and Adam Smith who believed that our moral actions and reactions were a given part of our human natures

(McKenna et al, 2008, p8). Hume was interested in explaining morality as an existing natural phenomenon and sought to displace *a priori* conceptions of human nature and morality with an approach in which everything about us is open to empirical investigation and to explanation in naturalistic terms. For Hume, moral judgments are essentially the deliverances of sentiment. We recognize moral good and evil by means of certain feelings: the calm pleasure of moral approval or the discomforting displeasure of moral disapproval (Lara, 2014). Smith proposed a theory of sympathy, in which the act of observing others, or more specifically, what we imagine they would feel in the circumstances, makes people aware of themselves and the morality of their own behaviour (Fleischacker, 2013).

For Strawson resentment is one of the most powerful reactive attitudes in moral decision making and he focuses on instances where this reactive attitude can be suspended. He comes up with a list of exceptions, or what he calls 'special considerations…..where the offended person might naturally or normally be expected to modify or mollify this feeling of resentment or remove it altogether' (Strawson, 1972, p7). The first exception is 'when the agent's conduct lacked any degree of ill will or disregard' (McKenna et al, 2008, p5). The action was accidental, inadvertent or unintentional. Strawson specifies cases where a moral wrong or injury was caused but the agent 'didn't mean to do it; hadn't realised; didn't know; couldn't help it; was pushed; had to do it; it was the only way and they had no alternative' (Strawson, 1972, p7-8). In these situations we still see the agent as a fully responsible moral agent, but we see the injury his action has caused as being one for which he/she was not fully, or at all responsible. It is the external circumstance that was inappropriate, rather than the agent's behaviour. The second group of exceptions is very different; the agent is an inappropriate source of moral responsibility, rather than his/her actions. Within this category of exceptions, there are two different stances we can take. A 'participant stance', where we believe the agent is generally a normal adult who is an appropriate target of our reactive attitudes, but who has performed an action out of character due to abnormal circumstances. The agent was under 'abnormal stresses; wasn't himself or he had been under very great strain recently' (Strawson, 1972, p8-9). In such cases we may temporarily suspend our normal reactive attitudes towards the agent's actions by feeling less resentment. An 'objective stance' occurs when the circumstances are normal but the agent is 'psychologically abnormal or morally underdeveloped' and when an agent is 'warped or

deranged, neurotic or just a child' (Strawson, 1972, p9). We modify our reactive attitudes profoundly when we see an agent in this light as we believe they are incapacitated for normal adult relationships and so our reactive attitudes are inappropriate for individuals of this kind (McKenna et al 2008, p6). The agent cannot be an agent and becomes a subject for a wide range of treatment (e.g., mental asylum) or as an object of social policy (e.g. children's rights). Strawson's objective attitude is congruent with traditional ethical views where mentally retarded adults and children are not considered to have the capabilities necessary for moral responsibility. He goes further, saying that our ability to adopt the objective attitude (to suspend or moral reactive attitudes) is a 'resource' available to us which we sometimes use with normal adults considered morally responsible. He is rather vague on why we would do this, but mentions we may 'as a refuge…from involvement or as an aid to policy or simply out of intellectual curiosity' (Strawson, 1972, p10). In such cases, we cannot suspend our reactive attitudes 'for long or altogether'…because we are 'human' (Strawson, 1972, p10). Although different, the participant and objective stances are not mutually exclusive and we can shift between them. Strawson uses the example of parents raising children to illustrate. 'Children are potentially, and increasingly as they mature, capable both of holding, and being objects of, the full range of human and moral attitudes, but are not yet truly capable of either' (Strawson, 1972, p20). To treat such entities, a type of compromise is needed, shifting between sometimes viewing children objectively being incapable of moral attitudes and sometimes viewing them as almost capable of these attitudes. The shifting between each stance leads to a type of moral duality, which Strawson reflects upon using the punishment of a child as an example, saying it is both like and unlike the punishment of an adult (Strawson, 1972, p20).

On Strawson's schema, the moral status of something is defined by whether we can feel resentment (or praise) towards it. We resent those who we perceive display ill-intention towards us that do not fit into Strawson's list of exclusions. The attitude of resenting is tied to our perception of the intention of the agent, and whether that agent has the capacity to be a moral agent, rather than directly to the agent's action. For example, if a dog bit me unprovoked, I would be angry as this is an initial human emotion towards being harmed. However, I would suspend my anger and resentment, as a moral reactive attitude, because dogs are not the subject of moral agency. There is subjectivity in some of Strawson's list of

exceptions, on when we suspend resentment (and thus moral judgement). His examples where people were not thought to be themselves or where they had been under pressure or great strain are difficult to determine objectively (Strawson, 1972, p8). If a mother was under tremendous pressure at work, came home stressed, tripped over her son's toy on the kitchen floor by mistake and then proceeded to beat her son in frustration; would I be able to adopt an objective attitude as an adult sibling to what she was doing to my younger brother because she was under extreme stress and therefore not herself at the time? I would find it difficult not to resent my mother and may, depending upon the intensity of the resentment I felt, be unable to put these feelings aside and adopt an objective stance. I would almost certainly resent a teenager on the eve of becoming an adult, who robbed and physically assaulted me. In both situations, I would be unable to get over my feeling of resentment and would thus, judge both my mother and the teenager as having committed a moral wrong towards me (and my brother). As an old aged pensioner, living in a nursing home, where I had no choice but to be looked after by mechanical nurse EthEl, would I resent it if it treated me like a child and called a supervisor because I decided not to take my medicine? In all likelihood, I would resent the situation, the supervisor, the programmers who made EthEl and probably EthEl itself. Would the fact that EthEl is a machine allow me to view my mechanical nurse objectively and be unresentful? I may be able to, but I may also not be able to suppress my feelings of resentment.

Machines currently may behave more like dogs or small children and are not yet at the stage where they can be said to behave like adult humans. Strawson's example of switching between treating children objectively because they are not yet adults capable of reactive attitudes and occasionally treating them as entities that are developing such attitudes, suggests that we may be able to view machines in a similar, developmental manner. What happens when robots advance to the point where they can interact with us in ways similar to our interpersonal interactions with each other? On Strawson's view, we may very well be able to resent such a machine if it does us harm. Even if we know it is just a machine, we may be unable to supress our feelings of resentment as we are emotional creatures. Say we were able to adopt an objective stance; we may slide back into resenting the robot, as we are unable to adopt this stance indefinitely. 'A sustained objectivity of inter-personal attitude, and the human isolation which that would entail, does not seem to be something

human beings would be capable of' (Strawson, 1972, p12). Therefore if we resent the robot, by so doing, we accord it moral agency and are able to sanction its actions. Resentment occurs when we adopt a participant attitude, or conversely, when we are unable to adopt an objective attitude towards something. Shabo believes that because reactive attitudes for Strawson are 'a condition of our humanity' (Strawson, 1985, p33; in Shabo, 2012, p131), suspending them is not a real possibility for us. While not strictly impossible, Strawson concludes that the wholesale renunciation of these reactions is 'practically inconceivable' given our actual psychological makeup (Strawson, 1962, p54, in Shabo, 2012, p132). Shabo calls this the 'Practical Inconceivability Claim' and refines Strawson's notion of cases where we are unable to adopt the objective stance in normal adult relationships, namely, in 'important kinds of personal relationships, such as mature friendship and reciprocal love' between consenting adults (Shabo, 2012, p133). Strawson says that whilst objective attitudes are 'emotionally toned in many ways…they are 'not in all ways' (Strawson, 1972, p10). Objective attitudes 'may include repulsion or fear, pity or even love, though not all kinds of love…but cannot include the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal relationships; it cannot include resentment, gratitude, forgiveness, anger' (Strawson, 1972, p10). Thus if we can resent something, we are unable to view it objectively and as such view it with a participant stance. It follows that the more we are able to interact with machines in ways in which we interact with each other; the more likely we are to be able to resent their wrongdoing and cultivate moral reactive attitudes towards their actions. Strawson would say that the objection on whether we should take this stance morally, just because we are prone towards taking it is irrelevant because he presupposes Kant's so-called 'ought implies can' principle: the action to which the 'ought' applies must indeed be possible under natural conditions (Kant in Guyer, 2004). According to Strawson, we would be unable to stop ourselves from having these feelings and thus, taking this stance would be inevitable and not a choice at all. Strawson sees 'the existence of the general framework of attitudes itself as something we are given with the fact of human society [these attitudes]…as a whole...neither calls for, nor permits, an external, rational justification' (Strawson, 1972, p25). He believes that suspending them to fit into for example, a theoretical conviction of the truth of determinism, is pointless, 'it is useless to ask whether it would not be rational for us to do what is not in our nature to (be able to) do' (Strawson, 1972, p20). Comments such as these by Strawson

create the impression that adopting reactive attitudes, exceptions notwithstanding, is an evitable part of being human. Thus, if we can resent technology we can adopt a moral reactive attitude towards it, regardless of whether we should do so, we would do so.

Ihde proposes a number of relational typologies between people and technology. In his 'alterity' relation we experience technology as a being that is different to us, for example, Sony's robotic dog AIBO (Idhe in Introna, 2011). When we interact with AIBO, it seems to exhibit a world of its own. As we interact our world withdraws into the background and AIBO's into the foreground. The technology becomes the focal entity with which we engage. It seems that as soon as behaviour achieves certain richness and we no longer understand the internal mechanisms of the agency, at some point we may resent a robot that harms us and therefore accord it agency on Strawson's reading.

3.1 Critique of Strawson

Strawson's *Freedom and Resentment* has elicited much commentary from supporters and critics alike. I review salient aspects thereof. A central critique focuses on the question of whether moral responsibility can be based on our feelings and resultant reactions without appeal to moral propositions, principles or rules to guide our moral judgements. Some raise 'concerns on the open or non-committal nature of Strawson's theoretical commitments as they relate to moral reasons and justification within the framework of reactive attitudes' (McKenna et al, 2008, p11). McKenna believes Strawson is unclear on whether reactive attitudes are mere feelings or whether they also have propositional content, involving beliefs of some relevant kind. As Strawson's account depends upon a 'feeling theory of emotion' it is impossible for us to judge reactive attitudes themselves as either reasonable or unreasonable (McKenna et al, 2008, p11). It may be the case that our reactive attitudes are inappropriate, for example, we may have the incorrect beliefs about an agent's intention or state of mind. As Strawson provides no guiding principles that can universally apply, we are left with the relativism of our own judgement in moral situations. Similarly, he does not seem to take into account the possibility that it may be possible to change certain of one's reactive attitudes through reflection, training and education. For example, racism as a basis of some reactive attitude might be subject to education, reflection, moral enlightenment and so forth. Bennett offers a qualified defence of Strawson, saying that moral accountability and the issue of whether moral blame or praise is appropriate are a matter of

degree. He interprets Strawson thus: 'my feeling of indignation at what you have done is not a perception of your objective blameworthiness, nor is it demanded of me by such a perception. It expresses my emotional make-up, rather than reflecting my ability to recognize a blame-meriting person when I see one' (Bennett in McKenna et al, 2008, p73e[7]). However is emotional make-up monolithic, unchangeable and impervious to reason? For Strawson, the reactive attitudes are a part of life and according to Magill are 'not the sort of thing that can be given an overall justification or stand in need of one' (Magill in McKenna et al, 2008, p206e). Magill believes that Strawson represents an important departure from 'traditional debates' which focus on trying to answer what justifies treating people as responsible for their actions (Magill in McKenna et al, 2008, p207e). This shift from beliefs or judgements as a qualifier of moral responsibility, to attitudes is an important challenge to traditional notions of moral responsibility. If blaming people and holding them responsible were based on beliefs or judgements, it would always be possible to ask what conditions are required for such beliefs or judgments to be true or warranted. Conversely, 'attitudes are neither true nor false, and are not warranted by anything over and above their standard conditions of applicability' (McGill in McKenna et al 2008, p207e). Strawson does say that we can suspend our reactive attitudes to achieve a sense of objectivity, but we do this when we view an agent as being incapable of moral responsibility, or when we see the situation the agent finds him/herself in as mitigating, rather than because we are adhering to an objective set of rules or principles. In fact Strawson believes that we are not capable of a universal objectivity of attitude (McGill in McKenna et al, 2008, p208e). Thus it is a mistake to provide any 'external, rational justification' for moral responsibility (Strawson, 1972, p25). But he leaves the question open on whether there may be objective sets of rules or principles for when situations are properly considered to be mitigating. Watson's quote below seems to say that Strawson does not, however it remains Watson's interpretation, as there are no direct quotes from Strawson to support or refute whether there are rules by which we can evaluate our reactive attitudes. Perhaps Watson's stance is too strong, as Strawson is non-committal on this point.

Watson interprets Strawson's positon as follows.

[7] I have used 'e' to denote an E-book reference, as page numbers in the E-book differ from the paper-based copy of Mckenna et al, which I also reference in this report.

'All traditional theories of moral responsibility acknowledge connections between these attitudes (gratitude and resentment, indignation, approbation, guilt, shame, some kinds of pride, hurt feelings, asking and granting forgiveness, and some kinds of love) and holding one another responsible. What is original to Strawson is the way in which they are linked. Whereas traditional views take these attitudes to be secondary to seeing others as responsible, to be practical corollaries or emotional side effects of some independently comprehensible belief in responsibility. Strawson's radical claim is that these are constitutive of moral responsibility; to regard oneself or another as responsible just is the proneness to react to them in these kinds of ways under certain conditions. There is no more basic belief which provides the justification or rationale for these reactions. The practice does not rest on a theory at all, but rather on certain needs and aversions that are basic to our conception of being human. The idea that there is or needs to be such an independent basis is where traditional views, in Strawson's opinion, have gone badly astray (Watson in McKenna et al, 2008, p133-134e).

Sneddon argues a different view that 'a natural development of Strawson's position…has gone unnoticed…that we should understand being morally responsible as having externally construed pragmatic criteria, not individually construed psychological ones' which he believes runs counter to contemporary ways of studying moral responsibility (Sneddon, 2005, p239). Sneddon equates the deployment of reactive attitudes, and thus moral responsibility, with a 'social competence'. 'To be morally responsible is to fit into the social practises governing the deployment of the reactive attitudes' (Sneddon, 2005, p241). Thus he views Strawson's criteria for moral responsibility as externally constructed or what he refers to as 'competenceE' (Sneddon, 2005, p242). He believes that 'almost all major contemporary theorists pursue the individualistic approach, citing, Watson (2001) and Fischer (1999) as examples, whilst acknowledging that more recent thought experiments by Putnam and Burge suggest that mental content is at least partly determined by aspects of an agent's environment (Sneddon, 2005, p242-243). Individualistic approaches equate moral responsibility with individuals possessing certain psychological states/processes/abilities. Strawson is opposed to such accounts. His explanation of moral responsibility rests on the importance of understanding how someone interacts with others, how they fit into the social

context in which they find themselves as this determines our reactive attitudes towards their behaviour (Sneddon, 2005, p245). McKenna says that Strawson's theory is ambiguous on the 'relationship between resentment and retribution' and suggests that a 'more detailed account of the implications of Strawson's naturalistic account of responsibility for a theory of retributive justice' is needed (McKenna et al, 2008, p13). For example, it is one thing to say that our emotional nature is fixed and incapable of fundamental alteration but quite another thing to say that the institutional practises associated with them are also incapable of being eliminated or removed from our society. As such, it is not clear on whether the institution of punishment (our justice system) requires some form of external rational justification or not. This provides some relief to my discussion on the implausibility of punishing machines. Perhaps it is sufficient that we feel resentment towards a machine without the corresponding, external justification of needing to punish it. However a question does arise, that once we see robots being like us, we may want to punish them like us. It is however conceivable that we could feel resentment (reliably, systematically) in the same kinds of situation and towards the same kinds of agent, without feeling an equally strong need for punishment.

A question Strawson does not discuss is whether adopting the objective stance is actually an emotional or a rational response in the first place. It is quite conceivable that whether, and how often, you adopt the objective stance is also determined by your value and political belief system. Political beliefs often affect attitudes towards punishment. For example, leftists would view the reason for crime more objectively, that it was correlated to poverty and that to fight crime the focus should lean more towards alleviating the causes. On this reading, the conditions rather than the person are blamed. Conservatives would see people who commit crime as personally responsible and take a reactive (i.e., emotional) stance that they should be punished.

Strawson can be accused of being relativist, as he dispels notions of appealing to moral concepts and thus ignores normative notions. What he does supply however, is a descriptive schema, a working-model of how our reactive attitudes towards each other, impact our moral judgements in practise (Strawson, 1972, p10). In a sense, Strawson is deterministic, the emotions we feel in our inter-personal interactions, determine how we react both non-morally and morally. Where we do exercise some form of restraint, and in a sense free-will,

is when we choose to adopt an objective stance towards entities who we believe do not qualify for full moral responsibility. Strawson does not dwell on the capabilities needed to qualify for such responsibility. A number of theorists critique him for omitting detail on what makes a person an appropriate target for reactive attitudes and thus moral responsibility (Nagel 1986; Watson 1987; Russell 1992; McKenna 1998 and Scalon 1998; in McKenna et al, 2008, p12). Although he discusses incapacity in some detail, namely, conditions under which we suspend our reactive attitudes and adopt an objective stance (e.g. with small children and severely mentally retarded adults); he does not detail those capacities that establish a person as a 'normal adult who is capable of full participation in the moral community' (McKenna et al, 2008, p12). McKenna believes that some relevant interpretation of moral capacity is 'a critical issue, since the pessimist/incompatibilist will argue that among the relevant capacities we must consider is the ability to act otherwise or (libertarian) free will' (McKenna et al, 2008, p12). McKenna concludes that 'unless Strawson can provide some form of plausible account of moral capacity then it would appear there is a significant gap or lacuna in his own version of compatibilism' (McKenna et al, 2008, p13). McKenna entertains an interesting example to illustrate this deficiency using the case of implants (McKenna et al, 2008, p14). Imagine that our basic dispositions, that shape and influence our attitudes and intentions towards others were implanted by means of an artificial technique, for example, genetic engineering or neuro-surgery (McKenna et al, 2008, p12-13). Per Strawson, our reactive attitudes would still continue to operate, if the agent is capable of showing good or ill will towards others. Practically McKenna does not explain how these implants could influence our dispositions. We could just as readily take a course in ethics, as an example. Critics could counter-argue that implantation eliminates the agent's moral responsibility and illustrate a weakness in Strawson's theory. However, on this critical reading, we may be committed to saying that ordinary social processes, such as the education system, also alter a person's character historically over time and thus negate their moral responsibility. They too amount to a form of implantation. Strawson's theory does not acknowledge that either such processes would materially impact our reactive attitudes. For example, consider an ISIS terrorist, who had been brainwashed via training since an early age to believe that beheading Christians was *Allah's* (Arabic word for God's) will. It may be the case that I could as a Christian take an objective stance towards the ISIS terrorist believing his actions to be the cause of brainwashing. Strawson does not give guidance on whether we can suspend

our reactive attitudes in such cases. His argument has flaws, it does not specify whether we can train ourselves to have better moral reactive attitudes, nor does it mention whether we can rationally evaluate some of these attitudes As such, Strawson's theory allows for 'the great variety of forms which these human attitudes may take at different times and in different cultures' and he even goes so far as to say that 'no doubt to some extent my own descriptions of human attitudes have reflected local and temporary features of our own culture' (Strawson, 1972, p26). He ends with a cautionary caveat that a variety of forms of reactive attitudes should not be a reason to dismiss their efficacy in shaping our sense of morality, 'in the absence of any forms [of reactive attitude] it is doubtful whether we should have anything that we could find intelligible as a system of human relationships, as human society' (Strawson, 1972, p26).

McKenna says that Strawson's discussion 'turns on several sharp dichotomies that may also be questioned…among them is the fundamental opposition between the objective and participant stances' (McKenna et al, 2008, p11). For example, could parents be said to develop an objective attitude towards their children and see them as subjects? True, we see children as not yet moral and therefore as not being subject to moral punishment, but surely we cannot fully suspend our range of human reactive attitudes to our children because they are not yet moral? We are still able to engage in other forms of personal, emotional response such as parental love, whilst disciplining our children for behaving badly. This relates to a wider, general difficulty on how it is possible to set-aside our reactive attitudes when we take up the objective stance as a 'resource' in normal adult cases without thereby discrediting our reactive attitudes in the first place (McKenna et al, 2008, p11). Zimmerman says that Strawson leaves open the question of 'what kind of tension exists between the stances' (Zimmerman in McKenna et al, 2008, p276e). The 'puzzle' for Zimmerman is 'how is it possible that a specifically cognitive or intellectual operation psychologically dispels an attitude without in so doing also rationally disqualifying the attitude? The psychological operation in question is, of course, the shift from the reactive to the objective stance' (Zimmerman in McKenna et al, 2008, p276-277e).

Various theorists critique the objective/participant dichotomy as being more problematic for Strawson than he acknowledges. Galen Strawson (son of PF Strawson) believes, unlike his father, that we can adopt an objective attitude towards ourselves and others, indefinitely

over time with much practise and experience. He uses the case of Buddhist monks and other mystics who he believes 'have succeeded in altering quite profoundly their experience of themselves (and others) as acting, thinking, and feeling beings…and by so doing have achieved what is in certain respects a more correct view of the world, precisely to the extent that they have ceased to regard themselves and others as truly self-determining sources of actions, and have thereby come to adopt the objective attitude' (Galen Strawson in McKenna et al, 2008, p130-131e). Sommers disagrees and defends PF Strawson's claim that exclusive objectivity would preclude meaningful relationships from being able to occur (Sommers 2007, in Shabo, 2012, p131) Shabo defends this claim saying we are unable to adopt an objective attitude if we are in a personal relationship with someone, denoted by a particular type of caring for them, for example, mature friendship and reciprocal love (Shabo, 2012, p133). Thus, when we adopt an objective attitude, we do not care in 'the required way' about that person's attitudes towards us and therefore 'someone who managed exclusive objectivity of attitude [which Shabo believes is impossible] would be excluded from personal relationships' (Shabo, 2012, p133). Some have even suggested that reactive attitudes are not as essential to human nature as Strawson suggests (Watson, 1987; Wallace, 1994; in McKenna et al, 2008, p10). These theorists assert that reactive attitudes are not permanent, unlike more basic human emotions such as love and fear. Rather they suggest that they are historically and socially specific (McKenna et al, 2008, p10).

To conclude, many of the objections raised to Strawson are valid but I believe that for my argument they are not strong enough to refute my usage of him. The objection that Strawson does not specify rules or principles to guide our moral judgements is valid, he does not. However, in some ways Strawson is Kantian as reactive attitudes are part of human nature, which impact moral principles. For Strawson it is easy for us to conform by displaying reactive attitudes; for Kant it is easy to ask whether all rational people would do likewise when making moral judgements. For both, the enforceability is easy because it is in our natures to do. Perhaps this is why Strawson is non-committal in terms of binding rules, they are unnecessary, the rule is our nature. It may well be that reactive attitudes towards a machine are unreasonable; however, the beauty of Strawson is that he does not judge this. Perhaps you cannot morally tell people not to feel resentment towards machines when they display this emotion?

In summary, Strawson has appeal - there is no innate essence to being a moral agent, if I am capable of entering into a certain type of relationship with a machine, I can accord it agency. Although Strawson's theory was modelled on the interpersonal relationships that people have between each other, we only have to perceive that machines have emotions for our reactive attitudes to kick-in; rather than machines having actual emotions.

I now look at the enhancement debate, as it provides insights into morality as a threshold concept which supports my developmental view. The debate also extends morality beyond humans, which gives credence to my argument for machine moral agency.

4. The enhancement debate

The following themes in the enhancement debate are useful in supporting my claim that moral agency can be extended to machines:

- Emerging technologies of human enhancement (to be smarter, have better memories, be stronger and quicker, have more stamina, live longer, be more resistant to diseases, and enjoy richer emotional lives) might result in a situation where some humans have a superior moral status to others, in the same way that humans are typically held to have a higher moral status than other species (in respect-based ethical systems such as Kantianism and contractualism). Such a situation would render false the assumption that all persons are necessarily morally equal, and it might have consequences for the assumption that machines do not have and cannot have moral status because they are not the same as humans.

- Many enhancement theorists view morality as a threshold concept meaning morality depends upon the properties of the agent and the degree to which these are possessed. This allows us to entertain a world where machines have a base level of qualities that can accord them moral agency, even if humans possess such qualities to a greater extent.

Enhancements are ubiquitous. In the broadest sense 'to enhance human beings is to expand their capabilities – to enable them to do what normal human beings have been hitherto unable to do' (Buchanan, 2013, p38). Computers can be perceived as enhancing us, for example, they help us process mathematical equations speedily (Buchanan, 2013, p38).

When people think enhancements most of us think about changing something inside our bodies, i.e., plastic surgery, genetic engineering and so forth. Buchanan makes the point that they can also result from certain, relevant forms of external or environmental changes. For example, the agricultural revolution of the mid-18th century afforded us better nutrition which led to physically stronger human bodies (Buchanan 2013, p39). Similarly, when we use computers they help us overcome biological limitations of our brains information processing and calculative abilities. They enhance our productivity and have 'profoundly changed our conception of ourselves and our world and shaped our most basic social relations' (Buchanan, 2013, p39).

The moral equality principle, understood as prescribing treatment of persons as equals, (i.e., with equal concern and respect, and not the principle of treating persons equally) is challenged by arguments from the enhancement debate. Some believe that human enhancements will lead to the development of a super-class of humans or 'post-humans' (Buchanan, 2013, p209). Thus moral agency can exist for both unaltered man and altered post-humans. Similarly, machines are advancing rapidly, and such technological enhancements may enable us to view machines analogously to post-humans, in the sense that neither is human. I claim that this moral duality would be necessary for us to live in a future world, increasingly intertwined with artificial forms of intelligence and altered post-humans. Such duality would allow man, post-man and machines to be treated with moral concern. One of the key objections to enhancement, the appeal to inequality, is that it will lead to a situation of physically or mentally enhanced haves and have not's. The concern is that such enhanced beings would have a higher moral status than that of normal human beings, leading to a 'morally bifurcated world' of persons and 'post-persons' (Buchanan, 2013, p209). Sandel contends that genetic engineering could augment our abilities to the degree where we start valuing ourselves in a more god-like than human like manner. It creates in his words a tendency towards 'hubris' (extreme pride/arrogance), which reduces three key features of our current sense of morality, namely, humility, responsibility and solidarity (Sandel, 2004, p9). In short, it starts eroding our morality and what we value as moral. A common thread emerges from debates in this arena, notably that changes in technology alter the way we consider moral (and other) matters. Some maintain that genetic enhancement actually erodes human agency (in the general sense of the term) by

overriding the effort it takes, for example, to excel at sport or be academically excellent. The 'admiration shifts from the [athlete/scholar] to the pharmacist' (Sandel, 2004, p5). Sandel believes the opposite that it leads to an explosion, rather than erosion, of moral responsibility as we leave less and less to choice. We can choose the sex of our children, change their height by giving them growth hormones, and increase the strength of our muscles and memory (Sandel, 2004, p9). In practise theorists like George Annas worry that these enhanced beings would think they were superior to us and would treat us as if we had lower moral status which could lead to a myriad of ethical evils such as exploitation, slavery and extermination (Annas in Buchanan, 2013, p224). Buchanan takes this concern seriously in relation to the large-scale human rights violations that have plagued our past and current histories (think Rwanda for example, or most recently Israel's serial bombing campaign in the Gaza strip). While this concern should be taken seriously, it provides a negative view of current human and post human nature. In a scenario where over time, the whole of humanity was enhanced so that there were no humans left and only post-persons remained, the moral equality assumption would be irrelevant. Is it then relevant where both persons and post-persons exist (Buchanan, 2013, p215)? Buchanan says it would still not be relevant, as the concept of moral rights (and obligations) is a 'threshold' concept not a scalar one (Buchanan, 2013, p215). On this view, moral rights (and agency) are ascribed if an individual (or in our argument machine) has certain capabilities or interests which qualify it for agency. The fact that some may have these capabilities to a greater degree is irrelevant, as is the fact that the agent in question may be non-human. Thus moral personhood, defined on the basis of qualities, rather than on any species-specific criteria, makes the accordance of moral agency to non-humans possible, both living and otherwise. A green Martian with the same capabilities would be accorded agency despite having a different morphology, biology and neurological structure to us. As morality is a threshold concept, this would apply even if people possessed such capabilities to a greater degree than Martians (or vice versa). Per MTT if we see machines behaving like us morally, we may see them as having some of our qualities for agency and using a threshold concept can accord them agency even if they possess such qualities to a lesser degree. Similarly, per Strawson, if we can resent or praise machines, like our Android father, we can accord them moral agency. In reality, morality is already a threshold concept. After all some people have better logic, and reason better than others (and so forth), but that does not mean they have a higher moral status. The threshold

concept actually accommodates many inequalities, including inequalities in the very characteristics that confer moral status (Buchanan, 2013, p217).

The answer to whether we can accord agency to machines and humans, in part relates to the type of meta-ethical principle selected. Respect-based ethical systems which accord consideration based on the belief that all beings have an intrinsic moral worth allow us to answer yes. This relates to the threshold concept of moral consideration. Kant stemmed from such a tradition and his definition implies a kind of inviolability of moral worth. On this reading, machines and humans can have equal moral worth. An interest-based reading implies that machines and man can both have moral worth, but that man due to his higher-order capabilities would have more moral worth than machines. Moor's scale of moral responsibility allows machines to have limited agency and humans full agency (Moor, 2006, p19-21). It is useful to view moral responsibility on such a continuum as it allows AI to currently be included as moral agents at least in a limited sense. The continuum aligns to the view of morality as being something one can develop or grow into.

To illustrate this developmental view we can see moral education as a type of moral enhancement. We educate our children on a sense of morality from a young age, and as they grow they get better at identifying moral right from wrong and their ability to morally reason grows. In this sense we are already enhancing the capabilities that give us moral agency. If we were able to enhance machines in a similar way we should be able to accord them agency. Learning algorithms exist in programming language today and it may be possible to 'train' machines in certain ways. Aristotle believed that training installed the right kind of virtues for morality. Could we enhance machines by installing moral virtues in them? This may be possible if morality constitutes a certain disposition to act in certain ways. For example, machines could be programmed to minimise harm to people. The movie *i-Robot* (2004) modelled on the collection of short stories by author (and Professor of biochemistry) Isaac Asimov, featured anthropomorphic robots in a futuristic world, used mostly as servants for various public services. These robots were programmed with three laws of robotics, a set of ethical directives, namely; to never harm a human or let a human come to harm; to always obey humans unless this violates the First Law; and to protect its own existence unless this violates the First or Second Laws. If we see morality as an external concept, which does not rely on qualities innate to humans but rather on the way we can

feel about an agent (per Strawson), then we can see i-Robots with their programmed laws of robotics as demonstrating moral good-will towards us. The resulting reactive attitude that we feel towards i-Robot is gratitude (the opposite of resentment) which allows us to accord the actions of the robot moral agency. Enhancements (and the enhancement debate) question whether human nature is unalterable. Evolutionary biologists believe it can be altered. The evolution of humans involves a series of adaptations (an ongoing process) at physical, mental and behavioural level. Enhancements help us to think of human nature (and moral agency) as part of an ongoing, evolutionary process and thus alterable. Enhancements 'challenge the idea that we have a fixed core of characteristics…which are unalterable' (Buchanan, 2013, p123). So if morality is an evolving concept for humans, can it evolve for machines?

In summary the enhancement debate has helped support my argument to extend agency beyond humans and it challenges the moral equality assumption. It allows us to view morality as a threshold concept where varying levels of qualification for agency are possible. This aligns well to my argument that morality is a developmental construct that can apply to non-human, machine entities. The threshold concept also supports my working definition of machine autonomy as providing a minimum set of standards for autonomy.

5. Conclusion

So post Strawson, has our working definition of machine autonomy gotten any better? Where does this leave us with machines? I take aspects of Strawson which are useful, namely, that morality is externally derived, based on the relationship we as humans are able to enter into with machines. Our reactive attitudes are emotional, we can take an objective stance, but this in itself might be based on either a rational or further emotive decision, as Strawson is non-committal here. In any event, it does not matter, on Strawson's relational definition the requirement is not that machines have actual emotions, but rather that we treat them as if they do.

I can thus use the best of Strawson for my machines, if they behave like us we react to them like us. Machines need to be autonomous for my definition to work, as we are autonomous. The more complex machine autonomous behaviour, the better, as the more removed we become as human operators from their actions, and the more likely we are to believe that

machines are acting independently like us and hence adopt a reactive attitude towards them. My working definition of machine autonomy as self-propelling, rational processors of information that can react to stimuli in the environment and tailor their actions accordingly, is a necessary step towards one day enabling us to adopt a reactive attitude towards machines.

The day when I miss my friend Paro, the robot that cares for me in my old age; or conversely, when I resent EthEl for inhibiting my autonomy to take care of myself, is the day when Strawson is proved empirically plausible. Until such time, the debate on whether morality is about our natural and irrevocable feelings in a social context; or whether our moral judgements are guided by metaphysical criteria which can be universally applied, will remain contested.

Reference list

2001: *A Space Odyssey*, 1968. [film] Directed by Stanley Kubrick. USA: Metro-Goldwyn-Mayer.

Allen, C., (2010). 'Artificial life, artificial agents, virtual realities: technologies of autonomous agency', in 'Information and Computer Ethics, Cambridge University Press, Cambridge. Ed. Floridi, L.

Allen, C., Wallach, W., and Smit, I., 2006. *Why Machine Ethics?* IEE Intelligent Systems, vol. 21, no. 4, July/August, pp12-17.

Anderson, M., Anderson, S. L. (2007). 'Machine ethics: creating an ethical intelligent agent', AI Magazine, volume 28, number 4, American Association for Artificial Intelligence.

Anon., 2014. Special Report: Rise of the Robots, *The Economist*, [online subscription], March 29th – April 4th, 2014. Available at: www.economist.com [Accessed 29 March, 2014].

Blade Runner, 1982. [film] Directed by Ridley Scott. USA: Warner Bros.

Bostrom, N., Sandberg, A., 'Cognitive Enhancement: Methods, Ethics, Regulatory Challenges', Scientific Engineering Ethics, 15, (2009), pp311-341.

Brey, P. (2010) Values in technology and disclosive computer ethics, in Floridi (Ed). The Cambridge handbook of Computer Ethics, Cambridge University Presss, Cambridge, p41-58.

Bringsjord S., Bello, P., and Ferrucci, D., 2001. Creativity, the Turing Test, and the (Better) Lovelace Test. *Minds and Machines*, 11, pp3-27.

Buchanan, A., 2013. *Beyond Humanity? The Ethics of Biomedical Enhancement.* Oxford:

Oxford University Press.

Bynam, T. W., (2006). 'Flourishing ethics', in Ethics and Information Technology', Springer, 8, pp157-173.

Christman, J., 'Autonomy in Moral and Political Philosophy', *The Stanford Encyclopaedia of Philosophy* (Spring 2015 Edition), Edward N. Zalta (ed.), [online] Available at: http://plato.stanford.edu/archives/spr2015/entries/autonomy-moral [Accessed 5 March, 2015].

Dennett, D. C. 1997.'When HAL Kills, Who's to Blame? Computer Ethics,' in HAL's Legacy: 2001's Computer as Dream and Reality, D. G. Stork (ed.), Cambridge, MA: MIT Press.

Deutsch, H., 'Relative Identity', *The Stanford Encyclopaedia of Philosophy* (Winter 2008 Edition), Edward N. Zalta (ed.), [online] Available at: http://plato.stanford.edu/archives/win2008/entries/identity-relative [Accessed 22 July 2014].

Dreyfus, L. and Dreyfus, S., 1988. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer.* New York: The Free Press, A Division of Macmillian Inc.

Fleischacker, Samuel, 'Adam Smith's Moral and Political Philosophy', *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition), Edward N. Zalta (ed.), [online] Available at http://plato.stanford.edu/archives/spr2013/entries/smith-moral-political [Accessed 1 March, 2015].

Floridi, L. (2010). 'Ethics after the information revolution', in Information and Computer Ethics, Cambridge University Press, Cambridge.

Floridi, L., Sanders, J. W., 'On morality of artificial agents', Minds and Machine, Kluwer Academic Publishers, volume 14, 2004, pp349-379.

Franklin-Hall, A., 2013. On becoming an adult: autonomy and the moral relevance of life's stages. *The Philosophical Quarterly*, vol.63, no. 251.

French, R. M., 2000. The Turing Test: the first 50 years. *Trends in Cognitive Sciences*, vol. 4, no. 3, pp115-122.

Guyer, P., (2004). Kant, Immanuel. In E. Craig (Ed.), *Routledge Encyclopaedia of Philosophy*. London: Routledge. Retrieved March 15, 2015, [online], Available at http://www.rep.routledge.com/article/DB047SECT11 [Accessed 15 March, 2015].

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, pp19-29.

Introna, L., 'Phenomenological Approaches to Ethics and Information Technology', *The Stanford Encyclopaedia of Philosophy* (Summer 2011 Edition), Edward N. Zalta (ed.), [online] Available at: http://plato.stanford.edu/archives/sum2011/entries/ethics-it-phenomenology [Accessed 8 September 2013].

i-Robot, 2004. [film] Directed by Alex Proyas. USA: 20th Century Fox.

Lara, D., 'Kant and Hume on Morality', The Stanford Encyclopedia of Philosophy (Winter 2014 Edition), Edward N. Zalta (ed.), [online] Available at: http://plato.stanford.edu/entries/kant-hume-morality [Accessed 17 December 2014].

McKenna, M. and Russell, P. eds., 2008. *Free Will and Reactive Attitudes: Perspectives on P. F. Strawson's 'Freedom and Resentment'*. Surrey, England: Ashgate Publishing Limited.

Moor, J. M., 'The nature, importance and difficulty of machine ethics', Machine Ethics, IEEE Computer Society, 2006.

Noorman, M., 'Computing and Moral Responsibility', *The Stanford Encyclopaedia of*

Philosophy (Summer 2014 Edition), Edward N. Zalta (ed.), [online] Available at: http://plato.stanford.edu/archives/sum2014/entries/computing-responsibility [Accessed 23 July 2014].

Sandel, M., 'The case against perfection: what's wrong with designer children, bionic athletes and genetic engineering', The Atlantic, April 2004.

Schmidt, C. T. A., and Kraemer, F., 2006. Robots, Dennett and the autonomous: a terminological investigation. *Journal for Artificial Intelligence, Philosophy, And Cognitive Science*, 16, 1, pp73-80.

Schneewind, J. B., 2011. Modern moral philosophy. A Companion to Ethics, P., Singer (ed), Blackwell Publishing, Oxford.

Shabo, S., 2012. Incompatibilism and personal relationships: another look at Strawson's objective attitude. *Australasian Journal of Philosophy,* vol. 90, pp131-147.

Shelley, M., 1818. *Frankenstein or the modern Prometheus.* London: Lackington, Hughes, Harding, Mavor and Jones.

Sneddon, A., 2005. Moral Responsibility: The difference of Strawson, and the difference it should make. *Ethical Theory and Moral Practise,* 8: pp239-264.

Star Wars, 1977. [film] Directed by George Lucas. USA: 20th Century Fox.

Strawson, P. F., (1972). Freedom and Resentment and other Essays, Milton Park: Routledge, pp1-28.

Sullins, J.P. 2006. 'When is a Robot a Moral Agent?' *International review of Information Ethics*, 6(12): 23-29.

Tavani, H. (2008). 'Floridi's Ontological Theory of Informational Privacy: Some Implications and Challenges', Ethics and Information Technology, 10 (155-166).

The Wizard of Oz, 1939. [film] Directed by Richard Thorpe. USA: Metro-Goldwyn-Mayer.

Turing, A. (1964), Computing machinery and intelligence, in A. R. Anderson, ed., *Minds and Machines*, Englewood Cliffs, NJ: Prentice-Hall, pp4-30.

Van Gulick, Robert, 'Consciousness', *The Stanford Encyclopaedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), [online] Available at: http://plato.stanford.edu/archives/spr2014/entries/consciousness [Accessed 15 February 2015].